ISSN 1847-3938

mipro 2015

organizer upro

38th international convention

May 25-29, 2015, Opatija – Adriatic Coast, Croatia

Lampadem tradere

mipro - path to knowledge and innovation

mipro proceedings

My profession.
My organization.
My IEEE.

Discover the benefits of IEEE membership.

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

Join today
www.ieee.org/join

IEEE

MIPRO 2015

38th International Convention

May 25 - 29, 2015
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Distributed Computing, Visualization and Biomedical Engineering /DC VIS

Telecommunications & Information /CTI

Computers in Education /CE

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

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

MIPRO Junior - Student Papers /SP

Edited by:
Petar Biljanović

International Program Committee

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

organized by

MIPRO Croatian Society

technical cosponsorship

IEEE Region 8

under the auspices of

Ministry of Science, Education and Sports of the Republic of Croatia
Ministry of Maritime Affairs, Transport and Infrastructure of the Republic of Croatia
Ministry of Entrepreneurship and Crafts of the Republic of Croatia
Ministry of Public Administration of the Republic of Croatia
Croatian Chamber of Economy
Primorsko-goranska County
City of Rijeka
City of Opatija
Croatian Regulatory Authority for Network Industries
Croatian Power Exchange - CROPEX

patrons

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

sponsors

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

All papers are published in their original form

For Publisher:

Petar Biljanović

Publisher:

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

Printed by:

GRAFIK, Rijeka

ISBN 978-953-233-083-0

Copyright © 2015 by MIPRO

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

CONTENTS

LIST OF PAPER REVIEWERS

LIST OF AUTORS

FOREWORD

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

INVITED PAPER

State-of-the-Art Semi-Classical Monte Carlo Method for Carrier Transport in Nanoscale Transistors 5
P. Palestri, E. Caruso, F. Driussi, D. Esseni, D. Lizzit, P. Osgnach, S. Venica, L. Selmi

PAPERS

Preparation and Characterization of SERS Substrates: from Colloids to Solid Substrates 15
L. Mikac, M. Ivanda, M. Gotić, D. Ristić, V. Đerek, H. Gebavi, P.G. Gucciardi, S. Trusso, C. D'Andrea

The Conductivity and TEMF of MoS_2 with Mo_2S_3 Additive 18
G.E. Yakovleva, A.S. Berdinsky, A.I. Romanenko, S.P. Khabarov, V.E. Fedorov

Strain-Sensing Element Based on Layered Sulfide $Mo_{0.95}Re_{0.05}S_2$ 21
V.A. Kuznetsov, A.S. Berdinsky, A.Y. Ledneva, S.B. Artemkina, M.S. Tarasenko, V.E. Fedorov

Electrical Properties of Composite Silicon Thin Films 25
V. Tudić

Impact of Different Gate Insulator Materials on the Electron Mobility in Ultra-thin (100) InGaAs-on-insulator MOS Devices 31
S. Krivec, M. Poljak, T. Suligoj

Impact of the Emitter Length Scaling on Electrical Characteristics of Horizontal Current Bipolar Transistor with Single Polysilicon Region 37
M. Koričić, J. Žilak, T. Suligoj

Optimization of Floating Guard Ring Parameters in Separate-Absorption-and-Multiplication Silicon Avalanche Photodiode Structure 43
I. Janeković, T. Knežević, T. Suligoj, D. Grubišić

Low-Power Low-Voltage CMOS Analog Signal Processing Circuits Using a Functional Core 48
C. Popa

Multifunctional Circuit Using Improved Performances Linearization Technique 52
C. Popa

Design of a High-Voltage Differential Amplifier Based on Stacked Low-Voltage Standard CMOS with Different Input Stages 56
S. Pashmineh, D. Killat

Opamp Based Bandgap Voltage Reference in 130 nm: Design and Schematic Level Simulation 62
A. Pajkanović, V. Malbaša

Design of a Power Supply and Voltage References for a Wirelessly Powered ADC in 0.18-µm CMOS 68
N. Bako, Ž. Butković, A. Barić

RFID UHF Protocol Implementation in Distributed Sensor Networks 74
I. Broz, N. Bako, Ž. Butković, A. Barić

Temperature Dependence of Injection-Locked Fabry-Pérot Laser Emission in WDM-PON Architectures 80
M. Šprem, D. Babic, M. Bosiljevac, Z. Šipuš

Fiber-Optic Vibration Sensor for High-Power Electric Machines 85
B. Igrec, M. Bosiljevac, S. Rudan, Z. Šipuš, D. Babić

Comparison of Electrical and Electromagnetic Properties of a Planar Inductor 89
J. Kundrata, A. Barić

Effects Influencing Repeatability of Direct Power Injection Measurements 93
F. Fajdetic, M. Mikic, T. Mandić, A. Barić

SPICE Analysis of RL and RC Snubber Circuits for Synchronous Buck DC-DC Converters 97
R. Blečić, R. Gillon, B. Nauwelaers, A. Barić

Dual-channel Programmable Precise Time-delay Pulse Signal Generator for Synchronous DC-DC Converter Applications 104
F. Hormot, J. Bačmaga, A. Barić

Effect of Dead Time and Semiconductor Device Voltage Drops of Output Voltage of Multilevel Converters 109
I. Mrčela, V. Šunde, M. Kovačić

Model of Three Level Neutral Point Clamped Converter (NPC) for Grid Connected Photovoltaic Systems 116
M. Novak, V. Šunde, Ž. Jakopović

Fast Development of a DSP-Based Control of Three-Phase Four-Leg Inverter for Auxiliary Power Supply Converters Using Hardware-in-the-Loop 121
I. Musulin, J. Babić, Ž. Jakopović

Noise Improvement Using SC Filters 127
I. Volarić, N. Stojković, S. Vlahinić

Phase Demodulation of Magnetic GMR Signal for Virtual Sensing Applications by using Hilbert Transform 133
M. Adnan, D. Hammerschmidt, M. Huemer

Thin-film CIGS Solar Modules for Design Driven Applications in the Frame of the FP7 NMP Project SolarDesign 137
W. Brenner, N. Adamović

Measurement System for Object Detection Based on Multielectrode Capacitive Sensor 143
D. Ćika, T. Martinović, H. Džapo

IBMS - Intelligent Building Management System Framework 149
S. Tepić, P. Pejić, J. Domšić, H. Mihaldinec, H. Džapo

A Hybrid Local Method for Stereo Correspondence Based on Three-Dimensional Recursive Search 155
M. Rožić, M. Đonlić, T. Pribanić

Two Approaches to Derive Approximate Formulae of NILT Method with Generalization 161
L. Brančík, N. Smith

Laboratory Exercise for Class B/AB Power Amplifier 167
Ž. Stojanović, S. Stojanović, I. Činikas

Integration of High-Level Synthesis to the Courses on Reconfigurable Digital Systems 172
I. Skliarova, V. Sklyarov, A. Sudnitson, M. Kruus

DISTRIBUTED COMPUTING, VISUALIZATION AND BIOMEDICAL ENGINEERING

PAPERS

DISTRIBUTED COMPUTING

Cloudflow – A Framework for MapReduce Pipeline Development in Biomedical Research 185
L. Forer, E. Afgan, H. Weißensteiner, D. Davidović, G. Specht, F. Kronenberg, S. Schönherr

Science Gateway for Distributed Multiscale Course Management in e-Science and e-Learning - Use Case for Study and Investigation of Functionalized Nanomaterials ... 191
Y. Gordienko, S. Stirenko, O. Gatsenko, L. Bekenov

Evaluation of POS Scheduling Algorithm for Distributed Computing Environment . 197
A. Shamakina, L. Sokolinsky

Performance evaluation on a Grid platform 201
A.A. Banka, M.A. Dedmari, M.T. Masoodi

Synergy of Volunteer Measurements and Volunteer Computing for Effective Data Collecting, Processing, Simulating and Analyzing on a Worldwide Scale..... 206
N. Gordienko, O. Lodygensky, G. Fedak, Y. Gordienko

Sucssessfull Implementation of L3B: Low Level Load Balancer 212
K. Cvetkov, S. Ristov, M. Gusev

Scalable and Parallel Machine Learning Algorithms for Statistical Data Mining - Practice & Experience..... 217
M. Riedel, M. Goetz, M. Richerzhagen, P. Glock, C. Bodenstein, A.S. Memon, M.S. Memon

Decomposition of Natural Join Based on Domain-Interval Fragmented Column Indices..... 223
E. Ivanova, L. Sokolinsky

Theoretical and Statistical Evaluation for Approximate Solution of Large, Over Determined, Dense Linear Systems 227
Ž. Jeričević, I. Kožar

DARIAH Meta Hosting: Sharing software in a distributed infrastructure 230
J. Rybicki, B. von St. Vieth

Building and Provisioning Bioinformatics Environments on Public and Private Clouds..... 236
E. Afgan, K. Krampis, N. Goonasekera, K. Skala, J. Taylor

Gathering Useful Programming Data: Analysis and Insights from Real-time Collaborative Editing **242**
M. Rantala, J. Soini, T. Kilamo

Microservices Validation: Mjolnirr Platform Case Study **248**
D.I. Savchenko, G.I. Radchenko, O. Taipale

Four Level Provenance Support to Achieve Portable Reproducibility of Scientific Workflows **254**
A. Bánáti, P. Kacsuk, M. Kozlovszky

Using the P-TOSCA Model for Energy Efficient Cloud **258**
B. Ivanovska, S. Ristov, M. Kostoska, M. Gusev

Achieving Dynamic Workflow Management System by Applying Provenance Based Checkpointing Method **263**
E. Kail, P. Kacsuk, M. Kozlovszky

Operating System Impact on CPU and RAM Utilisation when Introducing XML Security **267**
S. Ristov, M. Gusev

Evaluating the Impact of the Number of Processor Cores, Cache Memory and Threads During the Parallel Execution of Programs **272**
F. Dika, V. Xhafa

Performance Analysis of a New Cloud e-Business Solution **276**
K. Kolić, M. Gusev, S. Ristov

Interoperable Job Execution and Data Access through UNICORE and the Global Federated File System **282**
S. Memon, M. Riedel, C. Koeritz, A. Grimshaw

Multi-stream 2D Frequency Table Computation on Dataflow Architecture **288**
P. Škoda, V. Sruk, B. Medved Rogina

Porting an N-tier Application on Cloud using P-TOSCA: A Case Study **294**
M. Kostoska, A. Donevski, M. Gusev, S. Ristov

Cloud Portability Standardisation Overview **299**
E. Markoska, I. Chorbev, S. Ristov, M. Gusev

In-Situ Anonymization of Big Data **305**
T. Križan, M. Brakus, D. Vukelić

VISUALIZATION SYSTEMS

A GPU-based Simultaneous Real-Time EEG Processing and Visualization System for Brain Imaging Applications **315**
Z. Juhasz, G. Kozmann

Web Portals for Visualizing and Searching Spatial Data..... 321
L. Bećirspahić, A. Karabegović

Prototyping of Visualization Styles of 3D Scalar Fields Using POV-Ray Rendering Engine..... 328
J. Opiła

Freeform Spatial Modelling Using Depth-sensing Camera..... 334
M. Volarević, P. Mrazović, Ž. Mihajlović

Seismic Data Visualisation..... 340
M. Ivančić, Ž. Mihajlović, I. Ivančić

Situation and Location Awareness in Harsh Environment..... 344
M. Kozlovszky, D. Zavec Pavlinić, A. Oder, G. Fehér, P. Bogdanov

Error Correcting Codes and Their Usage in Steganography and Watermarking..... 349
H. Kostadinov, N.L. Manev

Model for Calculation of Post Service Efficiency by Using GIS..... 353
M. Kranjac, U. Sikimić, M. Vujaković

Generalization of Hypersphere Function..... 359
D. Letić, I. Berković

Visualization and Multimedia Presentation of Cultural Heritage..... 364
S. Rizvić, V. Okanović, A. Sadžak

Using Entropy Information Measures for Edge detection in Digital Images 368
D. Sušanj, V. Tuhtan, L. Lenac, G. Gulan, I. Kožar, Ž. Jeričević

BIOMEDICAL ENGINEERING

Wireless Electrocardiographic Monitoring in Veterinary Medicine..... 375
M. Brložnik, V. Avbelj

Software for Acid-Base Status Disorders Management 379
M. Mandžuka, E. Begić, Z. Begić, D. Bošković

Impact of Electrode Misplacement on the Accuracy of 12-lead ECG Synthesized from Differential Leads..... 384
I. Tomašić, R. Trobec, A. Rashkovska, M. Lindén

Extranodal Pacemaker Activity During Sleep - a Case Study of Wireless ECG Sensor Data..... 389
V. Avbelj

A New Color Assessment Methodology using Cluster-based Features for Skin Lesion Analysis 392
M.J.M. Vasconcelos, L. Rosado, M. Ferreira

Automatic Adaptation of Filter Sequences for Cell Counting 398
U. Čibej, J. Lojk, M. Pavlin, L. Šajn

Assessing Blood Pressure Unobrtusively by Smart Chair 404
D. Zazula, J. Kranjec, P. Kranjec, B. Cigale

Towards Noninvasive Bioimpedance Sensor Design Based on Wide Bandwidth Ring Resonator 409
M. Pavlin, F. Novak

Exploring Bacterial Biofilms with NAMD, a Highly Scalable Parallel Code for Complex Molecular Simulations 412
D. Tomić

Information System of the Institute for Blood Transfusion of Montenegro 417
A. Balota

TELECOMMUNICATIONS & INFORMATION

PAPERS

TELECOM MARKET AND REGULATIONS

Effective Strategies for Implementation of Customer Experience Management in MSPs in Bosnia and Herzegovina 429
N. Banović-Ćurguz , D. Ilišević

Regulatory Aspects of Child Online Protection 435
S. Hadžović, D. Šerval, S. Kovačević

Architecture and Functionality in M2M Standards 439
M. Ptiček, V. Čačković, M. Pavelić, M. Kušek, G. Ježić

Metodologija procjene kapaciteta RF spektra za potrebe implementacije mobilnih servisa u periodu do 2020. godine sa primjenom u BiH 445
A. Džebo, A. Krivošić

Uloga kvalitetnog dizajna poslovnih procesa u složenom okruženju telekomunikacijskog tržišta 450
D. Markulin, K. Musa

SERVICES

Semantic Approach to Accommodation & Booking Related Web Services 459
I. Pavković, I. Štajduhar

Voice over Wi-Fi **465**
D. Bilajbegović

An Example of Real Time QoE IPTV Service Estimator **468**
N. Goran, M. Hadžialić

Primjena mobilnih komunikacijskih tehnologija sa svrhom unapređenja usluga javnog prijevoza **472**
M. Dujak, Z. Civadelić, K. Vidović, P. Miletić

SIMULATIONS AND MEASUREMENTS

Computer Modeling and Simulation of Teletraffic Models for 5G Networks **479**
E.P. Ivanova, T. Iliev, G. Mihaylov, I. Stoyanov, F. Tsvetanov, E. Otsetova, D. Radev

Simulation of a Quasigroup Error-Detecting Linear Code **483**
N. Ilievska, D. Gligoroski

An Integrated Remotely Controlled Oceanographic Measuring Station **489**
H. Kalinić, D. Jelavić, D. Ivanković, J. Bašić, S. Muslim, S. Marasović, V. Dadić

Role of Performance Measurements in Process of Migration to IP Mobile Backhaul in 3G Networks **494**
D. Ilišević, N. Banović-Ćurguz

Comparative Analysis of GSM Coverage Prediction with Measurement Results for Urban Areas Using Statistical Nonparametric Mapping **499**
H. Lučkin, E. Lučkin, M. Škrbić

SOFTWARE ENGINEERING & DATA MANAGEMENT

Software Defined Networking Demands on Software Technologies **507**
T. Galinac Grbac, C.M. Caba, J. Soler

Data Collection for Software Defect Prediction – an Exploratory Case Study of Open Source Software Projects **513**
G. Mauša, T. Galinac Grbac, B. Dalbelo Bašić

Trendovi u izgradnji podatkovnih centara u Hrvatskoj **520**
M. Galac, D. Valenčić

Data Storage and Analysis System for Conducting Biotechnological Experiments **528**
D. Spahić, G. Mauša, S. Kraljević Pavelić, T. Galinac Grbac

CLOUD SYSTEMS

Automatizacija postavki mreže u *Cloud* sustavima preko samouslužnih portal..... 537
M. Maras, T. Kosanović

Data Storage and Synchronization in Private Cloud 541
D. Hodanić, N. Vrkić, M. Tomić

Testing in the Private Cloud 546
T. Pavić, I. Vrbovčan, M. Šoša Anić

NETWORKS & NETWORKS PERFORMANCE

The Methodology for the Optimal Choice of Access Technologies 553
Ž. Božinović, N. Hadžiahmetović

An Open Source TCP/UDP-Based Network Probing Tool for Real-Time Packet Loss Estimation 559
A.F. Khalifeh, A. Abbad, K.A. Darabkeh

Improvements in Failure Detection for Emergency Service Centers in IMS Network 564
I. Krevatin, Ž. Presečki, M. Gudelj

BER Analysis in Relay-Based DF Cooperative Diversity Systems with Relay & Destination Interferers 569
A.M. Magableh, M.F. Al-Mistarihi, R. Mohaisen, A.H. Sharaqa

Outage Probability Analysis in Relay-Based DF Cooperative Diversity Systems with Relay & Destination Interferers 574
R. Mohaisen, M.F. Al-Mistarihi, A.M. Magableh, A.H. Sharaqa

Closed Form Expression of Outage Probability in DCSK Cooperative Communication Systems over Nakagami-m Fading Channels 580
M.F. Al-Mistarihi, A.M. Magableh, M.M. Al-Khasawneh

Outage Probability of Dual-Hop Amplify-and-Forward Cognitive Relay Networks under Interference Power Constraints over Nakagami-m Fading Channels 584
M. Shurman, M.F. Al-Mistarihi, M. Alhulayil

Outage Probability of Amplify-and-Forward Underlay Cognitive Relay Networks with Selection Diversity over Nakagami-m Fading Channels 589
M. Alhulayil, M.F. Al-Mistarihi, M. Shurman

ENVIRONMENTAL PROTECTION

Simulation of Energy Efficient Telecommunication Network 597
G. Dimić, J. Seljan, D. Šimunić

Environmental Sustainability of Information and Communication Technologies for Advanced Metering and Home Area Networks 601
V. Mujan, S. Aleksić

Aspects of Remote Monitoring and Recording System of Non-ionizing Electromagnetic Radiation 607
D.S. Šuka, M.I. Simić, P.V. Pejović

Energetic Self-Regulation of eWALL System 612
Ž. Martinović, D. Budić, D. Šimunić

Investigation of Electromagnetic Compatibility of High Frequency Polymer-Welding Machines 617
L. Iliev, I. Stoyanov, T. Iliev, E.P. Ivanova

Energy Consumption Analysis of Real Metro-optical Network 621
J. Lorincz, E. Mujarić, D. Begušić

Doprinos informacijsko-komunikacijskih tehnologija smanjenju emisija CO_2 627
M. Berković, A. Kosovac

TELECOM INFRASTRUCTURE

A Low Energy Computer Infrastructure for Radio VOIP Supported Communication and SDR APRS in Education and Disaster Relief Situations 637
K. Hajdarević, S. Konjicija

Evaluation of Software Defined Radio Systems for Smart Home Environments 643
I. Vitas, D. Šimunić, P. Knežević

Smartwatch as a Remote Server Monitoring Device: Implementation and Interaction Design 647
A. Vega, S. Ljubić

Improving City Infrastructure Resilience Capacity: Water-supply Network Case Study 653
O. Jukić, I. Heđi, I. Špeh

Elektronička komunikacijska infrastruktura u zgradama 658
J. Bajt

ALGORITHMS & PROTOCOLS

Integrating Extreme Programming with Software Engineering Education 667
J.J.Y. Chen, M.M.Z. Wu

Sensors Definitions and Positioning for Load Balancing Algorithms 673
B. Radojević, M. Žagar, I. Golub

Design of the BCJR Decoding Algorithm with Reduced Space Complexity 678
D. Spasov

Towards a Self-organizing Sensor Network: An Algorithm 682
J. Seljan, D. Šimunić, G. Dimić, B. Vrandečić

Static Cluster and Dynamic Cluster Head (SCDCH) Adaptive Prediction-Based Algorithm for Target Tracking in Wireless Sensor Networks 686
M.A. Wahdan, M.F. Al-Mistarihi, M. Shurman

New Efficient On-Demand Multicast Routing Protocol over Mobile Ad hoc Networks 691
K.A. Darabkeh, O.A. Alsukour

Implementation of a HMIPv6 Extension in the INET and xMIPv6 Simulation Framework with Dynamic MAP Discovery for OMNeT 4.x 695
E. Kadušić, M. Škrbić, N. Živić, A. Žgalj, A. Hantalašević

NEW METHODS & SOLUTIONS

Matrix Inversion in the Design of Active Queue Management Mechanisms 703
P. Mrozowski, A. Chydzinski

An Overview of Various Types of Waveguide Grating Based Demultiplexors in WDM Systems 709
J. Tóth, Ľ. Ovseník, J. Turán

A Powerful Early-based Video Discarding Policy 714
K.A. Darabkeh, A.M. Awad, A.F. Khalifeh

Mobile Node Localization Using Infrared Angle of Arrival Sensor 719
D. Frantal, D. Arbula

A Profile- and Community-Driven Book Recommender System 724
I. Petrović, P. Perković, I. Štajduhar

Extending LARIS with Session Types: A Case Study in the Railway Interlocking ... 729
T. Kiss, F. Craciun, B. Parv

Monitoring Cerebral Processing of Gustatory Stimulation and Perception Using Emotiv Epoc 736
G. Madhale Jadav, M. Vrankić, S. Vlahinić

People Tracking Using Synthetically Generated Depth Maps 739
L. Vretenar, K. Lenac

Virtualizacija mrežnih funkcija u telekomunikacijama 745
O. Lovaković

COMPUTERS IN EDUCATION

PAPERS

The Relations Between Cultural Dimensions and Quality of Communication in Education 753
B. Grobenski, V. Vidaček-Hainš

Collaboration, Distribution and Culture – Challenges for Communication 758
H. Jaakkola, J. Henno, B. Thalheim, J. Mäkelä

Project- and Product-Oriented Course 766
J. Henno, H. Jaakkola, J. Mäkelä

Personalization of e-tivities using Web 2.0 tools and ELARS (E-learning Activities Recommender System) 770
J. Mezak, N. Hoić-Božić, M. Holenko Dlab

Model of Information System for Recognition of Non-Formal and Informal Learning in Higher Education 775
M. Balković, G. Đambić, Z. Jančić

Computer Game Angry Birds in Teaching Physics 781
D. Robič, R. Repnik

Towards Blended Learning: A Case Study 786
P. Linna, H. Keto, T. Mäkinen

The use of HILL Online e-learning Environment in Language Teaching 791
M. Linna, A. Valo

Information Literacy Assessment at the University of Zagreb: an Undergraduate Students' Perspective 796
M. Pavlovski, I. Dunđer

Tiny Operating System Kernel for Education Purposes 801
S. Ribić, A. Salihbegović

Usage of Compulsory Content in Secondary School Education 807
R. Barišić

Online Learning Efficiency in the Humanities 811
S. Gartner, M. Krašna

Designing m-learning Materials for Language Learning 816
D. Pukšič, M. Krašna

Comparison of ICT in Education Between Slovenia and Serbia 822
M. Popović, D. Cvetković, M. Krašna

The Application of Formative E-Assessment Data in the Final Exam Results Modeling Using Neural Networks 827
J. Gamulin, O. Gamulin, D. Kermek

Knowledge Generating House – a Tool for Education Based on Open Innovation Principle 832
I. Korobaničová, D. Paľová, N. Urbančíková

FASTER Platform – an On-Line Tool for EU Accountants Education 838
D. Paľová, M. Vejačka

User Knowledge Development in the Field of Electronic Services and Their Security in Slovakia 844
M. Vejačka, J. Bucko

A Course in Digital System Design using Unified E2LP Platform 850
I. Kaštelan, N. Pjevalica, M. Temerinac

Improving education experience with Augmented Reality (AR) 856
B. Kraut, J. Jeknić

Introduction of the Automated Assessment of Homework Assignments in a University-level Programming Course 862
M. Poženel, L. Fürst, V. Mahnič

Informatics Education and Requirements of Current Practice 868
L. Révészová

Integrating Learning Management System and Faculty Information System –Service Oriented Approach 874
P. Bjeljac, I. Zečević, B. Perišić

Evaluating the Effect of the Automatic Assessment Introduction in the ADS1 Course Practical Work 879
I. Rožanc, J. Mihelič

Conceptual Design, Development and Analyses of "E-Learn SEEU" Model and Framework 885
B. Fetaji, M. Fetaji, O. Rakipi, M. Ebibi

Hands-on System Programming Labs & Projects Based on E2LP Platform 891
B. Kordić, V. Marinković, M. Popović, I. Kaštelan

Wiki-based Books as Alternatives to Printed University Publications 895
M. Sakal, P. Tumbas, P. Matković, V. Pavličević

Analysis of Learning by Doing in the Creation of Instructional Video in Teaching Optics 901
A. Wirth, R. Repnik

Examples of Best Practice of University-Industry Collaboration in the Domain of Electro-CAD Application by using EPLAN P8 and E-learning Systems..... 905
D. Lukac

Self-paced Computer Aided Learning of Music Instruments 910
M. Konecki

Intelligent Tutoring Systems for Language Learning 915
V. Slavuj, B. Kovačić, I. Jugo

Exam Questions Consistency Checking 921
F. Škopljanac-Mačina, I. Zakarija, B. Blašković

An Experience in Integrating Learning Management System with User's Google Drive 925
L. Abazi-Bexheti, E. Jajaga, M. Apostolova-Trpkovska, B. Ismaili

Career and Skills Development for New Jobs 929
M. Lamza - Maronić, V. Jovanovski, J. Glavaš

The Implementation of QR Codes in the Educational Process 934
M. Filipović Tretinjak

Computers and Internet in Active Learning 937
A. Bednjanec, M. Filipović Tretinjak, M. Tretinjak

Project Teaching – is it Achievable Without the School Library? 940
D. Vincek

Benefits of Using Interactive Whiteboards at Kajzerica Elementary School 946
I. Rukljač, M. Draženović

Interactive Teaching with Socrative 949
M. Filipović Tretinjak, A. Bednjanec, M. Tretinjak

Using Cloud Computing in Stjepan Benceković Elementary School 953
B. Smolković Cerovski, M. Draženović

Project-based Learning : Students (and teachers) as Authors / co-authors of Interactive Multimedia Educational Content (Space Teacher) 956
K. Bedi

We Like Digital – Using Mobile and Web Applications in High-school Project Teaching 961
K. Bedi

Anonymity Ensurance in Creation of Satisfaction Surveys - Experience of Polytechnic of Šibenik 967
F. Urem, K. Fertalj, B. Šupe, I. Livaja

Implementation of Computer Programs in Visual Arts Education in Croatian Primary Schools 971
Z. Prohaska, S. Benković, I. Uroda

E-Learning of Students in Course „Fundamentals of Engineering Design“ Using the System of Evaluation and Maintaining the Quality of Education and Scientific Research 977
L.I. Baltova

The Developments in Mobile Learning and its Application in the Higher Education Including Libraries 982
R. Vrana

Open Science, Open Access and Open Educational Resources: Challenges and Opportunities 987
R. Vrana

The Emerging Strategy of Social Network Analysis: an Overview of Indicators, Concepts and Implications 992
A. Papić

Children's Interest in the Media - Empirical Research 997
M. Nadrljanski, V. Domitrović

Modeling and Analysis of the Polyhedral Pyramid Structures by Method of Modified Traveling Salesman Problem (TSP) 1003
S. Vukmirović, M. Čičin-Šain, I. Host

Geometric Programming in Designing of Mental Models on the Example of Strategic Thinking Between Synergies Competition and Cooperation 1009
S. Vukmirović, Z. Čapko, A. Babić

Valuation of Options by Using Excel 1016
Z. Prohaska, B. Olgić Draženović, I. Uroda

Exploring the Use of Computers in Classroom Teaching 1021
B. Kontošić, A. Kontošić, S. Biloslav

Intelligent Assistant for Helping Students to Learn Programming 1025
M. Konecki, N. Kadoić, R. Piltaver

Data Acquisition Systems - Examples of Laboratory Exercises 1030
O. Tasić, V. Tasić, D. Brodić, M. Pavlov-Kagadejev, M. Jevtić, M. Anđelić

The Gamification of Education 1034
S. Maravić Čisar, R. Pinter, P. Čisar,

Experiences and Practices in Blended Learning Environment 1039
T. Horvat, T. Alajbeg, S. Predanić

Application of ICT in Process of Informatization of Education System - Efficiency of Digital Log Book 1044
D. Glušac, I. Tasić, B. Radulović, Ž. Mušicki

Methodology for Online Study Program Development – Case Study ePITUP 1049
N. Kadoić, N. Begičević Ređep

Effective Knowledge Exchange with Modern Didactic Concepts 1055
C. Ungermanns, W. Werth

Implementation of Information Technology in Accounting Course: A Qualitative Analysis 1060
I. Bujan, A. Flac

Evaluating the Quality of Games Designed for Learning Programming by Students with Different Educational Background: An Empirical Study 1064
T. Orehovački, S. Babić

Approaches for Implementation of Distance Learning in Higher Education 1070
T.A. Hryhorova, V.P. Lyashenko, D.M. Kalinichenko

Usage of Databases and the Information Literacy of Students of the Faculty of Electrical Engineering and Computing 1075
N. Jelača

Comparing MOOCs in m-learning and e-learning Settings 1080
A. Bralić, M. Ćukušić, M. Jadrić

Mining Association Rules in Learning Management Systems 1087
M. Perušić Hrženjak, M. Matetić, M. Brkić Bakarić

Integrated Solution for On-line Environments Users' Learning and Support 1093
V. Riabtsev, K. Sinitsa, O. Voychenko

Istraživanje - mišljenje učenika, roditelja, učitelja i uprave škole o primjeni e-Dnevnika, unutar III. osnovne škole Čakovec 1097
N. Boj

Digitalne kompetencije za nastavnike 1102
V. Anđelić, H. Laušić Ramljak

Primjena internetskog foruma u prezentaciji i obrani seminarskih radova – studija slučaja kolegija Planiranje informacijskih sustava 1105
A. Sočković, A. Jakupović

Upravljanje Arduino platformom pomoću programskog jezika Processing 1110
A. Lacković

Učenje programiranja kroz robotiku 1114
P. Dobrić, M. Mačinko Kovač

Primjer implementacije e – učenja u nastavi Informatike 1117
J. Pažanin, M. Rosić, A. Grubišić

Digitalna zbirka zadataka s evaluatorom programskih rješenja 1121
V. Galešev, R. Šipek

Primjena računalne animacije u nastavi općeg obveznog tehničkog odgoja i obrazovanja 1127
D. Purković

Žene i informacijsko-komunikacijske tehnologije – zašto su studentice rijetkost na studiju računarstva? 1134
T. Babić, A. Ogrin, K. Šarkanj

Prema uspješnijem razvoju algoritamskog načina razmišljanja 1138
K. Blažeka

Kriteriji vrednovanja rada nastavnika od strane njihovih studenata 1144
M. Orlić Bachler, M. Marinović

Budućnost eUčenja 1147
M. Sertić

Analiza upotrebe programa u obrazovanju na primjeru sustava linearnih jednadžbi 1153
S. Čerkezović, H. Šišić, I. Jagodić

E-učenje podržano rješenjima u oblaku 1158
I. Livaja, F. Urem, A. Grubišić, I. Žaja, T. Radić Lakoš

Oblikovanje trodimenzionalnih objekata u sustavima virtualne stvarnosti prilagođenih učenju 1165
V. Filipašić, M. Kukec, N. Poljanec

Izborna ili redovna nastava u osnovnim školama 1170
D. Šokac, M. Antolović

E-učenje i europsko vojno obrazovanje 1172
V. Zekulić, D. Možnik

Integriranje Etwinning projekta u nastavu informatike i engleskog jezika..... 1178
B. Lamza, I. Naranđa

Stručno usavršavanje nastavnika putem e-učenja 1183
M. Mirković

Impress.js - open source prezentacijski okvir u nastavi uvoda u web tehnologije 1189
E. Ciriković, D. Suk

COMPUTERS IN TECHNICAL SYSTEMS

PAPERS

Design of a Domain Specific Language and IDE for Internet of Things Applications 1199
A. Salihbegović, T. Eterović, E. Kaljić, S. Ribić

Basic PHP Implementations, Opcodes and Internal Work 1205
A. Šimec, D. Lozić

WebRTC Technology Overview and Signaling Solution Design and Implementation 1209
B. Sredojev, D. Samardžija, D. Posarac

Methods for Geotargeting Redirection in Corporate Wide Area Networks 1213
D. Miljković

Geographically Dispersed Cluster of Web Application Servers on Virtualization Platforms 1219
D. Miljković

Geolocation and Data Type Impact on SQL Query Execution Time Using Different Database Approaches 1225
K. Šimenić, I. Lukić, M. Köhler

A Novel Versatile Architecture for Internet of Things 1229
L. Milić, L. Jelenković

Design and Development of the Software Solution for Analysis and Acquisition of the High Voltage Circuit Breakers Dynamic Resistance Measurement Results 1235
K. Obarčanin, A. Šečić, N. Hadžimejlić

Data Collector Service - Practical Approach with Embedded Linux 1240
M. Saari, P. Sillberg, P. Rantanen, J. Soini, H. Fukai

Overview of ICT Technologies for the SMILE concept of Adaptive Stimulating Environment 1245
A. Guseva, M. Gusev, S. Ristov

Android-Powered Web Server for Volunteer-Based Environment Sensing 1251
D. Škvorc, M. Oršić, S. Groš

Estimating Diver Orientation from Video Using Body Markers 1257
I. Rendulić, A. Bibulić, N. Mišković

Technologies for Distributed Flight Control Systems: a Review 1263
M. Šegvić, K. Krajček, E. Ivanjko

The Computer Architecture and Hardware Description Language 1269
A.S. Burlakov, A.E. Khmelnov

Heterogeneity Impact on MPSoC Platforms Performance 1274
N. Frid, D. Ivošević, V. Sruk

Problem-oriented on Tasks Data Mining Parallel Processor 1280
D. Adzinets, D. Lavnikevich, M. Lukashevich, Y. Seitkulov, M. Tatur

A Model of Speed Profiles for Urban Road Networks Using G-means Clustering ... 1284
T. Erdelić, S. Vrbančić, L. Rožić

Optimisation of Factory Floor Layout Using Force-directed Graph Drawing Algorithm 1290
T. Kanduč, B. Rodič

Review on Unit Commitment under Uncertainty Approaches 1296
K. Jurković, H. Pandžić, I. Kuzle

Describing Function Analysis of Systems with Fuzzy Compensated Odd Symmetric Nonlinear Actuators 1301
O. Kuljača, K. Horvat

Simple Day-Ahead Bidding Algorithm for a System with Microgrids and a Distributor 1306
L. Jelenković, T. Budrović

MIA: a Multi-Purpose Intelligent Assistant 1312
M. Konecki, M. Gams, N. Ivković

GUIDL IA: An Intelligent Assistant for Aiding Visually Impaired in Using GUIDL 1317
M. Konecki, R. Kudelić, H. Gjoreski

Anomaly Detection in Thermal Power Plant using Probabilistic Neural Network ... 1321
A. Hajdarević, I. Džananović, L. Banjanović-Mehmedović, F. Mehmedović

Neutral Grounding in Middle Voltage Power Systems: Simulation Model 1327
S. Ravlić, A. Marušić, J. Havelka

Utjecaj momenta tereta na brzinu vrtnje sinkronog elektromotornog pogona 1334
M. Pavetić, T. Bjažić

Funkcije naprednih mjernih sustava za električnu energiju 1340
Z. Lipošćak

Određivanje mjerne nesigurnosti 3D laserskog skenera 1345
N. Mijić

Analiza operativnih karakteristika javnog gradskog prijevoza temeljem prikupljanja podataka pokretnim osjetilima 1351
D. Budimir, D. Brčić, N. Jelušić

Optimizacija procesa naplate cestarine na Autocesti Rijeka-Zagreb 1357
J. Ostojić, E. Kalajdžić

Vođenje elektroenergetskih mikromreža na osnovi sinkroniziranih mjerenja..... 1363
S. Skok, M. Filipović, V. Kirinčić

INTELLIGENT SYSTEMS

PAPERS

SPECIAL SESSION ON BIOMETRICS, SOFT-BIOMETRICS AND DE-IDENTIFICATION

Binary Pattern Flavored Feature Extractors for Facial Expression Recognition: An Overview 1375
R. Lyngby Kristensen, Z.H. Tan, Z. Ma, J. Guo

Detecting Humans in Videos by Combining Heterogeneous Detectors 1382
K. Brkić, T. Hrkać, Z. Kalafatić

Identification of Authors of Documents Based on Offline Signature Recognition ... 1388
T. Marušić, Ž. Marušić, Ž. Šeremet

Comparing Affective Responses to Standardized Pictures and Videos: A Study Report 1394
M. Horvat, D. Kukolja, D. Ivanec

Retrieval of Multimedia Stimuli with Semantic and Emotional Cues: Suggestions from a Controlled Study 1399
M. Horvat, D. Kukolja, D. Ivanec

GWAT: The Geneva Affective Picture Database WordNet Annotation Tool 1403
M. Horvat, D. Duvnjak, D. Jug

Evaluation of the Quality of Voice over IP Telephony Using the Decision Support System "Qualimetric Expertise"..... 1408
R.S. Dorofeev, A.S. Dorofeev, O.A. Kostowskaya

INTELLIGENT SYSTEMS

A New Ensemble of Features for Breast Cancer Diagnosis..... 1415
İ. Işıklı Esener, S. Ergin, T. Yüksel

Preliminary Study on Detection of Atrial Fibrillation during Oscillometric Blood Pressure Measurement 1421
M. Meža

Hybride Neuro-fuzzy Expert System for Assessing Diabetes Risk 1426
Ž. Knok, Z. Avdagić, S. Omanović

Solving Strip-Packing Problem Using Sequence Pair 1430
D. Domović, T. Rolich

Predicting Stock Market Trends Using Random Forests: A Sample of the Zagreb Stock Exchange 1436
T. Manojlović, I. Štajduhar

GA-ANFIS Expert System Prototype for Detection of Tar Content in the Manufacturing Process 1441
L. Begić Fazlić, Z. Avdagić, I. Bešić

A Review of Feature Selection Methods with Applications 1447
A. Jović, K. Brkić, N. Bogunović

Using Similarity-based Selection in Evolutionary Design of Decision Trees 1453
L. Bošnjak, S. Karakatič, V. Podgorelec

A Semantic Markup Technique Based on Ontology Polysystem 1459
E. Cherkashin, K. Paskal, I. Bychkov

A Descriptive Specification Tools for Information System Design and Configuration 1465
A. Hmelnov, E. Fereferov, R. Fedorov, E. Cherkashin, I. Bychkov

Importance of Network Nodes for Navigation with Fractional Knowledge 1471
D. Helic, F. Geigl

Using Latent Features to Measure the Diversity of Recommendation Lists 1477
M. Kunaver, Š. Dobravec, A. Košir

Efficient Profiling and Distributed Synchronization of Statistical User Models for Content-Targeting and Real-time Applications 1482
Lj. Mercep

Text Skew Estimation Based on the Horizontal Entropy Calculation 1488
D. Brodić, M. Jevtić, Z.N. Milivojević, V. Tasić

Optimization of Ring-Star Transmission Problem in Telecommunication Systems Based on Ant Colony Algorithms 1492
V. Letić, Z. Avdagić, D. Bošković

Using String Similarity Metrics for Automated Grading of SQL Statements 1497
I. Štajduhar, G. Mauša

Multi-agent Modeling Methods for Massivley Multi-Player On-Line Role-Playing Games 1503
M. Schatten, I. Tomičić, B. Okreša Đurić

Temporal Dynamics of Changes in Group User's Preferences in Recommender Systems 1509
B. Karahodža, Dž. Đonko, H. Šupić

Predicting Sports Results Using Latent Features: A Case Study 1514
Š. Dobravec

Role-Playing as a Knowledge Management Tool 1520
L. Brajković, W. Afrić, V. Afrić

Nadogradnja inteligentnog sustava aerodromskog održavanja serijskih strujnih krugova 1526
Z. Balaž, I. Puhalović, G. Bjelić

INFORMATION SYSTEMS SECURITY

PAPERS

TECHNICAL TRACK

Disabling Same Origin Policy for Automatization of Webform Data Entry 1539
S. Ribić, A. Salihbegović, A. Huseinović

Towards a Rule-based Access Control Framework for Distributed Information Systems 1544
S. Afonin

Improving the Android Smartphone Security Against Various Malware Threats ... 1550
D. Hrestak, S. Picek, Ž. Rumenjak

Detection of DECT Identity Spoofing through Radio Frequency fingerprinting 1556
I. Sanchez, R. Satta, R. Giuliani, G. Baldini

Study of Two Security Constructs on Throughput for Wireless Sensor Multi-Hop Networks 1562
R.D. Sparrow, A.A. Adekunle, R.J. Berry, R.J. Farnish

A Practical NFC Relay Attack on Mobile Devices Using Card Emulation Mode 1568
D. Cavdar, E. Tomur

Honeytokens as Active Defense 1573
R. Petrunić

POLICY AND RISK TRACK

Artificial Intelligence in Network Intrusion Detection 1581
M. Štampar, K. Fertalj

Advanced Persistent Threats – Detection and Defense 1587
J. Vukalović, D. Delija

Security Risk Management in Complex Organization 1594
I. Sedinić, T. Perušić

PRIVACY TRACK

Violation of User Privacy by IPTV Packet Sniffing in Home Network 1603
S. Vukojević

Have we been Monetized and become Commodity without our Consent – Privacy in the Time of Big Data Technology 1609
M. Žagar, D. Mendeš Poljak

SECURITY AWARENESS AND EDUCATION TRACK

Impact of Security Education on Password Change 1617
V. Taneski, M. Heričko, B. Brumen

Empirical Study on ICT System's Users' Risky Behavior and Security Awareness 1623
K. Šolić, T. Velki, T. Galba

Moore's Curse on Textual Passwords 1627
B. Brumen, V. Taneski

DIGITAL FORENSICS TRACK

Concepts and Methodology in Mobile Devices Digital Forensics Education and Training 1635
D. Delija

RAM Data Significance in Digital Forensics 1641
K. Hausknecht, D. Foit, J. Burić

Competencies and Skills Needed for Digital Forensic Trainer 1645
D. Foit, J. Vukalović, K. Hausknecht

Challenges in Network Forensics 1651
J. Burić, D. Delija

An Approach to Digital Evidence Collection for Successful Forensic Application: An Investigation of Blackmail Case 1656
K. Hajdarević, V. Džaltur

BUSINESS INTELLIGENCE SYSTEMS

INVITED PAPER

Spatial OLAP for Agri-Environmental Data and Analysis: Lessons Learned 1667
S. Bimonte

PAPERS

Accelerating Time Series Subsequence Matching on the Intel Xeon Phi Many-core Coprocessor 1675
R. Miniakhmetov, A. Movchan, M. Zymbler

Outlier Detection as the Primary Step for Promotion Planning in Retail 1681
M. Banek, D. Osrečki, M. Vranić, D. Pintar

Data Warehouse in Era of Big Data Platforms 1687
T. Križan, M. Družetić

Delta View Generation for Incremental Loading of Large Dimensions in a Data Warehouse 1693
I. Mekterović, Lj. Brkić

SimBusPro: A Simulation-Based Decision Support Tool used for the Optimization of Business Processes running on the Cloud 1699
A. Özkök, A. Ercingöz, H.A. Dönmez, T. Dökeroğlu, V. İşler

System for Monitoring and Advanced Analysis of Handball Matches 1704
T. Hlupić, F. Jandrijević, J. Kovačev, L. Petricioli, T. Gracin, M. Baranović

Planiranje i nabava energije za pokriće gubitaka u distribucijskoj mreži 1710
I. Žarkić, Z. Lipošćak, M. Malovrh

Automatska monotona kategorizacija numeričkih varijabli 1715
V. Horvat, K. Bokulić

DIGITAL ECONOMY AND GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

INVITED PAPER

Public e-Procurement as a Key-Impact Factor in the UE Welfare Contest -The Italian Public e-Markets: Research Findings and e-Solutions Analisys..... 1725
F. Bof, G. Casella

PAPERS

Microeconomic and Macroeconomic Overview of ICT as a Primary Driver behind the Digital Economy 1738
S. Aksentijević, E. Tijan, G. Stumpf

Application of Service Oriented Architecture of Information Communication System in Nautical Ports 1744
P. Ristov, A. Mrvica, J. Žanić Mikuličić

Internet Challenges for Print Media Advertising Practice – Case of Belgrade Daily Newspapers 1751
M. Vidas-Bubanja, I. Bubanja

The Success of Open Source Software: A Review 1757
D. Margan, S. Čandrlić

Analysis of e-government Portal in Montenegro 1763
V. Lazović, T. Đuričković, B. Rondović, D. Kovačević, Z. Popović

Determinants of Electronic Commerce - the Factors of Acceptance in Montenegro 1769
B. Rondović, V. Lazović, Z. Popović, T. Đuričković, D. Kovačević

Laws of Network's Values and Values of Its Users 1775
E. Pirić, M. Bajrić, H. Bajrić

Open Issues of Open Data in Croatia 1780
D. Bužić, J. Đurić

Cloud Services and Marketing 1786
J. Jeknić, B. Kraut

RFID Based Access Control System 1793
P. Ristov, T. Mišković, A. Mrvica

Interoperablity of Bank Statements: A Case Study 1799
R. Chejkova-Nikolov, M. Gusev, M. Kostoska, S. Ristov

Social Networks in the Case of Tourism in Croatia 1805
Lj. Zekanović-Korona

Financial Agency FINA and Online Auctions in Croatian Amended Distraint Act .. 1811
A.J. Klasinc

Scalable Architecture of e-Ordering System in Cloud 1817
F. Dimitrievski, S. Ristov, M. Gusev, G. Armenski

The Potential of SEPA Credit Transfer Implementation in Croatia 1823
M. Ptiček, B. Vrdoljak, L. Humski, Z. Skočir, G. Bolanča, Ž. Gašparić

Business Process Orientation Construct Analysis in Companies in the Republic of Macedonia 1829
M. Mijoska, N. Levkov

The Role of the Public Sector in Defining the Communication Strategies in the Croatian Tourism Industry 1835
Đ. Bunja

Accounting Information Systems and their use in Local Governments Sector: Quality, Efficiency, Security and Control Procedures as (present) Challenges 1840
M. Boban, T. Šušak

Citybus Mobile Application 1846
J. Abou Aldan, M. Lončar, M. Ivašić-Kos

Analysis and Recommendations for e-business Development in Croatia 1851
R. Smokvina, S. Aksentijević, E. Tijan

Analiza sadržaja internetskih stranica županija, gradova i općina u Republici Hrvatskoj u prvoj polovici 2014. god. 1857
T. Lamut, A. Sarić

Implementacija informacijskog sustava za praćenje aktivnosti unutar jedinice javne uprave 1863
O. Jukić, I. Heđi, I. Špeh

Analiza sustav za nadzor potrošnje posebnih goriva 1868
S. Bojko, V. Kumpar-Tasevski, M. Mateljan, A. Topić

Social Media Marketing Activity in Western Balkan Banking Industry..... 1874
N. Levkov, M. Mijoska, D. Jovevski

MIPRO Junior – STUDENT PAPERS

PAPERS

Optical Add/Drop Multiplexers for All Optical Fiber Communication 1885
J. Ruzbarsky, J. Turan, Ľ. Ovseník

Korelacijska analiza dotoka u akumulacijske bazene 1890
E. Sutlović, K. Zoričić, T. Modrić

PLC/SCADA Systems in Automation Control Design for Individual Quick Freezing Process in Cooling Tunnels 1896
G. Jagetić, M. Habazin, T. Špoljarić

Sustav grijanja pogona za sušenje papira 1904
R. Bukvić, G. Malčić, D. Maršić

Uređaj za testiranje baterijskih ćelija 18650 1910
A. Jakoplić, D. Maršić, G. Malčić

Layout Optimizations to Decrease Internal Power and Area in Digital CMOS Standard Cells 1915
J. Innocenti, F. Julien, J.M. Portal, L. Lopez, Q. Hubert, P. Masson, J. Sonzogni, S. Niel, A. Regnier

A Simple Clock-Fault Detection Analog Circuit for High-Speed Crystal Oscillators 1921
A. Gamet, Y. Bacher, S. Meillère, P. Le Fevre, N. Froidevaux

Performance Analysis of Parallel Reduction and Matrix Multiplication on Parallella-16 Heterogeneous System 1925
N. Ćatić, K. Imamović, E. Jašarević, N. Nosović

Building Network-enabled Smart Sensors and Actuators 1930
I. Pavić, J. Puškar, I. Soldo, I. Spasić, H. Džapo, D. Ćika

Turning a Smartphone into Structured Light 3D Scanner 1935
M. Lelas, T. Pribanić

A Heuristics Approach to Hamiltonian Completion Problem (HCP) 1940
H. Petrić Maretić, A. Grbić

Security of End User Wireless Networks in Zagreb Area 1946
J. Redžepagić, D. Studen, V. Gavranić, A. Teković

Learning from Student Thesis Products: Examples and Experiences 1950
M. Vučićević, S. Predanić, T. Horvat

Extracting Domain Knowledge by Complex Networks Analysis of Wikipedia Entries 1955
N. Matas, S. Martinčić-Ipšić, A. Meštrović

LaNCoA: A Python Toolkit for Language Networks Construction and Analysis..... 1961
D. Margan, A. Meštrović

Parallelization of Two Popular Parallel Sorting Algorithms on GPU Using OpenCL 1967
D. Azinović, A. Hasanspahić, H. Hasić, N. Nosović

Mobile Application for Finding ATMs 1971
M. Lončar, J. Abou Aldan, M. Ivašić-Kos

LIST OF PAPER REVIEWERS

Abello Gamazo, A. (Spain)
Afgan, E. (United States)
Aksentijević, S. (Croatia)
Aleksić, S. (Croatia)
Amon, S. (Slovenia)
Andročec, D. (Croatia)
Aramo-Immonen, H. (Finland)
Arbula, D. (Croatia)
Ašenbrener Katić, M. (Croatia)
Avbelj, V. (Slovenia)
Babić, D. (Croatia)
Babić, S. (Croatia)
Bača, M. (Croatia)
Banek, M. (Croatia)
Banek Zorica, M. (Croatia)
Baotić, M. (Croatia)
Baranović, M. (Croatia)
Bartolić, J. (Croatia)
Basch, D. (Croatia)
Bažant, A. (Croatia)
Bellatreche, L. (France)
Bogdan, S. (Croatia)
Bojić, I. (Croatia)
Britvić, I. (Croatia)
Brkić, L. (Croatia)
Brkić, M. (Croatia)
Budin, A. (Croatia)
Budin, L. (Croatia)
Budiselić, I. (Croatia)
Bujan, I. (Croatia)
Buković, M. (Croatia)
Burić, J. (Croatia)
Butković, Ž. (Croatia)
Butorac, M. (Croatia)
Bužić, D. (Croatia)
Car, Ž. (Croatia)
Cihlar, D. (Croatia)
Črnjaric-Žic, N. (Croatia)
Čandrlić, S. (Croatia)
Čapko, Z. (Croatia)
Čereković, A. (Croatia)
Čičin-Šain, M. (Croatia)
Čubrilo, M. (Croatia)
Čupić, M. (Croatia)
Davidović, D. (Croatia)
Davidović, M. (Croatia)
Delač, G. (Croatia)
Denona-Bogović, N. (Croatia)
Depolli, M. (Slovenia)
Dešić, S. (Croatia)
Dobrijević, O. (Croatia)
Dobrišek, S. (Slovenia)
Džanko, M. (Croatia)
Džapo, H. (Croatia)
Đerek, V. (Croatia)
Erceg, I. (Croatia)
Filjar, R. (Croatia)
Fischer, D. (Croatia)
Galetić, V. (Croatia)
Galinac Grbac, T. (Croatia)
Gamberger, D. (Croatia)
Gamulin, O. (Croatia)
Garza, P. (Italy)
Gerić, S. (Croatia)
Glavaš, G. (Croatia)
Glavaš, J. (Croatia)
Glušac, D. (Serbia)
Golfarelli, M. (Italy)
Golub, M. (Croatia)
Golubić, S. (Croatia)
Granić, A. (Croatia)
Grgić, K. (Croatia)
Groš, S. (Croatia)
Grubišić, D. (United States)
Grubišić, I. (Croatia)
Grudenić, I. (Croatia)
Gulić, M. (Croatia)
Gulin, M. (Croatia)
Guna, J. (Slovenia)
Hajdarević, K. (Bosnia and Herzegovina)
Heimburger, A. (Finland)
Henno, J. (Estonia)
Hoić-Božić, N. (Croatia)
Holenko Dlab, M. (Croatia)
Horvat, G. (Croatia)
Horvat, M. (Croatia)
Hrabar, S. (Croatia)
Hrkać, T. (Croatia)
Huljenić, D. (Croatia)
Humski, L. (Croatia)
Ilić, Ž. (Croatia)
Inkret, R. (Croatia)
Ipšić, I. (Croatia)
Ivašić-Kos, M. (Croatia)
Ivković, N. (Croatia)
Ivošević, D. (Croatia)
Jakobović, D. (Croatia)
Jakupović, A. (Croatia)
Jardas, M. (Croatia)
Jelenković, L. (Croatia)
Jevtić, D. (Croatia)
Ježić, Z. (Croatia)
Ježić, G. (Croatia)
Job, J. (Croatia)
Jovanovic, V. (Netherlands)
Jović, A. (Croatia)
Jugović, A. (Croatia)
Juričić, V. (Croatia)
Jurišić, D. (Croatia)
Kalafatić, Z. (Croatia)
Kalpić, D. (Croatia)
Kaštelan, I. (Serbia)
Kaučič, B. (Slovenia)
Kecskeméti, G. (Hungary)
Keto, H. (Croatia)
Kirinić, V. (Croatia)
Kišasondi, T. (Croatia)
Kopčak, G. (Sweden)

Koričić, M. (Croatia)
Kosec, G. (Slovenia)
Kovač, M. (Croatia)
Kovačić, B. (Croatia)
Krajcar, S. (Croatia)
Krašna, M. (Slovenia)
Krevatin, I. (Croatia)
Križanović, K. (Croatia)
Krois, I. (Croatia)
Krznarić, M. (Croatia)
Kunaver, M. (Slovenia)
Kušek, M. (Croatia)
Lähdeniemi, M. (Finland)
Leino, M. (Finland)
Lenac, K. (Croatia)
Leppäniemi, J. (Finland)
Lerga, J. (Croatia)
Lešić, V. (Croatia)
Lipić, T. (Croatia)
Lučić, D. (Croatia)
Lugarić, T. (Croatia)
Lukac, D. (Germany)
Lutovac, B. (Montenegro)
Ma, Z. (China)
Magdalenić, I. (Croatia)
Mäkinen, T. (Finland)
Malarić, R. (Croatia)
Mance, D. (Croatia)
Mandić, T. (Croatia)
Maračić, M. (Croatia)
Marčetić, D. (Croatia)
Marinović, M. (Croatia)
Marjanović, T. (Croatia)
Marković, I. (Croatia)
Martinčić-Ipšić, S. (Croatia)
Mateljan, V. (Croatia)
Matetić, M. (Croatia)
Matijašević, M. (Croatia)
Mekovec, R. (Croatia)
Mezak, J. (Croatia)
Mihajlović, Ž. (Croatia)
Mikac, B. (Croatia)
Mikuc, M. (Croatia)
Milanović, I. (Serbia)
Milašinović, B. (Croatia)
Milišić, P. (Croatia)
Mišković, N. (Croatia)
Mitsa, V. (Ukraine)
Mlinarić, H. (Croatia)
Moguš-Milanković, A. (Croatia)
Mrković, B. (Croatia)
Nađ, Đ. (Croatia)
Nikitović, M. (Croatia)
Očko, M. (Croatia)
Palomäki, J. (Finland)
Pavešić, N. (Slovenia)
Pavlić, Z. (Croatia)
Pečar-Ilić, J. (Croatia)
Pejčinović, B. (United States)
Pelin, D. (Croatia)
Perić Hadžić, A. (Croatia)
Perttula, A. (Croatia)
Petošić, A. (Croatia)
Petrović, T. (Croatia)
Picek, S. (Croatia)
Pintar, D. (Croatia)
Pipić, N. (Croatia)
Pleslić, S. (Croatia)
Pobar, M. (Croatia)
Poljak, M. (Croatia)
Popović, M. (Serbia)
Požgaj, Ž. (Croatia)
Pripužić, K. (Croatia)
Radić, J. (Croatia)
Radić, N. (Croatia)
Rendulić, I. (Croatia)
Repnik, R. (Slovenia)
Resnik, D. (Slovenia)
Ribarić, S. (Croatia)
Rupčić, S. (Croatia)
Schatten, M. (Croatia)
Skala, K. (Croatia)
Skočir, Z. (Croatia)
Skorin-Kapov, L. (Croatia)
Slapničar, I. (Croatia)
Sluganović, I. (Croatia)
Sokele, M. (Croatia)
Sović, I. (Croatia)
Sprčić, V. (Croatia)
Sruk, V. (Croatia)
Stanič, U. (Slovenia)
Stapić, Z. (Croatia)
Staroniewicz, P. (Poland)
Stipčević, M. (Croatia)
Stupar, I. (Croatia)
Subašić, M. (Croatia)
Šupić, H. (Bosnia and Herzegovina)
Sužnjević, M. (Croatia)
Šarić, D. (Croatia)
Šarolić, A. (Croatia)
Šegvić, S. (Croatia)
Šikić, M. (Croatia)
Šimić, Z. (Croatia)
Šimunić, D. (Croatia)
Šipuš, Z. (Croatia)
Škvorc, D. (Croatia)
Štriga, D. (Croatia)
Štruc, V. (Slovenia)
Šunde, V. (Croatia)
Švedek, T. (Croatia)
Švogor, I. (Croatia)
Terzić, B. (Croatia)
Tomašić, I. (Slovenia)
Tomić, M. (Croatia)
Topič, M. (Slovenia)
Trobec, R. (Slovenia)
Trost, A. (Slovenia)
Tržec, K. (Croatia)
Tuomi, P. (Finland)
Turan, J. (Slovakia)
Varga, M. (Croatia)
Vasilijević, A. (Croatia)
Vidaček-Hainš, V. (Croatia)
Vinko, D. (Croatia)
Vladimir, K. (Croatia)
Vojković, G. (Croatia)

Vrančić, K.	(Croatia)	Vukmirović, S.	(Croatia)
Vranić, M.	(Croatia)	Vuković, M.	(Croatia)
Vrdoljak, B.	(Croatia)	Zonja, S.	(Croatia)
Vrtačnik, D.	(Slovenia)	Zulim, I.	(Croatia)
Vučić, M.	(Croatia)	Žgaljić Keko, A.	(Croatia)

AUTHOR INDEX

Abazi-Bexheti L. 925
Abbad A. 559
Abou Aldan J. 1846, 1971
Adamović N. 137
Adekunle A.A. 1562
Adnan M. 133
Adzinets D. 1280
Afgan E. 185, 236
Afonin S. 1544
Afrić V. 1520
Afrić W. 1520
Aksentijević S. 1738, 1851
Alajbeg T. 1039
Aleksić S. 601
Alhulayil M. 584, 589
Al-Khasawneh M.M. 580
Al-Mistarihi M.F. 569, 574, 580, 584, 589, 686
Alsukour O.A. 691
Anđelić M. 1030
Anđelić V. 1102
Antolović M. 1170
Apostolova-Trpkovska M. 925
Arbula D. 719
Armenski G. 1817
Artemkina S.B. 21
Avbelj V. 375, 389
Avdagić Z. 1426, 1441, 1492
Awad A.M. 714
Azinović D. 1967
Babic D. 80
Babić A. 1009
Babić D. 85
Babić J. 121
Babić S. 1064
Babić T. 1134
Bacher Y. 1921
Bačmaga J. 104
Bajrić H. 1775
Bajrić M. 1775
Bajt J. 658
Bako N. 68, 74
Balaž Z. 1526
Baldini G. 1556
Balković M. 775
Balota A. 417
Baltova L.I. 977
Bánáti A. 254
Banek M. 1681
Banka A.A. 201
Banović-Ćurguz N. 429, 494
Banjanović-Mehmedović L. 1321
Baranović M. 1704
Barić A. 68, 74, 89, 93, 97, 104
Barišić R. 807
Bašić J. 489
Bećirspahić L. 321
Bedi K. 956, 961
Bednjanec A. 937, 949
Begičević Ređep N. 1049
Begić E. 379
Begić Fazlić L. 1441
Begić Z. 379
Begušić D. 621
Bekenov L. 191
Benković S. 971
Berdinsky A.S. 18, 21
Berković I. 359
Berković M. 627
Berry R.J. 1562
Bešić I. 1441
Bibulić A. 1257
Bilajbegović D. 465
Biloslav S. 1021
Bimonte S. 1667
Bjažić T. 1334
Bjelić G. 1526
Bjeljac P. 874
Blašković B. 921
Blažeka K. 1138
Blečić R. 97
Boban M. 1840
Bodenstein C. 217
Bof F. 1725
Bogdanov P. 344
Bogunović N. 1447
Boj N. 1097
Bojko S. 1868
Bokulić K. 1715
Bolanča G. 1823
Bosiljevac M. 80, 85
Bošković D. 1492, 379
Bošnjak L. 1453
Božinović Ž. 553
Brajković L. 1520
Brakus M. 305
Bralić A. 1080
Brančík L. 161
Brčić D. 1351
Brenner W. 137
Brkić Bakarić M. 1087
Brkić K. 1382, 1447
Brkić Lj. 1693
Brložnik M. 375
Brodić D. 1030, 1488
Broz I. 74
Brumen B. 1617, 1627
Bubanja I. 1751
Bucko J. 844
Budić D. 612
Budimir D. 1351
Budrović T. 1306
Bujan I. 1060
Bukvić R. 1904
Bunja Đ. 1835
Burić J. 1641, 1651
Burlakov A.S. 1269
Butković Ž. 68, 74
Bužić D. 1780

Bychkov I. 1459, 1465
Caba C.M. 507
Caruso E. 5
Casella G. 1725
Cavdar D. 1568
Chejkova-Nikolov R. 1799
Chen J.J.Y. 667
Cherkashin E. 1459, 1465
Chorbev I. 299
Chydzinski A. 703
Cigale B. 404
Ciriković E. 1189
Civadelić Z. 472
Craciun F. 729
Cvetkov K. 212
Cvetković D. 822
Čačković V. 439
Čandrlić S. 1757
Čapko Z. 1009
Čerkezović S. 1153
Čibej U. 398
Čičin-Šain M. 1003
Činikas I. 167
Čisar P. 1034
Ćatić N. 1925
Ćika D. 143, 1930
Ćukušić M. 1080
Dadić V. 489
Dalbelo Bašić B. 513
D'Andrea C. 15
Darabkeh K.A. 559, 691, 714
Davidović D. 185
Dedmari M.A. 201
Delija D. 1587, 1635, 1651
Dika F. 272
Dimić G. 597, 682
Dimitrievski F. 1817
Dobravec Š. 1477, 1514
Dobrić P. 1114
Dökeroğlu T. 1699
Domitrović V. 997
Domović D. 1430
Domšić J. 149
Donevski A. 294
Dönmez H.A. 1699
Dorofeev A.S. 1408
Dorofeev R.S. 1408
Draženović M. 946, 953
Driussi F. 5
Družetić M. 1687
Dujak M. 472
Dunđer I. 796
Duvnjak D. 1403
Džaltur V. 1656
Džananović I. 1321
Džapo H. 143, 149, 1930
Džebo A. 445
Đambić G. 775
Đerek V. 15
Đonko Dž. 1509
Đonlić M. 155
Đuričković T. 1763, 1769
Đurić J. 1780
Ebibi M. 885
Ercingöz A. 1699
Erdelić T. 1284
Ergin S. 1415
Esseni D. 5
Eterović T. 1199
Fajdetic F. 93
Farnish R.J. 1562
Fedak G. 206
Fedorov R. 1465
Fedorov V.E. 18, 21
Fehér G. 344
Fereferov E. 1465
Ferreira M. 392
Fertalj K. 967, 1581
Fetaji B. 885
Fetaji M. 885
Filipašić V. 1165
Filipović M. 1363
Filipović Tretinjak M. 934, 937, 949
Flac A. 1060
Foit D. 1641, 1645
Forer L. 185
Frantal D. 719
Frid N. 1274
Froidevaux N. 1921
Fukai H. 1240
Fürst L. 862
Galac M. 520
Galba T. 1623
Galešev V. 1121
Galinac Grbac T. 507, 513, 528
Gamet A. 1921
Gams M. 1312
Gamulin J. 827
Gamulin O. 827
Gartner S. 811
Gašparić Ž. 1823
Gatsenko O. 191
Gavranić V. 1946
Gebavi H. 15
Geigl F. 1471
Gillon R. 97
Giuliani R. 1556
Gjoreski H. 1317
Glavaš J. 929
Gligoroski D. 483
Glock P. 217
Glušac D. 1044
Goetz M. 217
Golub I. 673
Goonasekera N. 236
Goran N. 468
Gordienko N. 206
Gordienko Y. 191, 206
Gotić M. 15
Gracin T. 1704
Grbić A. 1940
Grimshaw A. 282
Grobenski B. 753
Groš S. 1251
Grubišić A. 1117, 1158
Grubišić D. 43

Gucciardi P.G. 15
Gudelj M. 564
Gulan G. 368
Guo J. 1375
Gusev M. 1245, 212, 258, 267, 276, 294, 299, 1799, 1817
Guseva A. 1245
Habazin M. 1896
Hadžiahmetović N. 553
Hadžialić M. 468
Hadžimejlić N. 1235
Hadžović S. 435
Hajdarević A. 1321
Hajdarević K. 637, 1656
Hammerschmidt D. 133
Hantalašević A. 695.
Hasanspahić A. 1967
Hasić H. 1967
Hausknecht K. 1641, 1645
Havelka J. 1327
Heđi I. 653, 1863
Helic D. 1471
Henno J. 758, 766
Heričko M. 1617
Hlupić T. 1704
Hmelnov A. 1465
Hodanić D. 541
Hoić-Božić N. 770
Holenko Dlab M. 770
Hormot F. 104
Horvat K. 1301
Horvat M. 1394, 1399, 1403
Horvat T. 1039, 1950
Horvat V. 1715
Host I. 1003
Hrestak D. 1550
Hrkać T. 1382
Hryhorova T.A. 1070
Hubert Q. 1915
Huemer M. 133
Humski L. 1823
Huseinović A. 1539
Igrec B. 85
Iliev L. 617
Iliev T. 479, 617
Ilievska N. 483
Ilišević D. 429, 494
Imamović K. 1925
Innocenti J. 1915
Işıklı Esener İ. 1415
İşler V. 1699
Ismaili B. 925
Ivančić I. 340
Ivančić M. 340
Ivanda M. 15
Ivanec D. 1394, 1399
Ivanković D. 489
Ivanova E. 223
Ivanova E.P. 479, 617
Ivanovska B. 258
Ivanjko E. 1263
Ivašić-Kos M. 1846, 1971
Ivković N. 1312
Ivošević D. 1274
Jaakkola H. 758, 766
Jadrić M. 1080
Jagetić G. 1896
Jagodić I. 1153
Jajaga E. 925
Jakoplić A. 1910
Jakopović Ž. 116, 121
Jakupović A. 1105
Jančić Z. 775
Jandrijević F. 1704
Janeković I. 43
Jašarević E. 1925
Jeknić J. 856, 1786
Jelača N. 1075
Jelavić D. 489
Jelenković L. 1229, 1306
Jelušić N. 1351
Jeričević Ž. 227, 368
Jevtić M. 1030, 1488
Ježić G. 439
Jovanovski V. 929
Jovevski D. 1874
Jović A. 1447
Jug D. 1403
Jugo I. 915
Juhasz Z. 315
Jukić O. 653, 1863
Julien F. 1915
Jurković K. 1296
Kacsuk P. 254, 263
Kadoić N. 1025, 1049
Kadušić E. 695
Kail E. 263
Kalafatić Z. 1382
Kalajdžić E. 1357
Kalinichenko D.M. 1070
Kalinić H. 489
Kaljić E. 1199
Kanduč T. 1290
Karabegović A. 321
Karahodža B. 1509
Karakatič S. 1453
Kaštelan I. 850, 891
Kermek D. 827
Keto H. 786
Khabarov S.P. 18
Khalifeh A.F. 559, 714
Khmelnov A.E. 1269
Kilamo T. 242
Killat D. 56
Kirinčić V. 1363
Kiss T. 729
Klasinc A.J. 1811
Knežević P. 643
Knežević T. 43
Knok Ž. 1426
Koeritz C. 282
Köhler M. 1225
Kolić K. 276
Konecki M. 1025, 1312, 1317
Konecki M. 910
Kontošić A. 1021
Kontošić B. 1021

Konjicija S. 637
Kordić B. 891
Koričić M. 37
Korobaničová I. 832
Kosanović T. 537
Kosovac A. 627
Kostadinov H. 349
Kostoska M. 258, 294, 1799
Kostowskaya O.A. 1408
Košir A. 1477
Kovačev J. 1704
Kovačević D. 1763, 1769
Kovačević S. 435
Kovačić B. 915
Kovačić M. 109
Kozlovszky M. 254, 263, 344
Kozmann G. 315
Kožar I. 227, 368
Krajček K. 1263
Kraljević Pavelić S. 528
Krampis K. 236
Kranjac M. 353
Kranjec J. 404
Kranjec P. 404
Krašna M. 811, 816, 822
Kraut B. 1786
Kraut B. 856
Krevatin I. 564
Krivec S. 31
Krivošić A. 445
Križan T. 1687, 305
Kronenberg F. 185
Kruus M. 172
Kudelić R. 1317
Kukec M. 1165
Kukolja D. 1394, 1399
Kuljača O. 1301
Kumpar-Tasevski V. 1868
Kunaver M. 1477
Kundrata J. 89
Kušek M. 439
Kuzle I. 1296
Kuznetsov V.A. 21
Lacković A. 1110
Lamut T. 1857
Lamza - Maronić M. 929
Lamza B. 1178
Laušić Ramljak H. 1102
Lavnikevich D. 1280
Lazović V. 1763, 1769
Le Fevre P. 1921
Ledneva A.Y. 21
Lelas M. 1935
Lenac K. 739
Lenac L. 368
Letić D. 359
Letić V. 1492
Levkov N. 1829, 1874
Lindén M. 384
Linna M. 791
Linna P. 786
Lipošćak Z. 1710, 1340
Livaja I. 967, 1158
Lizzit D. 5
Lodygensky O. 206
Lojk J. 398
Lončar M. 1846, 1971
Lopez L. 1915
Lorincz J. 621
Lovaković O. 745
Lozić D. 1205
Lučkin E. 499
Lučkin H. 499
Lukac D. 905
Lukashevich M. 1280
Lukić I. 1225
Lyashenko V.P. 1070
Lyngby Kristensen R. 1375
Ljubić S. 647
Ma Z. 1375
Mačinko Kovač M. 1114
Madhale Jadav G. 736
Magableh A.M. 569, 574, 580
Mahnič V. 862
Mäkelä J. 758, 766
Mäkinen T. 786
Malbaša V. 62
Malčić G. 1904, 1910
Malovrh M. 1710
Mandić T. 93
Mandžuka M. 379
Manev N.L. 349
Manojlović T. 1436
Maras M. 537
Marasović S. 489
Maravić Čisar S. 1034
Margan D. 1757, 1961
Marinković V. 891
Marinović M. 1144
Markoska E. 299
Markulin D. 450
Maršić D. 1904, 1910
Martinčić-Ipšić S. 1955
Martinović T. 143
Martinović Ž. 612
Marušić A. 1327
Marušić T. 1388
Marušić Ž. 1388
Masoodi M.T. 201
Masson P. 1915
Matas N. 1955
Mateljan M. 1868
Matetić M. 1087
Matković P. 895
Mauša G. 1497, 513, 528
Medved Rogina B. 288
Mehmedović F. 1321
Meillère S. 1921
Mekterović I. 1693
Memon A.S. 217
Memon M.S. 217
Memon S. 282
Mendeš Poljak D. 1609
Mercep Lj. 1482
Meštrović A. 1955, 1961
Mezak J. 770

Meža M. 1421
Mihajlović Ž. 334, 340
Mihaldinec H. 149
Mihaylov G. 479
Mihelič J. 879
Mijić N. 1345
Mijoska M. 1829, 1874
Mikac L. 15
Mikic M. 93
Miletić P. 472
Milić L. 1229
Milivojević Z.N. 1488
Miljković D. 1213, 1219
Miniakhmetov R. 1675
Mirković M. 1183
Mišković N. 1257
Mišković T. 1793
Modrić T. 1890
Mohaisen R. 569, 574
Movchan A. 1675
Možnik D. 1172
Mrazović P. 334
Mrčela I. 109
Mrozowski P. 703
Mrvica A. 1744, 1793
Mujan V. 601
Mujarić E. 621
Musa K. 450
Muslim S. 489
Musulin I. 121
Mušicki Ž. 1044
Nadrljanski M. 997
Naranđa I. 1178
Nauwelaers B. 97
Niel S. 1915
Nosović N. 1925, 1967
Novak F. 409
Novak M. 116
Obarčanin K. 1235
Oder A. 344
Ogrin A. 1134
Okanović V. 364
Okreša Đurić B. 1503
Olgić Draženović B. 1016
Omanović S. 1426
Opiła J. 328
Orehovački T. 1064
Orlić Bachler M. 1144
Oršić M. 1251
Osgnach P. 5
Osrečki D. 1681
Ostojić J. 1357
Otsetova E. 479
Ovseník Ľ. 709, 1885
Özkök A. 1699
Pajkanović A. 62
Palestri P. 5
Paľová D. 832, 838
Pandžić H. 1296
Papić A. 992
Parv B. 729
Pashmineh S. 56
Paskal K. 1459
Pavelić M. 439
Pavetić M. 1334
Pavić I. 1930
Pavić T. 546
Pavković I. 459
Pavličević V. 895
Pavlin M. 398
Pavlin M. 409
Pavlov-Kagadejev M. 1030
Pavlovski M. 796
Pažanin J. 1117
Pejić P. 149
Pejović P.V. 607
Perišić B. 874
Perković P. 724
Perušić Hrženjak M. 1087
Perušić T. 1594
Petricioli L. 1704
Petrić Maretić H. 1940
Petrović I. 724
Petrunić R. 1573
Picek S. 1550
Piltaver R. 1025
Pintar D. 1681
Pinter R. 1034
Pirić E. 1775
Pjevalica N. 850
Podgorelec V. 1453
Poljak M. 31
Poljanec N. 1165
Popa C. 48, 52
Popović M. 822
Popović M. 891
Popović Z. 1763, 1769
Portal J.M. 1915
Posarac D. 1209
Poženel M. 862
Predanić S. 1039, 1950
Presečki Ž. 564
Pribanić T. 155, 1935
Prohaska Z. 1016
Prohaska Z. 971
Ptiček M. 439, 1823
Puhalović I. 1526
Pukšič D. 816
Purković D. 1127
Puškar J. 1930
Radchenko G.I. 248
Radev D. 479
Radić Lakoš T. 1158
Radojević B. 673
Radulović B. 1044
Rakipi O. 885
Rantala M. 242
Rantanen P. 1240
Rashkovska A. 384
Ravlić S. 1327
Redžepagić J. 1946
Regnier A. 1915
Rendulić I. 1257
Repnik R. 781, 901
Révészová L. 868
Riabtsev V. 1093

Ribić S. 801, 1199, 1539
Richerzhagen M. 217
Riedel M. 217, 282
Ristić D. 15
Ristov P. 1744, 1793
Ristov S. 1245, 212, 258, 267, 276, 294, 299, 1799, 1817
Rizvić S. 364
Robič D. 781
Rodič B. 1290
Rolich T. 1430
Romanenko A.I. 18
Rondović B. 1763, 1769
Rosado L. 392
Rosić M. 1117
Rožanc I. 879
Rožić L. 1284
Rožić M. 155
Rudan S. 85
Rukljač I. 946
Rumenjak Ž. 1550
Ruzbarsky J. 1885
Rybicki J. 230
Saari M. 1240
Sadžak A. 364
Sakal M. 895
Salihbegović A. 1539, 801, 1199
Samardžija D. 1209
Sanchez I. 1556
Sarić A. 1857
Satta R. 1556
Savchenko D.I. 248
Schatten M. 1503
Schönherr S. 185
Sedinić I. 1594
Seitkulov Y. 1280
Selmi L. 5
Seljan J. 597, 682
Sertić M. 1147
Shamakina A. 197
Sharaqa A.H. 569, 574
Shurman M. 584, 589, M. 686
Sikimić U. 353
Sillberg P. 1240
Simić M.I. 607
Sinitsa K. 1093
Skala K. 236
Skliarova I. 172
Sklyarov V. 172
Skočir Z. 1823
Skok S. 1363
Slavuj V. 915
Smith N. 161
Smokvina R. 1851
Smolković Cerovski B. 953
Sočković A. 1105
Soini J. 1240, 242
Sokolinsky L. 197, 223
Soldo I. 1930
Soler J. 507
Sonzogni J. 1915
Spahić D. 528
Sparrow R.D. 1562
Spasić I. 1930
Spasov D. 678
Specht G. 185
Sredojev B. 1209
Sruk V. 1274, 288
Stirenko S. 191
Stojanović S. 167
Stojanović Ž. 167
Stojković N. 127
Stoyanov I. 479, 617
Studen D. 1946
Stumpf G. 1738
Sudnitson A. 172
Suk D. 1189
Suligoj T. 31, 37, 43
Sušanj D. 368
Sutlović E. 1890
Šajn L. 398
Šarkanj K. 1134
Šečić A. 1235
Šegvić M. 1263
Šeremet Ž. 1388
Šerval D. 435
Šimec A. 1205
Šimenić K. 1225
Šimunić D. 597, 612, 643, 682
Šipek R. 1121
Šipuš Z. 80, 85
Šišić H. 1153
Škoda P. 288
Škopljanac-Mačina F. 921
Škrbić M. 499, 695
Škvorc D. 1251
Šokac D. 1170
Šolić K. 1623
Šoša Anić M. 546
Špeh I. 653, 1863
Špoljarić T. 1896
Šprem M. 80
Štajduhar I. 1436, 1497, 459, 724
Štampar M. 1581
Šuka D.S. 607
Šunde V. 109, 116
Šupe B. 967
Šupić H. 1509
Šušak T. 1840
Taipale O. 248
Tan Z.H. 1375
Taneski V. 1617, 1627
Tarasenko M.S. 21
Tasić I. 1044
Tasić O. 1030
Tasić V. 1030, 1488
Tatur M. 1280
Taylor J. 236
Teković A. 1946
Temerinac M. 850
Tepić S. 149
Thalheim B. 758
Tijan E. 1738, 1851
Tomašić I. 384
Tomičić I. 1503
Tomić D. 412
Tomić M. 541

Tomur E. 1568
Topić A. 1868
Tóth J. 709
Tretinjak M. 937, 949
Trobec R. 384
Trusso S. 15
Tsvetanov F. 479
Tudić V. 25
Tuhtan V. 368
Tumbas P. 895
Turán J. 709, 1885
Ungermanns C. 1055
Urbančíková N. 832
Urem F. 967, 1158
Uroda I. 971, 1016
Valenčić D. 520
Valo A. 791
Vasconcelos M.J.M. 392
Vega A. 647
Vejačka M. 838, 844
Velki T. 1623
Venica S. 5
Vidaček-Hainš V. 753
Vidas-Bubanja M. 1751
Vidović K. 472
Vincek D. 940
Vitas I. 643
Vlahinić S. 127
Vlahinić S. 736
Volarević M. 334
Volarić I. 127
von St. Vieth B. 230
Voychenko O. 1093
Vrana R. 982, 987
Vrandečić B. 682
Vranić M. 1681
Vrankić M. 736
Vrbančić S. 1284
Vrbovčan I. 546
Vrdoljak B. 1823
Vretenar L. 739
Vrkić N. 541
Vučićević M. 1950
Vujaković M. 353
Vukalović J. 1587, 1645
Vukelić D. 305
Vukmirović S. 1003, 1009
Vukojević S. 1603
Wahdan M.A. 686
Weißensteiner H. 185
Werth W. 1055
Wirth A. 901
Wu M.M.Z. 667
Xhafa V. 272
Yakovleva G.E. 18
Yüksel T. 1415
Zakarija I. 921
Zavec Pavlinić D. 344
Zazula D. 404
Zečević I. 874
Zekanović-Korona Lj. 1805
Zekulić V. 1172
Zoričić K. 1890
Zymbler M. 1675
Žagar M. 1609
Žagar M. 673
Žaja I. 1158
Žanić Mikuličić J. 1744
Žarkić I. 1710
Žgalj A. 695
Žilak J. 37
Živić N. 695

FOREWORD

The *38th International ICT Convention MIPRO 2015* was held from 25th until 29th of May 2015 in Opatija, the Adriatic Coast, Croatia. The Convention consisted of nine conferences under the titles: *Microelectronics, Electronics and Electronic Technology (MEET), Distributed Computing, Visualization and Biomedical Engineering (DC VIS), Telecommunications & Information (CTI), Computers in Education (CE), Computers in Technical Systems (CTS), Intelligent Systems (CIS), Information Systems Security (ISS), Business Intelligence Systems (miproBIS), Digital Economy and Government, Local Government, Public Services (DE/GLGPS)*. A special conference was dedicated to works of the students: *MIPRO Junior-Student Papers (SP)*.

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

The convention was organized by the Croatian ICT Society MIPRO with the help of numerous patrons and sponsors to whom we owe our sincere thanks. We specially single out our gold sponsors Ericsson Nikola Tesla, T-Croatian Telecom and Končar-Electrical Industries and silver sponsors HEP–Croatian Electricity Company and InfoDom. Our bronze sponsors are Hewlett Packard, IN2, Transmitters and Communications and Storm Computers.

To all who helped organize the 38th International ICT Convention *MIPRO 2015* as well as editing of this Book of Proceedings we extend our heartfelt thanks.

Prof. Petar Biljanović, PhD
International Program Committee
General Chair

MEET

International Conference on

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

Steering Committee

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

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

INVITED PAPER

State-of-the-art semi-classical Monte Carlo method for carrier transport in nanoscale transistors

P. Palestri, E. Caruso, F. Driussi, D. Esseni, D. Lizzit, P. Osgnach, S. Venica, L. Selmi
DIEG, University of Udine, via delle scienze 206, 33100, Udine, Italy
palestri@uniud.it

Abstract - We review the Monte Carlo method to model semi-classical carrier transport in advanced semiconductor devices. We report examples of the use of the Multi-Subband Monte Carlo method to simulate MOSFETs with III-V compound semiconductor channel. Monte Carlo transport modeling of graphene-based transistors is also addressed.

I. INTRODUCTION

The models implemented in commercial TCAD tools for semiconductor devices are derived from the Boltzmann Transport Equation (BTE) under assumptions that make them inaccurate when applied to nanoscale devices (L_G<100 nm), or to the study of hot carriers.

Exact numerical solutions of the BTE via the Monte Carlo (MC) method have been used since the 60s and the 70s to analyze uniform transport in semiconductors [1]. Monte Carlo simulators describe the carrier motion as a sequence of "free-flights" (i.e. collision-less motion according to Newton's law) interrupted by scattering events, which abruptly change the particle's momentum. The simulation is statistical in the sense that the duration of the free-flight, the selection of the scattering mechanism and of the state-after-scattering are randomly chosen according to the respective probability density functions.

During the 80s, MC models have become increasingly complex by including full-band descriptions of the E(**k**) relation and achieving the capability to describe realistic 2D devices [2],[3]. During the 90s, full-band MC simulators evolved by including comprehensive sets of scattering mechanisms and by improving the numerical efficiency. Mutual agreement between MC simulators developed by different groups was consolidated [4] and a wide range of successful applications to the study of I-V, RF and reliability properties of scaled MOSFETs [5][6], BJTs [7] and non-volatile-memories [8] was reported.

Quantization effects have become more and more relevant in scaled MOSFETs, and quantum corrections have been proposed for semi-classical MC simulators starting from the year 2000 [9]-[12]. These corrections describe the electrostatic effects due to the displacement of the inversion charge from the channel/dielectric interface, and sometimes also for the valley splitting [13]. However, other important aspects related to the quantized nature of the carrier gas (e.g. scattering rates modulated by size- and bias-induced quantization, strong degeneracy of the carrier gas) are not captured by these corrections. These limitations led to the development of the Multi-Subband Monte Carlo (MSMC) approach.

This paper reviews the main features of the MSMC method and reports sample applications to the study of MOSFETs with III-V channel materials and comparison with full-quantum transport simulations. The use of the MC method to study graphene-based devices is reported too. We thus demonstrate that the MSMC method is a very powerful and useful technique to assess the potential of new, technology-boosted, MOSFET architectures.

II. THE MULTI-SUBBAND MONTE CARLO APPROACH

The MSMC methodology has been initially proposed in [14]. Since then, MSMC simulators have been developed by various groups [15]-[20]. A typical simulation flowchart is reported in Figure 1.

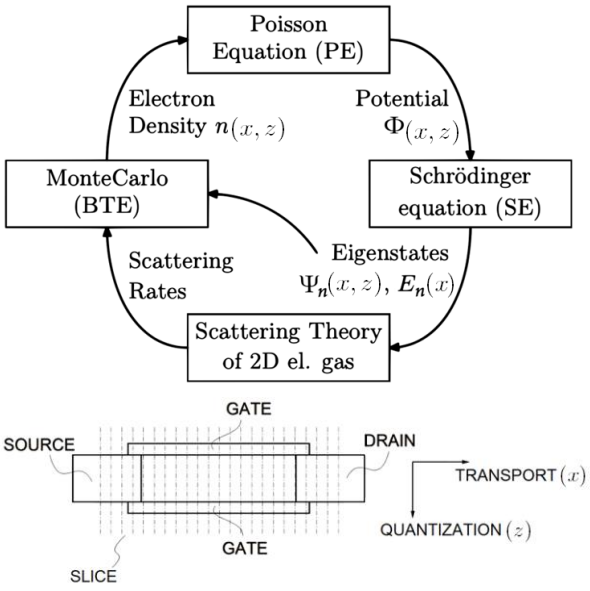


Figure 1 Top: Flowchart of a MSMC simulator; Bottom: sketch of a double-gate MOSFET with indication of the slices normal to the transport direction where the 1D quantization problem is solved.

The simulation proceeds as follows [21]. Starting from an initial guess of the 2D potential profile, the device is divided in sections along the channel (x) direction. In each section, the 1D Schrödinger equation (SE) is solved for all relevant valleys. The discrete energies levels in each section are used to build subband profiles along the

channel; the derivative of the subband energy along *x* provides the driving force to move particles in the MC simulation module. The solution of the SE also provides the wave-functions that are used to compute the scattering matrix elements for intra- and inter-subband transitions and, at a later stage, the charge distribution in the direction normal to transport.

Scattering rates and subband profiles feed the MC transport module that solves the multi-subband BTE [21]; the particle state contains only the position along the channel and the 2D wave-vector in the transport plane (normal to quantization). In other words, the MC module computes, for each subband and at each position, the occupation probability of the k-states in the transport plane. During the MC step, scattering events are rejected based on the availability of the final state [22], in order to enforce Pauli's principle that is of paramount importance to obtain accurate Fermi-Dirac occupation in highly doped regions at equilibrium and to account for the strong degeneration of the inversion channel. At the end of the MC step, the electron concentration profile is available. Then, the solution of the 2D Poisson equation provides a new estimate of the potential. The procedure is iterated until convergence is reached, i.e. until the drain current update among iterations is less than a few %.

In this paper, we show results using the MSMC tool developed at the University of Udine over the last 10 years. The simulator has been initially developed for silicon MOSFETs (bulk or SOI) [20]. Local phonon, Coulomb and surface roughness scatterings are commonly regarded as the dominant collision mechanisms in these devices. Only Δ valleys are relevant, and their weak non-parabolicity can be accounted for in the calculation as a correction of the in-plane dispersion relation, resorting to the conventional Effective Mass Approximation (EMA) when solving the 1D Schrödinger equation. Screening of the scattering events has been implemented following either the scalar or the matrix dielectric function approach [23]. Extension of the model to sSi [24] and Ge [25] required a more flexible treatment of multiple sets of valleys (CB minima) with arbitrary energy shift and inter-valley phonons. Alloy scattering was then added to enable the simulation of SiGe channels [26]. Furthermore, specific scattering models were added to simulate MOSFETs with high-k dielectrics [27]; in particular, remote polar phonon scattering (i.e. the scattering potential induced in the channel by the polar phonons in the dielectric) and remote Coulomb scattering (scattering induced by fixed charges in the gate stack). Models for metallic source and drain were also developed as described in [28]. Among possibly relevant but still missing scattering mechanisms, it is worth to mention that so far we do not include long-range Coulomb interactions [29][30] whose effect on the current of short MOSFETs is still debated [31].

The multi-valley energy relation and the wide variety of scattering mechanisms allow extending the use of the MSMC model to III-V MOSFETs. Due to the highly defective interfaces of III-V semiconductors, a model for interface states has been recently implemented [32]. Generation of e-h pairs by band-to-band tunneling have been added to enable the simulation of tunnel-FETs [33].

A MSMC for pMOSFETs has been developed too [34][35].

The rich portfolio of physical effects and scattering mechanisms increases the computational burden. In spite of this, competitive execution times and remarkably good numerical efficiency was achieved thanks to optimization and parallelization of the code [36].

III. Applications of MSMC transport modelling to III-V channel MOSFETs

As mentioned in the previous section, the MSMC method was initially developed for group IV semiconductors. Its extension to III-V compounds required additional models for polar phonon scattering [21] and interface states [32] and a quantization model taking into account the strong non-parabolicity of the Γ valley [37] and the layered structure of the channel.

The non-parabolic quantization model of ref. [37] not only corrects the subband energies, but also modifies the in-plane E(**k**) relation. As a result (see Figure 2), when the well thickness decreases, the effective mass (i.e. the inverse curvature of the E(**k**) relation) of the lowest subband increases. This result is consistent with the accurate quantization models based on DFT, tight-binding or k·p calculations [38].

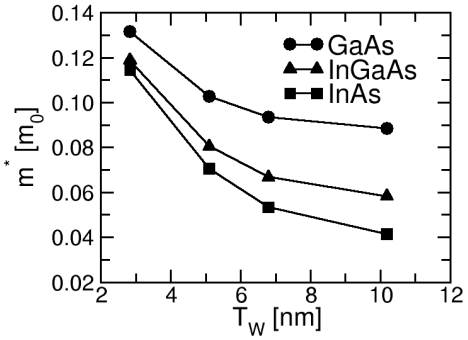


Figure 2: Effective mass (calculated from the inverse curvature of the in-plane E(**k**) of the lowest subband) vs. well thickness for different III-V compounds. The model of Ref.[37] has been used to account for the non-parabolicity of the Γ valley.

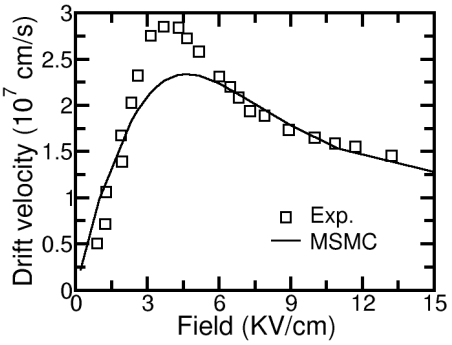


Figure 3: Comparison between simulated and measured [39] velocity-field curves in bulk $In_{0.53}Ga_{0.47}As$.

One of the advantages of the MC method is that it handles equally well long and short channel devices. The former are important for calibration purposes. As an example, Figure 3 compares the simulated and experimental velocity-field curves for bulk $In_{0.53}Ga_{0.47}As$

(the bulk structure is simulated in the MSMC as a thick well considering a large number of subbands), whereas Figure 4 compares MSMC results and measured mobility in $In_{0.53}Ga_{0.47}As$ inversion layers. The results in Figure 3 validate the bulk band structure and the phonon and alloy scattering models, whereas Figure 4 was used to calibrate the parameters of the surface roughness model. We employ here the "standard" linear model for surface roughness [43]. A more rigorous non-linear model has been developed [44][45] and is being implemented in the MSMC.

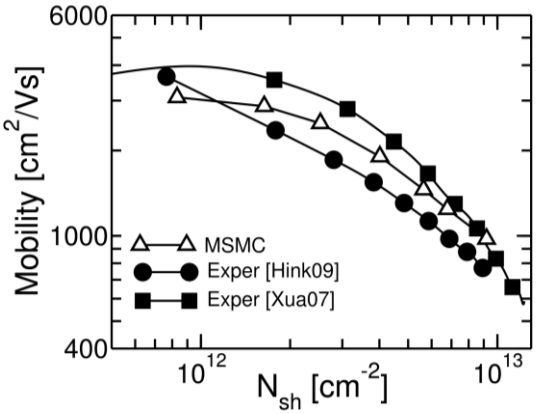


Figure 4: Comparison between simulated and measured [40][41] low-field mobility in bulk $In_{0.53}Ga_{0.47}As$ MOSFETs.

After calibration on long channel devices, the MSMC has been used to simulate the current drive of short channel III-V nMOSFETs [46],[42]. An example of transfer characteristics is reported in Figure 5 for a device with L_G=14nm, designed according to the ITRS roadmap for HP devices. These calculations account for all the relevant scattering mechanisms, namely: local and remote polar phonons, surface roughness, alloy scattering, intra- and inter-valley non-polar acoustic and optical phonons. Interface states have not been considered in these simulations. The on-current of the $In_{0.53}Ga_{0.47}As$ device is larger than in silicon but comparable to the one of highly strained (2GPa) silicon. The subthreshold swing is also improved in $In_{0.53}Ga_{0.47}As$ as discussed in details in [46].

The analysis of the internal variables in Figure 6 helps us understand the differences between $In_{0.53}Ga_{0.47}As$ and sSi channel. The profile of the first subband energy along the channel (right axis) shows that, due to the low density-of-states, the electron gas in the source and drain regions is strongly degenerate in $In_{0.53}Ga_{0.47}As$. The screening length of a highly degenerate electron gas is much longer than in a weakly degenerate one (as is the case of sSi, where the subband energy in the source and drain is close to the Fermi levels). As a result, the potential induced by the gate bias penetrates much deeper inside the source and drain in the $In_{0.53}Ga_{0.47}As$ device than in the sSi one. The "effective length" of the $In_{0.53}Ga_{0.47}As$ device is thus larger than in the sSi one [46] and this explains the better SS.

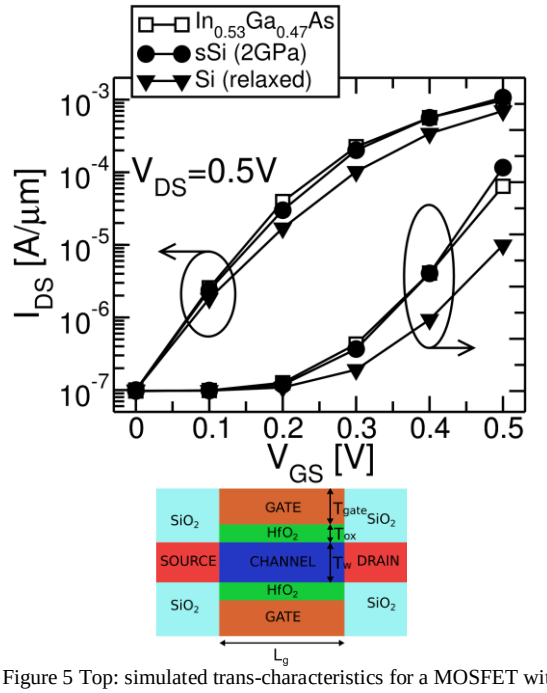


Figure 5 Top: simulated trans-characteristics for a MOSFET with L_G=14nm considering different channel materials. T_{OX}=4nm. The volumetric gate thickness is T_{GATE}=4nm. TW=5nm. The device is sketched in the lower panel. The source and drain junctions are abrupt and perfectly aligned to the gate edge.

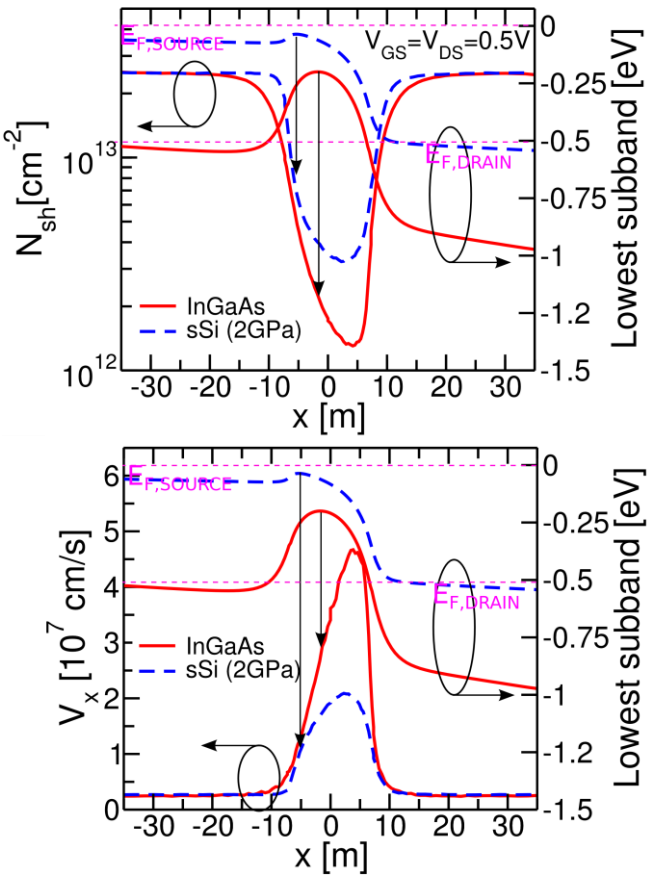


Figure 6: Internal quantities (subband profile, inversion density and average velocity) for the device of Figure 5 considering sSi and InGaAs channels. The energy reference (E=0) is the Fermi level in the source contact.

The inversion density profile ($N_{sh}(x)$, Figure 6, top graph) is, as expected, lower in $In_{0.53}Ga_{0.47}As$ than in sSi, due to the lower density of states. Conversely, the average velocity profiles along the channel (bottom graph) show that V_x is larger in $In_{0.53}Ga_{0.47}As$ than in sSi due to the lower effective mass in the transport direction. In particular, the velocity at the injection point (maximum of the potential barrier, indicated by the vertical arrows in Figure 6) is larger by a factor of about three, and this compensates the lower inversion charge, so that the current is almost the same in the two devices. The digital switching time can be estimated from the static simulations [46] and it is 0.327ps for the $In_{0.53}Ga_{0.47}As$ device vs. 0.362ps for the sSi one. This means that the charge modulation in the S/D regions due to the low DoS almost compensates the larger injection velocity of $In_{0.53}Ga_{0.47}As$.

IV. SEMI-CLASSICAL MC VS. FULL-QUANTUM MODELS

The MSMC describes quantization normal to the transport direction in a rigorous way. On the other hand, transport from source to drain is semi-classical, i.e. electrons are treated as point particles obeying Newton's law. This means that phenomena such as source-to-drain tunneling, that is relevant in the sub-threshold region of short channel devices are not accounted for. To overcome this limitation, following the effective potential approach already used for quantum corrections in semi-classical MC [9], we have proposed the subband smoothing procedure sketched in Figure 7 [47]: convolution between the subband profile and a Gaussian (envelope) function mimics the distributed nature of the electron wave-packet. In other words, the electron "feels" an average of the driving force in its surroundings, but it is still treated as a point particle in the MC transport core. The Gaussian shape is assigned to the electron when solving the Poisson equation [47].

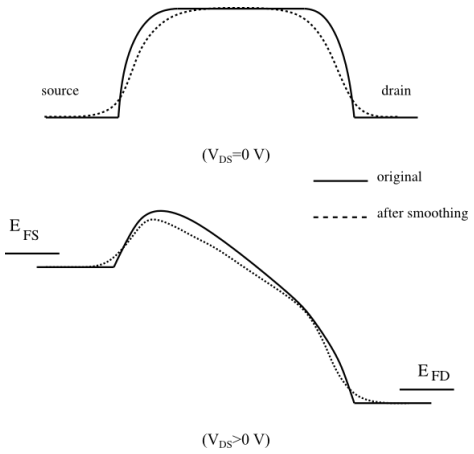


Figure 7 Sketch of the subband smoothing procedure applied in the MSMC. At low V_{DS} the smoothening leads to more penetration of the electrons from the S/D into the channel. At V_{DS}>0, the top of the potential barrier at the injection point is lowered: the enhanced thermionic emission mimics the source-to-drain tunneling.

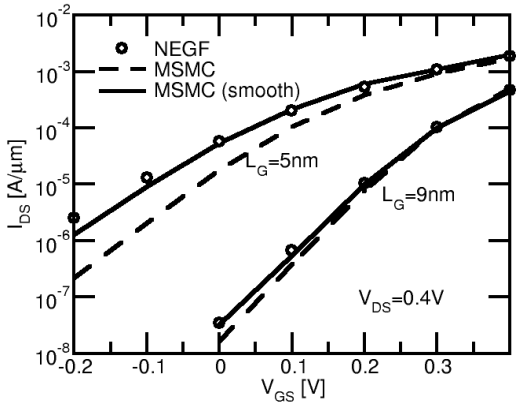


Figure 8 Comparison between MSMC and NEGF [48] simulations for Si MOSFETs with different gate length. Double-gate devices with 3nm film thickness and 0.9nm EOT.

The only parameter of this model is the standard deviation of the Gaussian function. Figure 8 compares the MSMC results with and without subband smoothening with NEGF results obtained by using "nanomos" [48]. Silicon MOSFETs are considered. A single value of 1.1nm for the standard deviation of the Gaussian function allows the MSMC to reproduce the NEGF results for devices down to 5nm gate length. Applicability of this approach to III-V channel MOSFETs is under investigation.

V. MONTE CARLO MODELING OF GRAPHENE DEVICES

Due to the large mobility and its monoatomic thickness [49], graphene has been widely investigated as material for electron devices, either as a replacement of the silicon channel in FETs or as part of alternative device structures able to exploit the graphene unique features.

Although quantum-mechanical effects are very relevant in graphene due to its gap-less nature and to the long coherence length, semi-classical transport and MC methods have been used as well to predict device performance, as shown in the following two examples.

A. Graphene FETs (GFETs)

The MC method has been used by several authors studied dissipative transport in GFETs [50]-[54]. When considering graphene foils, band-to-band-tunneling (BTBT) must be taken into account [52],[54] in order to describe the strong ambipolarity of the devices. We have shown in [54] that a simple local BTBT model is adequate to reproduce more accurate NEGF calculations.

As explained above, where we showed results for III-V MOSFETs, one of the main advantages of the MC approach is that the same model can be used for long channels (for calibration purpose) as well as for short channels (to predict the performance of ultimately scaled devices). In this spirit, Figure 9 compares MC simulations and experiments for mobility in graphene nano-ribbons. Here the dominant scattering mechanisms are edge roughness and remote polar phonons from the insulator [53]. A more extensive comparison for different

charge sheet densities and temperature can be found in [53].

Note that MC simulation of graphene foils does not require the MSMC approach, since there is no quantization-induced subband splitting. On the other hand, it also differs from a conventional 3D semi-classical MC since the k-space is 2D. Nano-ribbons, instead, may demand the description of quantization in the ribbon width [55]. In Figure 9, however, we use a 2D model with a ribbon width dependent energy dispersion relationship [51].

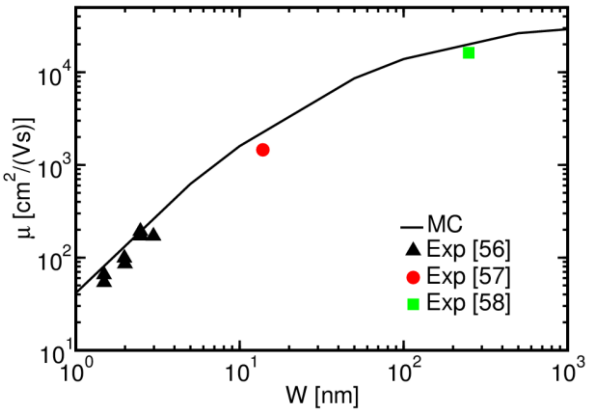


Figure 9: Simulated mobility as a function of the graphene ribbon width (W). Experiments are reported for comparison [56][57][58]. The simulated sheet charge density is low (10^{11}cm^{-2})

Having verified the correctness of the band structure model and of the scattering mechanisms in long channel devices, we have then run MC simulations of short channel devices. Sample results are illustrated in Figure 10 for a 50 nm long graphene FET. BTBT is responsible for the non-saturating output characteristic (circles): in fact, the simulation without BTBT (triangles) exhibits well behaved current saturation at large V_{DS}. The inclusion of scattering (filled circles) significantly reduces the current compared to ballistic transport.

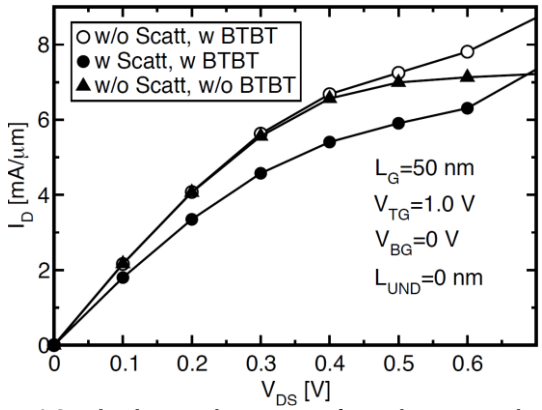


Figure 10 Simulated output characteristic of a graphene FET with 50nm channel. The top dielectric is 5nm of Al_2O_3, with the substrate consists of 300nm of SiO_2.

B. *Graphene Base Transistor (GBT)*

Ambipolarity strongly limits the performance of graphene FETs both in the digital (large off-current) and in the analog (high output conductance) domains [59]-[61]. For this reason, new device concepts exploiting graphene have been proposed. A relevant example is the Graphene Base Transistor (GBT) [62]-[64], sketched in Figure 11. A graphene base controls the tunnel injection of electrons from the (metal of silicon) emitter to an emitter/base insulator (EBI). The carriers then go through the graphene base, enter the conduction band of the base/collector insulator (BCI) and are eventually collected by the metal "collector".

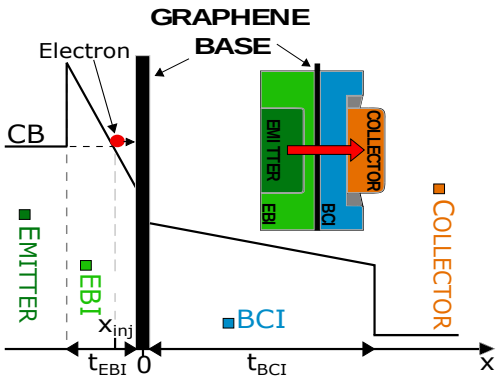


Figure 11: Conduction band diagram of the GBT. Electrons are injected in the EBI conduction band at x_{inj}. The inset shows the GBT structure.

Modeling the GBT mainly involves calculating tunneling currents [65]-[67] since this is the current limiting mechanism. Nevertheless, the MC approach has been useful to describe dissipative transport in the conduction band of the EBI and BCI layers and the energy distribution of the carriers. As an example, Figure 12 reports the average velocity (top) and the concentration profile (bottom) in a template GBT. The model considers electron scattering with the phonons of the dielectric (polar optical and non-polar acoustic) [68], and a non-parabolic description of the conduction band. We see that the velocity and concentration profiles in the BCI are essentially constant. This information was used to derive simplified electrical models to simulate the GBT under high injection conditions [67].

The MC simulation also provides the fraction of the electrons entering the BCI conduction band that are back-scattered into the base. These electrons are most likely captured by the graphene layer and contribute to the base current. Consequently, back-scattering in the BCI sets an upper limit to the common base current gain $\alpha_F = I_C/I_E$ (see values in the legend of Figure 12). According to our results, the simulated α_F is much higher than the experiments [63] (mainly because the assumption that the graphene base is fully transparent to electrons arriving from the EBI is not verified), but, nevertheless, small enough to pose a serious limitation to the device performance.

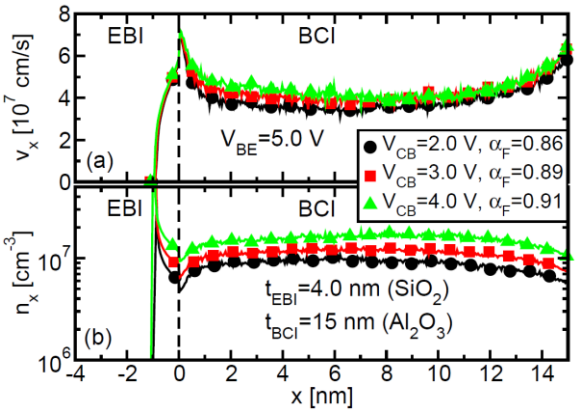


Figure 12 Average velocity v_x (a) and electron concentration n_x (b) along the GBT for a few values of V_{CB}. The concentration is low since the injected current is a few A/m^2.

VI. CONLUSIONS

In this paper, we have reviewed the Monte Carlo method for carrier transport in electron devices. Applications to MOSFETs with III-V channels and to graphene devices have been reported. This demonstrates that MC transport models are versatile tools, well suited to help in the assessment of the performance of modern nanoscale devices, and to support researchers in the evaluation of new materials and device architectures.

Compared to the conventional drift-diffusion model used in commercial TCAD, modern MSMC transport techniques provide a much more rigorous description of far-from-equilibrium transport and of carrier quantization. Compared to full-quantum transport methods, they can handle larger devices, allowing for an easier calibration with experimental data, and a more efficient and complete treatment of scattering. However, comparison between MC and full-quantum transport simulators is of paramount importance to verify and calibrate the many model ingredients; e.g., the quantization model with non-parabolic corrections used for III-V MOSFETs [38], the subband smoothening procedure [47] and the BTBT model for graphene [54].

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement III–VMOS Project No. 619326, Grant Agreement E2Switch project No. 619509, Grant Agreement GRADE project No. 317839 and by the Italian MIUR, through the Futuro in Ricerca project RBFR10XQZ8.

REFERENCES

[1] C. Jacoboni and L. Reggiani, "The Monte Carlo method for the solution of charge transport in semiconductors with applications to covalent materials," *Rev. Mod. Phys*.,vol. 55, pp. 645–705, 1983.

[2] M.V. Fischetti and S.E. Laux, "Monte Carlo analysis of electron transport in in small semiconductor devices including band-structure and space-charge effects," *Phys. Rev.B*, vol. 38, pp. 9721–9745, 1988.

[3] F.Venturi, R.K.Smith, E.Sangiorgi, M.R.Pinto, and B.Ricc`o, "A General Purpose Device Simulator Coupling Poisson and Monte Carlo Transport with Applications to Deep Submicron MOSFETs," *IEEE Trans. on Computer-Aided Design*, vol. CAD-8, p. 360, 1989.

[4] A. Abramo, et al.,"A comparison of numerical solutions of the Boltzmann transport equation for high-energy electron transport in silicon," *IEEE Transactions on Electron Devices*, vol.41, p. 1646, 1994.

[5] M.V. Fischetti and S.E. Laux, "Monte Carlo simulation of transport in technologically significant semiconductors of the Diamond and Zinc-Blende structures - Part II: submicrometer MOSFET's," *IEEE Trans. on Electron Devices*, vol. ED-38, p. 650, 1991.

[6] P. Palestri, D. Esseni, S. Eminente, C. Fiegna, E. Sangiorgi, and L. Selmi, "Understanding Quasi-Ballistic Transport in nano-MOSFETs. Part I: Scattering in the Channel and in the Drain," *IEEE Trans. on Electron Devices*, vol. 52, no. 12, pp. 2727–2735, 2005

[7] E. F. Crabbè, J. M. C. Stork, G. Baccarani, M. V. Fischetti, S. E. Laux., "The impact of nonequilibrium transport on breakdown and transit time in bipolar transistors," *in IEDM Tech. Digest*, p. 463, 1990

[8] A. Ghetti, L. Selmi, and R. Bez, "Low Voltage Hot Electrons and Soft-Programming Lifetime Prediction in Non-Volatile Memory Cells," *IEEE Trans. on Electron Devices*, vol. 46, no. 4, p. 696, 1999

[9] D.K. Ferry, R. Akis, and D. Vasileska, "Quantum Effects in MOSFETs: Use of an Effective Potential in 3D Monte Carlo Simulation of Ultra-Short Channel Devices," *in IEDM Tech. Digest*, pp. 287–290, 2000.

[10] H. Tsuchiya, B. Fischer, and K. Hess, "A Full-BAnd Monte Carlo Model for Silicon Nanoscale Devices with a Quantum Mechanical Correction of the Potential," *in IEDM Tech. Digest*, pp. 283–286, 2000.

[11] B. Winstead and U. Ravaioli, "A Quantum Correction Based on Schr¨odinger Equation Applied to Monte Carlo Device Simulation," *IEEE Trans. on Electron Device*s, vol. 50, pp. 440–446, 2003.

[12] P. Palestri, S. Eminente, D. Esseni, C. Fiegna, E. Sangiorgi, and L. Selmi, "An improved semiclassical Monte-Carlo approach for nano-scale MOSFET simulation," *Solid State Electronics*, vol. 49, no. 5, pp. 727–732, 2005.

[13] C. Sampedro-Matarin, F. Gamiz, A. Godoy, and F.J. Gracia Ruiz, "The Multivalley Effective Conduction Band-Edge Method for Monte Carlo Simulation of Nanoscale Structures," *IEEE Trans. on Electron Devices*, vol. 53, pp. 2703–2710, 2006.

[14] M.V. Fischetti and S.E. Laux, "Monte Carlo study of electron transport in silicon inversion layers," *Physical Review B*, vol. 48, pp. 2244–2274, 1993.

[15] C. Jungemann, A. Edmunds, and W.L. Engl, "Simulation of Linear and Nonlinear Electron Transport in Homogeneous Silicon Inversion Layers," *Solid State Electronics*, vol. 36, no. 11, pp. 1529–1540, 1993.

[16] F. Gamiz, J.A. Lopez-Villanueva, J.B. Roldan, J.E. Carceller, and P. Cartujo, "Monte Carlo Simulation of Electron Transoprt Properties in Extremely Thin SOI MOSFET's," *IEEE Trans. on Electron Devices*, vol. 45, no. 5, pp. 1122–1126, 1998.

[17] S.C. Williams, K.W. Kim, and W.C. Holton, "Ensemble Monte Carlo study of channel quantization in a 25nm n-MOSFET," *IEEE Trans. on Electron Devices*, vol. 47, no. 10, pp. 1864–1872, 2000.

[18] T. Ezaki, P. Werner, and M. Hane, "Self-Consistent Quantum mechanical Monte Carlo MOSFET device simulation," *Journal of Computational Electronics*, vol. 2, no. 2–4, pp. 97–103, 2003.

[19] J. Saint-Martin, A. Bournel, F. Monsef, C. Chassat, and P. Dollfus, "Multi sub-band Monte Carlo simulation of an ultra-thin double gate MOSFET with 2D electron gas ," *Semicond. Sci. Technol*., vol. 21, pp. L29–L31, 2006.

[20] L. Lucci, P. Palestri, D. Esseni, L. Bergagnini and L. Selmi, "Multisubband Monte Carlo Study of Transport, Quantization, and Electron-Gas Degeneration in Ultrathin SOI n-MOSFETs*", IEEE Trans. on Electron Devices*, Vol. 54, no. 5, pp. 1156-1164, 2007.

[21] D. Esseni, P. Palestri, L. Selmi, "Nanoscale MOS transistors: Semi-classical Transport and Applications", Cambridge, 2011.

[22] P. Lugli and D. K. Ferry, “Degeneracy in the ensemble Monte Carlo method for highfield transport in semiconductors,” *IEEE Trans. on Electron Devices*, vol. 32, no. 11, pp. 2431–2437, 1985

[23] P. Toniutti, D. Esseni, and P. Palestri. "Failure of the scalar dielectric function approach for the screening modeling in double-gate SOI MOSFETs and in FinFETs." *IEEE Trans. on Electron Devices*, vol. 57, n.11, pp. 3074-3083, 2010.

[24] D. Esseni, F. Conzatti, M. De Michielis, N. Serra, P. Palestri and L. Selmi, “Semi-classical transport modelling of CMOS transistors with arbitrary crystal orientations and strain engineering”, *Journal of Computational Electronics*, vol. 8, n..3-4, pp. 209–224, 2009.

[25] F. Conzatti, P. Toniutti, D. Esseni, P. Palestri and L. Selmi, “Simulation study of the on-current improvements in Ge and sGe versus Si and sSi nano-MOSFETs”, *in IEDM tech Digest*, pp. 363-366, 2010.

[26] D. Lizzit, P. Palestri, D. Esseni, A. Revelant, and L. Selmi, “Analysis of the Performance of n-Type FinFETs With Strained SiGe Channel”, *IEEE Trans. on Electron Devices*, vol.60, no.6, pp. 1884-1891, 2013.

[27] Toniutti, P.; Palestri, P.; Esseni, D.; Driussi, F.; De Michielis, M.; Selmi, L., "On the origin of the mobility reduction in n- and p-metal–oxide–semiconductor field effect transistors with hafnium-based/metal gate stacks," *Journal of Applied Physics* , vol.112, no.3, pp.034502,034502-12, 2012.

[28] V. Gudmundsson, P. Palestri, P.-E. Hellstrom, L. Selmi, M. Ostling, “Simulation of low Schottky barrier MOSFETs using an improved Multi-subband Monte Carlo model.” *Solid State Electonics*, vol. 79, p.172-178, 2013.

[29] M.V. Fischetti and S.E. Laux, “ Long-range Coulomb interactions in small Si devices. Part I: Performance and reliability,” *Journal of Applied Physics*, vol. 89, no. 2, pp. 1205–1231, 2001

[30] K. Nakanishi, T. Uechi, and N. Sano, “Self-Consistent Monte Carlo Device Simulations under Nano-Scale Device Structures: Role of Coulomb interaction, Degeneracy, and Boundary Condition,” *in IEDM Tech. Digest*, pp. 79–82, 2009.

[31] F. Monsieur, D. Rideau, V. Quenette et al., “The importance of the spacer region to explain short channel mobility collapse in28nm Bulk and FDSOI technologies”, *in ESSDERC 2014*, p.254

[32] P. Osgnach, E. Caruso, D. Lizzit, P. Palestri, D. Esseni, L. Selmi, “The impact of interface states on the mobility and drive current of InGaAs semiconductor n-MOSFETs”, *to appear on Solid-State Electronics*. DOI 10.1016/j.sse.2014.12.011

[33] A. Revelan, P. Palestri, P. Osgnach, L. Selmi, “Calibrated multi-subband Monte Carlo modeling of tunnel-FETs in silicon and III–V channel materials”, *Solid-State Electronics*, v. 88, p.54, 2013.

[34] F. Conzatti, M. De Michielis, D. Esseni, P. Palestri, "Drain current improvements in uniaxially strained p-MOSFETs: A Multi-Subband Monte Carlo study", *Solid State Electronics*, vol.53, pp.706-721, 2009.

[35] M. De Michielis, D. Esseni, P. Palestri, L. Selmi, "Semiclassical Modeling of Quasi-Ballistic Hole Transport in Nanoscale pMOSFETs Based on a Multi-Subband Monte Carlo Approach", *IEEE Transactions on Electron Devices*, vol.56, n.9, pp.2081-2091, 2009.

[36] P. Osgnach, A. Relevant, D. Lizzit, P. Palestri, D. Esseni and L. Selmi, “Toward computationally efficient Multi-Subband Monte Carlo simulations of nanoscale MOSFETs”, *SISPAD*, pp. 176-179, 2013.

[37] S. Jin, M. Fischetti, and T.-W. Tang, “Modeling of electron mobility in gated silicon nanowires at room temperature: Surface roughness scattering, dielectric screening, and band nonparabolicity,” *Journal of Applied. Physics*., vol. 102, no. 8, p. 083715, 2007.

[38] E. Caruso, G. Zerveas, G.Baccarani, L.Czornomaz, N.Daix, D.Esseni, E.Gnani, A.Gnudi, R.Grassi, M.Luisier, T. Markussen, P.Palestri, A.Schenk, L.Selmi, M.Sousa, K. Stokbro, M. Visciarelli, “Modeling approaches for band-structure calculation in III-V FET quantum wells”, *Proc. of EUSOSOI-ULIS*, pp. 101-104, 2015.

[39] Balynas V, Krotkus A, Stalnionis A, Gorelionok A T, Shmidt N M and Tellefsen J A, "Time-resolved, hot-electron conductivity measurement using an electro-optic sampling technique", *Applied Physics A*, vol. 51, n. 4, pp 357-360, 1990.

[40] C. Hinkle, A. Sonnet, R. Chapman, and E. M. Vogel, “Extraction of the effective mobility of In0.53Ga0.47As MOSFETs,” *IEEE Electron Device Lett*., vol. 30, no. 4, pp. 316–318, Apr. 2009.

[41] Y. Xuan, Y. Wu, T. Shen, T. Yang, and P. Ye, “High performance submicron inversion-type enhancement-mode InGaAs MOSFETs with ALD Al2O3 , HfO2 and HfAlO as gate dielectrics,” in IEDM Tech. Digest, pp. 637–640, 2007.

[42] E. Caruso, D. Lizzit, P. Osgnach, D. Esseni, P. Palestri and L. Selmi, “Simulation Analysis of III-V n-MOSFETs: Channel Materials, Fermi Level Pinning and Biaxial Strain”, *in IEDM Tech. Digest*, pp. 7.6.1– 7.6.4, 2014.

[43] D. Esseni, “On the Modeling of Surface Roughness Limited Mobility in SOI MOSFETs and its Correlation to the Transistor Effective Field ,” *IEEE Trans. on Electron Devices*, vol. 51, no. 3, pp. 394–401, 2004

[44] D. Lizzit, D. Esseni, P.Palestri, and L. Selmi, “Surface roughness limited mobility modeling in ultra-thin SOI and quantum well III-V MOSFETs”, *in IEDM Tech. Digest*, pp. 5.2.1–5.2.4., 2013.

[45] D. Lizzit, D. Esseni, P.Palestri, and L. Selmi,” A new formulation for surface roughness limited mobility in bulk and ultra-thin-body metal–oxide–semiconductor transistors”, *Journal of Applied Physics*, vol. 16, pp. 223702-1, 223702-13, 2014.

[46] D. Lizzit, D. Esseni, P.Palestri, Patrik Osgnach, and L. Selmi,” Performance Benchmarking and Effective Channel Length for Nanoscale InAs, In0.53Ga0.47As, and sSi n-MOSFETs”, *IEEE Trans. On Electron Devices*, Vol. 61, no. 6, pp. 2027-2034, 2014.

[47] P. Palestri, L. Lucci, S. Dei Tos, D. Esseni, and L. Selmi. “An improved empirical approach to introduce quantization effects in the transport direction in multi-subband Monte Carlo simulations”. *Semiconductor Science Technology*, vol. 25,.n.5, p. 055011, 2010.

[48] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M.S. Lundstrom, “nanomos 2.5: A two-dimensional simulator for quantum transport in double-gate mosfets,” *IEEE Trans. on Electron Devices*, vol. 50, no. 9, pp. 1914–1925, 2003.

[49] K.S. Novoselo, A.K.Geim,S.V. Morozov,D. Jiang, M.I. Katsnelso, I.V. Grigorieva, et al. “Two-dimensional gas of massless Dirac fermions in graphene”. *Nature*, v.438(7065) pp. 197–200, 2005.

[50] V. Perebeinos and P. Avouris, “Inelastic scattering and current saturation in graphene,” *Physical Review B*, vol. 81, p. 195442, 2010.

[51] M. Bresciani, P. Palestri, D. Esseni, and L. Selmi, “Simple and efficient modeling of the E-k relationship and low-field mobility in graphene nano-ribbons,” *Solid-State Electronics*, vol. 54, pp. 1015–1021, 2010.

[52] J. David, L. Register, and S. Banerjee, “Semiclassical Monte Carlo analysis of graphene fets,” *IEEE Trans. On Electron Devices*, vol. 59, no. 4, pp. 976–982, 2012.

[53] M. Bresciani, P. Palestri, D. Esseni, L. Selmi, B. Szafranek, and D. Neumaier, “Interpretation of graphene mobility data by means of a semiclassical Monte Carlo transport model,” *Solid-State Electronics*, vol. 89, pp. 161–166, Nov. 2013

[54] A. Paussa, G. Fiori, P. Palestri, M. Geromel, D. Esseni, G. Iannaccone, L. Selmi, “Simulation of the Performance of Graphene FETs With a Semiclassical Model, Including Band-to-Band Tunneling”, *IEEE Trans. on Electron Devices*, Vol. 61, no. 5, pp. 1567- 1574, 2014

[55] A. Betti, G. Fiori, G. Iannaccone, Y. Mao, “Physical insights on graphene nanoribbon mobility through atomistic simulations.”, *in IEDM Tech Digest*, pp. 897–900, 2009

[56] X. Wang, Y. Ouyang, X. Li, H. Wang, J. Gu, H. Dai, “Room-temperature all semiconducting sub-10-nm graphene nanoribbon field-effect transistors.” *Phys. Rev. Lett.*, v.100(20), p. 206803, 2008.

[57] L. Jiao, X. Wang, G. Diankov, H. Wang, H. Dai, “Facile synthesis of high-quality graphene nanoribbons.”, *Nature Nanotechnology*., v. 5(5), pp.321–5, 2010.

[58] M. Lemme, T. Echtermeyer, M. Baus, B. Szafranek, J. Bolten, M. Schmidt, et al. “Mobility in graphene double gate field effect transistors”, *Solid-State Electronics*, vol. 52(4), pp. 514–8, 2008.

[59] B. Obradovic, R. Kotlyar, F. Heinz, P. Matagne, T. Rakshit, M.D. Giles, M.A. Stettler, “Analysis of graphene nanoribbons as a

channel material for field-effect transistors", *Applied Physics Letters*, vol. 88, n.14, p. 142102, 2006

[60] I. Meric, M. Y. Han, A. F. Young, B. Ozyilmaz, P. Kim, and K. L. Shepard, "Current saturation in zero-bandgap, top-gated graphene field-effect transistors," *Nature Nanotechnology*, vol. 3, no. 11, pp. 654–659, Nov. 2008.

[61] I. Imperiale, S. Bonsignore, A. Gnudi, E. Gnani, S. Reggiani, and G. Baccarani, "Computational study of graphene nanoribbon FETs for RF applications," *in IEDM Tech Digest*, pp. 732–735, 2010

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

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

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

[65] F. Driussi, P. Palestri, and L. Selmi, "Modeling, simulation and design of the vertical graphene base transistor," *Microelectron. Eng.*, vol. 109, pp. 338–341, Sep. 2013.

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

[67] S. Venica, F. Driussi, P. Palestri, D. Esseni, S. Vaziri, L. Selmi, "Simulation of DC and RF Performance of the Graphene Base Transistor", *IEEE Trans. On Electron Devices*, Vol. 61, no. 7, pp. 2570- 2576, 2014.

[68] M. V. Fischetti, D. J. DiMaria, S. D. Brorson, T. N. Theis, J. R. Kirtley, "Theory of High-Field Electron Transport in Silicon Dioxide", *Physical Review B*, vol. 31, pp.8124–8142, 1985.

PAPERS

Preparation and characterization of SERS substrates: from colloids to solid substrates

L. Mikac[*], M. Ivanda[*], M. Gotić[**], D. Ristić[*], V. Đerek[*], H. Gebavi[*], P.G. Gucciardi[***], S. Trusso[***] and C. D'Andrea[***]

[*]*Center of Excellence for Advanced Materials and Sensing Devices, Ruđer Bošković Institute, Bijenička c. 54, Zagreb, Croatia*

[**]*Laboratory for Synthesis of New Materials, Ruđer Bošković Institute, Bijenička c. 54, Zagreb, Croatia*
[***]*CNR IPCF- Istituto per i Processi Chimico-Fisici, Viale F. Stagno D'Alcontres 37, I-98156, Messina, Italy*
ivanda@irb.hr

Abstract. – Surface enhanced Raman spectroscopy is an important technique for detection of wide number of analytes. Today, there is a growing need for the development of stable, sensitive, reproducible and portable SERS-active substrates for use in different fields. Some of the most extensively used SERS substrates are metal colloids. However, the use of the colloid solutions for SERS deals with important issues, like stability and reproducibility. Other popular types of substrates are rough or nanoporous surfaces, such as porous silicon, coated with noble metals. Porous silicon (pSi) is a semiconducting material typically obtained by electrochemical etching in hydrofluoric acid (HF). Due to its high surface-to-volume ratio pSi is an interesting material for photonic and sensing devices, drug delivery systems as well as for use in SERS. In this paper, attempts are made to prepare a stable and uniform noble metal colloidal solution as well as solid substrate using porous silicon (pSi) with deposited silver and gold nanostructures. SERS substrates obtained by different methods are evaluated for SERS efficiency using methylene blue and rhodamine 6G (R6G) at 514.5 and 633 nm. The prepared substrates showed good stability and reproducibility. For some substrates the concentrations detected were below nanomolar range.

I. INTRODUCTION

Raman spectroscopy is an important technique for identification as well as for structural characterization of wide number of analytes. Raman scattering is known to be a very inefficient process due to its low scattering cross-section. In order to improve the sensitivity of the process, surface enhanced Raman spectroscopy (SERS) has been used for several decades [1]. A lot has been written about the mechanism of the SERS but it is now widely accepted that there are two main factors affecting SERS: electromagnetic and chemical [2, 3]. In the electromagnetic mechanism, the laser light excites collective oscillations of the surface electrons in a metal which results in the enhancement of the local electromagnetic field. The molecule which is near the metal surface or adsorbed to it, experiences this enhancement and as a result the intensity of the Raman scattered light is enhanced. In the chemical mechanism, the polarizability of the molecule adsorbed on the metal surface increases. The total Raman signal enhancement is the product of both processes. Silver and gold are very good for fabrication of SERS substrates because of intense surface plasmon absorption in the visible region of the spectra.

Today, there is a growing need for the development of stable, sensitive, reproducible and portable SERS-active substrates for use in different fields. Some of the most extensively used SERS substrates are metal colloids [4, 5] because of their simplicity of preparation (in-situ reduction methods) and strong Raman enhancement. However, the use of the colloid solutions for SERS deals with important issues, like stability and reproducibility [6]. Nanostructures can also be fabricated electrochemically [7] and with methods that allow high organization in the structure, like lithography methods [8-10] and Langmuir-Blodgett technique [11]. Other popular types of substrates are rough or nanoporous surfaces, such as porous silicon, coated with noble metals [12]. Porous silicon (pSi) is a semiconducting material typically obtained by electrochemical etching in hydrofluoric acid (HF) [13, 14]. After the etching, material is consisted of a network of pores with different diameters and depth, depending on the etching conditions. Due to its high surface-to-volume ratio this is an interesting material for photonic and sensing devices, drug delivery systems as well as for use in SERS [15-18].

II. MATERIALS AND METHODS

The metal colloidal solutions were prepared using chemical reduction method. Ascorbic acid, tri-sodium citrate and ice-cold sodium borohydride were used was as reducing agents and citrate was used as stabilizer.

Silicon wafers (100) of p-type (boron-doped), with a resistivity of ~20 Ωcm and 525±25 μm thickness, were cut into ~2.5 x 2.5 cm pieces and degreased by sonication in acetone and ethanol and rinsed with ultra pure water. Electrochemical experiments were performed with home-made two-electrode Teflon cell in 12 %wt hydrofluoric acid solution. The constant current densities used were 4 and 10 mA/cm^2 for 15 or 30 minutes in 12 wt% HF acid. Silver was deposited on the surface of pSi by immersing the pSi

samples in 0.01 and 0.001 M aqueous silver nitrate ($AgNO_3$) solution, by drop casting or by pulsed laser ablation.

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

The morphology of silver nanoparticles was investigated with Jeol JSM 7000F field emission scanning electron microscope (FE SEM) coupled with Energy-dispersive spectroscopy (EDS). The particle size distribution was determined from corresponding SEM micrographs using ImageJ program and modified level set algorithm.

Surface-enhanced Raman spectra were recorded by Horiba Jobin Yvon HR800 and T64000 micro Raman systems. The exciting lines were 514.5 (Ar laser) and 632.8 (He-Ne laser). In order to reduce the photo degradation of the analyte, filters were used, and the laser excitation power was kept low. Prior to recording the Raman spectra, the samples were soaked in aqueous solutions of testing molecules at different concentrations for 30 minutes. As testing molecules pyridine, rhodamine 6G (R6G) and methylene blue (MB) were used.

III. RESULTS AND DISCUSSION

A. SERS in colloidal solutions

A series of silver colloids was prepared by using different reducing agents. They were investigated with respect to nanoparticles morphology and SERS activity [19]. The silver particles with 10–40 nm size and spherical shape were prepared by reducing silver nitrate with citrate, ascorbic acid and $NaBH_4$ in the presence of stabilizing agents. Thus obtained silver nanoparticles were used as SERS substrates with pyridine as analyte. The best enhancement of the Raman signal was achieved when sodium borohydride (0.1M) was used as aggregating agent in citrate reduced silver colloid (Fig. 1). The SERS enhancement factors for pyridine were calculated taking into account the number of molecules on the silver nanoparticles in laser activation volume and for the best results was estimated to be of the order 10^6.

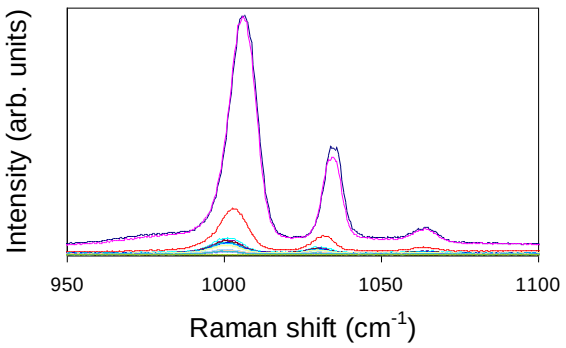


Figure 1. SERS spectra of pyridine in citrate-reduced Ag colloid sample recorded one after another during 25 minutes after addition of 0.1M $NaBH_4$ (the highest peak represents the longest time) [19].

B. SERS in solid substrates.

The electrochemical etching of silicon samples for both current densities and etching times resulted in the formation of macro pores with ~1 μm width and variable lengths. The etched samples were coated with noble metals in several ways: a) immersion in $AgNO_3$ solution, b) drop casting silver colloidal solution and c) laser ablation with silver and gold. The formation of the silver layer in all cases could be confirmed visually. A SEM micrographs of the SERS substrates obtained by immersion of pSi etched 30 min at 4 mA/cm^2 in 0.01 M $AgNO_3$ solution reveal that the silver layer is composed of homogeneously distributed nanoparticles that cover the surface leaving the macro pores open. The average size of the nanoparticles is around 100 nm but there are also some aggregates. On the SEM micrograph of the pSi with drop casted silver colloidal solution micron scale agglomerates of silver nanoparticles formed at the entrances of the pores are observed. The average size of the single silver nanoparticles is around 80 nm, but some smaller particles are present as well. The mean silver particle size on macroporous Si surface for both silver ablations and gold ablation was calculated to be in the range from 20 to 40 nm.

The efficiency of pSi templates with metal nanoparticles was tested using dye molecules, methylene blue and rhodamine 6G at different wavelengths.

In the case of immersion of pSi sample in silver nitrate solution the lowest R6G concentration detected at 514.5 nm was 10^{-9}M (Fig. 2). When silver colloidal solution was applied to pSi template, the lowest R6G concentration that was detected was 10^{-8}M. Considering samples ablated with silver and gold, R6G was detected only in one sample to the concentration of 10^{-8}M.

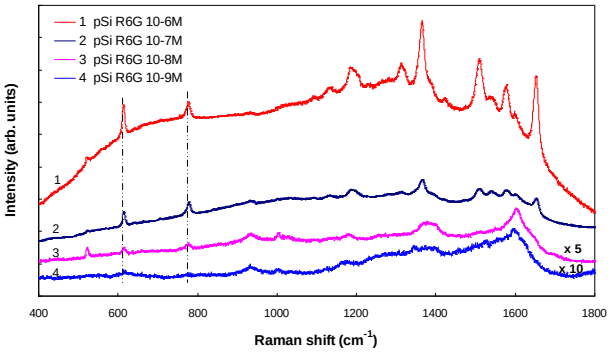


Figure 2. SERS spectra of of rhodamine 6G (concentrations 10^{-6}, 10^{-7}, 10^{-8} and 10^{-9} M from top to bottom) on Si (4 mA/cm^2; 30 min) with silver obtained by immersion plating; excitation line 514.5 nm; laser power 1 mW; accumulation time 10s.

Using 633 nm laser wavelength and MB the most pronounced SERS activity gave pSi samples ablated with 45000 pulses. For these samples nice spectra for 10^{-10}M concentration of MB were obtained (Fig. 3). The detection

limit of 10^{-10}M was also reached for 4 mA/cm^2 - 15 min pSi sample, silver ablated with 30000 pulses. For the gold ablated pSi samples, the lowest concentration detected was barely 10^{-8}M.

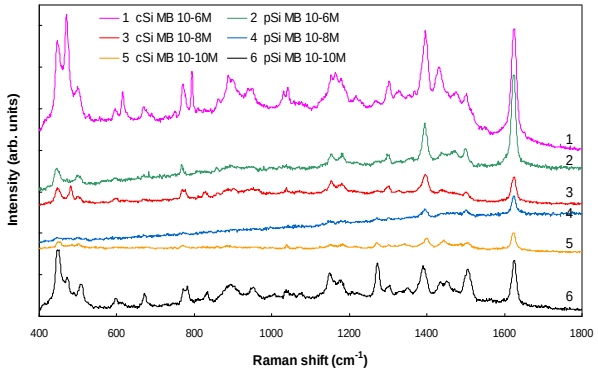


Figure 3. Raman spectra of methylene blue (concentrations 10^{-6}, 10^{-8} and 10^{-10} M from top to bottom) on macroporous Si (4 mA/cm^2; 30 min) and crystal Si (cSi) samples ablated with silver (45000 pulses); excitation line 633 nm; laser power 6µW; accum. time 10s.

IV. CONCLUSION

In conclusion, stable and reproducible SERS substrates were successfully prepared with different methods. First, silver colloids were used as SERS substrate with pyridine as analyte. The SERS enhancement factors for pyridine were estimated to be of the order 10^6. Second, porous silicon produced by anodisation of p-type silicon in hydrofluoric acid was used as a template for Ag and Au depositions. The SERS efficiency of the different pSi substrates was evaluated in terms of limit of detection (LOD) using methylene blue (MB) and rhodamine 6G (R6G) as probe molecules, and laser excitations at 514.5 and 633 nm. LODs below the nanomolar range are reproducibly obtained. Porous silicon proved to be a good base for the preparation of SERS substrates.

REFERENCES

1. Haynes, C.L., et al., Surface-enhanced Raman sensors: Early history and the development of sensors for quantitative biowarfare agent and glucose detection. J. Raman Spectrosc., 2005. 36(6-7): p. 471-484.
2. Schatz, G.C., M.A. Young, and R.P. Van Duyne, Electromagnetic mechanism of SERS, Top. Appl. Phys. 2006. p. 19-46.
3. Morton, S.M. and L. Jensen, Understanding the molecule-surface chemical coupling in SERS. J. Am. Chem. Soc., 2009. 131(11): p. 4090-4098.
4. Rivas, L., et al., Mixed silver/gold colloids: a study of their formation, morphology, and surface-enhanced Raman activity. Langmuir, 2000. 16(25): p. 9722-9728.
5. Guerrero-MartÃ-nez, A., et al., Nanostars shine bright for you. Colloidal synthesis, properties and applications of branched metallic nanoparticles. Curr. Opin. Colloid In. 16(2): p. 118-127.
6. Tantra, R., R.J.C. Brown, and M.J.T. Milton, Strategy to improve the reproducibility of colloidal SERS. J. Raman Spectrosc., 2007. 38(11): p. 1469-1479.
7. Finot, M.O., G.D. Braybrook, and M.T. McDermott, Characterization of electrochemically deposited gold nanocrystals on glassycarbon electrodes. J. Electroanal. Chem., 1999. 466(2): p. 234-241.
8. Abu Hatab, N.A., J.M. Oran, and M.J. Sepaniak, Surface-enhanced Raman spectroscopy substrates created via electron beam lithography and nanotransfer printing. ACS Nano, 2008. 2(2): p. 377-385.
9. Green, M. and F.M. Liu, SERS Substrates Fabricated by Island Lithography: The Silver/Pyridine System. J. Phys. Chem. B, 2003. 107(47): p. 13015-13021.
10. Le Ru, E.C., et al., Surface enhanced Raman spectroscopy on nanolithography-prepared substrates. Curr. Appl. Phys., 2008. 8(3-4): p. 467-470.
11. Tao, A., et al., Langmuir-Blodgett silver nanowire monolayers for molecular sensing using surface-enhanced Raman spectroscopy. Nano Lett, 2003. 3(9): p. 1229-1233.
12. Lin, H., et al., Surface-enhanced Raman scattering from silver-plated porous silicon. J. Phys. Chem. B, 2004. 108(31): p. 11654-11659.
13. Ramírez-Porras, F.B.D.a.A., Surface-Functionalized Porous Silicon Wafers: Synthesis and Applications, Advances in Chemical Sensors, ed. P.W. Wang. 2012: InTech.
14. Sailor, M.J., Porous Silicon in Practice: Preparation, Characterization, and Applications, ed. Wiley-VCH. 2012, Weinheim, Germany.
15. Stewart, M.P. and J.M. Buriak, Chemical and biological applications of porous silicon technology. Adv. Mater., 2000. 12(12): p. 859-869.
16. Schechter, I., M. Ben-Chorin, and A. Kux, Gas sensing properties of porous silicon. Anal. Chem., 1995. 67(20): p. 3727-3732.
17. Anglin, E.J., et al., Porous silicon in drug delivery devices and materials. Adv. Drug Deliver Rev., 2008. 60(11): p. 1266-1277.
18. Chursanova, M.V., et al., Optimization of porous silicon preparation technology for SERS applications. Appl. Surf. Sci., 2010. 256(11): p. 3369-3373.
19. Mikac, L., et al, Synthesis and characterization of silver colloidal nanoparticles with different coatings for SERS application. J. Nanopart. Res., 2014 (16): p.2748-2761

The conductivity and TEMF of MoS_2 with Mo_2S_3 additive

G.E. Yakovleva[1], A. S. Berdinsky [1], A. I. Romanenko[2], S. P. Khabarov[1] and V.E. Fedorov[2]
[1] Novosibirsk State Technical University/Semiconductor Devices & Microelectronics, Novosibirsk, Russia
[2] Nikolaev Institute of Inorganic Chemistry, Russian Academy of Sciences, Novosibirsk, Russia
e-mail address: berdinsky.alexander@gmail.com

Abstract - Transition-metal chalcogenides are prospective thermoelectric materials. One of them is a molybdenum disulfide MoS_2, which has a layered structure. MoS_2 has a good potential to have a high value of thermoelectric quality factor due to a high value of thermo-EMF and low value of thermal conductivity. But a low value of its electrical conductivity suppresses the thermoelectric quality factor ZT on the level of 0.1 at high temperatures. Present work shows the influence of metal addition to MoS_2 on electrical conductivity and Seebeck coefficient (SC) of final mixture. Mo_2S_3 was chosen as a metal addition. Mo_2S_3 has an electrical conductivity of 330 S/m and thermo-EMF of 10 μV/K at 300K. Bulk powder samples of MoS_2 with addition of 3, 6, 10, 30 and 60 wt% Mo_2S_3 were studied. An electrical conductivity of samples was measured in temperature range: 77 K – 423 K. All samples have shown semiconductor hopping conductivity with variable hopping length. The SC was measured in the temperature range of 300K - 500K. The addition of Mo_2S_3 decrease SC from 300μV/K to 75μV/K.

I. INTRODUCTION

Thermoelectric materials have a great interest for study. Environmental problems associated with non-renewable natural resources and environmental pollution are the cause of such interest [1].

The effectiveness of a thermoelectric material is characterized by dimensionless thermoelectric quality factor ZT, which is calculated by the formula "(1)".

$$ZT = S^2\sigma T/\lambda, \quad (1)$$

where S – Seebeck coefficient, σ – electrical conductivity, λ – thermal conductivity, T – temperature [2].

The search of new materials with high ZT consists of research of completely new composite materials.

Thermoelectric materials have a different temperature range where the quality factor reaches its maximum. Therefore thermoelectric materials are classified as low, middle or high temperature materials according to a temperature range with a highest value of quality factor.

Such materials as transition-metal chalcogenides are of great interest in recent time. One of the brightest representatives of such materials is a molybdenum disulfide. Molybdenum disulfide has its own distinctive electronic, optical and catalytic properties [3].

Molybdenum disulfide has a layered structure. Single-layer in disulfide molybdenum is a two-dimensional quasi-crystal, which consists of close-packed layer of molybdenum in between of two close-packed layers of sulfur S-Mo-S. Layers are held together due to weak Van der Waals forces, whereas atomic structures of layers are tied together by strong covalent forces. This material is structurally similar to graphene and hexagonal boron nitride (h-BN) [4].

Recent reports demonstrated thermopower of MoS_2 in a range of 600 to 700 μV/K at a high temperature. Electrical and thermal conductivity of MoS_2 have different values in a various sources. The value of the electrical conductivity is in a range of 0.01 to 10 S/m at room temperature. The value of thermal conductivity is in a range of 0.1 to 1 W/(m*K) [5-7].

Molybdenum disulfide has a potential to have a high value of thermoelectric quality factor due to high value of thermo-EMF and low value of thermal conductivity. But the low value of electrical conductivity suppresses its thermoelectric quality factor on the level of 0.1 at high temperatures [8].

Present work shows the influence of metallic addition to molybdenum disulfide on its electrical conductivity and Seebeck coefficient. Molybdenum(III) sulfide Mo_2S_3 was chosen as a metallic addition. Mo_2S_3 has an electrical conductivity of 330 S/m and thermopower of 10 μV/K at room temperature.

II. EXPERIMENTAL

A. Preparation of sample

MoS_2 and Mo_2S_3 were synthesized by a standard high temperature ampule method using high purity elements in stoichiometric ratio. The composite samples are prepared by means of intimately mixing molybdenum disulfide with 3, 6, 10, 30, 60 wt% Mo_2S_3.

The powdered composite materials were pressed to obtain 12 mm in diameter pellets-shape samples. The samples shape was further trimmed to rectangular form for study of thermoelectric and electronic properties. Such shape of the samples was obtained by means of diamond cutter. The ohmic contact to the samples was obtained using silver paste.

B. Electrical properties

The samples under study have n-type conductivity. Seebeck coefficient of the samples was measured at the temperature range of 300 to 500 K. The results on Seebeck coefficient are presented in Fig.1.

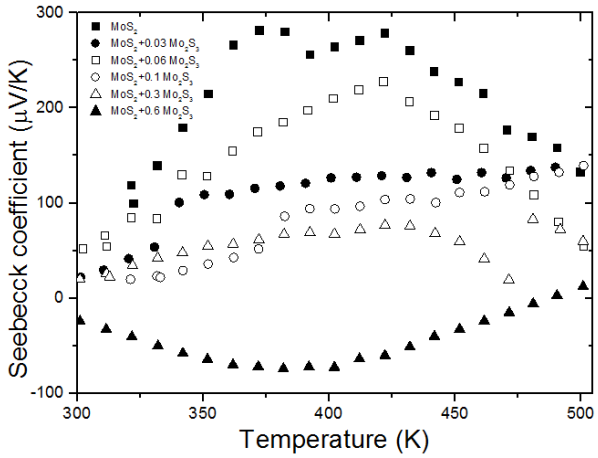


Figure 1. The Seebeck coefficient as function of temperature.

Seebeck coefficient of pure molybdenum disulfide has value of 100 to 300 μV/K over a temperature range of 300 to 500 K. It is observed that Seebeck coefficient is decreased with introduction of Mo_2S_3 to the MoS_2. Introduction of 60wt% of Mo_2S_3 has the highest effect on the Seebeck coefficient of the composite. At 60wt% of Mo_2S_3 Seebeck coefficient became negative which means than composite changed its conductivity to p-type.

Electrical conductivity was measured at the temperature in the range of 77 to 500 K. The measured results are displayed in Fig.2.

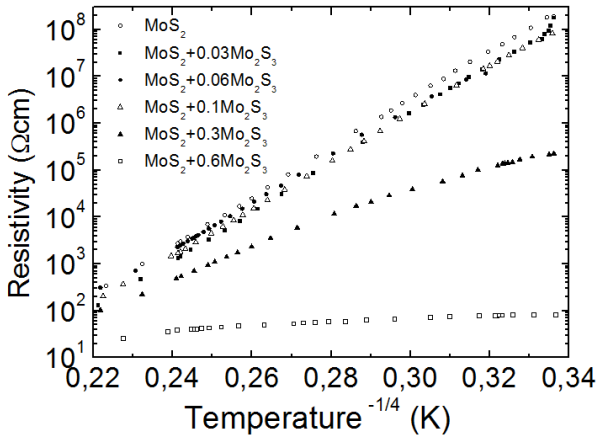


Figure 2. The resistivity as function of temperature.

The results of electrical conductivity measurement showed that the samples with addition of 60wt% of Mo_2S_3 have a higher electrical conductivity. The results are proved by the fact that Mo_2S_3 is a metal with high electrical conductivity. The results in Fig.2 clearly shown, that MoS_2 sample itself as well as all sample with Mo_2S_3 additive demonstrated 3D hopping conductivity with variable range hopping .

Thus all the samples belong to family of materials, where hopping conductivity with variable range hopping (VRH) is realized. Such type of conductivity was discussed by Mott in 1968 [9]. According to Mott's law, the resistivity varies with temperature by the formula "(2)".

$$\rho(T)=\rho_0 \cdot \exp[(T_0/T)^{1/4}] \quad (2)$$

where ρ_0 is the resistivity at infinite temperature and T_0 is the characteristic temperature

Thermoelectric power factor of the samples was calculated by the formula "(3)".

$$P = S^2\sigma \quad (3)$$

where S – the Seebeck coefficient, σ – electrical conductivity.

The results of the thermoelectric power factor calculation are presented in Fig. 3. One can see that addition of 60wt% Mo_2S_3 dramatically changes the thermoelectric power factor in comparison with MoS_2, e.g. thermoelectric power factor increased in 3.5 times in the average range of the temperatures. The smaller additions of Mo_2S_3 show the trend in decreasing of thermoelectric power factor.

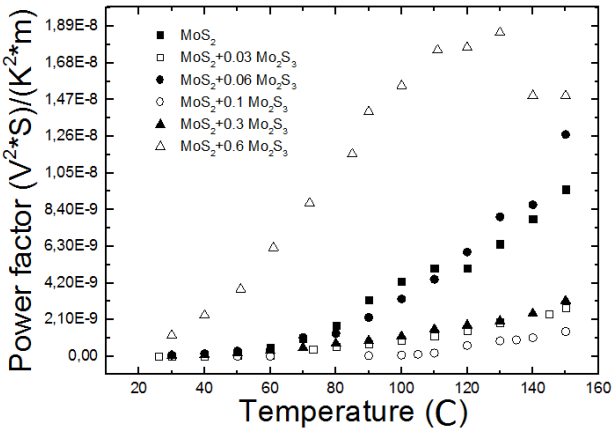


Figure 3. The power factor as function of temperature.

III. CONLUSION

We have measured MoS_2 pellet-shape pressed samples with different amount of Mo_2S_3 additive for analysis of TEMF and conductivity. It was shown that addition of Mo_2S_3 decreases Seebeck coefficient from 300μV/K to 75μV/K. The addition of 60wt% Mo_2S_3 gives the advantage in thermoelectric power factor.

It was shown that all samples have disordered structure with 3D hopping type of conductivity with variable hopping length. It was also shown that the temperature dependence of the samples omit Mott's law. By the variation in MoS_2/Mo_2S_3 composites is possible to change thermoelectric and conductive properties by controlling of Mo_2S_3 contents.

ACKNOWLEDGMENT

This study was supported by Russian Scientific Foundation, Project 14-13-00674. The authors are gratefully acknowledges to Dr. Kozeeva L.P. for the help in samples preparing.

REFERENCES

[1] Jong-Young Kim, Soon Mok Choi, Won-Seon Seo, and Woo-Seok Cho, " Thermal and electronic properties of exfoliated metal chalcogenides," Korean Chem. Soc. Korea, vol. 31, pp. 3225 – 3227, September 2010.

[2] V. L. Kalikhman and Ya. S. Umanskii, " Transition – metal chalcogenides with layer structures and features of the filling of their brillouin zones," Soviet physics uspekhi. Russia, vol. 15, pp. 728 -741, June 1973.

[3] Wen Huang, Haixia Da, and Gengchiau Liang, " Thermoelectric performance of MX_2 (M=Mo,W; X = S, Se) monolayaes," Journal of applied physics. Singapore, vol. 113, pp. 1 -7, March 2013.

[4] G. H. Lee, Y. J. Cui, N. Petrone, C. H. Lee, M. S. Choi, D. Y. Lee, C. Lee, W. J. Yoo, K. Watanabe, T. Taniguchi, C. Nuckolls, P. Kim and J. Hone, " Flexible and transparent MoS_2 field – effect transistors on hexagonal boron nitride – graphene heterostructures," ASCNANO, vol. 7, pp. 7931 -7936, August 2013.

[5] S. K. Srivastava and B. N. Avasthi, " Indium intercalation compounds of molybdenum disulfide; In_xMoS_2 ($0 \leq x \leq 1$)," Synthetic Metals. India, vol. 11, pp. 193-205, February 1985.

[6] S. K. Srivastava and B. N. Avasthi, " Preparation and characterization of molybbdenum disulphide catalysts," Journal of materials science. India, vol. 28, pp. 5032-5035.

[7] X. Liu, G. Zhang, Q. X. Pei, and Y. W. Zhang, " Phonon thermal conductivity of monolayer MoS_2 sheet and nanoribbons," Applied physics letters. Singapore, vol. 103, pp. 1 – 5, September 2013.

[8] Guo Huai-Hong, Yang Teng, Tao Peng, and Zhang Zhi-Dong, " Theoretical study of thermoelectic properties of MoS2," Chin. Phys. B. China, vol. 23, pp. 1 – 7, November 2013.

[9] N. F. Mott and E. A. Davis, " Electronic processes in non-crystalline materials," Clarendon, Oxford,1979.

Strain-sensing Element Based on Layered Sulfide $Mo_{0.95}Re_{0.05}S_2$

V.A. Kuznetsov[*,***a], A.S. Berdinsky[*], A.Yu. Ledneva[***], S.B. Artemkina[**,***], M.S. Tarasenko[***] and V.E. Fedorov[**,***b]

[*] Novosibirsk State Technical University, Novosibirsk, Russia

[**] Novosibirsk State University, Novosibirsk, Russia

[***] Nikolaev Institute of Inorganic Chemistry, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia

[a]e-mail address: vitalii.a.kuznetsov@gmail.com

[b]e-mail address: fed@niic.nsc.ru

Abstract – The work presents a study of piezoresistive effect of thin films of rhenium-doped molybdenum disulfide of composition $Mo_{0.95}Re_{0.05}S_2$ possessing layered structure of 2H-MoS_2 type. The compound was synthesized by high temperature ampoule method using stoichiometric mixture of elements. The thin films were formed by spraying colloidal dispersion produced by liquid exfoliation of solid phase in mixed solvent EtOH/H_2O. The phase $Mo_{0.95}Re_{0.05}S_2$ is semiconductor with effective activation energy of about 150 meV estimated from temperature dependences of resistance of the films. The dependence of resistance on deformation is presented. It has almost linear type and good reproducibility during a sinusoidal alternating load of 0.12% deformation amplitude. The strain gauge factor of 12 and 24 was obtained depending on preparation details from room temperature measurements using a beam of uniform strength (in bending).

I. Introduction

The piezoresistive effect is a change in the electric resistance of materials when mechanical strain is applied. The piezoresistive effect is pronounced one for semiconductors: for example, strain gauge factor of even polycrystalline and amorphous silicon can vary from -30 to +40, about a third that of single-crystal silicon, that are an order of magnitude larger than those observed in metals [1, 2]. Very high strain gauge factor of semiconductors determine heightened interest of investigators to new semiconductor materials.

MoS_2 is a bright representative of the layered family of transition metals dichalcogenides (TMDC) possessing semiconducting electronic properties as well as high thermal and mechanical stability [3]. The properties of 2H-MoS_2 were studied on the bulk state, but recently, a series of publications was devoted to exfoliated samples – nanosheets, mono- and bilayer MoS_2 that make them attractive for flexible electronic applications [4-7]. Because pure MoS_2 is characterized by rather high resistivity, in our work we used rhenium-doped sulfide $Mo_{0.95}Re_{0.05}S_2$. We showed recently [8] that substitution of 5-15% of molybdenum atoms by electron-rich rhenium ones results in changes of resistivity compared to non-substituted MoS_2. For instance, pressed powder sample of composition $Mo_{0.95}Re_{0.05}S_2$ revealed significant drop in the resistivity (by c.a. 10^4 times) and decrease of Δ_{eff} (the logarithmic derivative of electrical resistance with respect to the inverse temperature) in 4 times down to 0.086 eV at 300 K.

In this work we study a piezoresistive effect of $Mo_{0.95}Re_{0.05}S_2$ film samples prepared from its colloidal dispersion. The resistance temperature characteristics of the samples were measured as well.

This work was supported by the Russian Science Foundation (Grant № 14-13-00674).

II. Experimental

A. Preparation of $Mo_{0.95}Re_{0.05}S_2$ film samples

$Mo_{0.95}Re_{0.05}S_2$ was synthesized by high temperature ampoule method from stoichiometric mixture of Mo/Re/S and characterized by X-ray powder diffraction method. Thin films were formed by filtering of the $Mo_{0.95}Re_{0.05}S_2$ colloidal dispersion that was produced by exfoliation of solid phase in mixture solvent (45% EtOH + 55% H_2O) under ultrasonic treatment (Fig. 1). The XPD patterns of bulk material and thin film were indexed in hexagonal sp. gr. P63/*mmc* (No. 194), as 2H-MoS_2.

The particle sizes forming colloidal dispersions were of few layers thick with lateral dimensions of 50-200 nm (see Fig. 2). Six samples (from **1s** to **6s**, Table 1) for measuring the dependences of resistance on strain were prepared by spray method using preheated to 200°C steel beams coated with the polymer glue VL-931. Herewith **5s** and **6s** were influenced in situ with the ethanol-water solution in an attempt to form homogeneous films, because the films' forming usually proceeded with forming of bottom layers and islands not more than 0.3 mm in lateral dimension placed onto the layers. Such forming is distantly similar in growth to Stranski-Krastanov mechanism [9]. The beams with the samples were dried in air at 60°C for 3 hours. The samples **3s** – **6s** were covered with the polymer glue at room temperature followed by 80°C and 120°C dryings.

B. Resistance-strain measurements

A beam of uniform strength (in bending) was chosen for studying the piezoresistive effect (see Fig. 3). The feature of the beam lies in invariable strain ε throughout the trapezoidal section of the beam [10]. The beams were

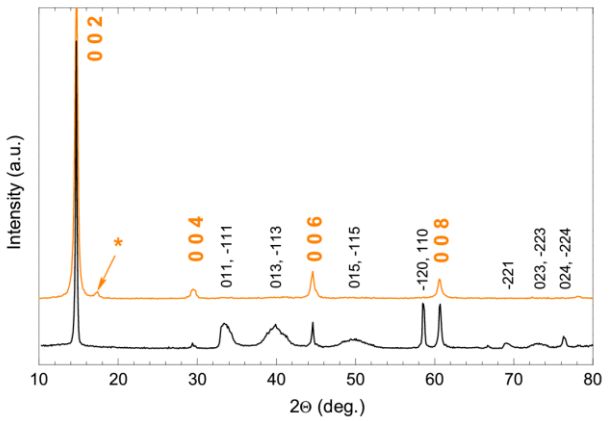


Figure 1. X-ray powder diffractogramms of bulk $Mo_{0.95}Re_{0.05}S_2$ (bottom line) and $Mo_{0.95}Re_{0.05}S_2$ film formed by filtering (upper line)

made of spring steel: thickness h = 0.5 mm; Young modulus E = $2.15 \cdot 10^{11}$ Pa; Poisson ratio μ = 0.3; linear temperature expansion coefficient α = $12.3 \cdot 10^{-6}$ K^{-1} (from 293 K to 393 K). The beams themselves had the following dimensions: the length of the place pinched – 12 mm; the width – 12 mm; the length of the trapezoidal section (section of uniform strain) – 43 mm; the length of the narrow part – 25.5 mm; the total length – 80.5 mm; the distance from the pinched line to the vertex of the triangle – 66 mm. The beams were bent by the pusher as it is shown on Fig. 3.

The dependences of resistance on strain in loading and unloading regimes as function of relative resistance variation on strain of the samples are shown in Fig. 4. The dependences for all samples are of similar shapes. One can see that the curves received are almost linear and the strain gauge factors for **1s** – **4s** and **5s** – **6s** samples have the same values within the error.

The strain gauge factor was defined as slope of connecting line between the two extreme values (values of relative resistance at the positive and negative strain maxima) by a well-known formula [11]:

$$K = \frac{\Delta R / R_{\varepsilon(\min)}}{\varepsilon}, \qquad (1)$$

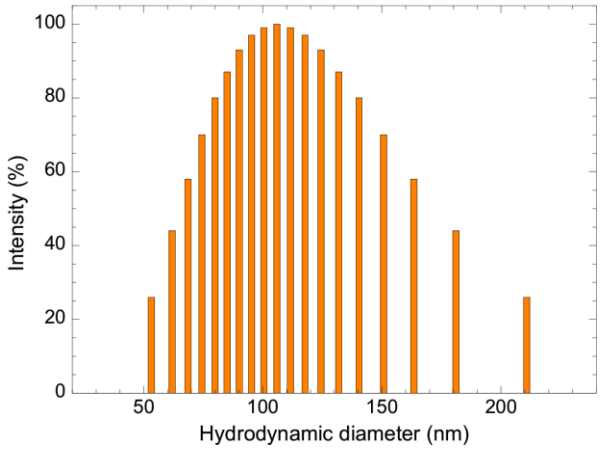


Figure 2. The size particles distribution in colloidal dispersion of $Mo_{0.95}Re_{0.05}S_2$ / 45% EtOH + 55% H_2O

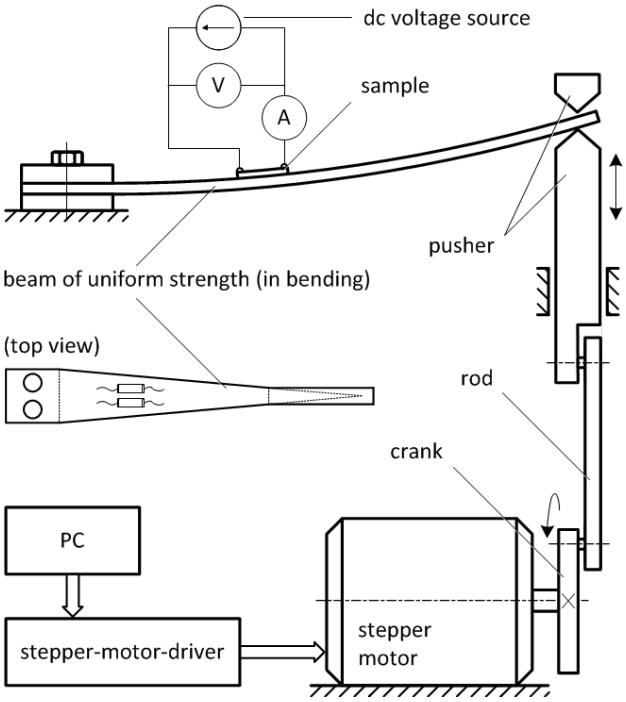


Figure 3. The device for measuring a strain gauge factor and for endurance test

where $\Delta R = R_{\varepsilon(\max)} - R_{\varepsilon(\min)}$; $\varepsilon = \varepsilon_{\max} - \varepsilon_{\min}$ – the strain leading to change of resistance ΔR; $R_{\varepsilon(\max)}$ and $R_{\varepsilon(\min)}$ are resistances at $\varepsilon_{\max}$ and $\varepsilon_{\min}$, respectively. To obtain the strain gauge factors the beams were loaded with the deflection increments about 2 mm. The strain values were calculated by ANSYS software.

The slope difference received for all samples might have appeared because of some reasons. First one probably is different forming technique, especially for **5s** and **6s** samples. The second reason may be various thicknesses of insulating bonding glue. Finally, the third reason may be that the samples **1s** and **2s** had not a covering in contrast to the other ones. In spite of the above-stated, one can see that neither the covering nor the possible glue thicknesses difference have affected the strain gauge factor, as the factors for **1s** – **4s** have the same values within the error (see Table 1). But at the same time the additional treatment of films of **5s** and **6s** samples has changed the sensitivity, and it will be discussed in the next subsection.

One can see the hysteresis for each sample in Fig. 4. The hysteresis may be concerned mainly with residual resistance changes due to geometrical and structural changes of the films in plastic strain range of them. One more reason may be hysteresis in the glue. The hysteresis for **5s** and **6s** samples may be essentially due to low accuracy rate because of their high resistances.

The error of the strain gauge factors was evaluated by the formula (2):

$$\Delta K = \sqrt{\sum_i \left(\frac{\partial K}{\partial X_i}\right)^2 \Delta X_i^2}, \qquad (2)$$

where X_i – are the current and voltage accuracy specifications of the measuring instruments used and strain accura-

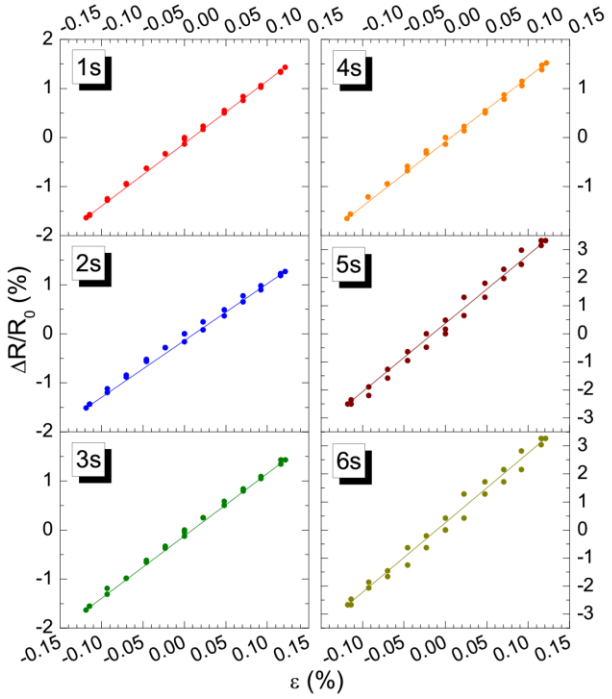


Figure 4. Resistance-strain characteristics of all samples at room temperature

cy. But at the same time, the errors for **3s** – **6s** were evaluated without considering of additive error of the current meter. The result is the strain gauge factor equal to 12.5±1.6 and 24.4±1.4 for **1s** – **4s** and **5s** – **6s** samples, respectively, with linear resistance-strain curves.

To examine the samples for endurance the beams were bent by sinusoidal alternating load of 0.12% deformation amplitude with 3 Hz the oscillation frequency. The samples were tested after 10, 30, 100, 300,… up to 100'000 load cycles [11]. The dependences of the strain gauge factors on the number of load cycles are shown in Fig. 5. As one can see there are not significant degradations of K along the range of number of cycles within the errors.

The piezoresistive effect of semiconductor and metallic resistors is related to changes of specific resistivity and geometric parameters respectively [1]. We cannot discuss here with confidence what makes the considered films sensing on strain because it requires an additional investigation which has not been carried out yet. The films of the samples consist of nanoparticles separated by insulating gaps and these films may be considered as disordered material. So we suppose the piezoresistive effect of the films may be concerned with a change of these gaps.

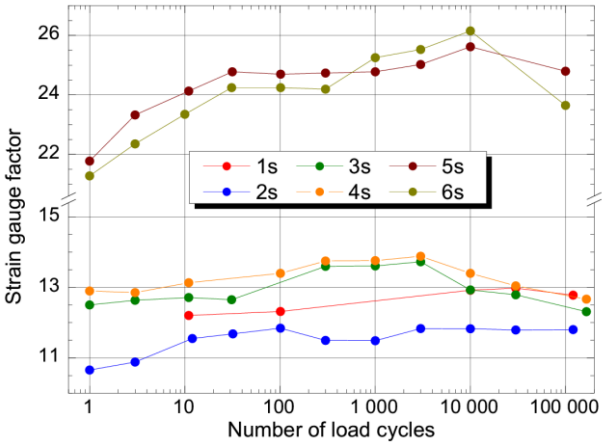


Figure 5. Dependences of the strain gauge factors on the number of load cycles. The oscillation strain amplitude is equal to 0.12%; the oscillation frequency is 3 Hz; $T \approx 296$ K

C. *Resistance temperature characteristics of the films*

To characterize the samples and especially the insulating gaps the temperature dependences of resistance of the samples have been measured (Fig. 6). These plots show that $R(T)$-dependences obtained are typical for semiconductors. The temperature dependences of resistance of **1s** and **2s** samples were measured over the range of temperature from 296 K to 453 K, but the dependences of the other samples were measured just to estimate the values of effective activation energies. The energies are defined as $\Delta_{eff} = d\ln(R)/d(1/T)$. The result values are shown in Table 1.

The greater effective activation energies of **5s** and **6s** samples correspond to larger average lengths of the insulating gaps between the particles in the films, as lengths of the insulating gaps in disordered material are characterized by energy barriers for electron tunneling [12-15]. Such difference in the energies may be also due to a deviation of average orientations of the nanoparticles relative to the beams surfaces that might be induced by the

TABLE I. SAMPLES' PARAMETERS

Sample	1s	2s	3s	4s	5s	6s
d, μm thickness	1.2	1.1	0.7	0.7	0.7	0.7
Covered with the polymer glue	–	–	+	+	+	+
Additional treatment with the solvent	–	–	–	–	+	+
$a \times b$, mm a – length between contacts b – width of the samples	2.0×2.6	2.5×2.5	8.3×2.5	8.5×2.5	7.7×2.5	7.8×2.5
Shape factor, a/b	0.77	1.00	3.32	3.40	3.08	3.12
R_S, MOhm/□ sheet resistance	10.0±0.4	19.0±0.6	64.5±1.8	66.7±2.0	258±21	330±35
K strain gauge factor	12.3±1.3	11.8±1.7	12.7±1.1	13.1±1.1	24.7±1.4	24.2±1.4
Δ_{eff}, meV effective activation energy	150	153	147	148	167	177

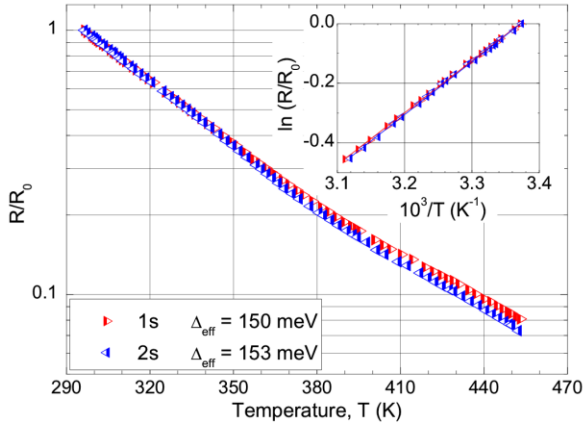


Figure 6. Resistance temperature characteristics of **1s** and **2s** samples

treatment of the samples **5s** and **6s** with the ethanol-water solution.

It is most likely in our case that just the orientation of the particles affected the factors, as electrical properties of layered materials strongly depend on crystal orientation (perpendicular and parallel to the *c* axis) [16]. Influence of the lengths of the insulating gaps on the factors is less likely than that of the orientation, because **5s** and **6s** have the different effective activation energies estimated and at the same time the same values of the factors within the error (see Table 1).

It is clear that the preparation details influence the properties of the resulting films including the strain sensitivity. We prepared six samples by spraying with different preparation details, but only the additional treatment of films of **5s** and **6s** samples has affected the strain gauge factor having made it higher. Anyway it needs an additional investigation.

III. Conlusion

The $Mo_{0.95}Re_{0.05}S_2$ thin films prepared by spraying colloidal dispersion have the strain gauge factors equal to 12 and 24 depending on the preparation details. The factor can be tuned by an additional treatment of films structure with the solvent.

Acknowledgment

The authors thank to Aleksey S. Cherkaev and Mikhail A. Chebanov (Novosibirsk State Technical University) for suggestions to measurements and calculating the beam strain by ANSYS software, respectively.

References

[1] M. Elwenspoek and R. J. Wiegerink, Mechanical microsensors. Berlin: Springer-Verlag Berlin Heidelberg, 2001.

[2] N. Maluf and K. Williams, An Introduction to Microelectromechanical Systems Engineering. Boston: Artech House Inc., 2004.

[3] C. N. R. Rao, H. S. S. R. Matte, and U. Maitra, "Graphene analogues of inorganic layered materials," Angewandte Chemie Int. Ed., vol. 52, pp. 13162-13185, 2013.

[4] M. Chhowalla, H. S. Shin, G. Eda, L.-J. Li, K. P. Loh, and H. Zhang, "The chemistry of two-dimensional layered transition metal dichalcogenide nanosheets," Nature chemistry, vol. 5, pp. 263-275, 2013.

[5] K.-G. Zhou, N.-N. Mao, H.-X. Wang, Y. Peng, and H.-L. Zhang, "A Mixed-Solvent Strategy for Efficient Exfoliation of Inorganic Graphene Analogues," Angewandte Chemie-International Edition, vol. 50, pp. 10839-10842, 2011.

[6] J. N. Coleman, M. Lotya, A. O'Neill, S. D. Bergin, P. J. King, U. Khan, et al., "Two-Dimensional Nanosheets Produced by Liquid Exfoliation of Layered Materials," Science, vol. 331, pp. 568-571, Feb 2011.

[7] V. Nicolosi, M. Chhowalla, M. G. Kanatzidis, M. S. Strano, and J. N. Coleman, "Liquid Exfoliation of Layered Materials," Science, vol. 340, pp. 1420-+, Jun 21 2013.

[8] V. E. Fedorov, N. G. Naumov, A. N. Lavrov, M. S. Tarasenko, S. B. Artemkina, and A. I. Romanenko, "Tuning electronic properties of molybdenum disulfide by a substitution in metal sublattice," in 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, CROATIA, 2013, pp. 11-14.

[9] J. A. Venables, Introduction to Surface and Thin Film Processes. Cambrige, UK ; New York: Cambridge University Press, 2000.

[10] E. A. Avallone, T. B. III, and A. M. Sadegh, Marks' standard handbook for mechanical engineers, 10 ed. New York: McGraw Hill Companies, Inc. , 2006.

[11] F. P. u. Strukturanalyse, "Experimental structure analysis - Metallic bonded resistance strain gages - Characteristics and testing conditions," in Characteristics to be tested vol. 2635, ed. Dusseldorf: Verein Deutscher Ingenieure, 2007, pp. 1-40.

[12] P. Sheng, "Fluctuation-Induced Tunneling Conduction in Disordered Materials," Physical Review B, vol. 21, pp. 2180-2195, 1980.

[13] P. Sheng, E. K. Sichel, and J. I. Gittleman, "Fluctuation-Induced Tunneling Conduction in Carbon-Polyvinylchloride Composites," Physical Review Letters, vol. 40, pp. 1197-1200, 1978.

[14] H. Xie and P. Sheng, "Fluctuation-induced tunneling conduction through nanoconstrictions," Physical Review B, vol. 79, p. 10, Apr 2009.

[15] Y.-R. Lai, K.-F. Yu, Y.-H. Lin, J.-C. Wu, and J.-J. Lin, "Observation of fluctuation-induced tunneling conduction in micrometer-sized tunnel junctions," AIP Advances, vol. 2, p. 6, 2012.

[16] J. A. Wilson and A. D. Yoffe, "The transition metal dichalcogenides discussion and interpretation of the observed optical, electrical and structural properties," Advances in Physics, vol. 18, pp. 193-335, 1969.

Electrical properties of composite silicon thin films

V.Tudić*

* Karlovac University of Applied Sciences /Department of Mechanical Engineering, Karlovac, Croatia

vladimir.tudic@vuka.hr

Abstract - Amorphous-nano-crystalline silicon composite thin film (a-nc-Si:H) sample were synthesized by Plasma Enhanced Chemical Vapor Deposition technique. The electrical properties of thin film and measurement of DC conductivities were accomplished using Dielectric spectroscopy (Impedance Spectroscopy) in wide frequency and temperature range. In analysis of impedance data, two approaches were tested: the Debye type equivalent circuit with two parallel R and CPE (constant phase elements) and modified one, with tree parallel R and CPE additionally interpreting crystal boundary effects. It was found that the later better fits to experimental results properly describing significant composite structure and crystal boundary dielectric effect with arbitrary hydrogen concentration indicating presence of strain. The amorphous matrix showed larger resistance and lower capacity than nano-crystal phase. Also it was found that composite silicon thin film cannot be properly describes by equivalent circuit only with resistors and constant phase elements in serial relation.

I. INTRODUCTION

It is natural that vast interest activation about silicon experimental and theoretical resources has been devoted to understanding its properties. Concerning third generation silicon thin film solar cells, its layers and device applications move inexorably to shorter length scales interface and surface characteristics gain in great importance. In analyze of silicon thin film properties it is desirable to explore the nature of chemical bonding not only the interface and surface of silicon layers but already include both static structural properties and kinetic aspects of its composites. A dark conductivity property of silicon thin film implies both aspects: static throe structural ordering issues and kinetic or dynamic properties revealed by impedance spectroscopy throe electrical properties.

The composite a-nc-Si:H thin films is mixed-phase material consisting of silicon nanocrystals embedded in amorphous silicon (a-Si:H) matrix. It has been the subject of intense research for variety of applications due to their unique optical and electrical properties [1]. Electrical and optical properties of a-nc-Si:H can be controlled by silicon nanocrystals size distribution and ratio of amorphous/nanocrystalline volume contribution. Measurements of the optical properties and structural analysis in previous published papers [2] showed that the spectral distribution of the absorption coefficient, in a wide range of crystal to amorphous fractions, can be maintained close to pure amorphous silicon in the visible part of the spectrum and showed square dependence on the photon energy. The average optical gap was larger for smaller nano-crystals and a higher crystal fraction *X*c just confirming the quantum size effects that correspond to quantum dots [3, 4].

Impedance spectroscopy (IS) has been widely used in last couple of years as a powerful technique for the study of the dielectric and conductivity behavior of ferroelectrics, ceramics, crystalline bulk and composite-crystalline silicon thin films [5, 6].

Because of silicon thin films composite form which is primarily characterized by the crystal fraction *X*c and crystal grains sizes, it is necessary to understand effects of complex micro structural features on the overall electrical properties of thin films. The resistance and capacitance of crystals grains and grain boundaries, which is frequency and temperature dependent, can be evaluated from IS spectra. This technique for determination of the starting values for the basic equivalent electrical circuit were used in the fitting procedure of experimental impedance data with simulated data in order to select most appropriate equivalent circuit. This technique enables us to separate the real and imaginary component of the complex impedance and related parameters, and hence provides information of the structure-property relationship in the rated sample.

The sample examined by IS technique with composite structures with presence of crystals grains and grain boundaries embedded in matrix demonstrate complex impedance common in the event of two phase materials with different conductivity or permittivity. At high temperatures may occur two successive semicircle arcs mostly as a rule. These properties can conventionally be displayed in a complex plane plots (Nyquist diagram) in terms of the same formalism, refer to [5, 7].

The micro-structural properties revealed by techniques like grazing incidence X-ray diffraction (GIXRD) described in [4], Raman spectroscopy (RS) or X-ray powder diffraction (XRD) [2] are fundamental in composite silicon thin films investigations but the correlation between microstructure and electrical properties is not always obvious or straightforward. Electrical properties may vary in intrinsic material and are intensive to minor variations in structure and composition or whether the properties are extrinsic since they could vary dramatically if impurities are present, especially if impurities to the grain boundaries [8]. Thanks to impedance spectroscopy and appropriate data analysis it is

possible to characterize the different electrically active regions in a material, qualitatively by demonstrating their existence and quantitatively, by measuring their individual electrical properties. Identification and adoption of the most appropriate equivalent circuit for representation of the electrical properties is essential as a further step towards proper understanding sample properties. It has to be done in materials that are electrically particularly heterogeneous and where the impedance response of one part of sample overlaps, in frequency domain, with the response of other regions, giving rise to a composite response. Therefore, from an equivalent circuit model analysis the responses of the different regions may then be deconvoluted and characterized separately. The adequate model subsumes that are acquired simulation results have to be consistent with experimental data. This is, in most cases, based on designers experience on various equivalent circuit modeling.

II. MATERIALS CHARACTERIZATION

A. Experimental

Sample of amorphous-nano-crystalline thin film with a thickness of 100 nm were deposited by the PECVD method using radio frequency glow discharge in a capacitively coupled parallel plate reactor, as described earlier (Gracin, Thin Solid Films). Used gas mixtures consisted of 5% silane and 95 % hydrogen and the substrate was glass.

The electrical conductivity of the thin film was measured by impedance spectroscopy (Novocontrol Alpha-N dielectric analyzer) in the frequency range 0.01 Hz – 10 MHz and in the temperature range -100–120 ^{o}C. For the electrical contact, two gold electrode pads separated by 4 mm were sputtered on the sample surface using SC7620 Sputter Coater, Quorum Technologies. The impedance spectra were analyzed by equivalent circuit modelling using the complex non-linear least squares fitting procedure.

B. Measurement Data

Typical GIXRD data for measured sample titled G01 is described in previous paper [4]. The volume fraction of crystalline/amorphous phase, estimated from the ratio of integrated intensities under crystalline diffraction peaks and broad amorphous contributions, was about 30%. The average crystal size was estimated from the width of crystalline peaks using Scherrer formula assuming that the size broadening due to strain could be neglected comparing to the size broadening. In that way obtained values are the same as estimated from HRTEM micrographs just verifying the procedure.

III. EQIVALENT CIRCUIT MODEL

There are various approaches to model non-ideality of electronically conducting materials such as amorphous nano-crystalline materials. Since non-ideality of impedance response in silicon samples are observed, introduction of CPS is mainly instead of ideal capacity used in equivalent circuit modeling. Equivalent circuit model described in earlier paper [9] consisted of to two R_1||CPE$_1$ and R_2||CPE$_2$ parallel elements in serial connection had provided correct picture and adoptable results. Calculated values adopted in fitting procedure represents well defined model with good alteration with experimental IS curves and defined equivalent circuit elements. But for better alteration of experimental results an improved equivalent circuit model must be presented to include proven strain in matrix according to arbitrary hydrogen concentration and therefore evidenced of crystal boundary effect – different dielectric region and prove of existing potential barrier. In that case an additional component will be added to basic structure and final impedance $\dot{Z}$ of modified complex equivalent circuit can be calculated according to literature [7] as:

$$\dot{Z} = \frac{A_1 R_1 (j\omega)^{n1}}{R_1 + A_1 (j\omega)^{n1}} + \frac{A_2 R_2 (j\omega)^{n2}}{R_2 + A_2 (j\omega)^{n2}} + \frac{A_3 R_3 (j\omega)^{n3}}{R_3 + A_3 (j\omega)^{n3}}, \quad (1)$$

Where $\dot{Z}(\omega) = \dot{Z}_1 + \dot{Z}_2 + \dot{Z}_3$ is impedance of tree R||CPE parallel elements in serial connection representing mixed-phase material with notable crystal boundary effect between phases and therefore different structural and individual electrical properties. Figure 1 shows a sketch of a a-nc-Si:H film where region 1 corresponds to the amorphous, region 2 to the nano-crystal and region 3 to the crystal boundary interface components.

Other formalisms as complex admittance $\dot{Y}(\omega)$, complex permittivity $\dot{\varepsilon}(\omega)$ complex modulus $\dot{M}(\omega)$ are describe and given in previous paper [7]. The relationship between them is given as

$$\dot{Z}(\omega) = 1/\dot{Y} = \dot{Y}^{-1}, \; \dot{M}(\omega) = 1/\dot{\varepsilon} = \dot{\varepsilon}^{-1} \quad (2)$$

And

$$\dot{M}(\omega) = j\omega C_0 \cdot \dot{Z}, \quad (3)$$

Where omega is angular frequency $(\omega = 2\pi f)$, $C_0 = \varepsilon_0 A / d$ presents the capacitance of the measuring cell in vacuum without presence of sample where A is cell area and d is plate distance or sample thickness, respectively. Capacitance C_0 of a condenser with the geometry of the sample is calculated $C_0 = 6.95 \cdot 10^{-15}$ F.

Serial resistor R_s presenting measuring electrode resistance reasonable can be excluded in presented equivalent circuit model since its parameter is a few orders of magnitude less in comparison to sample DC resistance. Therefore its influence in numerical analysis can be negligible. Capacitance C_0 of a condenser with the geometry of the sample is also excluded in many discussions made in previous papers analyze [9-11]. However in presented modified equivalent circuit model calculated value C_0 is presented with equivalent capacity named C_1 and as a notable capacitive value it is taken in the consideration in fitting procedure. As a result of fitting steps C_1 will be evaluated as well.

All paper IS data are publish according to courteously from Ruđer Bošković Institute, Zagreb.

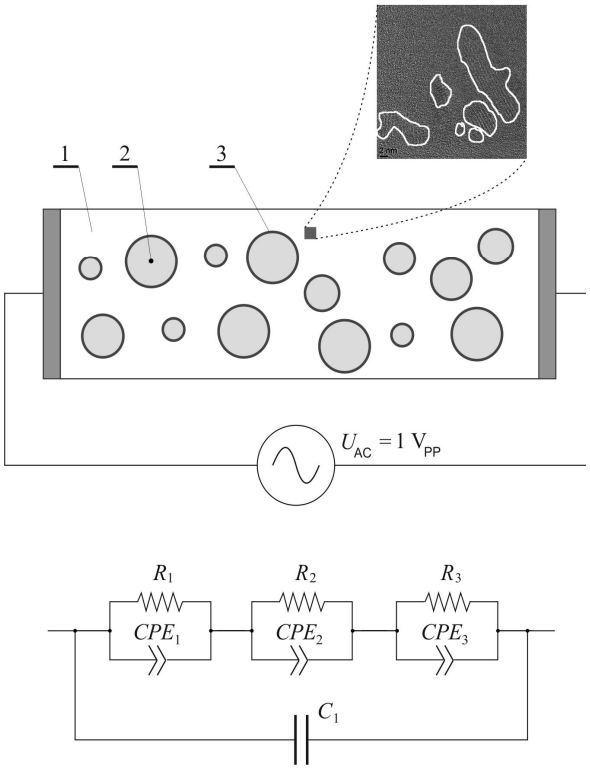


Figure 1. Sketch of the different components contributing to dark conductivity of composite sample connected to AC voltage and equivalent circuit model formed by impedances in series; 1–a-Si:H phase, 2–nc-Si:H phase, 3–Crystal boundary phase.

IV. EXPERIMENTAL RESULTS

Amorphous-nanocrystalline composite silicon thin film named G01 have been investigated in the frequency range 0.01 Hz – 10 MHz using 80 steps and $1V_{pp}$ of modulation in the temperature range from -100 °C to 120 °C . Experimental impedance data of named sample shows semicircle arcs in tree temperature cases: 80 °C , 100 °C and 120 °C [9]. Interpreting the results of AC IS experiment to give information on electrical properties (DC conductivity) or microstructure picture as existence of phases, especially for complex materials like a-nc-Si:H requires a kind of complex approach.

Predominate opinion interprets experimental IS results of composite silicon thin film in terms of series or parallel combinations of resistors and capacitors and special elements like the CPE. It is not proven jet that the complex topology of real nano-structures can be described by using other elements in combination with R-CPS series and parallel ideas. Physically is not possible to consider the complex microstructure as either a parallel or a series arrangement of phases. Rather, since the current must flow according to some arbitrary frequency through amorphous lower-conducting phase in some percolation effective path and through higher-conducting nanocrystalline phase, there is a degree of tortuosity present, which gives a more complicated, IS response. This anisotropic structure influences electronic transport, since carriers flowing between crystals (conduction path in Figure 2) will encounter more crystals boundaries then when flowing throe high resistance a-Si:H matrix.

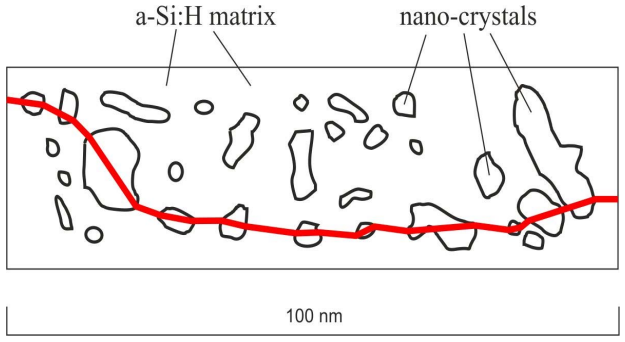


Figure 2. Geometrical model of sample G01structure derived from HRTEM [9]. Red line represents arbitrary conduction path - effective current path on some arbitrary frequency throe composite structure.

The total resistance of the circuit can easily be read from the $Z''(Z')$ curve. The presented spectrum allows clearly determination only dominant resistance amorphous phase R_1 at the beginning of the fitting procedure. Different relaxation time or time constants are detected based on peaks plots of $Z'' = \log(f)$ and electric modulus $M'' = \log(f)$, $(1/\omega_{max})_1 = \tau_1 = A_1 R_1$ and further on [7]. $-Z''$ and M''/C_0 plotted versus frequency f show two peaks for thin film sample G01. Since $-Z''_{max}$ correspond to R_2 and M''/C_0 correspond to $1/2C_1$, the conductivity and the capacitance values were determined [6, 7].

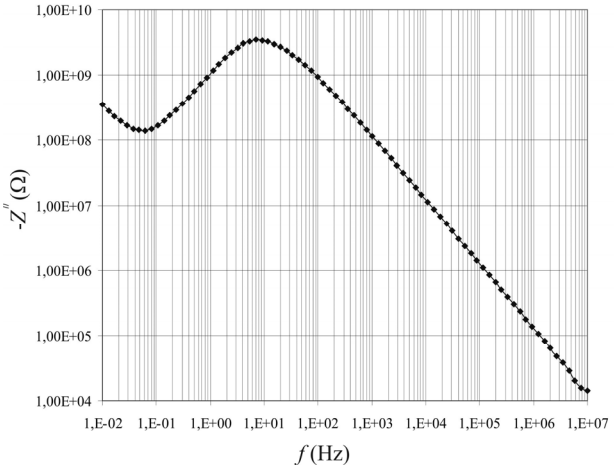


Figure 3. $-Z''$ plot versus frequency (f) of a-nc-Si:H sample G01 at temperature 120 °C .

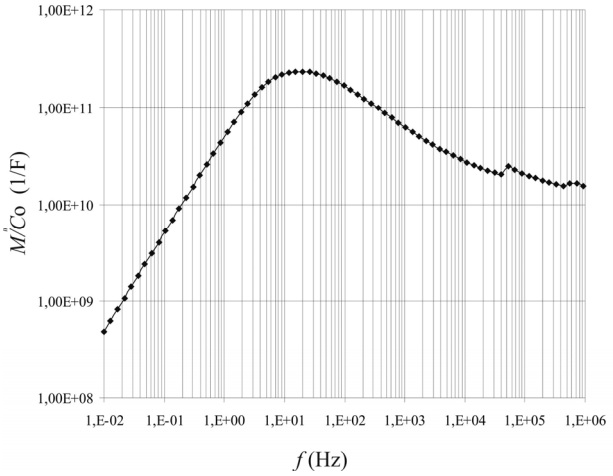


Figure 4. M''/C_0 plot versus frequency (f) of a-nc-Si:H sample G01 at temperature 120 °C .

Depending on the frequency one of the components dominates in serial connection giving us possibility to obtain its resistance R and CPE value - capacitance A.

It is possible that inclusion geometry capacitance of metal-contact interface and porous oxide influence contact/film layer also contributes to determining complexity of arcs in the impedance spectrum, and not just materials series or parallel character. But this influence is not discussed of this study.

In the IS data the dominant amorphous phase is presented by parallel R_1 and CPE_1 element and evaluated with data from Table I, especially amorphous resistance R_1=7.27 GΩ and amorphous capacitance A_1=1.5 pF as it was shown earlier in [9]. Calculated capacitance of A_1=1.25 pF from theoretical relation M''/C_0 which corresponds to value $1/2C_1$ is in good agreement with calculated one. Exponent index n_1 as no-ideality index of impedance semicircle arcs at investigated temperature 120 °C is calculated as 0.937. This index points up the inhomogeneous amorphous structure reach of dangling bonds and hydrogen atoms earlier revealed by techniques like grazing incidence X-ray diffraction (GIXRD).

Nanocrystalline-phase resistance R_2=0.877GΩ is order of magnitude lower than amorphous resistance as expected. Additional nc-Si:H phase capacitance A_2 evaluated as 3.4 pF (shown in Table II) can be expressed only according to M''/C_0 , $M'' = f(M')$ and $M'' = \log_{10}(f)$ plots. This capacitance suggests existence of second present phase: reach nanocrystalline phase. Its existence is previously evaluated thru volume fraction index Xc=33 % [11, 12]. Consequently this is not just one particular crystal size or distribution influence to overall capacitance. Nanocrystalline phase calculated capacitance A_2 =3.4 pF of a-nc-Si:H sample suggests complex connection of nano-sizes individual crystals whit substantially higher capacity than amorphous phase (A_1=1.5 pF). Furthermore, no-ideality index n_2 is calculated as 0.89. This value reveals reach capacitive nature of its phase sins its value is significantly lower than one.

TABLE I. CALCULATED VALUES EXTRACTED FROM EXPERIMENTAL IMPEDANCE DATA

Impedance plot data	Data index / units	Values
Total equivalent resistance	R_p / GΩ	8.11
Specific sample resistance	ρ_p / Ωm	811
Specific sample conductivity	σ_p / μScm^{-1}	12.33
Amorphous-phase resistance	R_1 / GΩ	7.27
Amorphous-phase specific resistance	ρ_1 / Ωcm	72700
Amorphous-phase specific DC conductivity	σ_1 / μScm^{-1}	13.85
Amorphous-phase dark pre-factor	σ_{01} / kScm^{-1}	12.2
log (freq.) plot maximum	$\log_{10} f_{2max}$ (Z)	0.875
Frequency maximum	f_{1max} / Hz	7.499
Amorphous-phase Relaxation time	$1/\omega_{1max}$ / sec.	0.0212233
Amorphous-phase non-ideality index	n_1	0.937
Amorphous-phase capacitance	A_1 / pF	1.5
Fitted Measuring cell capacitance	C_1 / pF	0.005

TABLE II. CALCULATED VALUES EXTRACTED FROM EXPERIMENTAL MODULUS DATA

Modulus plot data	Data index / units	Values
Nanocrystalline-phase resistance	R_2 / GΩ	0.877
Nanocrystalline-phase specific resistance	ρ_2 / Ωcm	87.7
Nanocrystalline-phase specific DC conductivity	σ_2 / μScm^{-1}	114.2
Nanocrystalline-phase dark pre-factor	σ_{02} / kScm^{-1}	304
log (freq.) plot maximum	$\mathrm{Log}_{20} f_{1max}$ (Z)	1.25
Frequency maximum	F_{2max} / Hz	17.78
Nanocrystalline-phase Relaxation time	$1/\omega_{2max}$ / sec.	0.00895
Nanocrystalline-phase non-ideality index	n_2	0.89
Nanocrystalline-phase capacitance	A_2 / pF	3.4
Boundary-phase resistance	R_3 / GΩ	0.013
Boundary-phase capacitance	A_3 / pF	2
Boundary-phase non-ideality index	n_3	0.83

Based on introduced improved equivalent circuit model which consists of three R||CPE elements, existence of additional capacitance A_3=2 pF suggests presence of third dielectrically phase in investigated composite material. More likely it is a boundary effect of existing potential barriers arising from the resulting space charge layer at the silicon crystals boundaries crowded with hydrogen atoms. Also another parameter which proves the specific material dielectrically properties is no-ideality index of this phase – n_3. In fitting procedure it is estimated as 0.83. This index clearly indicates different dielectric properties of its phase - more likely a quarter of nanometer wide boundary regions around formed silicon crystals formed by hydrogen atoms and therefore a weak dangling bonds presence. Another evidence of different dielectric region inherent as silicon crystals boundaries can be annotate throe boundary phase resistance R_3. Evaluated value R_3=13 MΩ is significantly lower then amorphous resistance R_2=877 MΩ proving thin hydrogen dielectric layer. Also, effective capacitance of measuring electrodes is fitted as $C_1 = 5 \cdot 10^{-15} F$ despite of its nominal value $C_0 = 6.95 \cdot 10^{-15}$ F as result of the sample geometry calculation published in [9]. It means that equivalent capacity of measuring electrodes is little bit smaller than expected. It is also prove of effective current path between electrodes despite of its geometrical shape.

It is common opinion that hydrogen presence in this region can be overcome by thermal treatment - annealing. These phenomena can be noted according to hydrogen atom redistribution in porous matrix and weak dangling bonds crystal boundary regions and predictable structural change phenomena in composite silicon thin films described elsewhere [13, 14].

All calculations based on experimental data and fitting curves on impedance and modulus data according to equations (1), (2) and (3) are shown in TABLE I and TABLE II.

V. CONCLUSION

The electric properties of amorphous-nanocrystalline silicon composite thin film were studied in a wide frequency range at different temperatures. The impedance response of a-nc-Si:H sample G01 were interpreted in terms of model of equivalent circuit with tree parallel R and CPE elements connected in serial. By the analysis of equivalent circuit impedance, combined with the analysis of experimental data of electrical modulus, it was possible to extract the resistance and capacitance contribution of all regions to the overall electrical properties of silicon composite thin films. Better model of equivalent circuit is established in the fitting process of experimental impedance data. Multi-arc behavior of experimental results had proved presumption of multi-phase existence, series arrangement of dominant amorphous and nanocrystalline phase with clearly evidence of crystal boundary effects. The non-ideality parameter $n_2 = 0.89$ highlights the capacitive nature of composite structure reach of nano-crystals embedded in amorphous matrix while another index $n_3 = 0.83$ highlights different dielectric properties of boundary regions around formed silicon crystals. The amorphous phase resistance is higher than the same of nano-crystals while a-Si:H capacitance is smaller than capacitance in nc-Si:H phase. Impedance peak against frequency is proportional to the resistance of the composite material, while the peak of electrical modulus against frequency is inversely proportional to the existing capacitance, determining equivalent circuit parameters.

So far, the observed additional specificity in the electrical and structural properties can be explained as a consequence of hydrogen effusion and redistribution in the porous film forming boundary regions around crystals with arbitrary concentrations. The analysis of the IS data confirmed that crystals and its boundary effect has great contribution in the silicon thin film sample resistance and therefore consequently its conductivity.

However, the goal of this work is quite ambitious in its choice of equivalent circuit model to be represented, and the simplicity of model investigated here may prove to be a limitation on its ultimate quantitative success. In any case, a detailed comparison of dynamical results extracted from a well-defined model, such as one derived here, and carefully measured experimental properties on acceptable temperatures will inevitably lead to a deeper understanding of silicon-hydrogen surface chemistry, interactions and coexistence in silicon thin films.

REFERENCES

[1] D. Gracin, K. Jurajić, I. Djerdj, A. Gajović, S. Bernstorff, V. Tudić, M. Čeh, "Amorphous-nanocrystalline silicon thin films for single and tandem solar cells", 14th Photovoltaic Technical Conference - Thin Film & Advanced Silicon Solutions, June, 2012, Aix en Provence, France.

[2] D. Gracin, J. Sancho-Parramon, K. Juraić, A. Gajović, M. Čeh, "Analysis of amorphous-nanocrystalline multilayer structures by optical, photo-deflection and photocurrent spectroscopy", Micron, 2009, vol. 40, Issue 1, pp. 56-60.

[3] D. Gracin, B. Etlinger, K. Jurajić, A. Gajović, P. Dubček, S. Bernstorff, "The DC conductivity and structural ordering of thin silicon films at the amorphous to nano-crystalline phase transition", Journal Vacuum, 2007, vol. 82, Issue 2, pp. 205-208

[4] K. Juraić, D. Gracin, I. Djerdj, A. Lausi, M. Čeh, D. Balzar, "Structural analysis of amorphous-nanocrystalline silicon thin films by grazing incidence X-ray diffraction", Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms, 2012, vol. 284, pp. 78-82.

[5] A.R. West, D.C., Sinclair, N. Hirose, "Characterization of electrical materials, especially ferroelectrics, by impedance spectroscopy", Journal of Electroceramics, 1997, vol. 1, pp. 65-71.

[6] S.K. Barik, R.N.P. Choundhary, A.K. Singh, AC impedance spectroscopy and conductivity studies of Ba0.8Sr0.2TiO3 ceramics", Advanced Material Letter, 2011, vol. 2, Issue 6, pp. 419-424. DOI: 10.5185/amlett.2011.2228.

[7] E. Barsoukov, J.R. Macdonald, Impedance spectroscopy: theory, experiment, and applications, Wiley, 2005.

[8] D. Kohen, J.C. Trully, F.H. Stillinger, Modeling the interaction of hydrogen with silicon surfaces, Surface Science, 1998, vol. 397, pp. 225-236.

[9] V. Tudić, "AC Impedance Spectroscopy of a-nc-Si:H Thin Films", Scientific Research Engineering, July, 2014, vol. 6, No. 8, 449-461. doi: 10.4236/eng.2014.68047.

[10] I. Ferreira, L. Raniero, E. Fortunato, R. Martins, "Electrical properties of amorphous and nanocrystalline hydrogenated silicon

films obtained by impedance spectroscopy“, Thin Solid Films, 2006, vol. 511 – 512, pp. 390 – 393.

[11] K. Shimakawa, “Photo-carrier transport in nanocrystalline silicon films”, Journal of Non-Crystalline Solids, 2006, vol. 352, Issues 9-29, pp. 1180–1183.

[12] D. Gracin, B. Etlinger, K. Jurajić, A. Gajović, P. Dubček, S. Bernstorff, “DC conductivity of amorphous-nanocrystalline silicon thin films”, Journal Vacuum, 2010, vol. 84, Issue 1, pp. 243–246.

[13] D. Gracin, V. Tudić, A. Šantić, K. Jurajić, A. Gajović, D. Meljanac, et. al, “The influence of thermal annealing on the structural and electrical properties of amorphous-nano-crystalline thin Si films”, 16th International Conference on Thin Films, October, 2014, Dubrovnik, Croatia.

[14] D. Gracin, V. Tudić, A. Šantić, K. Jurajić, A. Gajović, D. Meljanac, et. al, “The influence of thermal annealing on the structural and electrical properties of amorphous-nano-crystalline thin Si films”, Special Issue of Thin Solid Films Journal, 2015, in press.

Impact of Different Gate Insulator Materials on the Electron Mobility in Ultra-thin (100) InGaAs-on-insulator MOS Devices

Sabina Krivec, Mirko Poljak and Tomislav Suligoj

Department of Electronics, Microelectronics, Computer and Intelligent Systems
Faculty of Electrical Engineering and Computing, University of Zagreb
HR-10000 Zagreb, Croatia
E-mail: {skrivec, mpoljak, tom}@zemris.fer.hr

***Abstract* - In this paper, we studied the influence of the parameters of technologically relevant dielectrics (SiO_2, Al_2O_3 and HfO_2) and channel thickness downscaling on the electron mobility in InGaAs-on-insulator channels with (100) surface orientation. The study of mobility properties is enabled by a self-consistent Schrödinger-Poisson solver that is coupled with a scattering simulator based on the momentum relaxation time approximation, in order to calculate the scattering rates of electrons with acoustic, optical, remote and polar phonons, surface roughness, Coulomb impurities and alloy disorder. Simulations revealed that SiO_2 is the optimum gate dielectric material irrespective of the body thickness, followed by Al_2O_3 or HfO_2, depending on InGaAs-to-oxide surface roughness.**

Keywords* - *InGaAs-on-insulator; ultra-thin body; mobility; scattering; (100); SiO_2; Al_2O_3; HfO_2.

I. Introduction

III-V channel MOS devices draw attention due to high electron mobility and acceptable band gap. $In_{0.53}Ga_{0.47}As$, as a member of that group of materials, is one of the candidates for low-power and high-frequency CMOS applications [1]. On the other hand, InGaAs has a high dielectric constant, which makes it more susceptible to short-channel effects (SCEs). Consequently, to suppress SCEs, ultra-thin body (UTB) structures have to be employed [2-4]. The InGaAs band gap of 0.75 eV is lower than in silicon, but high enough to results in acceptable band-to-band tunneling, and consequently OFF-state current consumption. Due to scaling of bulk CMOS, the gate oxide thickness was reduced well below 5 nm, causing leakage currents due to tunneling, which lead to a high power consumption and a reduced device reliability [5]. Therefore, SiO_2 was replaced by high-*k* materials, such as Al_2O_3 and HfO_2, which provide the same capacitance as SiO_2 but with thicker oxide layers, thus eliminating the associated leakage effects [6].

This work is sponsored by the Croatian Science Foundation under the project „High-performance Semiconductor Devices for Wireless Circuit and Optical Detection Applications" (contract no. 9006)

Recently, first demonstration of InGaAs-OI with wafer size scalability up to Si wafer size of 300 mm and larger, with Al_2O_3 as gate (GOX) and buried (BOX) oxide has been reported [7]. The experimental data has shown the high electron mobility of 1700 cm^2/Vs and mobility enhancement as much as 3× against Si MOSFETs. Based on this measurement, we present in this paper the results of physics-based modeling of electron mobility in ultra-thin body (100)-oriented InGaAs-OI devices with technologically relevant materials (SiO_2, Al_2O_3 and HfO_2) as gate dielectrics. Electron mobility in UTB InGaAs-OI is investigated by examining the impact of body thickness (T_B) downscaling and of all relevant scattering mechanisms, especially the mechanisms that are inherently present in those materials, such as soft-phonon scattering. The goal is to reveal the optimum device structure considering different gate dielectric materials.

II. Physics-based Mobility Modeling

InGaAs has complex bandstructure consisting of Γ, Δ and *L* energy valleys, just like other III-V compound semiconductors. Regardless of the fact that Γ valley has the lowest conduction band energy, the energy difference between valleys is rather small which means that Δ and *L* valleys can also be filled in strong inversion. Therefore, the electrons from all three valleys are included in simulations. The equi-energy ellipsoid related to each valley is generally described with transversal (m_t) and longitudinal (m_l) effective mass. Knowing m_t and m_l, we can calculate confinement (m_z), density-of-states (m_d) and conductivity masses (m_c) based on expressions reported in [8]. By employing T_B-dependent effective masses [9] and nonparabolicity factors $\alpha_\Gamma = 1.1$, $\alpha_\Delta = 0.55$, $\alpha_L = 0.44$, the bandstructure effects appearing in extremely thin III-V devices are taken into account.

Using the self-consistent Schrödinger-Poisson solver, eigenstates and eigenfunctions of 2-D electron gas (2DEG) in the sub-20 nm-thick InGaAs channel are calculated. The solver is coupled with a scattering rate simulator that is based on the momentum relaxation time approximation (MRTA) [9, 10]. The scattering mechanisms included in

TABLE I. PARAMETERS OF POLAR PHONONS IN DIELECTRICS EXAMINED IN THIS PAPER. DATA IS TAKEN FROM [9].

Parameter	SiO_2	Al_2O_3	HfO_2
ε_0	3.9	12.53	22
ε_{int}	3.05	7.27	6.58
ε_∞	2.50	3.20	5.03
$\hbar\omega_{TO1}$ [meV]	55.60	48.18	12.40
$\hbar\omega_{TO2}$ [meV]	138.10	71.41	48.35

our model are: non-polar acoustic and optical (NPP), polar optical (POP), surface optical phonons (SOP), which must be taken into account when using high-k materials as dielectrics [11]; surface roughness (SR) scattering using Prange-Nee approximation with wave-functions that penetrate into the GOX and BOX layers; Coulomb (CO) scattering which includes scattering with ionized impurities (N_{II}) and interface charge (N_{int}); and alloy disorder (AD) scattering. Equations for each of the scattering mechanisms can be found in [9], accept for SOP scattering. For the intravalley scattering of electrons by SOPs from the i-th to j-th subband in valley v, the matrix element is given by [12]:

$$M^{i,j}_{SO,s}(q) = 4\pi^2 e^2 \left[\left| \int_0^\infty \Psi_i(z)\, \Psi_j(z)\, \varphi_{FI,s} dz \right|^2 + \left| \int_0^\infty \Psi_i(z)\, \Psi_j(z)\, \varphi_{FI,s} dz \right|^2 \right], \quad (1)$$

where $\Psi_{i,j}$ are normalized wave-functions, $\varphi_{FI,s}$ and $\varphi_{BI,s}$ are scattering potential induced by SO phonons at the front and back interface defined as [12]:

$$\varphi_{FI,s} = \left[\frac{\hbar\omega_{SO,s}}{2\pi\varepsilon_S q} \right]^{1/2} e^{-qz},$$
$$\varphi_{BI,s} = \left[\frac{\hbar\omega_{SO,s}}{2\pi\varepsilon_S q} \right]^{1/2} e^{-q(T_B - z)}, \quad (2)$$

where q is a wave-vector and s denotes one of the two SO modes included in the model. Phonon energies are weakly dependent on the wave-vector and are given by [12]:

$$\hbar\omega_{SO,1,2} = \left[\frac{\varepsilon^{(ox)}_{0,i} + \varepsilon_{InGaAs}}{\varepsilon^{(ox)}_{i,\infty} + \varepsilon_{InGaAs}} \right]^{1/2} \hbar\omega_{TO,1,2}, \quad (3)$$

where $\hbar\omega_{TO,1,2}$ are the energies of the bulk oxide optical modes, ε_0 is static dielectric constant, ε_∞ is high-frequency value for dielectric constant and ε_{int} is dielectric permittivity at intermediate frequency between $\omega_{TO,1}$ and $\omega_{TO,2}$. The values of parameters mentioned above for dielectrics of intrest are listed in Table I. Knowning the matrix element, the scattering rate is obtained by Fermi Golden rule [10]. Since multiple scattering mechanims were included in the model, total scattering rate for each subband in each valley is obtained by summing up the contributions from all mechanisms in order to find subband mobility using the Kubo-Greenwood formula [10]. The total mobility is calculated by summing up all obtained subband mobilities weighted by the corresponding subband carrier densities [9].

TABLE II. MATERIAL AND SCATTERING PARAMETERS USED IN SELF-CONSISTENT SCHRÖDINGER-POISSON AND MRTA MOBILITY SIMULATIONS.

Bulk Effective Masses			
Material		GaAs	InAs
Valley	Γ	$m_l = m_t = 0.067$	$m_l = m_t = 0.023$
	Δ	$m_l = 1.90$, $m_t = 0.19$	$m_l = 1.90$, $m_t = 0.19$
	L	$m_l = 1.90$, $m_t = 0.075$	$m_l = 1.90$, $m_t = 0.075$
Bandstructure and Scattering Parameters			
$E_\Gamma = 0.75$ eV, $E_\Delta = 1.33$ eV, $E_L = 1.21$ eV $D_{AP,\Gamma} = 7$ eV, $D_{AP,\Delta} = 9.2$ eV, $D_{AP,L} = 9.0$ eV $E_{OP} = 32$ meV, $D_{OP} = 10\times10^8$ eV/cm $E_{POP} = 32$ meV, $\varepsilon_0 = 13.94$, $\varepsilon_\infty = 11.64$ $v_{s,T} = 2974$ m/s, $v_{s,L} = 4253$ m/s, $\rho = 5506$ kg/m^3 $a = 5.868$ Å, $V_0 = 0.5$ eV, $N_{II} = 4\times10^{15}$ cm^{-3}			
InGaAs-insulator Interface Parameters			
$L = 2.6$ nm, $\Delta = 2.3$ nm, $N_{int} = 6.5\times10^{12}$ cm^{-2}			

III. MODEL CALIBRATION

The calibration of the model is carried out on the experimental data for 9 nm-thick InGaAs-OI MOSFET from [7] with 10 nm thick Al_2O_3 as gate oxide (T_{ox}) and Al_2O_3 as buried oxide (T_{BOX}) by keeping the material and phonon scattering parameters fixed, while adjusting interface charge density (N_{int}) and surface roughness parameters. Fig. 1 shows good agreement between simulation and experiment for InGaAs channel between 10 nm thick Al_2O_3 gate oxide and 15 nm thick Al_2O_3 BOX, which proves that all relevant scattering mechanisms are taken into account in the model. The discrepancy between the simulated and experimental mobility at higher N_{inv} is attributed to strong carrier screening included in the model.

In order to match the experimental data $N_{int} = 6.5\times10^{12}$ cm^{-2}, $\Delta = 2.3$ nm and $L = 2.6$ nm were used for both interfaces. Other parameters used in simulations are listed in Table II. Rather high values of surface roughness scattering parameters (Δ and L) are extracted due to the fact that only intra-subband matrix elements are used for modeling the surface roughness scattering. Effective masses required for solving Schrödinger-Poisson system are obtained by linear interpolation from GaAs and InAs effective masses. Since the initial fitting of the model was performed for 9 nm thick InGaAs-OI device, additional simulations for body thickness up to 30 nm were performed to examine mobilities for thicker body devices. Fig. 2 shows total electron mobility dependence on inversion charge density for body thicknesses from 5 to 30 nm with Al_2O_3 as gate oxide and BOX material. Also, as the body thickness is decreased, mobility is decreasing too, e.g. at $N_{inv} = 2\times10^{12}$ cm^{-2}, the decrease is equal to 84% and 38% at $N_{inv} = 10^{13}$ cm^{-2} when T_B scales down from 30 to 5 nm. The strong T_B-dependence in Fig. 2 is the result of strong volume inversion effect, resulting in a wide distribution of the inversion charge [9]. This effect is more pronounced for thicker devices. Wider distribution of the inversion charge is a result of light effective mass in Γ valley. For a thicker device, the steepness of the wave-function is smaller because of the light $m_{z,\Gamma}$ of InGaAs, whereas the light $m_{d,\Gamma}$ enables the filling of higher

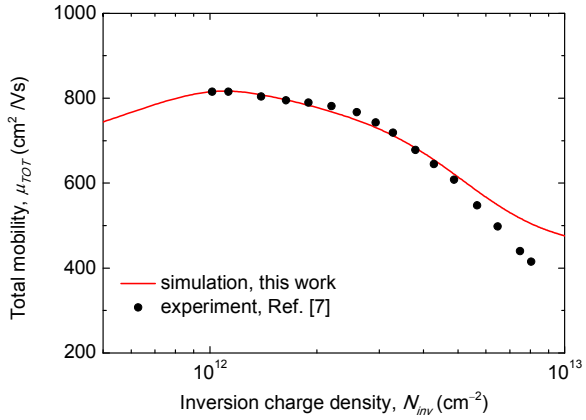


Fig. 1. Comparison between experiment and simulated electron mobility for (100)-oriented InGaAs-OI MOSFET with body thickness T_B = 9 nm, T_{ox} = 10 nm and T_{BOX} = 15 nm. Al_2O_3 is used as GOX and BOX material. Experimental data is taken from [7].

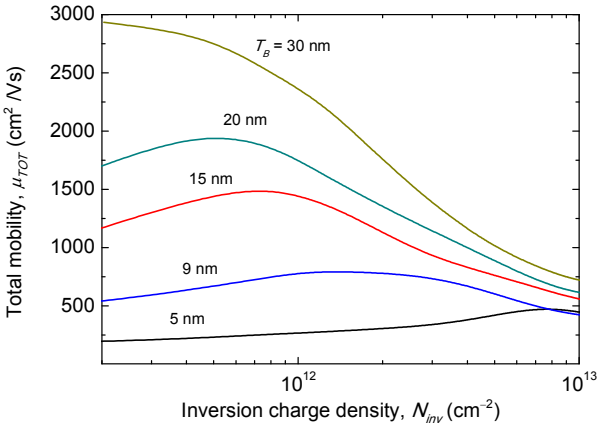


Fig. 2. Mobility in (100)-oriented InGaAs-OI MOSFET with T_B ranging from 5 to 30 nm. Al_2O_3 is used as gate oxide and BOX material (T_{ox} = 10 nm, T_{BOX} = 15 nm).

subbands in Γ valley. Due to the wider distribution and thin body, electrons experience scattering potential caused by SR and the interface charge of the bottom interface as well, although the device has single gate.

IV. Simulation Results and Discussion

The impact of using different gate dielectrics on the total mobility of the 20 nm-thick device is shown in Fig. 3. While performing simulations for different gate dielectrics, the channel material parameters and scattering parameters listed in Table II. were not modified. Depending on the type of material used as dielectric, different dielectric functions and electron affinities were used. That allows us to compare the influence of different gate dielectrics only by comparing their inherent characteristics. As shown in Fig. 3, the total mobility is significantly higher for the structure with SiO_2 gate dielectric compared to the structures with Al_2O_3 and HfO_2

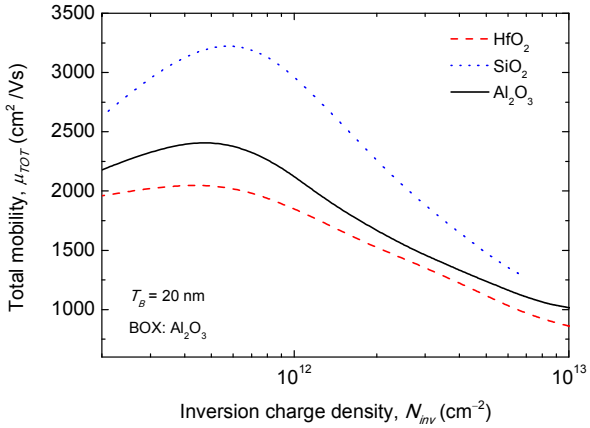


Fig. 3. Total electron mobility in (100)-oriented InGaAs-OI channel with different gate oxides and Al_2O_3 BOX for T_B = 20 nm.

as gate dielectrics. The difference is mostly pronounced at N_{inv} = 5×10^{11} cm^{-2}, where mobility of 3250 cm^2 /Vs for device with SiO_2 as gate oxide is 24% and 35% higher than mobility of devices where gate oxides are Al_2O_3 and HfO_2, respectively. Increasing the inversion charge density, the difference is smaller. At N_{inv} = 6×10^{12} cm^{-2}, mobility of the device with SiO_2 as gate oxide is 13% and 18% higher than mobility where gate oxides are Al_2O_3 and HfO_2, respectively.

The contribution of all scattering mechanisms on total mobility for each structure, is shown in Fig 4. In the case of SiO_2 as gate dielectric the mobility is dominantly determined by CO scattering at lower N_{inv}, as shown in Fig. 4. (b), while for the structures with Al_2O_3 and HfO_2, the scattering with SOP has a strong impact on the mobility both at lower and higher inversion densities (Figs. 4. (a) and (c)). The cumulative effect of CO and SOP scattering for the structures with Al_2O_3 and HfO_2 as gate dielectrics results in a slower increase of total mobility at lower inversion charge densities as compared to the structures with SiO_2 as GOX. The use of alternative high-*k* dielectrics with large difference between the static and optical permittivities, especially when used not only as gate dielectric, but also as buried oxide as in our case, increase the contribution of soft optical phonons scattering in the degradation of total mobility. As shown in Figs. 4. (a) and (c) for Al_2O_3 and HfO_2, respectively, SOP scattering has generally a stronger influence on total mobility in comparison with SiO_2 as a gate dielectric, where basically only SR and CO scattering determinate the total mobility. Also, stronger influence can be observed for HfO_2 because of lager difference in dielectric constants in comparison with Al_2O_3.

The scattering of electrons on surface roughness is more prominent for structures with high-*k* gate oxides, as can be seen by comparing Figs. 4. (a) and (b) with Fig. 4. (c). The explanation lies in the slope of the wave-function at semiconductor-insulator interface since this parameter determines the matrix element for defining

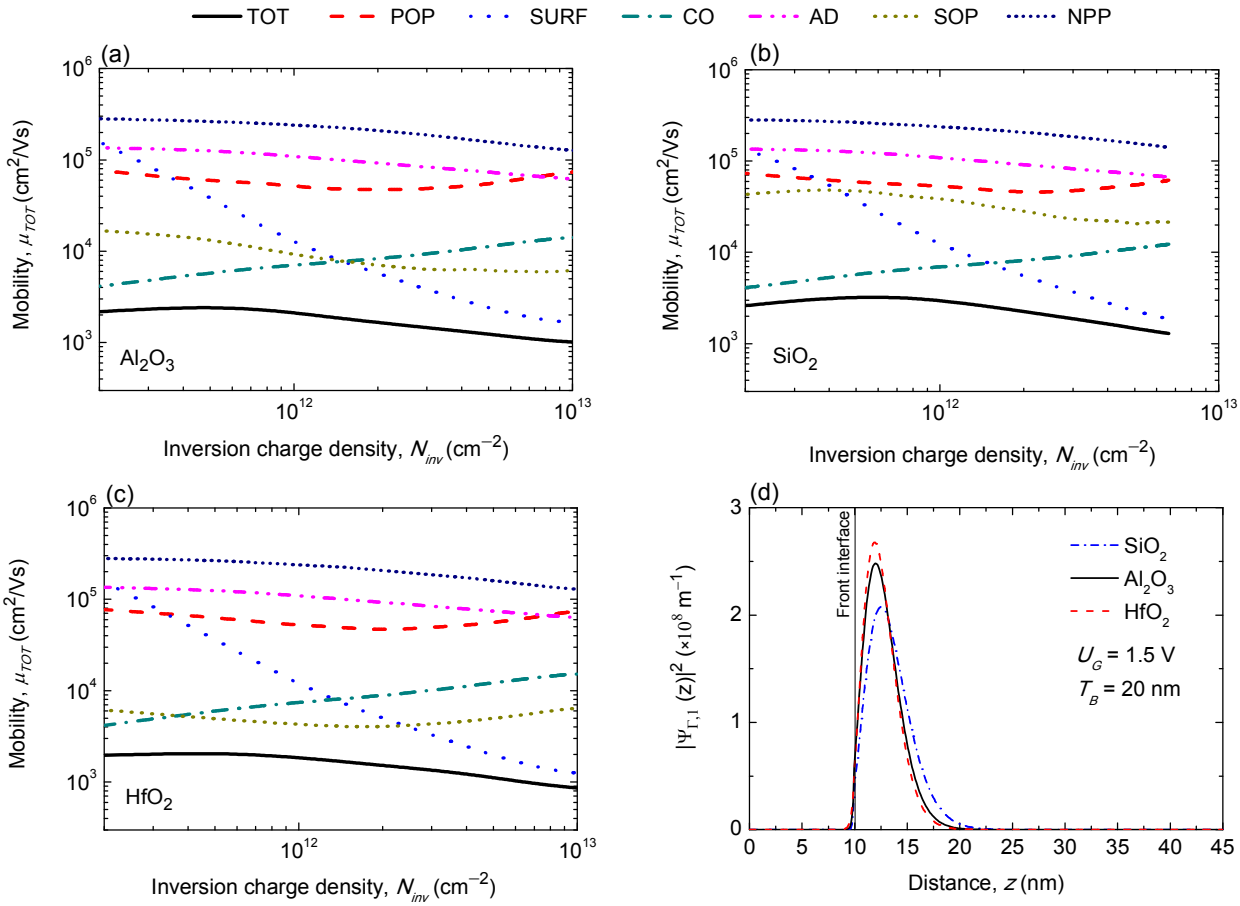


Fig 4. Dependence of mobility on inversion charge density in InGaAs-OI devices with: (a) Al_2O_3, (b) SiO_2, (c) HfO_2 as gate dielectric; (d) Γ-valley ground-state wave-function for $T_B = 20$ nm at $U_G = 1.5$ V for structures with different gate dielectric materials.

MRT in case of SR scattering [9]. Hence, Γ-valley ground-state wave-function for the structure with T_B = 20 nm at U_G = 1.5 V is shown in Fig 4. (d). The smaller wave-function slope at the front interface for the SiO_2-InGaAs-Al_2O_3 structure, in contrast to structures with high-k gate materials, is consequence of smaller dielectric constant. In general, from the flux continuity at the semiconductor-insulator interface we have:

$$F_s = \frac{\varepsilon_{ox}}{\varepsilon_s} F_{ox}, \qquad (4)$$

where $\varepsilon_{ox}, \varepsilon_s$ is the dielectric function and F_{ox}, F_s is the electric field in the oxide and the semiconductor, respectively. In accordance with (4), electric field exhibits a stronger influence in semiconductor when the gate insulator has a higher dielectric constant.

In Fig. 5, the effect of body thickness downscaling is reported for structure with Al_2O_3 as gate dielectric and BOX, i.e. the structure that was used to calibrate the model. Total electron mobility is extracted for three inversion charge density values ranging from $N_{inv} = 10^{12}$ cm^{-2} to relatively strong inversion $N_{inv} = 6\times10^{12}$ cm^{-2}. As the body thickness is scaled down, total electron mobility decreases for all three values of inversion charge density. The decrease of mobility is mostly pronounced for a weak inversion, and least for strong inversion. For the strong inversion, a sharper drop of mobility is visible for $T_B < 6$ nm.

Qualitatively, the explanation of that behaviour can be found by examining the properties of electron wave-functions in the channel. For this purpose, Γ-valley

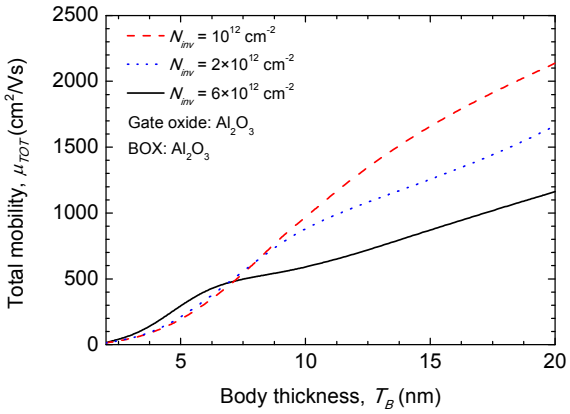


Fig. 5. Dependence of total electron mobility on body thickness for different inversion densities. Al_2O_3 is used as GOX and BOX.

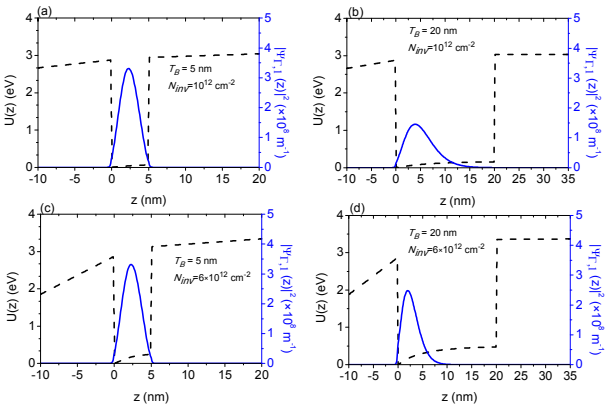


Fig. 6. Potential profiles and Γ-valley ground-state wave-functions for $T_B = 5$ nm [in (a) and (c)] and $T = 20$ nm [in (b) and (d)]. For each T_B, two cases are shown, i.e. for N_{inv} of 10^{12} cm^{-2} and 6×10^{12} cm^{-2}. Al_2O_3 is used as GOX and BOX.

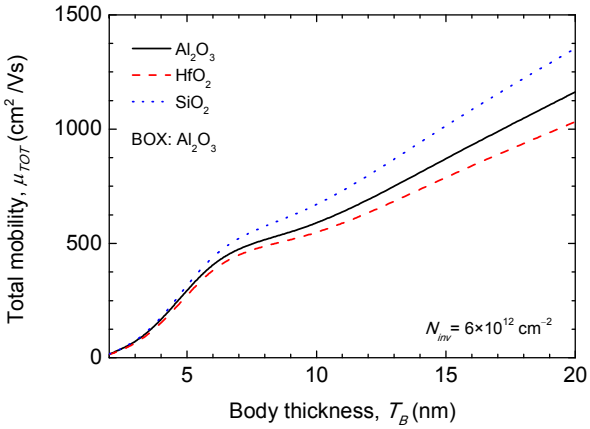


Fig. 7. Dependence of total electron mobility in InGaAs-OI devices on body thickness for different gate oxides at $N_{inv} = 6\times10^{12}$ cm^{-2} ($\Delta = 2.3$ nm, $L = 2.6$ nm).

ground-state wave-functions are shown in Fig. 6 for T_B of 5 and 20 nm, at $N_{inv} = 10^{12}$ cm^{-2} and $N_{inv} = 6\times10^{12}$ cm^{-2}. In weak inversion, due to less pronounced field-induced quantization, the Al_2O_3-InGaAs-Al_2O_3 quantum well acts more like a square well with finite walls as shown in Figs. 6. (a) and (b). As seen in Fig. 6. (b), the wave-function spreads more into the channel resulting in an increased sensitivity of scattering mechanisms to body thickness. In strong inversion (N_{inv} = 6×10^{12} cm^{-2}), the field-induced confinement is dominant and the wave-functions are positioned closer to the top interface. Thus, the impact of body thickness downscaling is less pronounced which can be seen in Figs. 6. (c) and (d).

Fig. 7 shows dependence of total mobility extracted at $N_{inv} = 6\times10^{12}$ cm^{-2} in InGaAs-OI devices on body thickness for different materials as gate dielectrics. For $T_B < 6$ nm, a sharp decrease of mobility can be observed for every examined structure. Also, for body thicknesses less than 6 nm, the values of mobility for each structure are approximately the same. Since SR scattering parameter Δ is equal to 2.3 nm, which is comparable to body thickness in this range, SR scattering is a dominant scattering mechanism defining mobility for T_B < 6 nm for each structure. As body thickness increases, mobility difference between structures is more pronounced, especially between structures with SiO_2 as gate dielectric in comparison to other two structures. The reason for this is a weaker influence of SOP scattering for higher T_B for device with SiO_2 as gate dielectric compared to the other two structures. Total mobility for the structure with SiO_2 as gate dielectric at $T_B = 6$ nm is 450 cm^2/Vs which is 8% and 13% higher than total mobilities of structures with Al_2O_3 and HfO_2, respectively. Moreover, total mobility for the structure with SiO_2 as gate dielectric at $T_B = 15$ nm is 1017 cm^2/Vs which is 14% and 22% higher than total mobilities of the structures with Al_2O_3 and HfO_2, respectively. Also, due to the stronger influence of SR scattering for device with HfO_2 as gate dielectric, mobility for that device is lower than mobility for device with Al_2O_3 as gate dielectric. With body thickness increasing, the difference between these two mobilities is increasing following the SR scattering limited mobility trend.

The impact of all scattering mechanisms for different structures is reported in Fig. 8. For every structure, the SR scattering dominantly determines the total mobility, especially for extremely thin bodies (T_B < 6 nm). Dominant influence of SR scattering is expected because of choosing quite high values of SR scattering parameters (Δ and L) already explained in Section II. The second relevant mechanism, particularly for structures with Al_2O_3 and HfO_2 as gate dielectrics is SOP scattering. Since the scattering potential of the SO phonons is defined with exponential function for front and back interface (2), thus a considerable body thickness dependence is expected, as can be seen in Fig. 8.

Assuming a smaller Δ that could be achieved in modern III-V manufacturing, the total mobilities for each structure are obtained and reported in Fig. 9. Compared to the results for higher values of Δ (see Fig. 7), mobility difference between structures with SiO_2 as gate dielectric in comparison to the other two structures is more pronounced. Also, mobility for HfO_2-InGaAs-Al_2O_3 device has slightly higher values of mobility for T_B

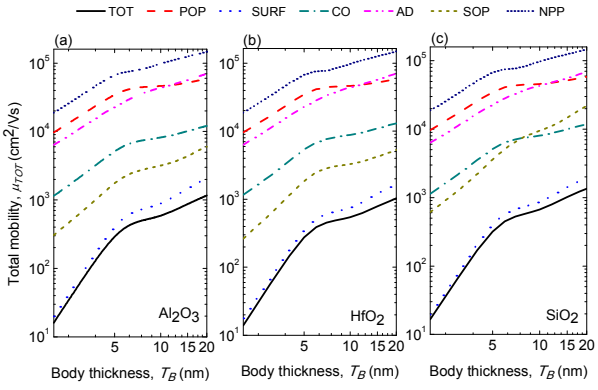


Fig. 8. Impact of different scattering mechanisms in InGaAs-OI devices with Al_2O_3 as BOX and different gate oxides at $N_{inv} = 6\times10^{12}$ cm^{-2} ($\Delta = 2.3$ nm, $L = 2.6$ nm).

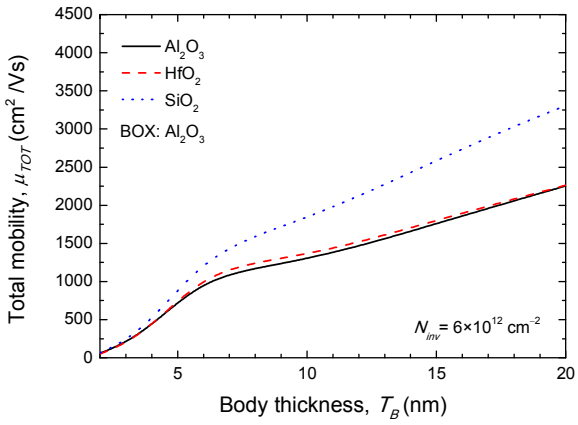


Fig. 9. Dependence of total electron mobility in InGaAs-OI devices on body thickness for different gate oxides at $N_{inv} = 6\times10^{12}$ cm^{-2} with decreased influence of SR scattering ($\Delta = 1$ nm, $L = 2.6$ nm).

between 6 and 15 nm in comparison to mobility for Al_2O_3-InGaAs-Al_2O_3 device. With smaller Δ, total mobility for structure with SiO_2 as gate dielectric at $T_B = 6$ nm is 1223 cm^2/Vs thus 21% and 17% higher than total mobilities of structures with Al_2O_3 and HfO_2, respectively. Additionally, the total mobility for SiO_2-InGaAs-Al_2O_3 device at $T_B = 15$ nm is 2589 cm^2/Vs which is 32% and 31% higher than total mobilities of the structures with Al_2O_3 and HfO_2, respectively.

As shown in Fig. 10, for the structures where Al_2O_3 and HfO_2 are gate dielectrics, mobility is dominantly determined both with SOP and SR scattering. Cumulative effect of these two scattering mechanisms at lower N_{inv} and effect of SOP scattering at higher N_{inv} are producing more pronounced mobility difference in comparison with the mobility of SiO_2-InGaAs-Al_2O_3 device where total mobility is still determinated only by SR scattering. Higher mobility for device with HfO_2 as gate dielectric compared to Al_2O_3-InGaAs-Al_2O_3 device is result of stronger influence of SOP scattering and weaker influence of SR scattering for $T_B > 6$ nm.

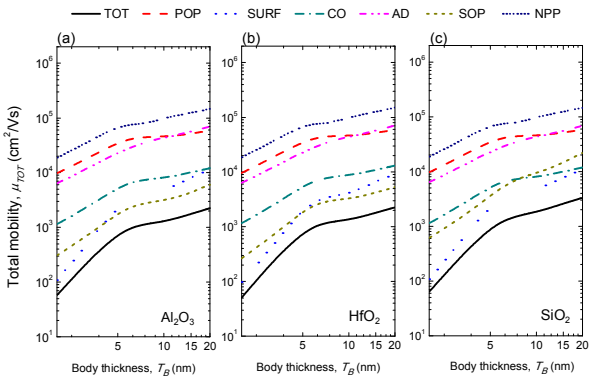


Fig. 10. Impact of different scattering mechanisms in InGaAs-OI devices with Al_2O_3 as BOX and different gate oxides on body thickness at $N_{inv} = 6\times10^{12}$ cm^{-2} with decreased influence of SR scattering ($\Delta = 1$ nm, $L = 2.6$ nm)

V. Conclusion

In this paper, the results of physics-based modeling of electron mobility in InGaAs-OI devices with technologically relevant gate dielectrics (SiO_2, Al_2O_3 and HfO_2) are presented. With body thickness downscaling, the total electron mobility decreases monotonically in all the examined structures and for all values of inversion charge density. The simulations have shown that SiO_2 is the optimum gate dielectric material in terms of carrier transport as it exhibits the highest mobility values in the entire T_B range investigated in this work. For $T_B > 6$ nm, mobility difference between the SiO_2-InGaAs-Al_2O_3 devices and alternative structures is more pronounced due to a weaker influence of SOP scattering resulting in mobility of 450 cm^2/Vs for SiO_2-InGaAs-Al_2O_3 device at $T_B = 6$ nm (8% and 13% higher than in structures with Al_2O_3 and HfO_2 as gate oxide, respectively). Assuming smaller Δ, mobility difference between structures with SiO_2 in comparison to the other two structures is even more pronounced due to the cumulative effect of SOP and SR scattering in devices with high-k oxide materials. The resulting mobility is 1223 cm^2/Vs at $T_B = 6$ nm for SiO_2-InGaAs-Al_2O_3 structure, i.e. 21% and 17% higher than in devices with Al_2O_3 and HfO_2, respectively.

References

[1] S. Takagi *et al*., "MOS interface and channel engineering for high-mobility Ge/III-V CMOS", Tech. Digest of International Electron Devices Meeting (IEDM) 2012, p. 505.

[2] S. Takagi *et al*., "Carrier-transport-enhanced channel CMOS for improved power consumption and performance", IEEE Trans. Electron Devices 55, p. 21, 2008.

[3] M.Yokoyama *et al*., "Ultrathin body InGaAs-on-insulator metal-oxide-semiconductor field-effect transistor with InP passivation layers on Si substrates fabricated by direct wafer bonding", Appl. Phys. Exp. 4, p. 54202-1, 2011.

[4] M. Radosavljevic *et al*., "Non-planar, multi-gate InGaAs quantum well field effect transistor with high-k gate dielectric and ultra-scaled gate-to-drain/ gate-to-source separation for low power logic applications", Tech. Digest of International Electron Devices Meeting (IEDM) 2010, p. 126.

[5] J. H. Stathis *et al*., "Reliability of ultra-thin oxides in CMOS circuits," Microelectronics Reliability 43, p. 1353, 2003.

[6] M. Radosavljevic *et al*., "Advanced high- k gate dielectrics for high-performance short-channel $In_{0.7}Ga_{0.3}As$ quantum well field effect transistors on silicon substrate for low power logic applications", Electron Devices Meeting (IEDM) 2009, p. 1-4.

[7] S. H. Kim *et al*., "High performance InGaAs-on-insulator MOSFETs on Si by novel direct wafer bonding technology applicable to large wafer size Si", Symposium on VLSI Technology 2014, p. 1-2.

[8] F. Stern *et al*., "Properties of semiconductor surface inversion layers in the electric quantum limit", Phys. Rev. 163, p. 816, 1967.

[9] M. Poljak *et al*., "Assessment of electron mobility in ultra-thin body InGaAs-on-insulator MOSFETs using physics-based modeling", IEEE Trans. Electron Devices 59, p. 1636, 2012.

[10] D. Esseni, P. Palestri, L. Selmi, Nanoscale MOS Transistors: Semi-Classical Transport and Applications, Cam. Univ. Press, 2011.

[11] T. P. O'Regan *et al*., "Calculation of the electron mobility in III-V inversion layers with high-k dielectrics", J. Appl. Phys. 108, p. 103705, 2010.

[12] D. Esseni *et al*., "Physically based modeling of low field electron mobility in ultrathin single- and double- gate SOI n-MOSFETs", IEEE Trans. Electron Devices 50, p. 2445, 2003.

Impact of the Emitter Length Scaling on Electrical Characteristics of Horizontal Current Bipolar Transistor with Single Polysilicon Region

M. Koričić, J. Žilak and T. Suligoj
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
marko.koricic@fer.hr, josip.zilak@fer.hr, tomislav.suligoj@fer.hr

Abstract - Emitter length scaling of HCBT with single polysilicon region is investigated by 3D device simulations with the emphasis on the high frequency characteristics. It is shown that collector current and junction capacitances have linear dependence on the emitter length. Collector resistance can be represented by two components which appear in parallel, one that scales proportionally with the emitter length and a constant part which describes lateral portion of the extrinsic transistor. Cut-off frequency is improved for small emitter area devices due to current spreading in the collector region which reduces the current density and causes the base push-out at higher collector currents. The effect cannot be captured by scalable transistor model and separate set of model parameters is needed for transistors with very small emitter size.

I. INTRODUCTION

Scalability of transistor models used in circuit simulators is a highly desirable property. With the technology progress and scaling, small geometry effects arise requiring separate set of model parameters for devices of different scale, often referred to as parameter binning. This modeling approach is implemented in modern circuit simulators and is found as an option in all models suitable for submicron CMOS technologies such as BSIM3 [1] and BSIM4 [2].

Horizontal Current Bipolar Transistor (HCBT) is a lateral bipolar transistor which is integrated with standard 180 nm CMOS with small number of additional process steps [3]. Low-cost and high-performance characteristics make HCBT suitable for fabrication of RF circuits for wireless applications. RF mixer reported in [4] was designed by using Gummel-Poon model extracted for unit size single polysilicon region HCBT [5]. Unit transistors were added in parallel where higher collector current drive was required. HCBT emitter is defined by polysilicon etching process and its height is given by the technology. Therefore, the collector current in the circuit design can be tuned by changing the emitter length or by adding transistors of unit size in parallel, the former being the more flexible solution. In order to address the ability to make scalable HCBT models, impact of lateral scaling of the emitter length and lateral portion of the extrinsic transistor is investigated by 3D device simulations.

This work has been supported by the Croatian Science Foundation under the Project no. 9006.

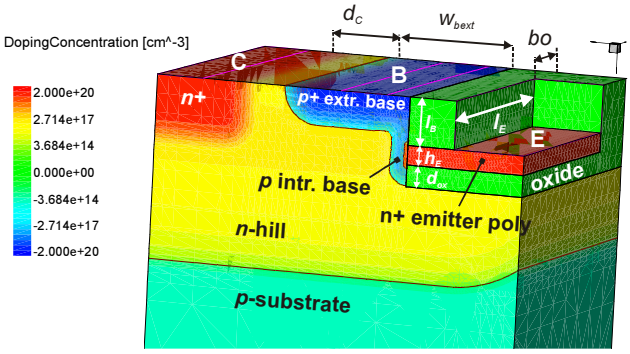


Fig. 1. Simulated HCBT structure with marked geometrical parameters.

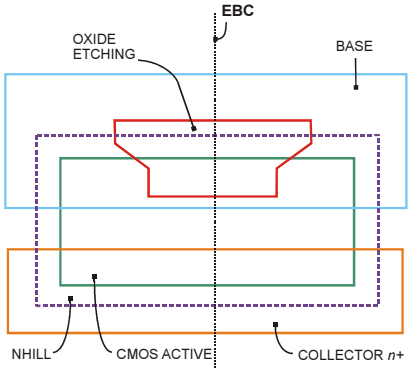


Fig.2. Lithography masks important for the fabrication of HCBT.

II. DEVICE STRUCTURE

3D model of the single polysilicon region HCBT used in the simulations is presented in Fig. 1 and the lithography masks important for the fabrication of the HCBT regions are shown in Fig. 2. HCBT is fabricated in silicon island defined by the CMOS shallow-trench-isolation (STI) using a *CMOS active* mask (Fig. 2). Intrinsic collector marked as *n*-hill in Fig. 1 is defined by dedicated implantation using a *nhill* mask (Fig. 2). Extrinsic base is implanted on the top surface of the silicon island using a *base mask* (Fig. 2) and is extended laterally across the whole active area to accommodate for mask misalignment. Emitter trench is defined by oxide wet etching using an *oxide etching mask* (Fig. 2), which has to be narrower than the CMOS active exposing only a

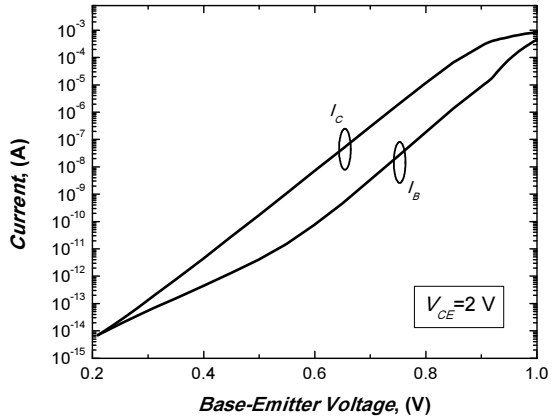


Fig.3. Simulated Gummel characteristics of HCBT structure with l_E=1 μm and bo=0.3 μm.

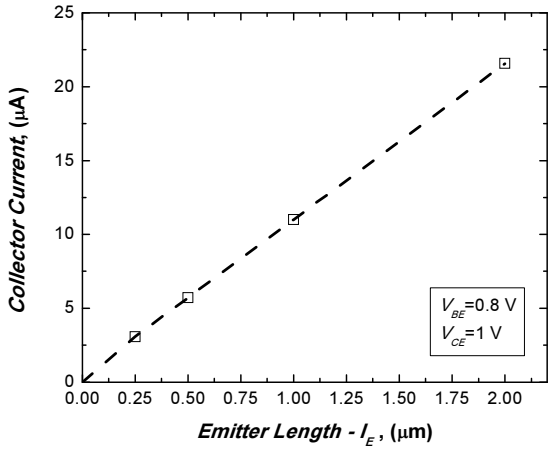


Fig.4. Simulated collector currents (I_C) for different emitter length (l_E) at V_{BE}=0.8 V and V_{CE}=1 V. Dashed line presents linear interpolation of simulated results showing linear scaling of I_C with l_E.

portion of STI sidewall, where intrinsic transistor is fabricated. The intrinsic base is fabricated by angled ion implantation using the *base mask* again (Fig. 2). Implant is done on the sidewall exposed by oxide etching with the remaining oxide serving as a hard mask, self-aligning the intrinsic base to the shape of the trench. The emitter is fabricated by deposition and etch-back of *in-situ* doped amorphous-silicon, which crystallizes during the emitter drive-in annealing. In this way, the emitter and the intrinsic base are self-aligned resulting in the state-of-the-art electrical performance. Transistor has a very small emitter height (h_E in Fig. 1) of around 80 nm, which is considerably smaller than the minimum lithography resolution. Due to the small h_E, one-sided base contact can be tolerated achieving the maximum oscillation frequency (f_{max}) comparable to the best reported vertical-current bipolar transistors with implanted base [6]. An extrinsic collector (i.e. collector contact region) is formed during CMOS source/drain implantation by using a *collector n+* mask, which is available from the CMOS.

Since the emitter trench has to be narrower than the active region, a certain lateral overlap of the extrinsic base with respect to the intrinsic base exists (bo in Fig. 1). The emitter trench defines the emitter length (l_E), which can be changed in the circuit design in order to tune the transistor operating current. These two parameters are changed by the lateral scaling. Additional geometrical parameters that can be changed by transistor mask layout are an extrinsic base width (w_{bext}) and a distance between the extrinsic base and the collector contact region (d_c). These parameters should be kept as small as possible since w_{bext} affects collector-base capacitance (C_{jBC}) and d_c affects collector series resistance (R_C). The lower limit is set by the formation of reliable base contact in case of w_{bext} [7] and by the base-collector junction breakdown in case of d_c. Geometrical parameters are summarized in Table 1. All doping profiles in the transistor are defined as analytical Gaussian profiles. Vertical collector doping profile is a steep profile reported in [8], which is optimized for maximum f_TB_{VCEO} product. It should be noted that the intrinsic base profile in fabricated devices is steeper than the Gaussian resulting in slightly higher cut-off frequency (f_T) obtained by the measurements reported in [8] compared to the simulations presented in this paper. Nevertheless, the results are within 20 % margin and only relative comparison of simulation results was important for this study. Since the transistor is symmetrical with respect to the middle of the emitter (*EBC* line in Fig. 2), only half of the transistor is simulated in order to save computational time.

TABLE I SUMMARY OF THE GEOMETRICAL PARAMETERS OF SIMULATED HCBT STRUCTURES

l_B, (nm)	h_E, (nm)	d_{OX}, (nm)	w_{bext}, (nm)	d_c, (nm)
200	80	90	400	400

III. ANALYSIS OF ELECTRICAL CHARACTERISTICS

Simulated Gummel characteristics for the structure with l_E=1 μm and bo=0.3 μm at V_{CE}=2 V are shown in Fig. 3. Scaling of the collector current (I_C) is investigated by simulating transistors with different l_E and bo=0.3 um and with the same terminal voltages applied, i.e. at V_{BE}=0.8 V and V_{CE}=1 V. Results are represented by symbols in Fig. 4. Dashed line shows the linear interpolation of simulated points indicating linear scaling of I_C with the l_E.

Simulated cut-off frequency (f_T) for structures with different l_E and bo=0.3 μm are shown in Fig. 5. It can be seen that maximum value of f_T falls from 31.2 GHz to 28.3 GHz or by 9.3 % if l_E is increased from 0.25 μm to 2 μm. It can also be observed that maximum f_T settles to constant value for larger l_E. For example, if l_E is increased from 1 to 2 μm, f_T is reduced by only 2 %. Additionally, peak f_T value is shifted to higher collector currents for structures with larger l_E, which is expected since fall-off of f_T is caused by the Kirk effect [9] which starts at the certain collector current density. Cut-off frequency depends on the time constants associated with the charges stored in the transistor regions and is given by [9]:

$$\frac{1}{2\pi f_T} = \tau_F + \frac{kT}{I_C}\left(C_{jBE} + C_{jBC}\right) + \left(R_C + R_E\right)C_{jBC} \qquad (1)$$

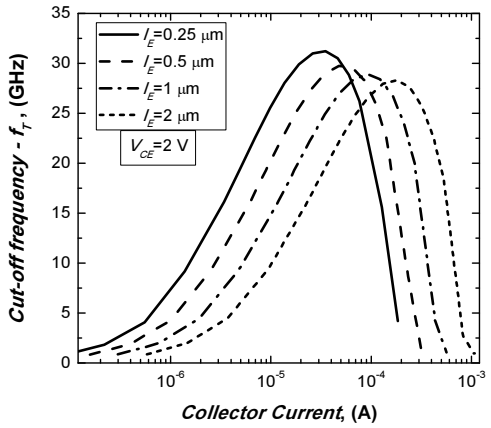


Fig. 5. Simulated cut-off frequency (f_T) dependence on collector current for structures with different emitter length (l_E) and base overlap bo=0.3 μm.

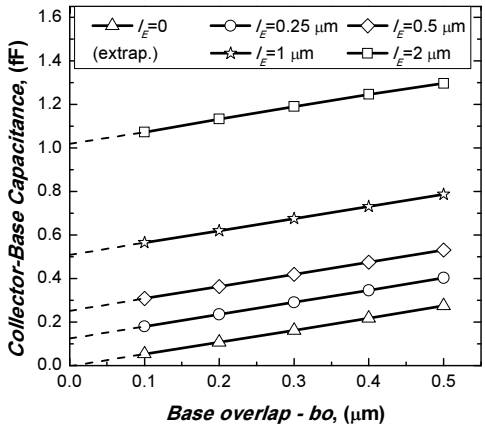


Fig.7. Extracted collector-base junction capacitance for structures with different base overlap(bo) and emitter length (l_E). Extrapolated values for l_E=0 are added to demonstrate linear scaling of capacitances.

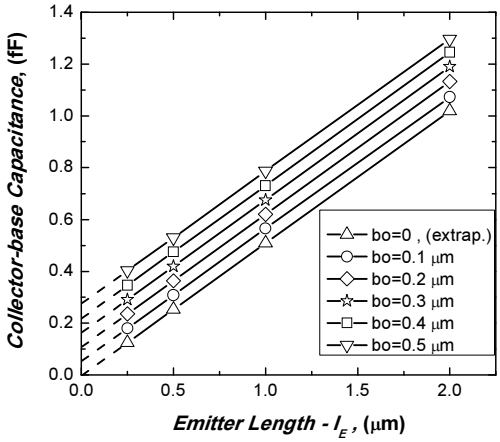


Fig.6. Extracted collector-base junction capacitance for structures with different emitter length (l_E) and base overlap(bo). Extrapolated values for bo=0 are added to demonstrate linear scaling of capacitances.

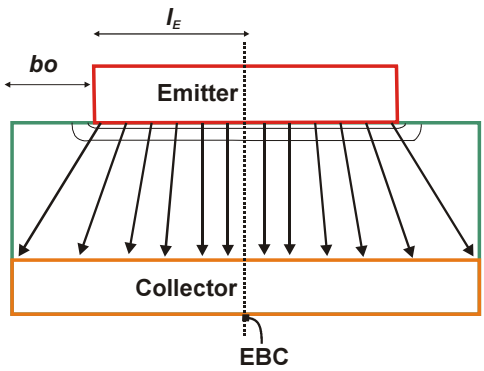


Fig. 8. Current spreading in the collector. Due to larger collector length (l_E+bo) compared to the emitter length (l_E) current spreads laterally. The effect is more pronounced for transistors with smaller l_E. Cross-section parallel to wafer surface is shown. Parameters l_E and bo of simulated half-structure are marked for reference.

where τ_F is forward transit time consisting of base and base-collector depletion region transit times, C_{jBE} and C_{jBC} are base-emitter and base-collector depletion region capacitances and R_C and R_E are collector and emitter series resistances, respectively. Since the intrinsic doping profiles are the same in all simulated structures, the forward transit time should be the same for all of them. In order to address influences of terms in (1), C_{jBC} and R_C are extracted from the simulated electrical characteristics of HCBT structures with different geometries. The emitter length (l_E) and the extrinsic base overlap (bo in Fig. 1) are varied in simulations.

C_{jBC} is extracted from AC simulation performed at low frequency (f=50 MHz) at the bias point with V_{BE}=0.8 V and V_{CE}=2 V. The dependence of C_{jBC} on l_E and bo are shown in Figs. 6 and 7. The results are shown twice in order to demonstrate that capacitances scale linearly both with l_E and bo. Also, extrapolated C_{jBC} for bo=0 (in Fig. 6) and l_E=0 (in Fig. 7) are shown to demonstrate that capacitance extrapolate to zero when both bo and l_E tend to zero. As a result we can conclude that C_{jBC} scales linearly with lateral dimensions and that the same behavior can be expected for C_{jBE}. Therefore, the second term in (1) is the same for all structures regardless of l_E since both I_C (see Fig. 4) and C_{jBC} (Figs. 6 and 7) rise linearly with l_E. Component R_EC_{jBC} from the third term in (1) is very small since R_E is order of magnitude smaller than R_C. Furthermore, R_E depends on the emitter polysilicon sheet resistance and scales linearly with l_E, similar as in case of integrated resistors. Therefore, component R_EC_{jBC} should be independent on l_E, since C_{jBC} rises and R_E falls linearly with l_E. On the other hand, R_C cannot be expected to scale linearly with l_E since electrons can spread when they leave collector-base depletion region and enter the neutral collector (see Fig. 8). This depends on the relative difference in collector and emitter length [10]. The collector length is equal to l_E plus bo. For smaller l_E and larger bo, the collector current spreading should be more pronounced. R_C is extracted from the saturation region of the output characteristics [11] of simulated HCBT structures with different l_E and bo. Results are shown in

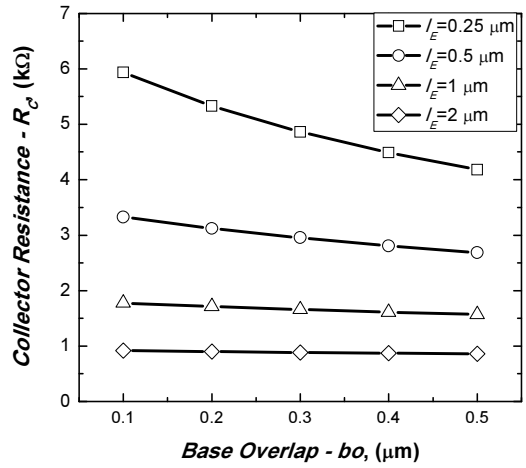


Fig. 9. Extracted series collector resistances (R_C) for structures with different overlap(bo) and emitter length (l_E).

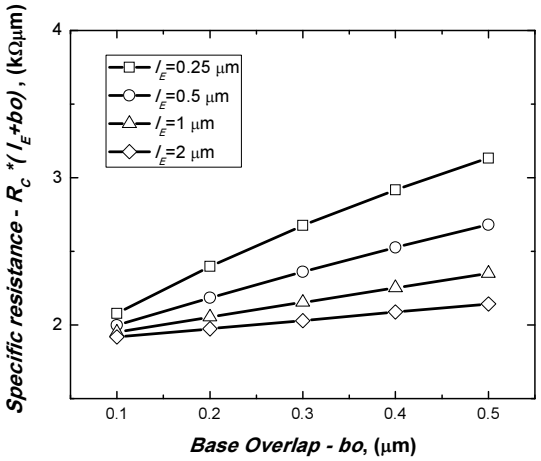


Fig. 10. Calculated collector specific resistance for structures with different overlap(bo) and emitter length (l_E).

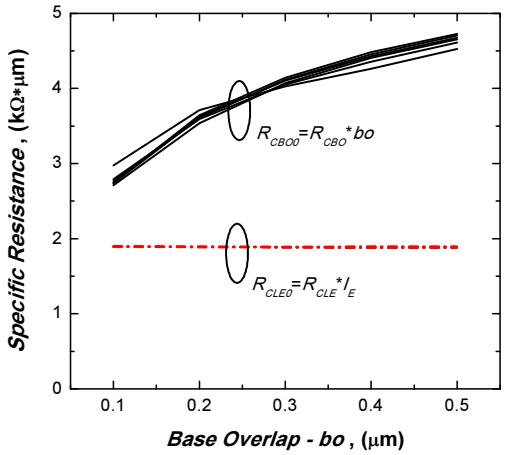


Fig. 11. Calculated collector specific resistance components for structures with different overlap (bo). Resistances are calculated by solving 6 sets of two equations obtained by permuting equations (3) for different l_E. There are 6 curves for R_{CLE0} basically overlapping.

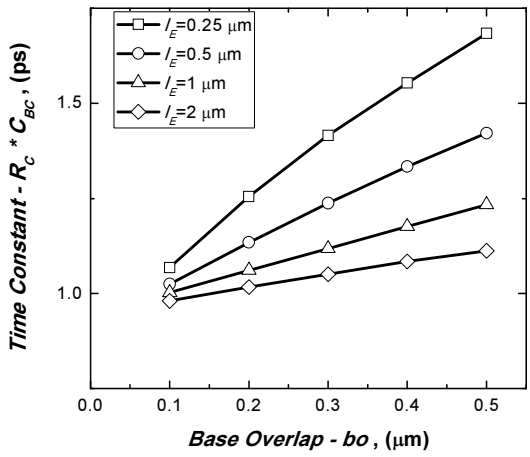


Fig. 12. Transistors' R_CC_{jBC} time constants calculated from the extracted transistor parameters for structures with different base overlap(bo) and emitter length (l_E)

Fig. 9. It can be seen that R_C doesn't have linear dependence on bo, which is more pronounced for transistors with small l_E.

This is confirmed by plotting specific collector resistance (in Fig. 10), calculated as a product of the extracted R_C and the collector length (l_E+bo). Since the lines are not flat, dependence of R_C on l_E+bo is not linear. It can also be seen that lines representing specific resistance are tending to a single point if bo tends to zero. This means that component of collector series resistance arising from the portion of collector overlapped by the emitter window (defined by l_E), scales linearly with l_E. The non-linear dependence of R_C on l_E+bo arises from the nonlinear dependence of the R_C component determined by the bo overlap of collector. Therefore, R_C could be decomposed into two components, R_{CLE} which scales linearly with l_E and R_{CBO} which has a non-linear dependence on bo. Those two components appear in parallel and a set of equations for different l_E and fixed bo can be written as:

$$\frac{1}{R_{Ci}} = \frac{1}{R_{CLEi}} + \frac{1}{R_{CBO}} \tag{2}$$

where we have assumed that component R_{CBO} is the same for all structures with the same bo and different l_E. Next, since we postulate that R_{CLE} scales linearly with l_E, we can write:

$$\frac{1}{R_{Ci}} = \frac{l_{Ei}}{R_{CLE0}} + \frac{1}{R_{CBO}} \tag{3}$$

where R_{CLE0} is specific resistance. We have extracted R_{Ci} of 4 simulated structures with 4 different l_{Ei} and we can write i=4 equations. There are only two unknowns (i.e. specific resistance R_{CLE0} and resistance R_{CBO}) and the system is redundant since we can construct 6 sets of two equations by permuting those 4 equations. Solutions of all 6 sets for each bo are shown in Fig. 11. It can be observed that specific resistances R_{CLE0} calculated from all 6 equation sets overlap, showing that the same result is

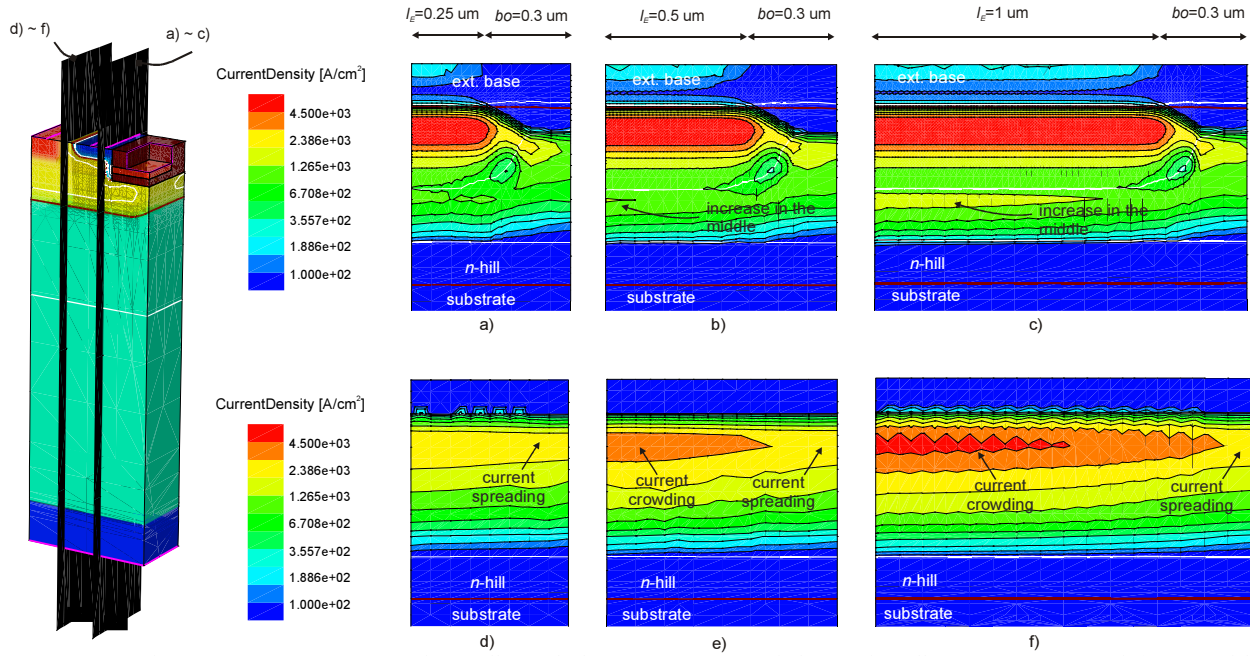


Fig.13. Current density at cross sections across base-collector depletion region (a~c) and close to the collector contact region (d~f). 3D model with marked cross-sections is given for reference.

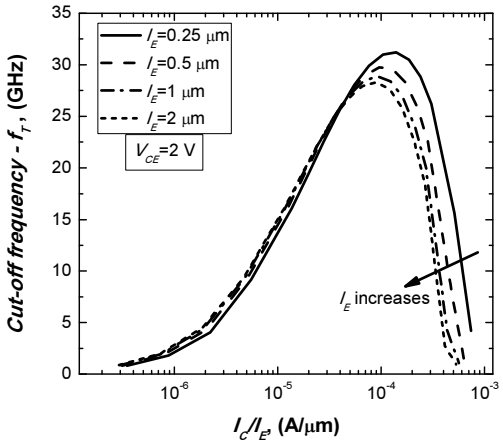


Fig. 14. Simulated cut-off frequency (f_T) dependence on collector current normalized to the emitter length (l_E) for structures with different emitter length (l_E) and base overlap bo=0.3 μm.

obtained for any combination of l_E meaning that R_{CLE0} specific resistance component is constant and independent on l_E. Therefore the assumption that R_{CLE} scales linearly with l_E is valid since it fits well to simulated data. Furthermore, the lines representing R_{CLE0} are flat indicating linear scaling of R_{CLE} resistance component with l_E regardless of bo. The results obtained for specific resistance R_{CBO0} component are calculated as a product of R_{CBO} obtained as a solution from the 6 equation sets with bo. Results also overlap within reasonable margin indicating that it is independent on l_E, but dependence on bo is non-linear. This leads us to the conclusion that in order to obtain scalable model, R_C can be modeled by two components - R_{CLE} which scales linearly with l_E and R_{CBO} which is independent on l_E and is constant for fixed bo.

Time constant R_CC_{jBC} calculated from extracted R_C (Fig. 9) and C_{jBC} (Fig. 6 and 7) is shown in Fig. 12. The largest time constant is observed for transistor with smallest l_E. This is contrary to the highest f_T result from Fig. 5 meaning that R_CC_{jBC} is small enough for all simulated structures and that it has minor influence on f_T compared to other terms in (1). It can also be seen that as bo increases, R_CC_{jBC} increases for all structures meaning that the increase of C_{jBC} is stronger than the decrease of R_C, which can be observed as a saturation of extracted R_C in Fig. 9 in case of larger bo.

In order to identify the reason for improved f_T of small l_E transistors, current distributions in the intrinsic base-collector depletion region and near the collector contact region are plotted for structures with different l_E at the same bias point (i.e. V_{BE}=0.8 V and V_{CE}=2 V). The results for transistors with l_E=0.25, 0.5 and 1 μm are shown in Fig. 13. For transistors with larger l_E increase of current in a cross-section close to the intrinsic base-collector junction (Fig. 13a~c) appears below the white line which marks the depletion region edge. This means that portion of current is pushed downwards. In case of smaller l_E transistors, this increase is not observed meaning that current spreads laterally. This is also observed in a cross section close to the collector contact region (Fig. 13d~f) where we can see that for larger l_E, region of higher current density appears near the middle of the transistor. This increase of the current density can cause the onset of base push-out (i.e. Kirk effect) and reduce the maximum f_T. To confirm this, the results from Fig. 5 are plotted against collector current normalized to l_E in Fig. 14. Since current is scaled linearly with l_E (see Fig. 3) we can conclude about local current density from the onset of

Kirk effect. It can be observed that peak f_T is shifted to smaller normalized I_C/l_E for structure with larger l_E. f_T fall-off and therefore Kirk effect starts earlier meaning that local current density is higher. Current spread for small l_E is more pronounced and current density is locally reduced causing the Kirk effect at larger I_C/l_E.

IV. CONCLUSION

Impact of the emitter length scaling of HCBT with single polysilicon region is investigated by 3D device simulations. The collector current and junction capacitances have linear dependence on the emitter length. Collector resistance can be modeled by the parallel combination of the constant resistance representing the extrinsic transistor and the one seen from the emitter window, which scales linearly with the emitter length. Cut-off frequency (f_T) of HCBT with single polysilicon region is improved for transistors with small emitter length. Current spreading in the collector reduces the collector current density and postpones the onset of base push-out to higher collector currents. It was shown that f_T is improved by approximately 10 % compared to long emitter devices. From the perspective of device modeling this imposes the constraint on the usage of HCBT models in circuit simulator. Separate set of model parameters should be given for transistors with small emitter length.

REFERENCES

[1] Y. Cheng, C. Hu, "MOSFET Modeling and BSIM3 User Guide," Kluwer Academic Publishers, 1999.

[2] W. Liu. C. Hu,"BSIM4 and MOSFET Modeling for IC Simulation," World Scientific Publishing, 2011.

[3] T. Suligoj, M. Koričić, H. Mochizuki, S. Morita, K. Shinomura, H. Imai, "Horizontal Current Bipolar Transistor (HCBT) for the Low-cost BiCMOS Technology", European Solid-State Device Research Conference ESSDERC 2009, pp. 359-362

[4] T. Suligoj, M. Koričić, J. Žilak, H. Mochizuki, S. Morita, K. Shinomura and H. Imai, Optimization of Horizontal Current Bipolar Transistor (HCBT) Technology Parameters for Linearity in RF Mixer," in *Proc. IEEE Bipolar / BiCMOS Circuits and Technology Meeting,* 2013, pp. 13-16.

[5] T. Suligoj, M. Koričić, H. Mochizuki, S. Morita, K. Shinomura, and H. Imai, " Horizontal Current Bipolar Transistor (HCBT) with a Single Polysilicon Region for Improved High-Frequency Performance of BiCMOS ICs," *IEEE Electron Device Lett.*, vol. 31, no. 6, pp. 534-536, June 2010.

[6] J. Böck, H. Knapp, K. Aufinger, T. F. Meister, M. Wurzer, S. Boguth, L. Treitinger, "High-Performance Implanted Base Silicon Bipolar Technology for RF Applications," *IEEE Transactions on Electron Devices*, vol. 48, no. 11, pp. 2514-2519, November 2001.

[7] M. Koričić, T. Suligoj, H. Mochizuki, S. Morita, K. Shinomura and H. Imai "Design Considerations for Integration of Horizontal Current Bipolar Transistor (HCBT) with 0.18 µm Bulk CMOS Technology," *Solid-state electronics*, Vol. 54; pp. 1166-1172, 2010.

[8] T. Suligoj, M. Koričić, H. Mochizuki, S. Morita, K. Shinomura, and H. Imai, "Collector Region Design and Optimization in Horizontal Current Bipolar Transistor (HCBT)," in *Proc. IEEE Bipolar / BiCMOS Circuits and Technology Meeting,* 2010, pp. 212-215.

[9] Y. Taur, T. H. Ning, "Fundamentals of Modern VLSI Devices," Cambridge University Press, 1998.

[10] T. Onai, K. Nakazato. Y. Kiyota, T. Nakamura, "FRACS (Fully Radiative Current Path Structure) - A High-speed Bipolar Transistor with Sub-0.1 µm Emitter, " *Symposium on VLSI Technology, Tech. Digest*., 1992., pp. 54-55.

[11] I. Getrau, "Modeling the Bipolar Transistor", Elsevier Scientific Publishing Company, 1978.

Optimization of Floating Guard Ring Parameters in Separate-Absorption-and-Multiplication Silicon Avalanche Photodiode Structure

I. Janeković*, T. Knežević*, T. Suligoj*, D. Grubišić**
* Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
** Laser Components DG, Inc., Tempe, Arizona, USA
tomislav.suligoj@fer.hr

Abstract - In this paper, the effect of the floating guard ring (FGR) parameters, such as number of FGRs and the distances between the FGRs, on the prevention of the breakdown at the multiplication layer edge for a given avalanche photodiode (APD) structure has been analyzed. Four structures are examined: APD without FGR and APDs with one, two and three FGRs. The width and depth of the boron-doped FGRs is 1 µm with peak concentration of 10^{18} cm^{-3}. The intrinsic breakdown voltage in the active region of the APD is 371 V and the epitaxial layer is fully depleted at 14 V. The breakdown occurs in the active area of the APD structure by using three FGRs. The optimal position of three FGRs are determined and the breakdown voltage of 409 V is achieved at the multiplication edge, which is higher by 38 V than in the active region. In contrast, the edge breakdown voltage is 170 V, 270 V and 348 V for structures without floating guard rings, with one FGR and with two FGRs, respectively. Spectral responsivity of the structure is analyzed in the range of 400 to 1000 nm showing that such APD can be effectively used for detection of the visible light.

I. INTRODUCTION

Avalanche photodiodes (APDs) are commonly used devices for photon detection in optical communication systems because of their high gain, high sensitivity and high quantum efficiency [1], [2]. Additionally, the APDs are also exploited in laser systems, ceilometers and astronomical observations. Mostly exploited structures are beveled edge diode, reach through diode and guard ring diode [3]. The diodes with guard rings are compatible with planar technologies such as CMOS [3], which make them attractive for numerous applications.

Separate-absorption-and-multiplication (SAM) APDs with thin pn-junction have lower noise and higher operating speed as compared to APDs with a single absorption-multiplication region [4]. SAM APD structures have an accumulation region with low electric field with the main purpose of photon absorption, and a multiplication region with high electric field, where the avalanche multiplication occurs [4].

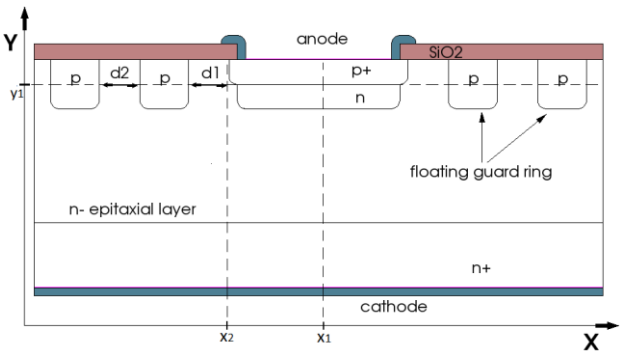


Figure 1. Cross section of simulated APD.

SAM APDs have a lower multiplication noise because either holes or electrons are causing impact ionization. Multiplication gain is not uniform and it depends on the wavelength of the incoming radiation.

APD structure has to be carefully designed such that the highest impact ionization is located in the active part of the device, where the incident radiation is absorbed. The critical part of the photodiode structure is the multiplication layer edge, where the premature breakdown can occur. The built-in electric field is higher at the edge of the multiplication region due to junction curvature, which can increase the noise of the APD, reduce the gain, and also reduce the desired breakdown voltage (V_{BR}). There are several techniques used to solve this problem, including metal plates located above the diode [5], shallow trench isolation (STI) [6], diffused guard rings and floating guard rings [6], [7]. Adding the guard ring close to the multiplication region edge will decrease the electric field near the multiplication junction, due to the reduced curvature, which will increase the local breakdown voltage. Choosing the optimum parameters of the FGRs such as the number of guard rings, position and the doping profile is critical. The goal of the optimization of the floating guard ring parameters is that the breakdown voltage of the multiplication region edge should be higher than the breakdown voltage of the active part of the structure. It must be kept in mind that the number of the

This work is sponsored in part by the Croatian Science Foundation under the project „High-performance Semiconductor Devices for Wireless Circuit and Optical Detection Applications" (contract no. 9006).

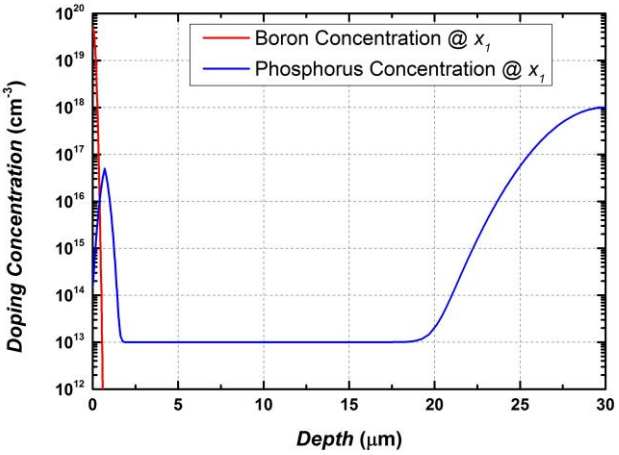


Figure 2. Doping profile in the active area part of the APD structure at the lateral distance of 75 μm (x_1 in Fig. 1).

FGRs should be minimal since every FGR adds a "dead region" to the whole photodiode structure, leading to a smaller photon absorption efficiency due to lower fill factor.

In this paper, we analyze how the change of the FGR parameters such as the number of FGRs and the distances between the FGRs prevents the breakdown at the multiplication region edge for a given APD structure. Four structures are analyzed: structure without FGR and structures with one, two and three FGRs. The optimal position and the parameters of FGRs are determined using the device simulation software. The spectral responsivity of the proposed structure is also analyzed.

II. APD STRUCTURE

Cross section of the simulated silicon photodiode is shown in Fig. 1. All the simulations have been performed using Sentaurus TCAD software [8]. Active part of the APD consists of the highly doped boron region with the thickness of 0.4 μm and the peak doping concentration of $5 \cdot 10^{19}$ cm^{-3} at the surface. Multiplication region is formed by defining an n-type Gaussian doping profile with the peak concentration of $5 \cdot 10^{16}$ cm^{-3} at the depth of 0.7 μm. Consequently, the pn-junction depth is 0.4 μm. Epitaxial layer is defined as uniformly doped layer with the concentration of 10^{13} cm^{-3}. Back side of the APD is doped by phosphorus with the peak concentration of 10^{18} cm^{-3}. Doping profile in the active part of the structure is shown in Fig. 2. Dimensions of the structure cross-section are 150 x 30 μm.

III. ELECTRICAL SIMULATIONS

Electrical simulations are performed using the following physical models. *Masetti* model for mobility degradation due to the impurity scattering is used. *Canali* model is used to simulate carrier velocity saturation due to high electric field. The *Slotboom* model is used for the bandgap narrowing model. Three generation-recombination models are used in the simulations: *SRH*, *Auger* and *van Overstraeten-de Man* for modeling the impact ionization generation [9].

The APD structure described in the section II has V_{BR} of 170 V. Comparison of the electric field at the edge of the multiplication region and in the active part of the structure is shown in Fig. 3. The electric field at the edge is higher than the electric field in the active part of the structure due to the junction curvature and a higher electric field. Consequently, the breakdown occurs at the edge of the multiplication region, leading to a higher optical sensitivity at the edge instead in the active region and preventing photogenerated carriers from obtaining sufficient energy in multiplication region, which could limit the gain.

The V_{BR} of the active region is obtained by 1D simulations of the intrinsic part of APD and equals 371 V, which is considerably higher than V_{BR} of 170 V resulting from 2D simulations. In order to improve APD characteristics and achieve the breakdown in the multiplication region, in the value obtained by 1D simulations, the guard rings should be added to the periphery of the structure. In this work we used the

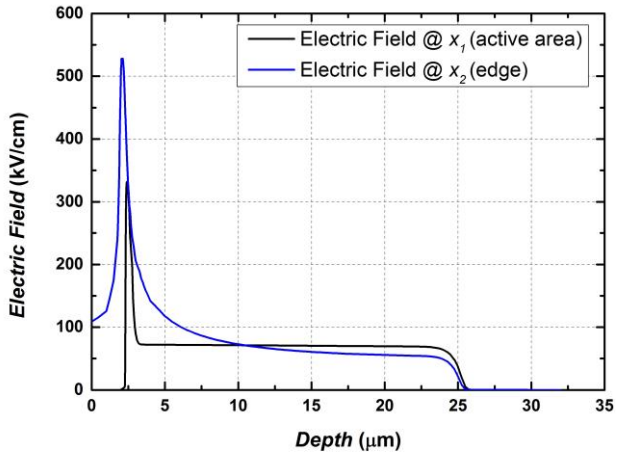


Figure 3. Electric field in the center of the active part of the structure (black line) and at the edge of the multiplication region (blue line) of the photodiode at the reverse bias V_D = 170 V.

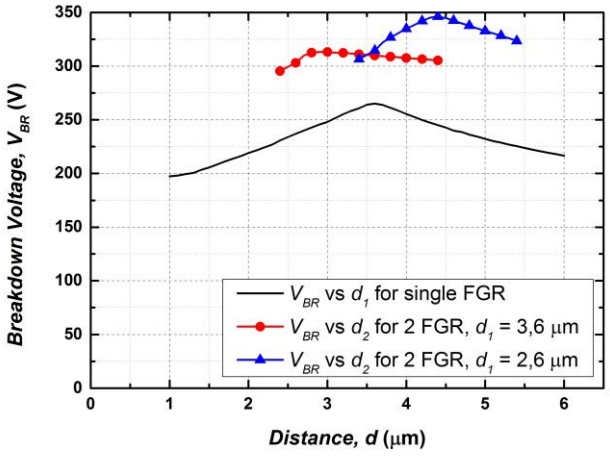


Figure 4. Breakdown voltage vs distance between the FGR and the multiplication region (black line) and breakdown voltage vs. the distance between the FGRs for d_1 = 3.6 μm (circles) and d_1 = 2.6 μm (triangles).

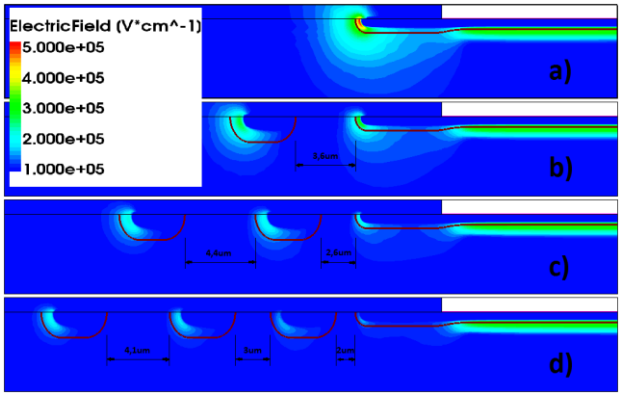


Figure 5. Electric field distribution at V_D = 160 V for the structure (a) without FGR, (b) with single FGR, (c) with two FGRs and (d) with three FGRs.

floating guard rings. The on-mask width of the FGRs is 1 µm, the junction depth is 1 µm, whereas the surface concentration is 10^{18} cm^{-3}.

Firstly, the addition of a single FGR to the periphery of the structure will be analyzed. In order to achieve the highest V_{BR}, the position of the first FGR has been varied with respect to the edge of the multiplication region (d_1). The maximum value of V_{BR} of 270 V is achieved for d_1 of 3.6 µm. For d_1 < 3.6 µm, the breakdown occurs at of the periphery of the FGR. For d_1 > 3.6 µm, the breakdown occurs at the edge of the multiplication region.

The next step is the addition of another FGR since only one FGR was unable to prevent the premature breakdown. The distance of the second FGR with respect to the first FGR (d_2) is also varied to determine the maximum V_{BR}. After adding the second FGR, it is necessary not only to find the optimal position of the second FGR, but also the optimal position of the first FGR. The highest V_{BR} of 348 V is achieved for d_1 = 2.6 µm and d_2 = 4.4 µm, which is still below the V_{BR} in the active region obtained by 1D simulation. Breakdown voltage versus the FGRs distances is depicted in Fig. 4. For the structure with two FGRs, choosing d_1 at the maximum V_{BR} does not result with the highest V_{BR} for varied d_2. As already mentioned, the position of the first

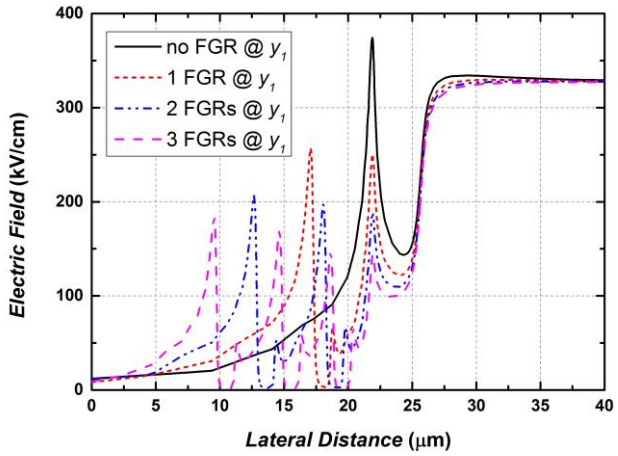


Figure 6. Electric field distribution at the edge of the multiplication region at the depth of 0.4 µm, V_D = 160 V.

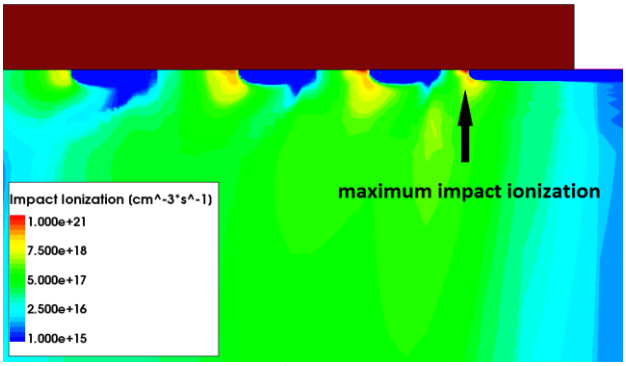


Figure 7. Impact ionization of the structure without multiplication region with three FGRs, d_1 = 2 µm, d_2 = 3 µm, d_3 = 4.1 µm. V_D = 409 V.

FGR needs to be optimized as well to achieve the highest V_{BR}. Finally, the third FGR has been added to the structure. The distance between the third FGR and the second FGR (d_3) is also varied to determine the maximum V_{BR}. After the optimization of d_1, d_2 and d_3, the highest breakdown voltage of 409 V has been achieved for the parameters d_1 = 2 µm, d_2 = 3 µm and d_3 = 4.1 µm. In order to reach the breakdown at the edge, higher than V_{BR} in the active region, such simulations has to be performed without the multiplication region in the active part of the structure. Distribution of the electric field for structures without FGRs, and with one, two and three FGRs at V_D = 160 V is depicted in Fig. 5. For the structure without FGRs, the electric field at the edge is larger comparing to the electric field in the multiplication region. In the structure with a single FGR the electric field is distributed between the FGR and the edge of the multiplication layer resulting in the reduced peak value. However, for lower V_D, the electric field increases faster at the edge of FGR as compared to the electric field in the multiplication region. Eventually, the breakdown will occur at the same place. Adding the second FGR causes the electric field to spread over the FGRs more uniformly, which decreases the field at the edge of the multiplication region. By adding the third FGR, the electric field has the highest value in multiplication region, while the field at the edge is redistributed between the FGRs. By adding FGRs, a better electrical isolation in photodiode's periphery is achieved, which means that the current is directed toward the active area of multiplication region. Electric field distribution at the depth of 0.4 µm, where its peak in the multiplication region is expected, is

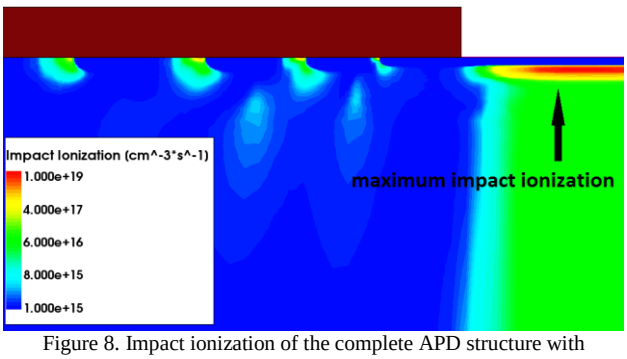


Figure 8. Impact ionization of the complete APD structure with multiplication region. V_D = 371 V.

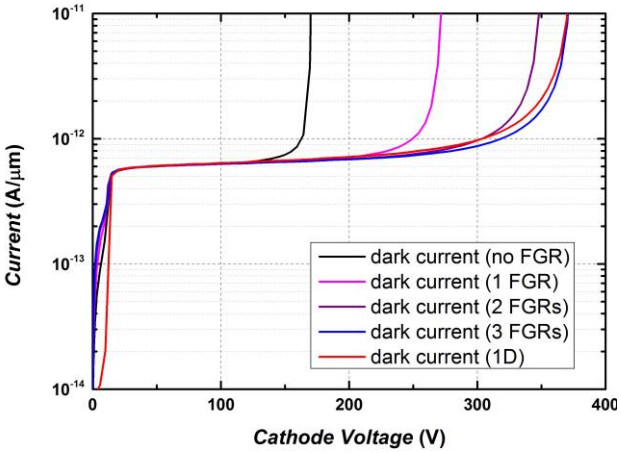


Figure 9. I-V characteristics of the structures without and with one, two and three FGRs.

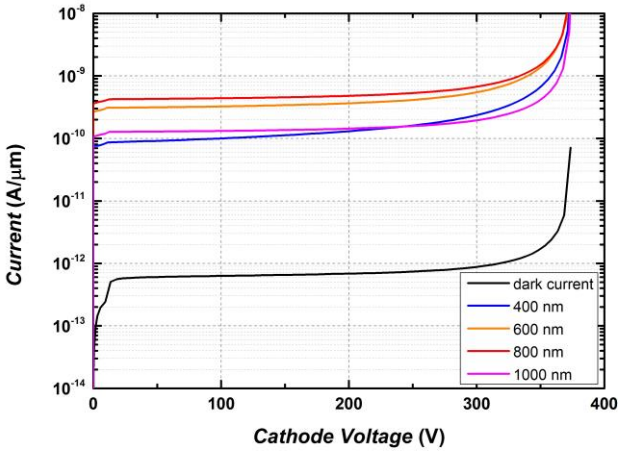


Figure 10. I-V characteristic of the optimized structure with three FGRs for different wavelengths. Light source intensity is 10^{-3} W/cm^2. TMM model is used in the simulations.

shown in Fig. 6. The diode voltage is V_D = 160 V, which is below V_{BR} of the structure without FGRs. Although the electric field at the edge of the multiplication region and guard rings does not have the maximum value at a given depth, it can be observed how FGRs reduce the electric field. Impact ionization at V_D = 409 V of the optimized structure with three FGRs and without the multiplication region is depicted in Fig. 7. Maximum impact ionization is at the edge of the structure, showing the location of the breakdown. The breakdown for structure without FGRs occurs at the same place. The V_{BR} at FGRs is 38 V higher than the V_{BR} of the multiplication region. Such a difference should be sufficient to compensate the process variations in reality.

Impact ionization for the structure with multiplication region is shown in Fig. 8. By adding the multiplication region, the maximum impact ionization is now shifted to the active part of the structure indicating the position of the breakdown. The I-V characteristics for the aforementioned structures are shown in Fig. 9. I-V characteristics show that the full depletion of the structure is at V_D = 14 V. Also, the increase of dark currents in the structures with added FGRs is negligible. The simulations were performed neglecting the silicon/oxide interface traps.

IV. OPTICAL SIMULATIONS

The optical simulations have also been performed for the optimized structure with 3 FGRs. Light source with intensity of 10^{-3} W/cm^2 and with varied wavelengths is assumed. The I-V characteristics of the dark and illuminated structures are shown in Fig. 10. Wavelengths of the light source are 400 nm, 600 nm, 800 nm and 1000 nm. In simulations silicon band gap is taken into account for describing how many of the absorbed photons are converted to generated electron-hole pairs. Wavelength dependent refractive and absorption index has been used. An illumination window is defined only over the active area. The gain dependency on diode voltage is depicted in Fig. 11 for APD illuminated with wavelength of 600 nm and light source intensity of 10^{-3} W/cm^2.

The avalanche multiplication does not have a significant impact on the APD performance for voltages below 350 V. The higher diode voltages cause a higher impact ionization and a higher internal gain of the structure caused by the avalanche multiplication. In the subsequent optical analysis, the diode is biased at 372.26 V, which corresponds to the gain of 50.

Spectral responsivity of the optimized APD structure for gain of 50 is shown in Fig. 12 using two simulation models. Black curve represents responsivity simulated with model where only the Beer-Lambert law for light absorption is included. Responsivity has values over 20 A/W for wavelengths from 415 to 830 nm, which means that the analyzed structure can be effectively used for detection of visible spectrum light. Red curve represents responsivity with simulated Transfer Matrix Method (TMM) model, where the reflection, transmission and total absorption is included in the simulations. This gives more realistic model for the optical detection. Responsivity with the TMM model is 35% lower compared to the first model at the wavelength of 600 nm. This can be improved if an anti-reflecting coating is applied on APD surface.

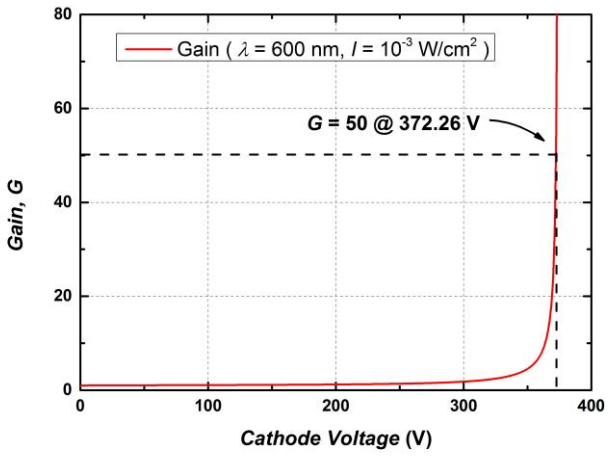


Figure 11. Gain vs the diode voltage for illuminated APD. Light source intensity is 10^{-3} W/cm^2 and the wavelength is 600 nm.

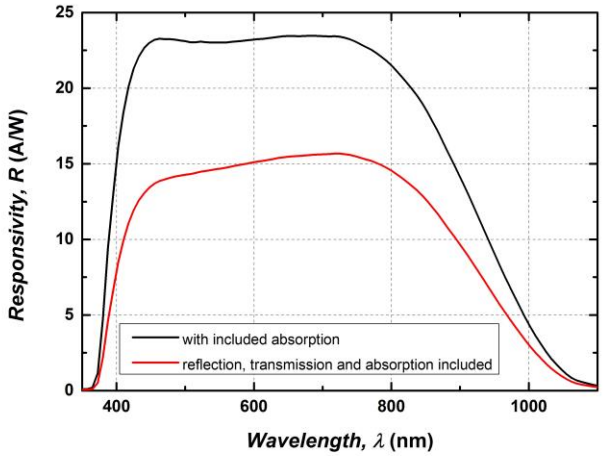


Figure 12. Spectral responsivity of the structure at V_D = 372.26 V for light source intensity of 10^{-3} W/cm^2.

V. CONCLUSION

It has been shown in the paper that APDs can suffer from premature breakdown at the edge of multiplication region. For that reason it is necessary to implement guard rings around photosensitive area. Edge breakdown has been successfully prevented by using three floating guard rings. After adding FGRs, the highest value of breakdown voltage of 409 V is achieved for structure without multiplication region and with the parameters d_1 = 2 µm, d_2 = 3 µm and d_3 = 4.1 µm. The breakdown of APD with the multiplication region is 371 V as predicted in 1D model, and it occurs in the multiplication region. The optical responsivity dependency on wavelength has shown that the simulated structure is suitable for the visible light detection.

REFERENCES

[1] K.M. Johnson, "High-speed photodiode signal enhancement at avalanche breakdown voltage", IEEE Transactions on Electron Devices, 1965.

[2] J.R. Biard, W.N. Shaunfield Jr, "A model of the avalanche photodiode", IEEE Transactions on Electron Devices, Vol. ed-14, No. 5, May 1967.

[3] M.A. Othman, S.N. Taib, M.N Husain, Z.A.F.M. Napiah, "Reviews of avalanche photodiode for optical communication technology", ARPN Journal of Engineering and Applied Sciences, Vol. 9, No.1, January 2014.

[4] K.A. Anselm et al., "Performance of thin separate absorption, charge, and multiplication avalanche photodiodes", IEEE Journal of Quantum Electronics, Vol. 34, No. 3, pp. 482–490, March 1998.

[5] F. Conti, M. Conti, "Surface breakdown in Si planar diodes equipped with field plate", Solid State Electronics 15 (1972) 93.

[6] E. Kamrani, F. Lesage, M. Sawan, "Premature edge breakdown prevention techniques in CMOS APD fabrication", New Circuits and Systems Conference, IEEE 10th International, pp. 345–348, Montreal, QC, June 2012.

[7] H. Finkelstein, M.J. Hsu, S.C. Esener, "STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology", IEEE Electron Device Letters, Vol. 27, No. 11, pp. 887–889, November 2006.

[8] Sentaurus Workbench, H-2013.03, 2013. Synopsys, Inc.

[9] R. van Overstraeten, H. de Man, "Measurement of the ionization rates in diffused silicon p-n junctions", Solid-state Electronics, Pergamon Press, Vol. 13, pp. 583-608, 1970

Low-Power Low-Voltage CMOS Analog Signal Processing Circuits Using a Functional Core

Cosmin Popa
Faculty of Electronics, Telecommunications and Information Technology
University Politehnica of Bucharest
e-mail: cosmin_popa@yahoo.com

Abstract **- The paper presents an original approach of designing analog signal processing circuits, based on the re-using of the same functional core for implementing a multitude of circuit functions: signal gain with theoretical null distortions, signal squaring, voltage multiplying with very good linearity and simulation of a perfect linear resistor with both positive and negative equivalent resistance. The great advantages of the increased modularity and controllability and of the reduced design costs represent an immediate consequence of the multiple functions realized by the proposed structures. Because the most important circuit complexity is concentrated in implementing the core of the multifunctional structure, both circuit area and power consumption per each realized function can be strongly reduced using this method, especially for multifunctional structures that are able to implement an important number of functions. The overall error of the linear designed computational structures is *0.4%* and the approximation error for the squaring circuit is *0.27%*, in the condition of a low-voltage low-power operation (a supply voltage of *1.5V* and a medium current consumption of *50μA* for each implemented circuit function).**

Keywords - analog signal processing, functional core, linearity, multifunctionality, low-voltage low-power operation

I. INTRODUCTION

Analog signal processing (ASP) structures are extensively used in the latest VLSI designs. Even their capability of generating circuit functions is very wide, most of these functions are developed starting from a relatively limited number of primitive functions. From this point of view, it could be identified two important classes of ASP structures: with linear and nonlinear behavior. The nonlinear components from the operation of an ASP structure are usually related to polynomial functions or, in most cases, to squaring function (because of the extensively use of the squaring characteristic specific to a MOS transistor biased in saturation).

We will present in this paper an intuitive design of ASP circuits that is able to generate using an original method a multitude of linear and nonlinear circuit functions.

An important goal in ASP is represented by the possibility of reusing the same functional core for implementing a multitude of circuit functions, the increased modularity that could be achieved being reflected in an important reduction of the design costs per implemented function.

The proposed ASP circuit realizes the following circuit functions:

- Signal gain with theoretical null distortions using a transconductance amplifier circuit;
- Signal squaring with increased accuracy using a squaring circuit;
- Simulation of a perfect linear resistor with both positive and negative resistance using an active resistor circuit;
- Voltage multiplying with very good linearity using a multiplier circuit.

The transconductance amplifier is an important stage in VLSI designs, having many applications in a large area of high-performances analog/mixed ICs, such as operational amplifiers, voltage comparators, voltage regulators, video amplifiers, modulators and demodulators or A/D and D/A converters. The linearity of the circuit is limited because of the squaring characteristic of MOS transistors biased in saturation, being possible to achieve a relatively good linearity only for a limited input voltage range. Thus, the transconductance amplifiers require the implementation of a linearization technique for decreasing the superior-order nonlinearities and for increasing the maximal range for the input voltage amplitudes. There were reported in literature many techniques [1]-[4] for improving the MOS transconductance amplifier linearity.

Squaring circuits [5]-[8] are very important CMOS analog signal processing structures, finding multiple applications in analog and mixed-signal designs. Their implementations are relatively simple using the squaring characteristic of the MOS transistor biased in saturation.

CMOS active resistors [9]-[18] are widely used in VLSI analog designs for replacing the large value passive resistors, having the great advantages of the controllability of the equivalent resistance. Amplitude control in low distortion oscillators, voltage controlled amplifiers and active RC filters represent their usual utilisation domains. The first generation of MOS active resistors [11], [12] exploits the quasi-linear characteristic of MOS transistors working in the linear region. The disadvantage of this approach of designing active resistor structures is that the circuit operation is inherently nonlinear and the distortion components depend on MOS technological parameters. An improved method for obtaining CMOS active resistors exclusively uses MOS transistors biased in saturation. A multitude of linearization techniques [12]-[17] have been implemented in order to minimize the nonlinearity of active resistor circuits. The active resistor structures offer the

possibility of obtaining a controllable negative equivalent resistance, finding very large domains of applications.

Multiplier circuits represent intensively used blocks in telecommunication applications, medical equipments, hearing devices and disk drives. In order to improve the linearity of a classical multiplier structure, specific techniques [18]-[21] have been reported in literature.

II. THEORETICAL ANALYSIS

A. Design of the functional core

A functional core represents a computational circuit having a reasonable complexity that is able to generate a multitude of circuit functions.

The most important advantage of the proposed method of designing analog signal processing circuits using a functional core is related to the multifunctional capability. As the most important design effort is concentrated for improving the performance of the functional core (which is common for all the implemented circuit functions) and because this core has the largest complexity, all designing costs per each implemented function will be strongly reduced comparing with classical approaches of designing analog signal processing circuits. Additionally, the power consumption for an implemented function can be reduced because the functional core consumes the most important power of the circuit.

The proposed implementation of the functional core is presented in Fig. 1. It consists in a differential amplifier, having a composing MOS transistor biased at a constant reference current, I_O.

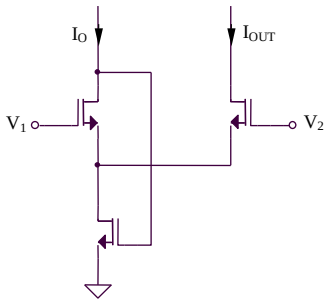


Fig. 1. Functional core

The differential input voltage of the functional core can be expressed as follows:

$$V_1 - V_2 = V_{GS}(I_O) - V_{GS}(I_{OUT}). \quad (1)$$

Biasing in saturation all MOS transistors from Fig. 1, it results:

$$V_1 - V_2 = \sqrt{\frac{2}{K}}\left(\sqrt{I_O} - \sqrt{I_{OUT}}\right). \quad (2)$$

The expression of the functional core output current will be:

$$I_{OUT} = I_O - \sqrt{2KI_O}(V_1 - V_2) + \frac{K}{2}(V_1 - V_2)^2. \quad (3)$$

The advantage of this implementation is given by the availability of both linear and squaring terms with respect to the differential input voltage variable, $V_1 - V_2$.

B. Design of a linear transconductance amplifier

In order to obtain a linear transconductance amplifier, two parallel-connected functional cores from Fig. 1 will be used, resulting the transconductance amplifier circuit presented in Fig. 2.

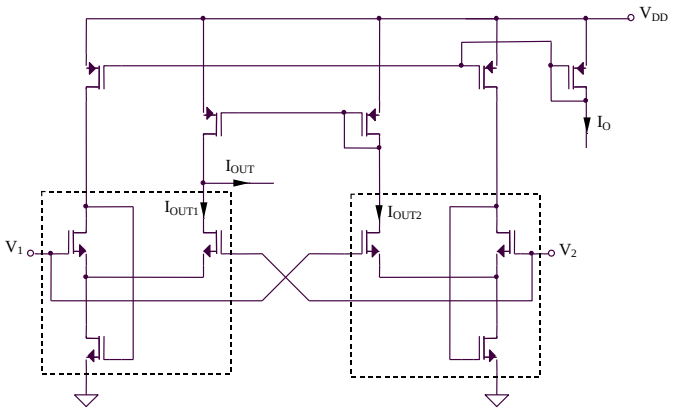


Fig. 2. Transconductance amplifier

The expressions of I_{OUT1} and I_{OUT2} currents are:

$$I_{OUT1} = I_O - \sqrt{2KI_O}(V_1 - V_2) + \frac{K}{2}(V_1 - V_2)^2 \quad (4)$$

and:

$$I_{OUT2} = I_O + \sqrt{2KI_O}(V_1 - V_2) + \frac{K}{2}(V_1 - V_2)^2. \quad (5)$$

The output current of the transconductance amplifier will be expressed by:

$$I_{OUT} = I_{OUT2} - I_{OUT1} = \sqrt{8KI_O}(V_1 - V_2). \quad (6)$$

illustrating a linear dependence of the I_{OUT} output current on the $V_1 - V_2$ differential input voltage.

C. Design of a squaring circuit

The squaring function can be obtained using the last term from the expression (3) of the functional core output current. Considering the sum of I_{OUT1} and I_{OUT2} output currents from Fig. 3, expressed by (4) and (5), the linear term with respect to the differential input voltage will be cancel out:

$$I_{OUT} = I_{OUT1} + I_{OUT2} - 2I_O = K(V_1 - V_2)^2. \quad (7)$$

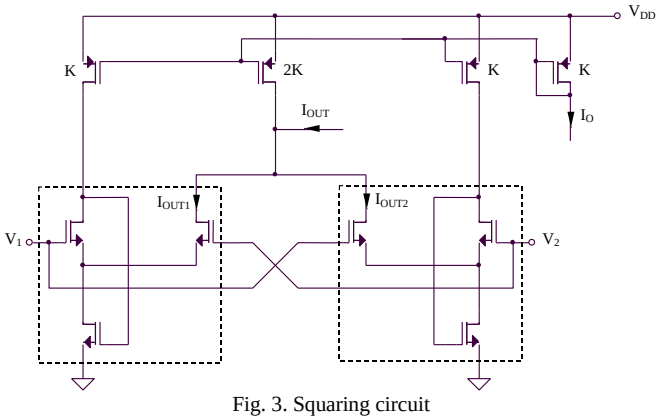


Fig. 3. Squaring circuit

D. Design of a linear multiplier circuit

The original idea for obtaining the multiplying function (Fig. 4) is to use an anti-parallel connection of two functional cores from Fig. 1, M1-M3 and M4-M6, respectively. The I_{O1} and I_{O2} currents are generated using two parallel-connected differential amplifiers, M7-M10.

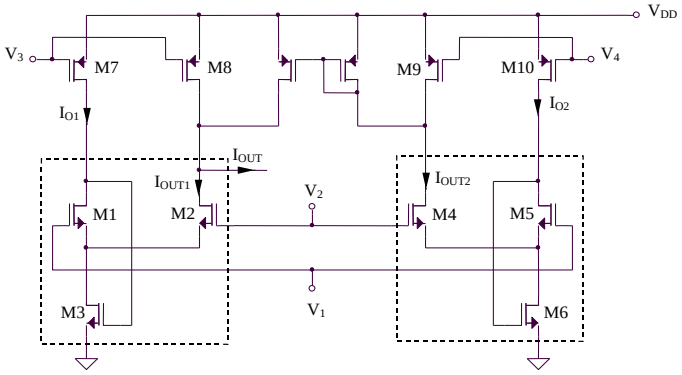


Fig. 4. Multiplier circuit

The expressions of I_{OUT1} and I_{OUT2} output currents are:

$$I_{OUT1} = I_{O1} - \sqrt{2KI_{O1}}(V_1 - V_2) + \frac{K}{2}(V_1 - V_2)^2 \qquad (8)$$

and:

$$I_{OUT2} = I_{O2} - \sqrt{2KI_{O2}}(V_1 - V_2) + \frac{K}{2}(V_1 - V_2)^2, \qquad (9)$$

while I_{OUT} current can be expressed as follows:

$$I_{OUT} = (I_{OUT2} - I_{O2}) - (I_{OUT1} - I_{O1}) = \\ = \sqrt{2K}\left(\sqrt{I_{O1}} - \sqrt{I_{O2}}\right)(V_1 - V_2) \qquad (10)$$

Replacing the squaring dependencies of I_{O1} and I_{O2} currents on the source-gate voltages of M7-M10 transistors, it results:

$$I_{OUT} = K(V_1 - V_2)(V_4 - V_3). \qquad (11)$$

E. Design of linear positive and negative active resistor circuits

In order to simulate an active resistor circuit with improved linearity, the proposed method is to consider two functional cores from Fig. 1 and to pass between the input terminals the difference of their output currents, $I_{OUT2} - I_{OUT1}$, expressed by (6) (Fig. 5). Because of a linear current-voltage characteristic, it is possible to define an equivalent resistance between V_1 and V_2 as follows:

$$R_{ECH} = \frac{V_1 - V_2}{I_{OUT2} - I_{OUT1}} = \frac{1}{\sqrt{8KI_O}}. \qquad (12)$$

The equivalent resistance can be controlled by the I_O reference current.

Changing the sense of the current passing between the input terminals (Fig. 6), it is possible to obtain a controllable negative equivalent resistance:

$$R_{ECH}' = \frac{V_1 - V_2}{I_{OUT1} - I_{OUT2}} = -\frac{1}{\sqrt{8KI_O}}. \qquad (13)$$

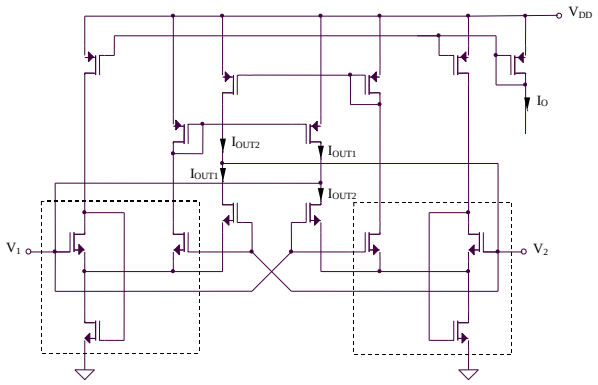


Fig. 5. Active resistor with positive equivalent resistance

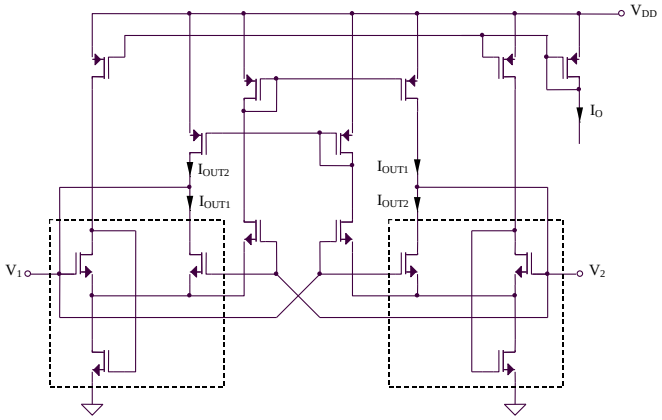


Fig. 6. Active resistor with negative equivalent resistance

III. SIMULATIONS

The simulation of the transconductance amplifier transfer characteristic (6) is presented in Fig. 7, illustrating an excellent linearity (a simulated linearity error of *0.4%*).

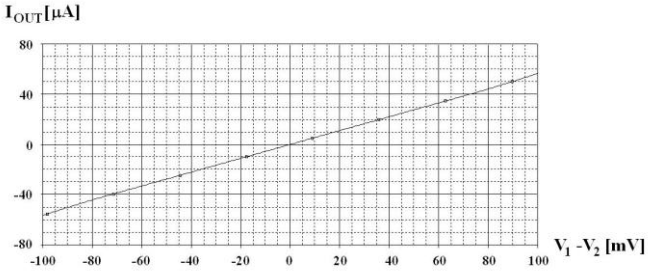


Fig. 7. Simulated transfer characteristic of the transconduactance amplifier

The simulation of the squaring circuit characteristic (7) is presented in Fig. 8, resulting a simulated approximation error $\varepsilon = 0.27\%$.

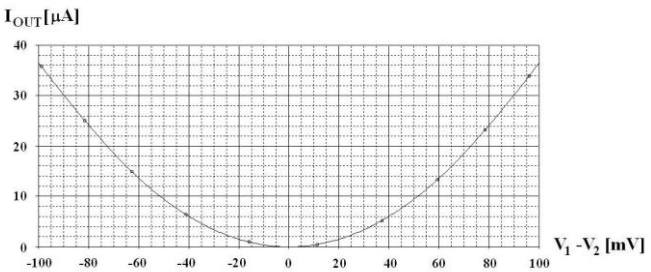


Fig. 8. Simulated characteristic of the squaring circuit

The simulations were made using SPICE, for a 0.35 μm CMOS process.

IV. CONCLUSIONS

An original design technique for obtaining analog signal processing circuits able to generate a multitude of circuit functions has been presented. The paper analyses and implements a large number of analog signal processing circuits, exploiting the specific characteristic of a single functional core. In this way, it is possible to obtain, with minor changes in the design, five important functions: signal gain with theoretical null distortions, signal squaring, voltage multiplying with very good linearity and simulation of a perfect linear resistor with both positive and negative equivalent resistance. As a consequence of the multiple functions implemented using the same functional core, the proposed structure presents the advantages of an increased modularity and controllability and of the associated reduced design costs. Comparing with similar designs from literature, the linearity of the proposed structure is strongly increased by implementing original techniques, while the circuit complexity per function is reduced as a result of the proposed multifunctionality. The linear designed computational structures has an overall error of 0.4%, while the proposed squaring circuit has an approximation error of 0.27%, all of them being designed for a low-voltage low-power operation (the supply voltage is 1.5V and the medium current consumption for each implemented circuit function is 50μA).

REFERENCES

[1] Grasso, A.D., Pennisi, S., „High-Performance CMOS Pseudo-Differential Amplifier", Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on, 23-26 May 2005, pp. 1569 - 1572 Vol. 2.

[2] Jongchan Kang, Kyungho Lee, Jehyung Yoon, Yisun Chung, Sungbo Hwang, Bumman Kim, „Differential CMOS Linear Power Amplifier with 2nd Harmonic Termination at Common Source Node", Radio Frequency integrated Circuits (RFIC) Symposium, 2005. Digest of Papers. 2005 IEEE, 12-14 June 2005, pp. 443-446.

[3] Jongchan Kang, Jehyung Yoon, Kyoungjoon Min, Daekyu Yu, Joongjin Nam, Youngoo Yang, Bumman Kim, „A Highly Linear and Efficient Differential CMOS Power Amplifier with Harmonic Control", Solid-State Circuits, IEEE Journal of, Volume 41, Issue 6, June 2006, pp. 1314-1322.

[4] Mohieldin, A.N., Sanchez-Sinencio, E., Silva-Martinez, J., „A Fully Balanced Pseudo-Differential OTA with Common-Mode Feedforward and Inherent Common-Mode Feedback Detector", Solid-State Circuits, IEEE Journal of, Volume 38, Issue 4, April 2003, pp. 663-668.

[5] "A novel two quadrant MOS translinear Squarer-divider cell", De La Blas, C.A., Lopez, A., Electronics, Circuits and Systems, 2008. ICECS 2008. 15th IEEE International Conference on, 2008, Pp.5 – 8

[6] Naderi, A., Mojarrad, H., Ghasemzadeh, H., Khoei, A., Hadidi, K., "Four-quadrant CMOS analog multiplier based on new current squarer circuit with high-speed", EUROCON 2009, IEEE, 2009, Pp.282 – 287

[7] De La Cruz-Blas, C.A., Lopez-Martin, A.J., Carlosena, A., "1.5-V square-root domain second-order filter with on-chip tuning", , Circuits and Systems I: Regular Papers, IEEE Transactions on, Vol. 52, Issue: 10, 2005, Pp.1996 – 2006

[8] Garofalo, V., Coppola, M., De Caro, D., Napoli, E., Petra, N., Strollo, A.G.M., "A novel truncated squarer with linear compensation function", Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, 2010, Pp.4157 – 4160

[9] C. Popa, "Synthesis of computational structures for analog signal processing," Springer Publishing House, 449 pp., 2011

[10] C. Popa, "Superior-order curvature-correction techniques for voltage references," Springer Publishing House, 179 pp., 2009

[11] Z. Wang, "Current-controlled Linear MOS Earthed and Floating Resistors and Application," IEEE Proceedings on Circuits, Devices and Systems, 1990, pp. 479-481.

[12] L. Sellami, "Linear Bilateral CMOS Resistor for Neural-type Circuits," Proceedings of the 40th Midwest Symposium on Circuits and Systems, vol. 2, 1997, pp. 1330-1333.

[13] H. Zouaoui-Abouda, and A. Fabre, "New high-value floating controlled resistor in CMOS technology", Instrumentation and Measurement, IEEE Transactions on, vol. 55, Issue 3, June 2006, pp. 1017-1020.

[14] K. Vavelidis, Y. Tsividis, "R-MOSFET Structure Based on Current Division," Electron. Lett., vol. 29, 1993, pp. 732–733

[15] K.Vavelidis, Y. P. Tsividis, F. Op't Eynde, Y. Papananos, "Six-Terminal MOSFET's: Modelling and Applications in Highly Linear, Electronically Tunable Resistors," IEEE J. Solid-State Circuits, vol. 32, 1997, pp. 4–12

[16] C. Muniz-Montero, R. Gonzalez-Carvajal, A. Diaz-Sanchez, J.M. Rocha, "Low Frequency, Current Mode Programmable KHN Filters Using Large-Valued Active Resistors", Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on, 2007

[17] T. Sanchez-Rodriguez, C. Lujan-Martinez, R. Carvajal, J. Ramirez-Angulo, A. Lopez-Martin, "CMOS Linear Programmable Transconductor Suitable for Adjustable Gm-C Filters", Electronics Letters, Volume 44, Issue 8, April 10 2008 Pp. 505 – 506

[18] Sawigun, C., Mahattanakul, J., "A 1.5V, wide-input range, high-bandwidth, CMOS four-quadrant analog multiplier", Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on, 2008, Pp. 2318 – 2321

[19] Sawigun, C., Demosthenous, A., Pal, D., "A low-voltage, low-power, high-linearity cmos four-quadrant analog multiplier", Circuit Theory and Design, 2007. ECCTD 2007. 18th European Conference on, 2007 , Pp. 751 - 754

[20] Akshatha, B.C., Akshintala, V.K., "Low Voltage, Low Power, High Linearity, High Speed CMOS Voltage Mode Analog Multiplier", Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on, 2009, Pp. 149 – 154

[21] Mahmoud, S.A., "Low voltage low power wide range fully differential CMOS four-quadrant analog multiplier", Circuits and Systems, 2009. MWSCAS '09. 52nd IEEE International Midwest Symposium on, 2009, Pp. 130 – 133

Multifunctional Circuit Using Improved Performances Linearization Technique

Cosmin Popa
University Politehnica of Bucharest,
Faculty of Electronics, Telecommunications and Information Technology, Romania
cosmin_popa@yahoo.com

Abstract **- Original low-power low-voltage multifunctional structure with improved performances will be presented, allowing to implement (with minor changes in the design) two important functions: signal gain with theoretical null distortions and voltage multiplying with very good linearity. The great advantages of the increased modularity and controllability and of the reduced design costs associated represent an immediate consequence of the multiple functions realized by the proposed structures. The linearity is strongly increased by implementing original techniques, while the silicon occupied area per function is reduced as a result of the proposed multifunctionality. The structures are supplied at $\pm 3V$, presenting a very good linearity (in the worst case, $THD<0.4\%$ for differential amplifiers and $THD<0.5\%$ for multiplier), correlated with an extended range of the input voltage (at least $\pm 0.5V$).**

Keywords - Linearity, multifunctionality, computational circuits, VLSI design

I. INTRODUCTION

An important goal in VLSI designs is represented by the possibility of a multiple use for the same cell, the increased modularity that could be achieved being reflected in an important reduction of the design costs per implemented function.

The differential amplifier is an important stage of a large area of applications, including high-performances analog/mixed ICs, such as operational amplifiers, voltage comparators, voltage regulators, video amplifiers, modulators and demodulators or A/D and D/A converters. The linearity of the circuit is relatively poor because of the fundamental nonlinear characteristic of MOS transistors, resulting the possibility of achieving a relatively good linearity only for a restricted input voltage range (the amplitude of the input voltage for the classic differential amplifier using MOS transistors in saturation have to be below a few hundreds of mV). In conclusion, it results the necessity of implementing a linearization technique for decreasing the superior-order nonlinearities of the MOS differential stage and for increasing the available range for the input voltage amplitudes. It exists in literature many circuit techniques used to improve the MOS differential amplifier linearity. It was presented in [1], [2] a third and fifth-order harmonics cancellation with good results and a relatively simple circuit implementation.

A constant-sum of the gate-source voltages circuit connection was described in [3] and it allows an important reduction of the total harmonic distortions coefficient of the circuit. In [4], it was presented and implemented in CMOS technology a simple technique based on square-root circuits for improving the CMOS differential stage linearity, which compensates the quadratic characteristic of the MOS transistor in saturation.

Computational circuits and particularly multiplier circuits are important building blocks for telecommunication applications, medical equipments, hearing devices and disk drives. In CMOS technology, analog circuits based on the square-algebraic identity can be easily realized from the well-known square-law model of the MOS transistors in the saturation region. Based on this principle, several basic multiplier circuits have been developed [5]-[9], having the important disadvantage of a relatively poor linearity.

The original idea proposed in this paper is to use a linear CMOS differential amplifier for obtaining (with minor changes in the design) two important functions:

- The signal gain with theoretical null distortions;
- Voltage multiplying with very good linearity.

II. THEORETICAL ANALYSIS

A. Analysis of the classical MOS differential amplifier

The analysis of the large signal operation for the classical MOS differential structure (Fig. 1) can quantitatively evaluate the circuit's nonlinearity, being possible to determine the weight of each superior-order distortion introduced by the structure nonlinearity.

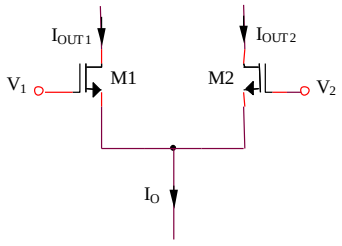


Fig. 1. Classical MOS differential structure

The output currents of the classical differential amplifier can be expressed as a function of the differential input voltage as follows:

$$(I_{OUT1})_{1,2} = \frac{I_O}{2} \pm \frac{I_O}{2}\sqrt{\frac{KV^2}{I_O} - \frac{K^2V^4}{4I_O^2}} \qquad (1)$$

so:

$$I_{OUT1} = \frac{I_O}{2} + \frac{I_O}{2}\sqrt{\frac{KV^2}{I_O} - \frac{K^2V^4}{4I_O^2}} \quad (2)$$

$$I_{OUT2} = \frac{I_O}{2} - \frac{I_O}{2}\sqrt{\frac{KV^2}{I_O} - \frac{K^2V^4}{4I_O^2}} \quad (3)$$

The differential output current will be:

$$I_{OUT2} - I_{OUT1} = -\frac{V}{2}\sqrt{4KI_O - K^2V^2} \quad (4)$$

The $(I_{OUT2} - I_{OUT1})(V)$ function is strongly nonlinear, the quantitative evaluation of the nonlinearity being possible using the Taylor series expansion. The Taylor series expansion of the transfer function (4) can be written as follows:

$$(I_{OUT2} - I_{OUT1})(V) = -K^{1/2}I_O^{1/2}V + \frac{K^{3/2}}{8I_O^{1/2}}V^3 + \frac{K^{5/2}}{128I_O^{3/2}}V^5 + \ldots \quad (5)$$

or:

$$(I_{OUT2} - I_{OUT1})(V) = a_1V + a_3V^3 + a_5V^5 + \ldots \quad (6)$$

a_k being constant coefficients with respect to V differential input voltage. So, as a result of squaring characteristics of MOS transistors biased in saturation, the classical MOS differential amplifier behavior is extremely nonlinear. The first term from (6) is linearly dependent on the input voltage, while the following two terms model the third-order and fifth-order nonlinearities of the differential structure. In conclusion, it is obviously the necessity of implementing a linearization technique for improving the performances of the classical MOS differential stage (good linearity for a relatively extended range for the amplitudes of the input voltage).

B. The analysis of the classical MOS Gilbert multiplier

The classical MOS Gilbert multiplier schematic is presented in Fig. 2. The differential output current of the circuit can be expressed as follows:

$$I_{OUT1} - I_{OUT2} = (I_{D1} - I_{D4}) - (I_{D2} - I_{D3}) \quad (7)$$

where:

$$I_{D1} - I_{D4} = (V_1 - V_2)\sqrt{\frac{KI_O}{2} - \frac{K^2}{4}(V_1 - V_2)^2 - \frac{K}{2}(V_3 - V_4)\sqrt{KI_O - \frac{K^2}{4}(V_3 - V_4)^2}} \quad (8)$$

and:

$$I_{D2} - I_{D3} = (V_1 - V_2)\sqrt{\frac{KI_O}{2} - \frac{K^2}{4}(V_1 - V_2)^2 + \frac{K}{2}(V_3 - V_4)\sqrt{KI_O - \frac{K^2}{4}(V_3 - V_4)^2}} \quad (9)$$

So, analyzing (7), (8) and (9) relations, it results that the classical MOS Gilbert multiplier presents an important nonlinearity.

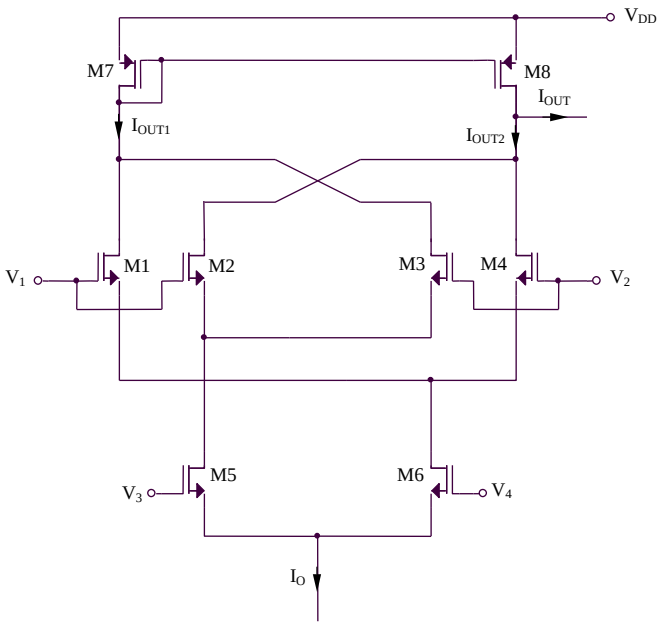


Fig. 2. MOS Gilbert multiplier

C. Improved linearity MOS differential structure

The original improved linearity MOS differential structure is proposed in Fig. 3. "SQ" block represents a squaring circuit, having the original implementation presented in Fig. 5.

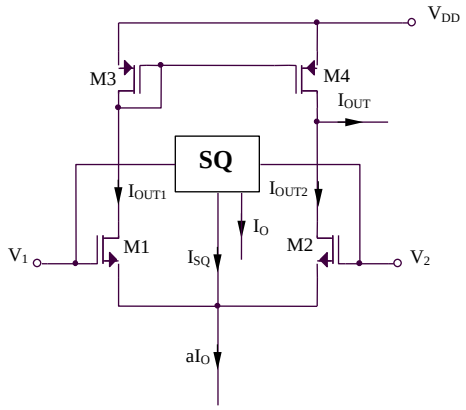


Fig. 3. Improved linearity MOS differential structure

The differential output current for this circuit has the following expression:

$$I_{OUT} = (V_1 - V_2)\sqrt{K(aI_O - I_{SQ}) - \frac{K^2}{4}(V_1 - V_2)^2} \quad (10)$$

In order to obtain a linear behavior of the proposed differential structure, the I_{SQ} output current of the squaring circuit "SQ" must be the sum of a constant term and a term proportional with the square of the differential input voltage:

$$I_{SQ} = bI_O - \frac{K}{4}(V_1 - V_2)^2 \quad (11)$$

b being a positive constant, depending on the particular implementation of the squaring circuit. Replacing (11) in (10), it results:

$$I_{OUT} = \sqrt{(a-b)KI_O}\,(V_1 - V_2) = G_m(V_1 - V_2) \quad (12)$$

$G_m = \sqrt{(a-b)KI_O}$ being the circuit equivalent transconductance.

D. Implementation of the squaring circuit "SQ"

The proposed method for designing the required voltage squaring circuit is based on a differential amplifier (Fig. 4) [9] having a controllable asymmetry between the geometries of its composing transistors. This difference between the aspect ratios of MOS transistors will introduce in the output currents of the differential amplifier a term proportional with the square of the differential input voltage.

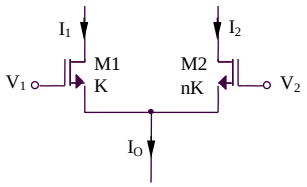


Fig. 4. Asymmetrical differential structure

The differential input voltage, V, can be expressed as:

$$V = V_{GS1} - V_{GS2} = \sqrt{\frac{2I_1}{K}} - \sqrt{\frac{2(I_O - I_1)}{nK}} \tag{13}$$

resulting:

$$\frac{K}{2}V^2 = I_1 + \frac{I_O - I_1}{n} - 2\sqrt{\frac{I_1(I_O - I_1)}{n}} \tag{14}$$

The expression of I_1 current can be obtained solving the following second-order equation, derived from (14):

$$I_1^2\left[\left(\frac{n-1}{n}\right)^2 + \frac{4}{n}\right] + \left(\frac{I_O}{n} - \frac{KV^2}{2}\right)^2 + I_1\left[2\frac{n-1}{n}\left(\frac{I_O}{n} - \frac{KV^2}{2}\right) - \frac{4I_O}{n}\right] = 0 \tag{15}$$

So:

$$I_1 = \frac{I_O}{n+1} + \frac{n(n-1)}{2(n+1)^2}KV^2 + \frac{nV}{(n+1)^2}\sqrt{2KI_O(n+1) - K^2nV^2} \tag{16}$$

and:

$$I_2 = I_O - I_1 = \frac{nI_O}{n+1} - \frac{n(n-1)}{2(n+1)^2}KV^2 - \frac{nV}{(n+1)^2}\sqrt{2KI_O(n+1) - K^2nV^2} \tag{17}$$

The complete realization of a voltage squaring circuit, based on the previous proposed method, uses a cross-coupling of two differential amplifiers having controllable asymmetries between their geometries, M1-M2 and M3-M4 (Fig. 5) [9].

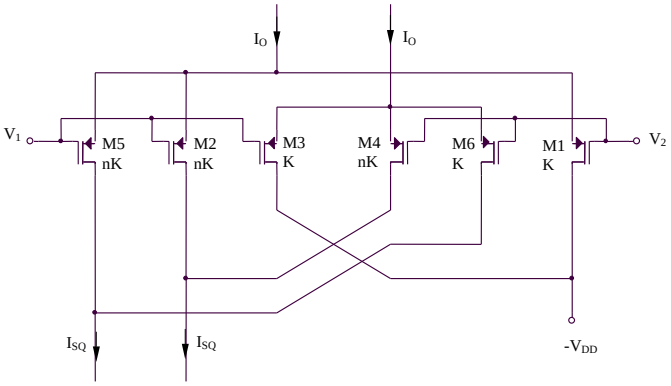


Fig. 5. Squaring circuit

Using (16) and (17), it results:

$$I_{SQ} = I_{D2} + I_{D4} = \frac{nI_O}{n+1} - \frac{n(n-1)}{2(n+1)^2}KV^2 - \frac{nV}{(n+1)^2}\sqrt{2KI_O(n+1) - K^2nV^2} + \frac{nI_O}{n+1} - \frac{n(n-1)}{2(n+1)^2}KV^2 + \frac{nV}{(n+1)^2}\sqrt{2KI_O(n+1) - K^2nV^2} = \frac{2nI_O}{n+1} - \frac{n(n-1)}{(n+1)^2}KV^2 \tag{18}$$

E. Improved linearity MOS multiplier circuit

The proposed multiplier circuit with improved linearity is presented in Fig. 6. The squaring circuit "SQ" has the previous presented implementation (Fig. 5).

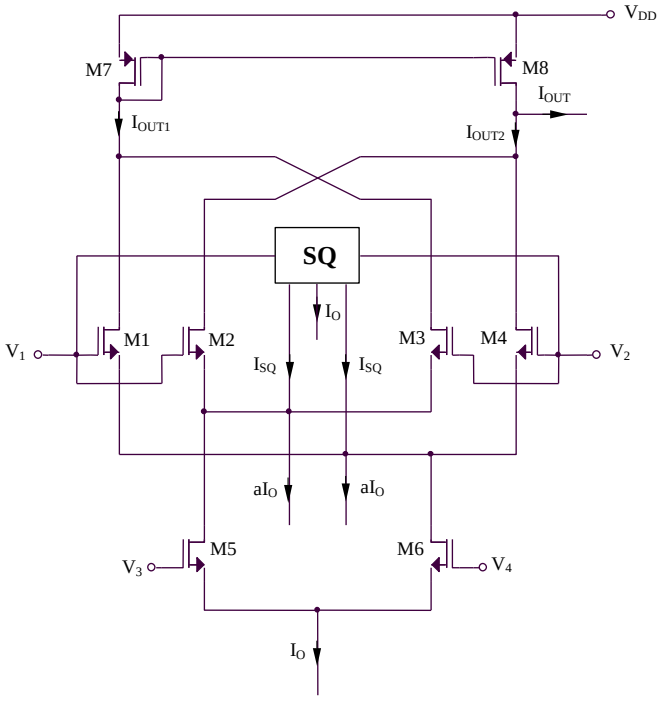


Fig. 6. Improved linearity multiplier circuit

The differential output current of the circuit shown in Fig. 6 can be expressed as follows:

$$I_{OUT1} - I_{OUT2} = (I_{D1} - I_{D4}) - (I_{D2} - I_{D3}) \tag{19}$$

The expression of the M1-M4 differential amplifier output current is:

$$I_{D1}-I_{D4}=(V_1-V_2)\sqrt{\begin{matrix}KI_{D6}+KI_O\left(a-\dfrac{2n}{n+1}\right)+\\ +K^2(V_1-V_2)^2\left[\dfrac{n(n-1)}{(n+1)^2}-\dfrac{1}{4}\right]\end{matrix}} \quad (20)$$

For obtaining a linear behavior of the multiplier, it is necessary that $\frac{n(n-1)}{(n+1)^2}=\frac{1}{4}$, resulting $n_{1,2}=1\pm\frac{2}{\sqrt{3}}$ and $a_{1,2}=\frac{\sqrt{3}\pm 2}{\sqrt{3}\pm 1}$. So, choosing these proper pairs (a_1,n_1) or (a_2,n_2), it results:

$$I_{D1}-I_{D4}=(V_1-V_2)\sqrt{KI_{D6}} \quad (21)$$

Similar conditions impose:

$$I_{D2}-I_{D3}=(V_1-V_2)\sqrt{KI_{D5}} \quad (22)$$

resulting:

$$I_{OUT1}-I_{OUT2}=\sqrt{K}(V_1-V_2)\left(\sqrt{I_{D6}}-\sqrt{I_{D5}}\right) \quad (23)$$

equivalent with:

$$I_{OUT1}-I_{OUT2}=\frac{K}{\sqrt{2}}(V_1-V_2)(V_3-V_4) \quad (24)$$

IV. CONCLUSIONS

An original low-power low-voltage multifunctional structure with improved performances was presented, allowing to implement (with minor changes in the design) two important functions: signal gain with theoretical null distortions and voltage multiplying with very good linearity. The great advantages of the increased modularity and controllability and of the reduced design costs associated represent an immediate consequence of the multiple functions realized by the proposed structures. Comparing with other similar designed reported in literature, the linearity is strongly increased by implementing original techniques, while the silicon occupied area per function is reduced as a result of the proposed multifunctionality. The structures are supplied at $\pm 3V$, they presenting a very good linearity (in the worst case, $THD<0.4\%$ for differential amplifiers and $THD<0.5\%$ for multiplier), correlated with an extended range of the input voltage (at least $\pm 0.5V$).

REFERENCES

[1] Grasso, A.D., Pennisi, S., „High-Performance CMOS Pseudo-Differential Amplifier", Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on, 23-26 May 2005, pp. 1569 - 1572 Vol. 2.

[2] Jongchan Kang, Kyungho Lee, Jehyung Yoon, Yisun Chung, Sungbo Hwang, Bumman Kim, „Differential CMOS Linear Power Amplifier with 2nd Harmonic Termination at Common Source Node", Radio Frequency integrated Circuits (RFIC) Symposium, 2005. Digest of Papers. 2005 IEEE, 12-14 June 2005, pp. 443-446.

[3] Jongchan Kang, Jehyung Yoon, Kyoungjoon Min, Daekyu Yu, Joongjin Nam, Youngoo Yang, Bumman Kim, „A Highly Linear and Efficient Differential CMOS Power Amplifier with Harmonic Control", Solid-State Circuits, IEEE Journal of, Volume 41, Issue 6, June 2006, pp. 1314-1322.

[4] Mohieldin, A.N., Sanchez-Sinencio, E., Silva-Martinez, J., „A Fully Balanced Pseudo-Differential OTA with Common-Mode Feedforward and Inherent Common-Mode Feedback Detector", Solid-State Circuits, IEEE Journal of, Volume 38, Issue 4, April 2003, pp. 663-668.

[5] C. Kuan-Hung, and C. Yuan-Sun, "A Low-Power Multiplier With the Spurious Power Suppression Technique", Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, vol. 15, Issue 7, July 2007, pp. :846-850

[6] S. Desai, P. Trivedi, and V. Von Kanael, "A Dual-Supply 0.2-to-4GHz PLL Clock Multiplier in a 65nm Dual-Oxide CMOS Process", *in* Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International, 11-15 Feb. 2007, pp. 308-311.

[7] E. Bergeret, J. Gaubert, and P. Pannier, "Standard CMOS voltage multipliers architectures for UHF RFID applications: study and implementation", in *RFID, 2007. IEEE International Conference on*, 2007, pp. 115 – 120.

[8] C. Kuan-Hung, and C. Yuan-Sun, "A Low-Power Multiplier With the Spurious Power Suppression Technique", Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, vol. 15, Issue 7, July 2007, pp. :846-850

[9] S. Desai, P. Trivedi, and V. Von Kanael, "A Dual-Supply 0.2-to-4GHz PLL Clock Multiplier in a 65nm Dual-Oxide CMOS Process", *in Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 11-15 Feb. 2007, pp. 308-311.

Design of a high-voltage differential amplifier based on stacked low-voltage standard CMOS with different input stages

Sara Pashmineh, Dirk Killat, *Member, IEEE*
Chair Microelectronics, Brandenburg University of Technology, Cottbus, Germany
{pashmine, killat}@tu-cottbus.de

Abstract - This paper presents the design of a high-voltage differential amplifier using six different pre-input stage circuits to reduce high-voltage input levels to low-voltage signals. The proposed circuits are designed using 65 nm CMOS process technology with a nominal voltage of 2.5 V and a supply voltage of 5 V. The designs are based on stacked low-voltage standard CMOS transistors. The different pre-input stage circuits are compared to each other in terms of their circuit description, drawbacks, advantages and simulation results. The principle of the five designed pre-input stage circuits can be applied to higher voltage range input signals as well.

I. INTRODUCTION

The development of MOS technology scaling has increased the challenge of designing high-voltage circuits. A high-voltage integrated circuit is defined as a circuit with a supply voltage higher than the nominal voltage of its transistors. One method of designing these circuits is to use high-voltage compatible transistors [1]. Unfortunately, such transistors are technology dependent and not available in recently developed processes such as the 28nm TSMC technology. Another way of designing these high-voltage circuits compatible with scaled technologies is to use stacked low-voltage standard CMOS transistors [2]–[11].

Power management, which plays an increasing role in electronic systems for consumer, sensor and automotive applications, requires high-voltage circuits such as low-dropout voltage regulators or switching converters that use error amplifiers in the feedback loop. To meet all operating and load conditions, these amplifiers must operate with high-voltage input and output signals and high supply voltages. The objective of this paper is the design of a differential amplifier based on stacked low-voltage standard transistors without loading the input with a dc current.

This paper introduces the design of a high-voltage differential amplifier having a high-voltage input Vin_{HV} and a high supply voltage Vdd_{HV} of 5 V, which can be between 2.6 V and 5 V (= $2\times V_n$). The input voltages can vary between 0 V and Vdd_{HV}.

In order to keep the voltage between the terminals of each transistor within technology limits, the amplifier is designed using stacked low-voltage standard CMOS transistors (Fig. 1). The gate-terminals of the nMOS transistors *ML4* and *MR4* are regulated by input signals. Both transistors are low-voltage standard transistors. Therefore, to drive the amplifier, the input signals must not exceed the nominal voltage V_n, which is 2.5 V in this design. Pre-input stage circuits are needed to reduce the high-voltage inputs into low-voltage signals and, such a circuit is termed here as a **H**igh to **L**ow **V**oltage **L**evel **S**hifter (HLV-Level Shifter).

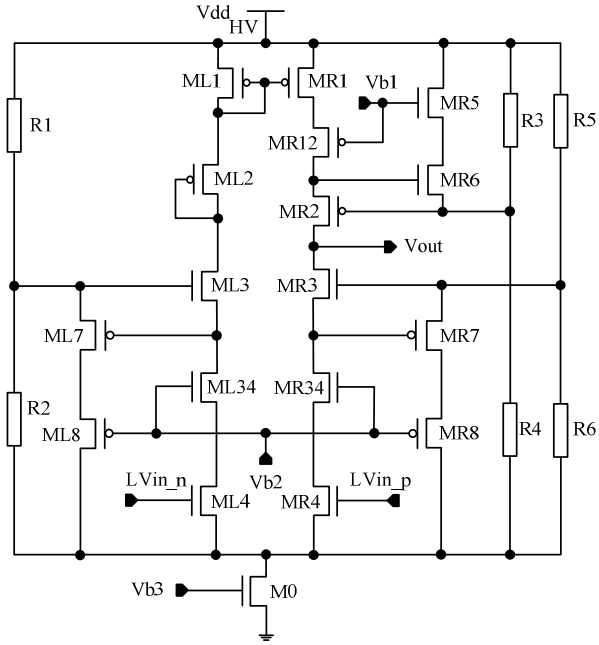


Fig. 1. The structure of the proposed high-voltage differential amplifier

This paper presents six different designs for a pre-input stage HLV-Level Shifter circuit. These designs are compared to each other in terms of their circuit description, drawbacks and advantages by simulation.

This work is organized as follows: Section II describes the proposed high-voltage differential amplifier and section III introduces six different designs of a pre-input stage to shift the level of high-voltage input signals of the proposed HV-differential amplifier down to a low voltage. Section IV presents the simulation results together with a comparison of the proposed HV-amplifiers using the different high to low input voltage converters. Finally, the conclusions are given.

II. SYSTEM DESCRIPTION

Fig. 1 shows the proposed high-voltage differential amplifier with single-ended output. The circuit is based on stacked low-voltage standard CMOS transistors.

The circuit is supplied with 5 V, which is twice the nominal voltage 2.5 V. To restrict the voltage between the terminals of every transistor to stay within the technology limits, both differential stages need to be based on at least 2-stacked CMOS transistors. Because of diverse gate-voltages, gate-drain and gate-source capacitances, the unregulated charge and discharge of stage-nodes can cause an overvoltage between the terminals of each transistor. To prevent this overvoltage, extra transistors *MR12*, *ML34* and *MR34* are included to the 2-stacked CMOS transistors (*MR1*, *ML1*, *MR2*, *ML2*, *MR3*, *ML3*, *MR4* and *ML4*), as can be seen in Fig. 1. The nMOS transistors *MR4* and *ML4* are controlled by low-voltage input signals.

When the input signal V_{LVin_p} is low enough to turn off the transistor *MR4*, the pull-down nodes of the right stage(drain nodes of *MR3*, *Mr34* and *MR4*) become charged, causing *MR7* to turn off. In this case, the gate of *MR3* is regulated only by the voltage divider of *R5* and *R6*, which turns it off. With a high input signal, the gate-source of *MR3* starts discharging and the pMOS transistor *MR7* turns on. As a result, the gate voltage of *MR3* decreases in order to maintain its gate-drain and gate-source voltages within the technology limit, as will be shown in section IV, Fig. 8 and 9.

In a similar way, the nMOS transistor *ML3* is regulated by the pMOS *ML7* and the voltage divider consisting the resistors *R1* and *R2*. The pMOS transistor*MR6* and the voltage divider consisting of resistors *R3* and *R4* control the pMOS transistor *MR2*. The transistors *MR12* and *MR5* are regulated by the reference voltage *Vb1* in order to avoid the overvoltage.

The designed differential amplifier is regulated by the input signals, which are fed to the gate terminals of both low-voltage nMOS transistors *MR4* and *ML4*. With regard to the technology limit, the input signals have to be equal to or less than the nominal voltage *Vn* (2.5 V).

Therefore, a pre-input stage circuit is required to reduce both high-voltage inputs into low-voltage signals. In the next section, the design and comparison of six different **H**igh into **L**ow **V**oltage Level Shifter (HLV-Level Shifter) pre-input stage circuits are presented.

III. DESIGN AND COMPARISON OF PRE-STAGE INPUT CIRCUITS

A. HLV-Level Shifter A

Fig. 2 demonstrates a level shifter, which is based on a voltage divider. The circuit contains two resistors *Rx* and *Ry* whose values are determined according to the high voltage input signals. However, this method has two major disadvantages; the voltage divider draws a current from the input, the dynamic range is reduced, and the signals are subject to resistor variations.

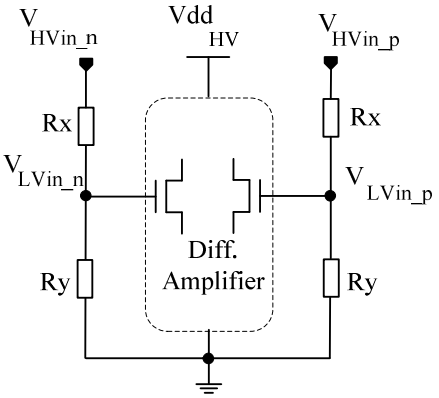


Fig. 2. HLV-Level Shifter A

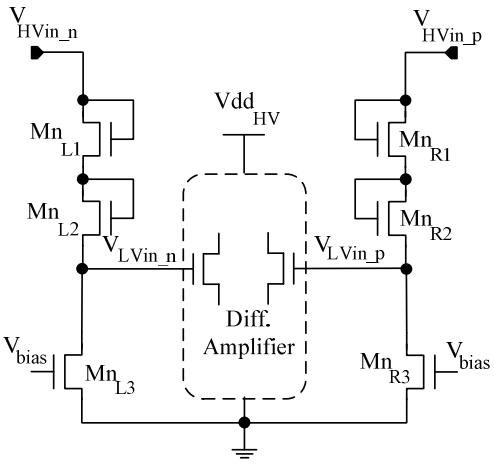


Fig. 3. HLV-Level Shifter B

B. HLV-Level Shifter B

Fig. 3 shows the second HLV-Level Shifter. To stabilize and reduce variations of the shifted voltage, nMOS transistors are used rather than resistors as found in HLV-Level Shifter (A). The node between the two series-connected nMOS diodes (*Mn1* and *Mn2*) and the nMOS transistor Mn3 generates the required low input voltage for the proposed differential amplifier. The transistor *Mn3* is biased by a reference voltage *Vbias*.

To reduce a higher voltage input signal in the range of $(N–Vn)\times Vn$ to $N\times Vn$, it would be necessary to have $2\times(N–1)$ nMOS diodes connected in series.

This level shifter has the same disadvantage as HLV-Level Shifter (A), as the input source must feed a current through the nMOS transistor *Mn3*, whose source is ground connected.

C. HLV-Level Shifter C1 using current mirrors

Fig. 4a shows the design of a new HLV-Level Shifter using current mirrors. The circuit generates low-voltage signals from high-voltage inputs in two processes: (1) converting the voltage V_{HVin} to the current I_1 (V/I-converter) and (2) the mirrored current to the voltage V_{LVin} (I/V-converter). The circuit is supplied by a low-voltage Vdd_{LV}, which is equal to the nominal voltage (2.5 V). On the first path of each side, the circuit contains a resistor *R*, which is connected in series with a gate-drain connected nMOS transistor Mn_{R3} (Mn_{L3}). Due

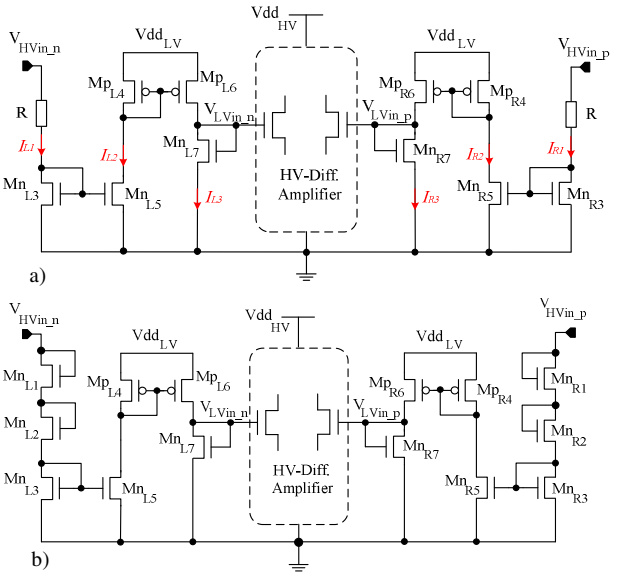


Fig. 4. HLV-Level Shifter a) C1 b) C2

to the two current mirrors, the current I_1 generated from the HV-signal through the resistor R, is mirrored in the second path (I_2) and in the third path (I_3). The generated voltage V_{LVin} is the required voltage for the proposed differential amplifier:

$$I_3 = \frac{W_6 \cdot L_4}{W_4 \cdot L_6} \cdot I_2 = \frac{W_5 \cdot L_3}{W_3 \cdot L_5} \cdot I_1 \quad (W_6 = W_4, L_6 = L_4) \qquad (1)$$

$$Vg_{Mn7} = V_{LVin} = \sqrt{\frac{2 \cdot I_3}{\beta}} + V_{nTh} \qquad (2)$$

The equations (1) and (2) show that V_{LVin} is generated by a current I_1, which depends on the high voltage input and the resistor R. However, the disadvantage of this circuit is that the input is loaded by a current.

Fig. 4b shows an improved circuit of this HLV-Level Shifter (C1). To avoid fluctuations of the resistors, two gate-drain connected nMOS transistors are used in series instead of the resistor R. An additional drawback of both level shifters (C1 and C2) is that the generated signal V_{LVin} is limited to the threshold voltage of Mn7, as can be obtained from (2).

To avoid input loading HLV-Level Shifters D, E and F have been designed.

D. HLV-Level Shifter D

Fig. 5 shows a HLV-Level Shifter using a common-drain stage. The high-voltage input signal, which varies between 0 V and 5 V controls the gate of the nMOS transistor *MnR* (*MnL*). The supply voltage of this circuit is 2.5 V. When the gate-source voltage of the nMOS exceeds the threshold voltage, the node between the transistor *Mn* and the resistor R_{R1} (R_{L1}) starts to charge from 0 V until 2.5 V, which can then be applied to the inputs of the proposed HV-differential amplifier. The disadvantage of this method is that a high-voltage input signal with only a short range of 5 V can be applied (Table 1), otherwise an overvoltage between the transistor's node would occur.

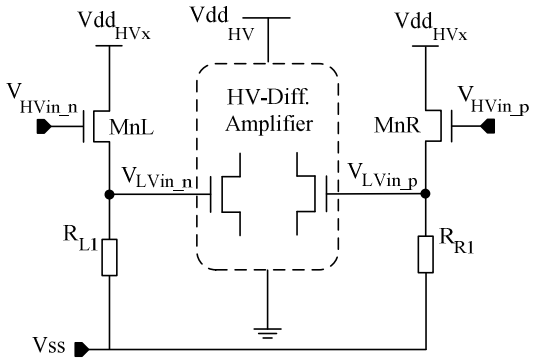


Fig. 5. HLV-Level Shifter D

TABLE 1. The voltage range of the HLV-Level Shifter D (V)

Vdd_{HV}	Vss	V_{HVin}		Vdd_{HVx}	V_{LVin}	
		min.	max.		min.	max.
5.0	0	0.0	5.0	2.5	0.0	2.5
7.5	2.5	2.5	7.5	5.0	2.5	5.0
10.0	5.0	5.0	10.0	7.5	5.0	7.5
$N \times V_n$	$Vdd_{HV} - 5$	$(N-1) \times V_n$	$N \times V_n$	$(N-1) \times V_n$	$(N-2) \times V_n$	$(N-1) \times V_n$

To keep the gate-source voltage (Vgs) of *Mn* in the range of technology limit (2.5 V), Vss in Fig. 5 has to be the minimum value of the input signal V_{HVin}, as can be seen in (Table 1). The problem is that the source node of *Mn*, which generates the low-voltage input signal V_{LVin}, can be charged when Vgs of *Mn* exceeds the threshold voltage.

E. HLV-Level Shifter E

To extend the range of the applied high input voltage for the HLV-Level Shifter D of Fig. 5, an extra nMOS transistor *MnL1* (*MnR1*) is included in series between the input transistor *MnL2* (*MnR2*) and the supply voltage node Vdd_{HV} in Fig. 6 To maintain the voltage between the terminals of each transistor in the range of technology limit, the gate voltage of *Mn1* needs to be varied from 2.5 V according to the high-input signal V_{HVin}. Therefore, a second stage containing series-connected pMOS transistors (*Mp1*, *Mp2* and *Mp3*) is designed to each side of the circuit in Fig. 6. This stage also has a supply voltage of Vdd_{HV} (5 V). The gates of *Mp2* and *Mp3* are controlled by the nodes of the first stage according to the high-voltage input signal V_{HVin}.

The gate of *MP1* is biased by a reference voltage of 2.5 V to maintain the gate-source voltage of *Mp2* in the range of 2.5 V, as can be expressed with the following equations:

$$I_{Mp1} = \frac{\beta}{2}(|Vgs_{Mp1}| - |V_{pTh}|)^2 = I_{MP2} \qquad (3)$$

$$Vgs_{Mp1} = 2.5\,V \approx Vgs_{Mp2} = Vg_{Mn1} - V_{LVin} \qquad (4)$$

Due to this, the gate voltage of *Mn1* is shifted to the gate of *MP2* (V_{LVin}) about 2.5 V. In the on-state, the transistor *Mp2* operates in both regions (saturation and triode), therefore its gate-source voltage cannot be exactly 2.5 V, as can be seen in the simulation results in Fig. 11a.

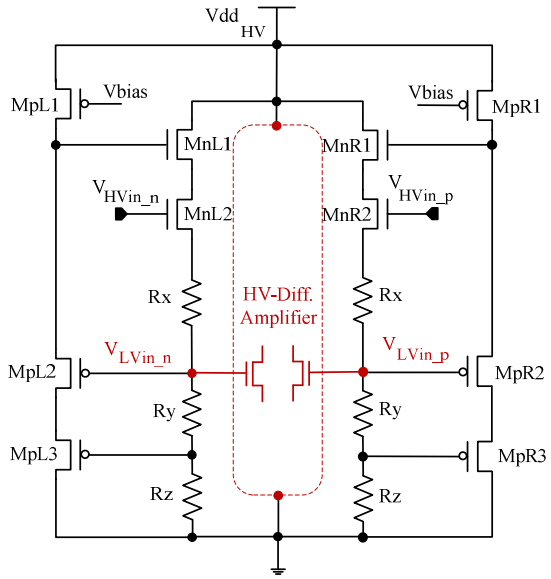


Fig. 6. HLV-Level Shifter E

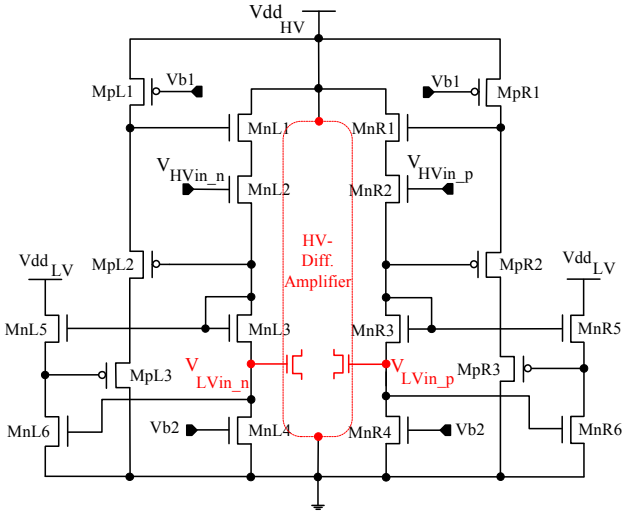


Fig. 7. HLV-Level Shifter F

For a higher input voltage ($N \times Vn$), ($N–1$) nMOS transistors are needed to be connected in series to the input nMOS transistor *Mn2*.

F. HLV-Level Shifter F

However, HLV-Level Shifter (E) suffers from fluctuations of the resistors. Fig. 7 shows an optimized HLV-Level Shifter using two nMOS transistors with *Mn4* and gate-drain connected *Mn3*, instead of the resistors. The node voltage between both transistors (Mn3 and Mn4) is not low enough to drive the pMOS transistor Mp3 of the second stage. Therefore, the third stage containing series-connected nMOS transistors (*Mn5* and *Mn6*), which forms the current mirror *Mn3* and *Mn4,* is designed. The node between both transistors *Mn5* and *Mn6* provides a voltage for the gate of *Mp3* of the second stage according to the high-voltage input signal.

IV. SIMULATION RESULT

The simulation results in Fig. 8 show the dc and transient voltage characteristics of *MR3* and *MR34* for the proposed HV-differential amplifier (Fig. 1). The pMOS transistor *MR7* and the voltage divider (*R5* and *R6*) control the gate of *MR3* according to the input signal. In the off-state, the gate voltage is about 4.5 V and in the on-state approx. 2.45 V. The dc voltage characteristics of the pull-up nodes according to the high-voltage V_{HVin}, the low-voltage input V_{LVin} and the output V_{out} can be seen in Fig. 9. The transistors *MR6*, *MR7* and the voltage dividers (*R5*, *R6*) and (*R3*, *R4*) control the right-hand stage of the proposed differential amplifier and maintain the voltage between the terminals of each transistor to be within the technology limit with 5% tolerance using extra transistors *MR12*, *MR5*, *MR34* and *MR8*.

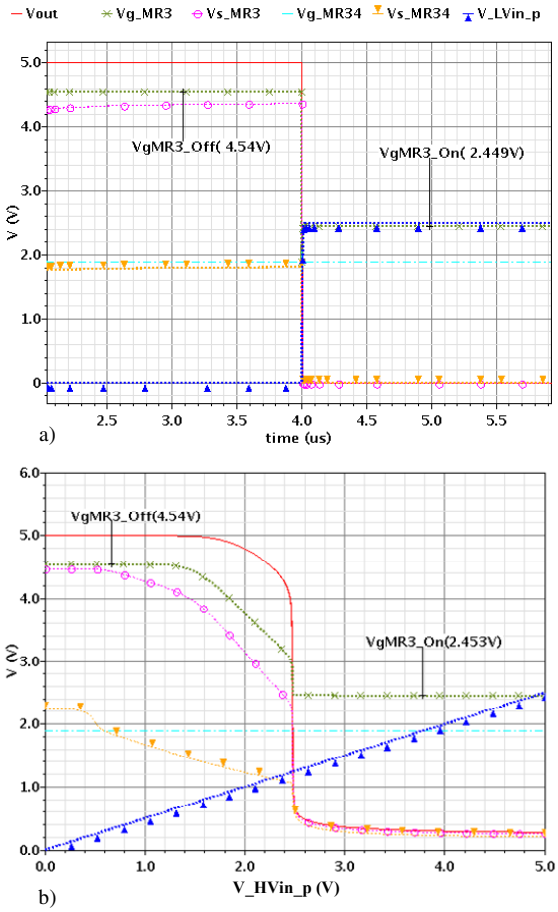


Fig. 8. a)Tr- and b) DC-simulation voltage characteristics of the pull-down nodes (right-stage of the diff. amp.) vs. V_{HVin_p}

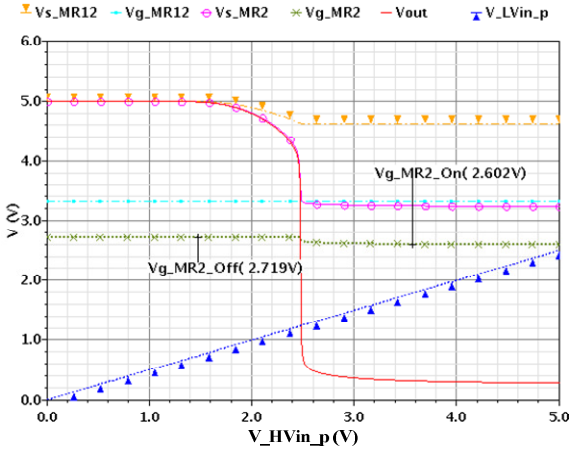


Fig. 9. Voltage characteristics of the pull-up nodes (right-stage of the diff. amp.) vs. V_{HVin_p}

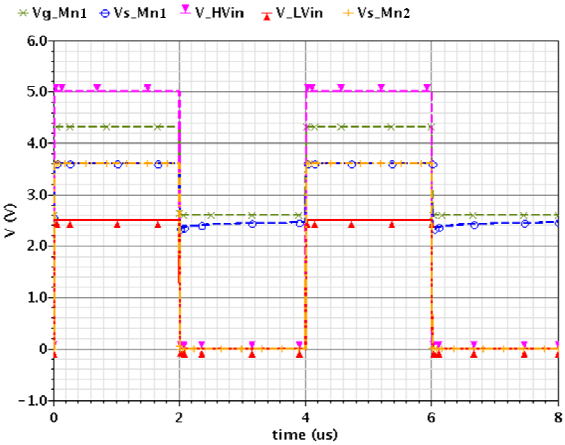


Fig. 10. Node voltage characteristics of HLV-level shifter E

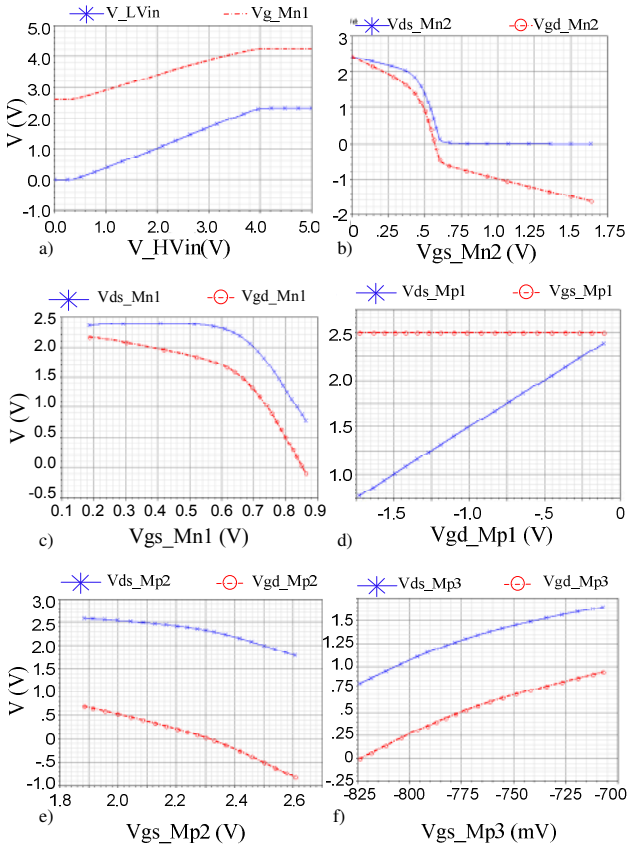


Fig. 11. a) voltage characteristics of Vg_{Mn1} and V_{LVin} vs. V_{HVin}
b)-f) voltage between transistors' terminals (HLV-Level Shifter E)

Fig. 10 demonstrates the node voltage characteristics of the transistors *Mn1* and *Mn2* for the HLV-Level Shifter (E). The HV-input signal V_{HVin} is a square-wave and varies between 0 V and 5 V. The gate voltage of *Mn1* is generated by the second stage (*Mp1*, *Mp2* and *Mp3*) and varies between 4.3 V and 2.6 V to keep the voltage between the terminals of each transistor within the technology limit. The generated desired input signal V_{LVin} varies between 0 V and 2.5 V.

The characteristic of the gate voltage of *Mn1* generated by the second stage can be seen in Fig. 11a according to V_{LVin}, which is generated by the first stage and controls the gate voltage of *Mp2*.

The simulation results in Fig. 11 b-f and 12 show that the voltage between the terminals of each transistor for HLV-Level Shifters (E and F) is within the technology limit with 5% tolerance.

Fig. 13 shows that the first three HLV-Level Shifters (A, B and C1/C2) lose input current. In comparison, the first level shifter is more favorable because of its lower current loss (max. 5 µA).

Fig. 14 demonstrates the generated voltages of the designed HLV-Level Shifters, which are less than or equal to the nominal voltage of 2.5 V. According to the generated LV-signals, the output of the proposed HV-differential amplifier can be charged and discharged between 0 V and 5 V.

Table 2 gives a comparison between delays, rise (*RT*) and fall (*FT*) times of the HV-amplifier output voltage with all designed HLV-Level Shifters. The HLV-LS A dissipates less current than the other two level shifters (B and C), but the output of the proposed amplifier using this level shifter has the highest delay, rise (*RT*) and fall times (*FT*). With HLV-Level Shifter (F), the output has a satisfied delay, rise and fall times in comparison to the other level shifters. Transistor dimensions and resistor values of the designed circuits are given in Table 3.

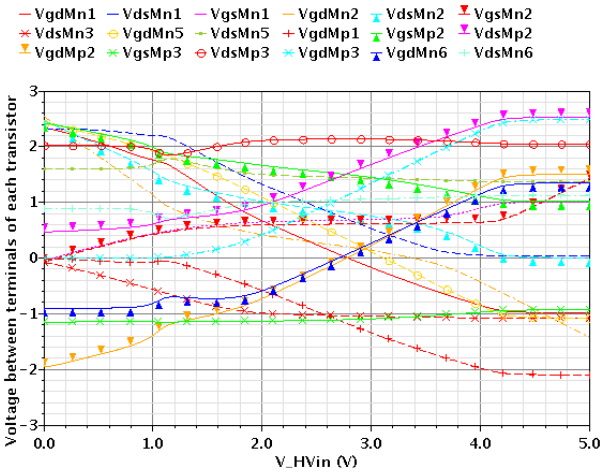


Fig. 12. Voltage between terminals of transistors (HLV-Level Shifter F)

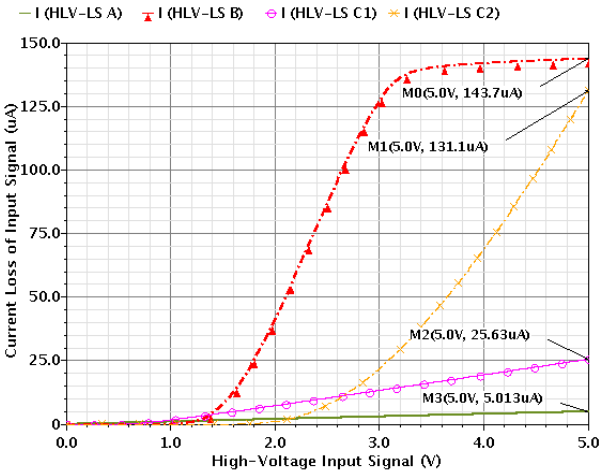


Fig. 13. Current loss of the input signals (HLV-LS A, B, C1 and C2)

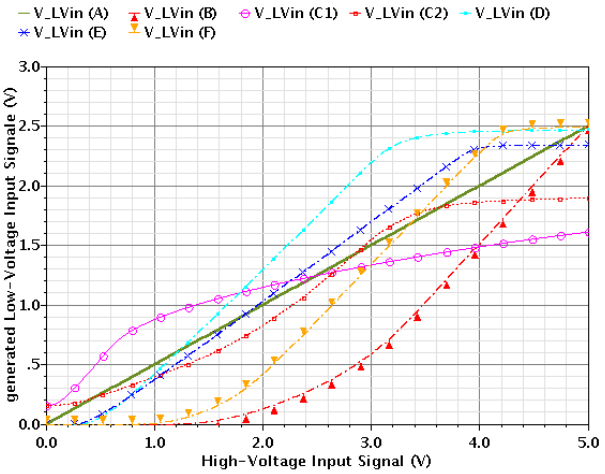


Fig. 14. Output voltages of the designed HLV-Level Shifters (A-F)

TABLE 2. Comparison results of Vout with different HLV-LS

HLV-LS	max. Iin (µA)	RT (ns)	FT (ns)	delay (ns)
A	5.01	2.12	2.82	13.78
B	143.7	1.74	0.26	1.1
C1	25.63	1.64	0.41	2.3
C2	131.1	1.83	0.51	1.7
D	0	1.59	0.50	1.2
E	0	1.90	2.46	7.9
F	0	1.63	0.54	1.15

TABLE 3. Transistor dimensions length/width (nm/µm) and resistor values (Ω) of the designed circuits

HV-differential amplifier					
ML/R1	*MR12*	*MR2*	*ML2*	*MR/L3*	*MR/L34*
280/1	500/4	280/10	500/1	280/10	500/3
MR5	*MR6*	*ML/R7*	*ML/R8*	*ML/R4*	*M0*
500/6	500/2	280/1	280/	500/3	280/3
R2/R6		*R3*		*R1/R5/R4*	
652.3K		150K		1.5M	

HLV-LS A	**HLV-LS B**
MnL/R1, MnL/R2	*MnL/R1, MnL/R2, MnL/R3*
500/1	500/1

HLV-LS C1 & HLV-LS C2			**HLV-LS D**		
R	*Mn1, Mn2, Mn3*	*Mp4, Mn5, Mnp5*	*Mn7*	*Mn1*	*R1*
150K	500/1	280/10	280/2	280/1	80K

HLV-LS E		**HLV-LS F**		
Mn1	*Mn2*	*Mn1*	*Mn2*	*Mn3*
500/1	280/6	280/3.4	280/8	280/0.4
Mp1/Mp2	*Mp2*	*Mn4*	*Mn5,Mn6*	*Mp1*
280/1	280/10	280/0.45	280/0.4	280/0.4

Rx	*Ry*	*Rz*	*Mp2*	*Mp3*
80K	110K	74K	280/1	500/9

V. CONLUSION

In this paper, a HV differential amplifier is presented with a supply voltage of 5 V. As the design uses standard transistors with nominal voltage of 2.5 V, low input voltages are required. Six types of pre-input stage circuits are designed for reducing the high-voltage signal down to the nominal voltage. Five of the presented HLV-Level Shifters (A, B, C, E and F) are extendable for use with higher input voltages. The first three HLV-Level Shifters (A, B and C) draw an input current. When comparing input current, delays, rise and fall times of the HV-amplifier output with the different level shifters, it is concluded that the HLV-Level Shifter Type-F is the most favorable due to low delay, rise and fall times without loading the input with a dc current and without the need of resistors.

The presented HV-differential amplifier and the pre-input stage HLV-Level Shifter circuits are technology independent and can be used with every process technology.

REFERENCES

[1] S. Bandyopadhyay, Y. Ramadass, and A. Chandrakasan, “20 µA to 100 mA DC-DC converter with 2.8 to 4.2 V battery supply for portable applications in 45 nm CMOS”, ISSCC 2011, pp. 386-388

[2] S. Monga, V. Kumar, “A 73µW 400Mbps stress tolerant 1.8V-3.6V driver in 40nm CMOS”, ESSCIRC 2011, pp. 187 – 190, 2011.

[3] C. Menolfi, T. Toifl, M. Rueegg, M. Braendli, P. Buchmann, M. Kossel, T. Morf, “A 14Gb/s high-swing thin-oxide device SST TX in 45nm CMOS SOI”, ISSCC 2011, pp. 156 – 15, 2011.

[4] Hyouk-Kyu Cha ; Dongning Zhao ; Jia Hao Cheong ; Bin Guo ; Hongbin Yu ; Minkyu Je, “A CMOS High-Voltage Transmitter IC for Ultrasound Medical Imaging Applications”, Circuits and Systems II, TCSII,V. 60 , TCSII.2258260, pp. 316 – 320, 2013

[5] B. Serneels, E. Geukens, B. De Muer, T. Piessens, “A 1.5W 10V-output Class-D amplifier using a boosted supply from a single 3.3V input in standard 1.8V/3.3V 0.18µm CMOS”, ISSCC 2012, pp. 94-96, 2012

[6] Y. Ahn, J. Roh, “5-V Buck Converter Using 3.3-V Standard CMOS Process With Adaptive Power Transistor Driver Increasing Efficiency and Maximum Load Capacity”, ITPE, vol. 27, NO. 1, pp. 463-471, 2012

[7] S.R. Bradburn, H.L. Hess, “An integrated high-voltage buck converter realized with a low-voltage cmos process”, 53rd Int. MWSCAS, pp. 1021-1024, 2010

[8] B. Serneels, M. Steyaert, “Design of High Voltage xDSL Line Drivers in Standard CMOS”, Springer, 2008

[9] P. Swaroop, A. Vasani, M. Ghovanloo, “A High-Voltage Output Driver for Implantable Biomedical Stimulators and I/O Applications”, 49th Int. MWSCAS'2006, pp. 566-569

[10] E. Mentze, H. Hess, K. Buck, T. Windley, “A Scalable High-Voltage Output Driver for Low-Voltage CMOS Technologies” IEEE Trans., VLSI Systems, vol.14, NO.12, pp.1347-1353, 2006

[11] B. Serneels, E. Geukens, B. De Muer, T. Piessens, “A 1.5W 10V-output Class-D amplifier using a boosted supply from a single 3.3V input in standard 1.8V/3.3V 0.18µm CMOS” ISSCC, pp.94-96, 2012

Opamp Based Bandgap Voltage Reference in 130 nm: Design and Schematic Level Simulation

Aleksandar Pajkanovic
Faculty of Technical Sciences
University of Novi Sad
apaj@uns.ac.rs

Veljko Malbasa
Faculty of Technical Sciences
University of Novi Sad
malbasa@uns.ac.rs

***Abstract*—A bandgap voltage reference circuit employing an operational amplifier is designed using the unified current-control model transistor model and 130 nm CMOS standard process. Schematic level simulations in Cadence Design System Spectre Simulator show temperature variation of the output voltage within 0.40 mV over a temperature range of 165 K (from - 40°C up to 125°C) in the nominal case, i.e. 1.06 ppm/°C. The supply voltage (3.3 V) change of ±10 % causes a disturbance in the output voltage of only 2.37 mV, i.e. 0.10 % of the output nominal value. The process variations, examined through corner and Monte Carlo analysis, are also presented and discussed.**

Keywords-CMOS, integrated circuits, bandgap, voltage reference, temperature stability

I. Introduction

In this paper we continue the development of a process, voltage and temperature (PVT) stable bandgap voltage reference, started in [1]. Namely, the design of DC current and voltage sources internal to the chip is an essential step in monolithic integrated circuits design, because the performance of nonlinear analog components depends strongly on the current flowing through them, i.e. biasing conditions. The task that these references should accomplish is to properly bias analog components of integrated circuits. Among a variety of solutions, efficient, simple and easy-to-design structures are the priority. These reference circuits should be autonomous when the supply voltage is switched on. Further, they should be based on intrinsic physical properties or on reproducible technology parameters [2], [3].

A PVT invariant reference voltage within an integrated circuits brings stable biasing conditions. A circuit showing promising results in the domain of temperature stability was demonstrated in [1]. Nevertheless, the supply voltage variance consequences were not tackled as successfully in the same circuit. Therefore, the main motivation for this work is obtaining a circuit with a better response to the supply voltage variations.

Since the simplest possible topology employed in [1] was only able to provide temperature stability to an acceptable degree (compared to similar circuits found in literature [1]), a somewhat more complex circuit is synthsized in this paper. Namely, the bandgap voltage reference is designed using the unified current-control model (UICM) of MOSFET [3] and the topology is based on an operational amplifier [3], [4]. The circuit is simulated using the Cadence Design System Spectre Simulator in 130 nm CMOS standard process. The schematic level simulation results show temperature variation of the output voltage within 0.40 mV over a temperature range of 165 K (from - 40°C up to 125°C) in the nominal case, i.e. 1.06 ppm/°C. The supply voltage (3.3 V) change of ±10 % causes a disturbance in the output voltage of only 2.37 mV, i.e. 0.10 % of the output nominal value. Besides the achieved goal of improving the supply voltage stability, the temperature stability is also improved. Such advances of the circuit characteristics are paid by higher power consumption and larger chip area. The process variations, examined through corner and Monte Carlo analysis, are also presented and discussed.

II. Model Used

The design methodology is based on the UICM model [3], [5]. It is a current-based MOSFET model that uses the concept of inversion level. According to the UICM model, the drain current can be split into the forward (I_F) and reverse (I_R) currents:

$$I_D = I_F - I_R = I_S(i_f - i_r), \tag{1}$$

where I_S is the normalization specific current, and i_f and i_r are inversion levels, forward and reverse, respectively. The forward and reverse currents depend on the gate to source and gate to drain voltages, respectively. If the transistor operates in saturation, we have:

$$I_F \gg I_R, \text{ thus: } I_D \approx I_F = I_S \cdot i_f. \tag{2}$$

The normalization current is a function of the technology:

$$I_S = \frac{\mu C_{ox} \phi_T^2 n}{2} \frac{W}{L}, \tag{3}$$

where μ represents the charge mobility, C_{ox} gate-oxide capacitance per area, ϕ_T thermal voltage, n slope factor and W/L transistor aspect ratio.

The inversion level value signifies the transistor operation region in the following way: if $i_f < 1$, the transistor operates in weak inversion and if $i_f > 100$ the transistor operates in strong inversion region. If $1 < i_f < 100$, the transistor operates in moderate inversion region.

According to this model, the voltage and current are related in the following manner [3]:

$$\frac{V_P - V_{S(D)}}{\phi_T} = \sqrt{1 + i_{f(r)}} - 3 + \ln\left(\sqrt{1 + i_{f(r)}} - 1\right), \quad (4)$$

where V_P is the pinch-off voltage, given as the difference of the gate potential and the zero bias threshold voltage (modified by the slope factor):

$$V_P \approx \frac{V_G - V_{T0}}{n}, \text{ if } V_G \approx V_{T0}. \quad (5)$$

Further detailed description of the model can be found in [3] and [5].

III. Bandgap Voltage Reference Theory of Operation

A bandgap voltage reference operation is based on adding two voltages having equal and opposite temperature coefficients, i.e. a voltage proportional to absolute temperature (PTAT) and a voltage complementary to absolute temperature (CTAT) [3], [4]. The logic behind this idea is to cancel out the temperature dependence, so that the output voltage, i.e. reference voltage is independent of temperature variation.

The reference voltage, V_{REF} we can write as:

$$V_{REF} = \alpha \cdot V_{CTAT} + \beta \cdot V_{PTAT}. \quad (6)$$

where α and β are coefficients of proportionality.

The temperature coefficient of a voltage is numerically expressed by the first derivative of that voltage as a function of temperature. Thus, in order to accomplish that V_{REF} is independent of temperature, its derivative needs to be zero:

$$\frac{\partial V_{REF}}{\partial T} = \alpha \frac{\partial V_{CTAT}}{\partial T} + \beta \frac{\partial V_{PTAT}}{\partial T} = 0. \quad (7)$$

To achieve this, a CTAT and a PTAT voltages need to be designed and, then, the coefficients α and β need to be set in such a way that slopes of these two voltages are equal.

The easiest way to obtain a CTAT voltage is the forward-biased pn junction - a diode. The diode current is given by [3]:

$$I_D = I_S \cdot e^{\frac{V_D}{\phi_T}}, \quad (8)$$

where I_S represents a diode saturation current and V_D is the diode voltage. From this relation, we extract the diode voltage as:

$$V_D = \phi_T \ln\left(\frac{I_D}{I_S}\right). \quad (9)$$

The two members comprising the expression for the diode voltage are of opposite temperature coefficient, but the PTAT (the thermal voltage) is negligible when compared to CTAT (the logarithm expression). The diode voltage is a CTAT in its nature [3], [4]. The temperature coefficient of the diode voltage can be calculated around -1.5 mV/K at room temperature [1].

To obtain a PTAT voltage, a pn junction is again used, i.e. several pn junctions are used. Namely, the idea is to subtract two V_D voltages, thus canceling the CTAT members and extracting the thermal voltage, i.e. the PTAT member. Such an effect can be obtained if there are different currents flowing through physically identical pn junctions, Fig. 1. Assuming an array of n identical diodes (Fig. 1), according to (9), the voltage of each one of them is given as:

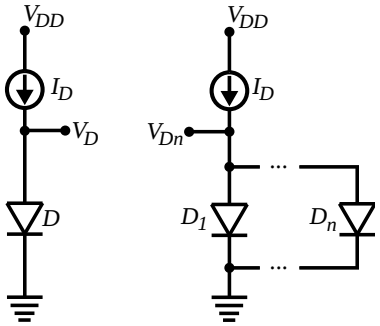


Figure 1. A PTAT voltage can be obtained as the difference of V_D and V_{Dn}

$$V_{D1} = V_{D2} = ... = V_{Dn} = \phi_T \cdot \ln\left(\frac{I_D}{n \cdot I_S}\right). \quad (10)$$

Observing the difference of V_D and V_{Dn}, yields:

$$\begin{aligned} V_D - V_{Dn} &= \phi_T \cdot \ln\left(\frac{I_D}{I_S}\right) - \phi_T \cdot \ln\left(\frac{I_D}{n \cdot I_S}\right) \\ &= \phi_T \cdot \ln(n), \end{aligned} \quad (11)$$

which means that this voltage difference is PTAT in its nature.

The subtraction can be performed by inserting a resistor, as shown in Fig. 2. The voltage over R_1 is given as:

$$V_{R1} - V_{Dn} = I_D \cdot R_1. \quad (12)$$

If V_D and V_{R1} are equal, we can write:

$$V_{R1} - V_{Dn} = I_D \cdot R_1 = \phi_T \cdot \ln(n). \quad (13)$$

Therefore, in order for this idea to work, the current generators in Fig. 2 must be replaced by a circuit which will make sure that $V_D = V_{R1}$. In this circuit, the current I_D is completely determined by the resistor R_1.

The temperature coefficient of $V_{R1} - V_{Dn}$ is that of the thermal voltage:

$$\frac{\partial \phi_T}{\partial T} = \frac{k}{q} = 86.25 \frac{\mu \text{V}}{\text{K}}. \quad (14)$$

In this paper the negative feedback connected operational amplifier is employed to provide equal voltage levels at points V_D and V_{R1}. Namely, if an operational amplifier is in negative feedback configuration its inputs are considered to be at the same voltage level [3], [4]. Thus, the ideal current generators in Fig. 2 are replaced by the adequate circuit, shown in Fig. 3.

Now, both CTAT and PTAT voltages are provided, even though of unequal temperature coefficients. This issue is solved by setting the coefficients of proportionality, α and β, in the

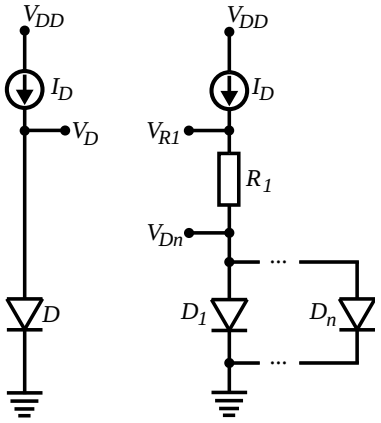


Figure 2. To obtain a PTAT, voltages V_D and V_{R1} must be equal

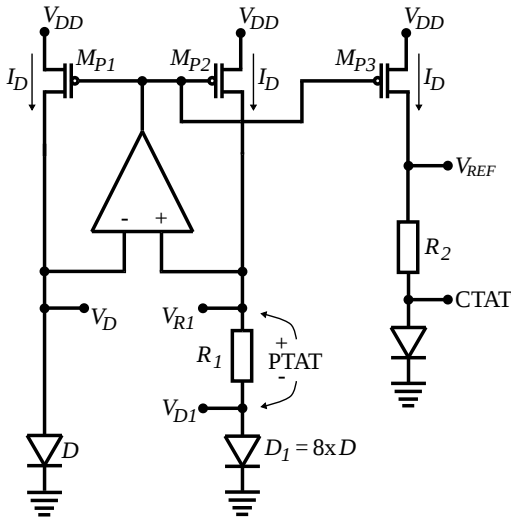


Figure 3. Proposed circuit schematic

correct way. The last step is to sum these two voltages so that the whole circuit yields an output voltage constant to temperature variation. The easiest way to sum voltages is to simply connect them in series. This is what the third branch in Fig. 3 actually does. Namely, the output transistor, M_{P3}, copies the current from the current mirror. Thus, the voltage V_{REF} is given as:

$$V_{REF} = V_D + V_{R2} = V_D + I_D \cdot R_2. \tag{15}$$

From (13) we can write:

$$I_D = \frac{\phi_T \cdot \ln(n)}{R_1}, \tag{16}$$

which, used in (15), yields:

$$V_{REF} = V_D + \frac{R_2}{R_1} \cdot \ln(n) \cdot \phi_T. \tag{17}$$

Comparing this expression with (6), we realize the following identities: $V_{CTAT} = V_D$, $V_{PTAT} = \phi_T$ and $\alpha = 1$, $\beta = (R_2/R_1) \cdot \ln(n)$. Substituting according to these relations into (7), we obtain:

$$\frac{\partial V_D}{\partial T} + \frac{R_2}{R_1} \cdot \ln(n) \cdot \frac{\partial \phi_T}{\partial T} = 0. \tag{18}$$

Temperature coefficients of pn junction and thermal voltages are known from [1] and (14), respectively, whereas R_1, R_2 and n remain to be determined during circuit design.

IV. Circuit Design Using UICM Model

The operational amplifier and its design and operation are not discussed within this paper. It is further investigated in [6], while here only the important characteristics are given in Tab. I. The amplifier provides the open loop gain of 75 dB and gain-bandwidth product of almost 90 MHz, which means that we can assume equal voltage level at the input contacts, positive and negative, when the amplifier is in negative feedback configuration.

TABLE I. Operational Amplifier Characteristics

Parameter	Value
open loop gain A_{v0dB}	66 dB
bandwidth ω_{3dB}	40 kHz
gain-bandwidth GBW, ω_u	90 MHz
phase-margin PM	60°
input common-mode range maximum $ICMR+$	3 V
input common-mode range minimum $ICMR-$	1.3 V
slew rate SR	50 [V/μs]
power consumption P_{DD}	2.46 mW

The information on the input common-mode range of this amplifier influences further design of the bandgap voltage reference. Namely, for the operational amplifier to operate as intended, the minimum DC voltage at the input contacts is 1.3 V. Observing the schematic in Fig. 3, we conclude that the maximum voltage at the inputs is V_D, which is of the order of 0.75 V, therefore $V_D < ICMR-$. This means that the proposed schematic needs to be modified so that it allows somewhat higher voltages at the inputs of the operational amplifier. The simplest way is to add an additional diode for the two branches connected to the positive and the negative contacts, as shown in Fig. 4. Now, the voltage at the inputs is $V_D^{'} = 2V_D$, which is of the order of 1.5V, therefore $2V_D > ICMR-$.

Since the proposed schematic was modified, some of the equations given in section III also need to be modified. First of all, the DC voltage at the negative input of the operational amplifier is now:

$$V_D^{'} = 2V_D \tag{19}$$

while the voltage $V_{Dn}^{'}$ is given by:

$$V_{D1}^{'} = V_{D2}^{'} = ... = V_{Dn}^{'} = 2 \cdot \phi_T \cdot \ln\left(\frac{I_D}{n \cdot I_S}\right). \tag{20}$$

while the PTAT voltage is obtained by subtraction:

$$\begin{aligned} V_D^{'} - V_{Dn}^{'} &= 2 \cdot \phi_T \cdot \ln\left(\frac{I_D}{I_S}\right) - 2 \cdot \phi_T \cdot \ln\left(\frac{I_D}{n \cdot I_S}\right) \\ &= 2 \cdot \phi_T \cdot \ln(n), \end{aligned} \tag{21}$$

Assuming that $V_D^{'} = V_{R1}^{'}$, as discussed at the beginning of this section, the voltage over the resistor in the second branch is given by:

$$V_{R1}^{'} - V_{Dn}^{'} = I_D \cdot R_1 = 2 \cdot \phi_T \cdot \ln(n). \tag{22}$$

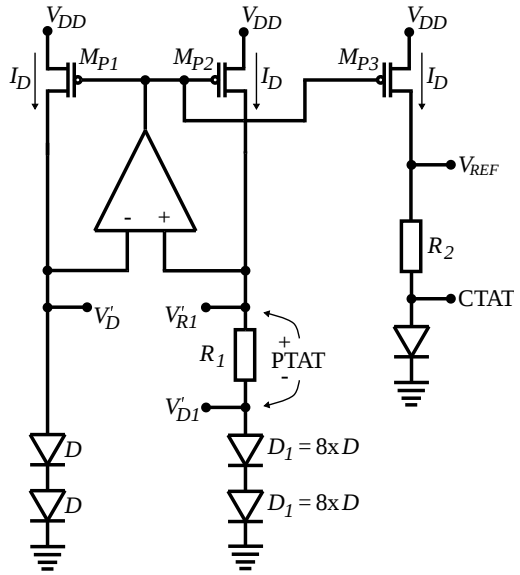


Figure 4. Modified circuit schematic

Finally, taking into account all the modifications, the output voltage previously given by 17 is now:

$$V_{REF} = V_D + 2 \cdot \frac{R_2}{R_1} \cdot \ln(n) \cdot \phi_T. \tag{23}$$

which means that now $V_{CTAT} = V_D$, $V_{PTAT} = \phi_T$ and $\alpha = 1$ and $\beta = 2 \cdot (R_2/R_1) \cdot \ln(n)$ The last equation to be modified is 18. Since in the third current mirror branch the number of diodes is not doubled, the first member of 18 remains as it is. The second member is modified as follows:

$$\frac{\partial V_D}{\partial T} + 2 \cdot \frac{R_2}{R_1} \cdot \ln(n) \cdot \frac{\partial \phi_T}{\partial T} = 0. \tag{24}$$

The remaining circuits components to be calculated are the transistors of the current mirror and the resistors. In order to calculate the needed values for all of the components shown in Fig. 3, certain parameters must be assumed. For a current mirror, it is of primary importance that all of the transistors be as equal as possible. Since mismatch is inversly proportional to the transistor size [7], larger transistor should be used - therefore, $W = 5$ μm. For the same reason, the number of parallel diodes in the second branch is chosen as $n = 8$. Namely, this means that there will be 18 diodes in the PTAT circuit, which can be positioned in two matrices of 3x3. In that way, matching conditions are better during the layout phase [4]. For the same reason, a somewhat larger resistor is chosen, $R_1 = 1$ kΩ.

The current mirror transistors MP1, MP2 and MP3 need to operate in saturation and in strong inversion. According to the UICM model, that means: $i_f \gg i_r$. To make sure that they will all remain saturated and in strong inversion in all cases, we assume $i_f = 800$ for all three of them.

Using the transistor model approximate parameters, it is possible to calculate the sheet normalization current:

$$I_{SHP} = \frac{I_S}{W/L} = \frac{\mu C_{ox} \phi_T^2 n}{2} = 13.4 \text{ nA}. \tag{25}$$

According to (16), the drain current, i.e. the current of each of the branches, is determined by the resistor R_1. After the modification of (16) in accordance with (19)-(23), it can be calculated as $I_D \approx 108\mu$A. Now it is possible to calculate the transistor ratio:

$$\frac{W}{L} = \frac{108 \cdot 10^{-6}}{800 \cdot 13.4 \cdot 10^{-9}} = 10.08 \approx 10, \tag{26}$$

which means that $L = 0.5$ μm.

The final unknown value of this circuit is the resistivity of R_2, which is determined from the temperature invariance condition, (24). Even though an approximate value for $\partial V_D/\partial T$ is already mentioned in section III of this paper, it varies significantly for each technology. Therefore, the actual value of this coefficient for this particular technology process is obtained through an experiment: a simple DC simulation yields $\partial V_D/\partial T \approx -1.2$ mV/K, for the range of temperatures from -40 up to 125°C.

Substituting now the obtained temperature coefficients and the rest of the known values in 24, we write an equation with only one unknown:

$$-1.2 \cdot 10^{-3} + 2 \cdot \frac{R_2}{10^3} \cdot \ln 8 \cdot 86.25 \cdot 10^{-6} = 0, \tag{27}$$

which provides a solution to the unknown value of the final circuit component, the resistor $R_2 \approx 3300$ Ω.

V. Simulation Results

Since we used approximate values for the calculations performed in section IV of this paper, some fine tuning is expected. After several iterations of simulations, these are the final parameter values of all the components used in circuit in Fig. 4 all transistor channel widths are 5 μm and all lengths are 0.5 μm, $n = 8$ $R_1 = 1$ kΩ and $R_2 = 2.28$ kΩ.

In the following simulations, the characteristics of the designed bandgap reference voltage circuit are demonstrated. First, the output of the circuit, i.e. the reference voltage, is shown as a function of temperature and of supply voltage. Then, process variations are simulated, observing the worst case scenarios through corner analysis and the statistical mismatch through Monte Carlo analysis. In the end, transient analysis is performed to show that the circuit's operating point is stable.

From Fig. 5 we see that the output voltage varies only ± 0.2 mV around the mean value of 1.14 V over a temperature range

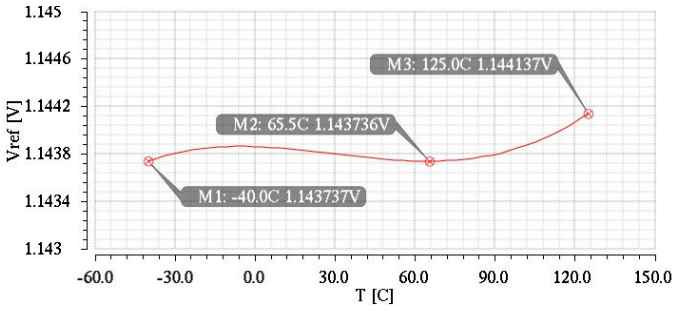


Figure 5. Temperature dependence of the output voltage at nominal supply voltage

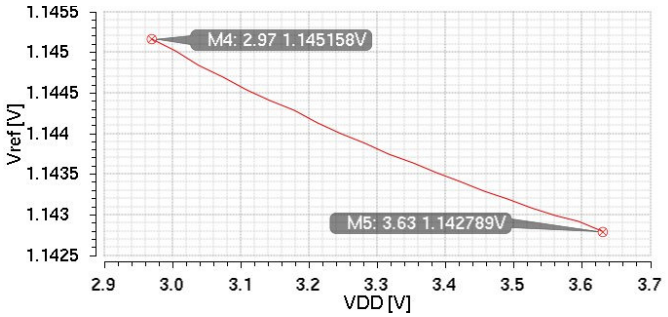


Figure 6. Output voltage dependence on the supply voltage

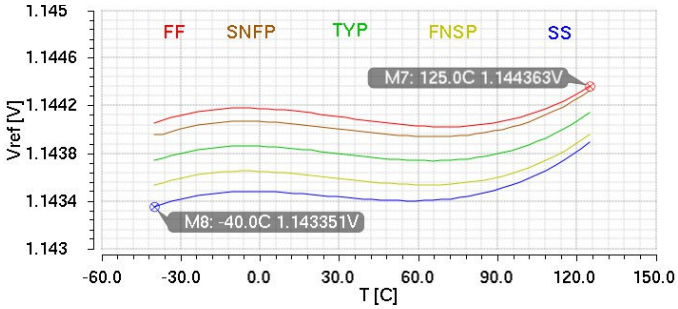


Figure 7. Corner analysis performed on top of the output voltage temperature dependence analysis

160 K wide, i.e. from -40 to 125 °C. This is an error of less than 0.02 % and of ≈1.06 ppm/°C.

Supply voltage variations are demonstrated in Fig. 6. The nominal value of V_{DD} is 3.3 V and here the circuit behavior at ±10 % is shown. The simulation results show that, with the supply voltage change from 2.97 V to 3.63 V, the reference voltage changes its value within the limit of ±1.18 mV around the mean value of 1.14 mV. Relative error is 0.1 %, i.e. 1.57 ppm/mV.

In Figs. 7-10 process variations of the circuit in question are analyzed. Corner and Monte Carlo analysis are performed on top of temperature (Figs. 7 and 9) and on top of supply voltage (Figs. 8 and 10) variation analysis. Through corner analysis, worst case scenarios are treated, showing that the maximum possible change of the output voltage taking into account both temperature and process variations is around 1 mV, which is ±0.04 %. Also, the maximum possible variation of the output voltage as a consequence of both supply and process variations is around 3 mV, i.e. ± 0.13 %.

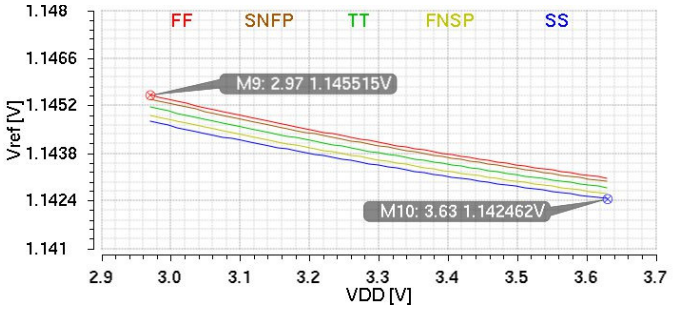


Figure 8. Corner analysis performed on top of a supply voltage analysis of the output voltage

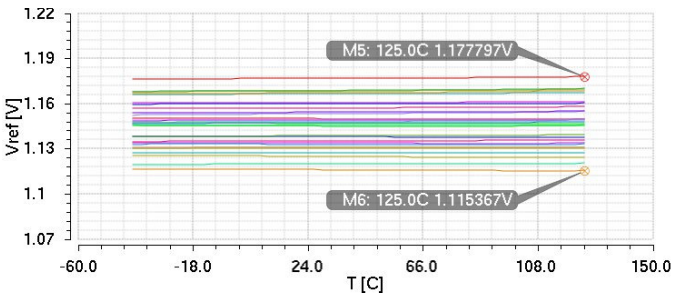


Figure 9. Statistical process variation observed through the Monte Carlo analysis of the temperature dependence of the output voltage

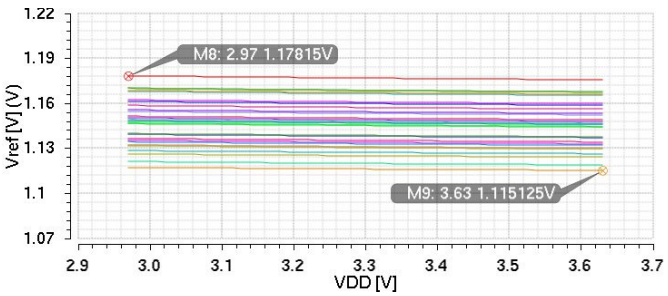


Figure 10. Statistical process variation observed through the Monte Carlo analysis of the output voltage dependence on the supply voltage

If the circuit operates correctly for 30 iterations of Monte Carlo simulation, there is 99 % probability that over 80 % of all possible component values operate correctly - according to the technology process documentation. Therefore, the statistical significane of 30 iterations is acceptable. In this paper, transistor mismatch is analyzed through Monte Carlo simulations, each with the mentioned number of iterations. The statistical analysis is performed on top of temperature (Fig 9) and supply power variation analysis (Fig. 10). In the first case (temperature and mismatch), the analysis shows the reference voltage maximum variation of 62.43 mV, i.e. ±2.72 %. When mismatch is analyzed over the supply voltage variation analysis, the output voltage maximum variation is 63.02 mV, i.e. ±2.75 %.

In Fig. 4 the start-up circuitry is ommited for simplicity. Nevertheless, in order for this bandgap voltage reference to work at the intended stable operation point, such subcircuit is mandatory. Further discussion on the start-up circuit can be found in [3]. In this paper, only the time domain analysis is given in Fig. 11, which shows that the circuit provides nominal output voltage within 62.5 ns.

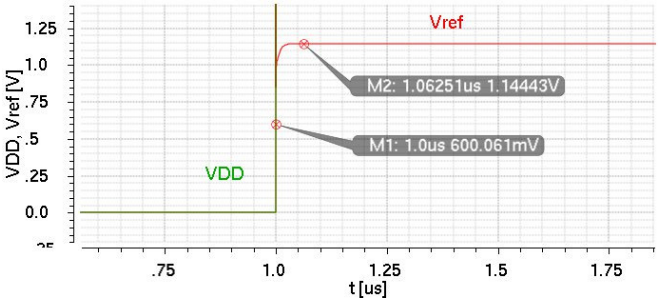


Figure 11. Time domain analysis, showing that the circuit starts-up in 112 ns

VI. Discussion

In Tab. II the characteristics of several circuits of the same purpose are compared. In [8] a bandgap voltage reference is implemented through a topology employing a current-voltage mirror instead of an amplifier, while in [9] a two-stage amplifier is used. These two are characterized by a somewhat larger temperature coefficient and lower power consumption, but at a narrower temperature range. Circuits presented in [10] and [11] show very weak variance of the reference voltage as a function of temperature over a wider temperature range. Also, these are characterized by higher power consumption, up to 1.3 mW. The topologies implemented use operational amplifier, with the addition of a 4-bit trimming circuit in [11]. The purpose of this additional circuit is to reduce the influence of the process variations.

The circuit presented in this paper shows the lowest output voltage temperature variation of the circuits in Tab. II, but it also shows the highest power consumption. There is an obvious trade-off between these characteristics, so increased power consumption is an expected price to pay for severly improved temperature coefficient of only 1.06 ppm/°C. One of the main goals of this research is the improvement of our work presented in [1]. Namely, in that paper a voltage bandgap reference based on a simpler topology is discussed. There, we concluded that the circuit shows satisfactory behavior concerning the temperature coefficient (11.5 ppm/°C), but also that its supply voltage caused output variation is not acceptable ($\pm$16.98 mV, i.e. 21.53 ppm/mV). The circuit presented in this paper is characterized by the improved performance. Namely, temperature variation is improved 10 times, while the supply voltagae caused variation is improved more than 13 times and it is 1.57 ppm/mV. Besides increased power consumption, this circuit is also characterized by a more complex topology and, consequently, higher demand for chip area.

The purpose of this comparison in no way is a claim which of the circuits is better, since each one of them is designed with the author having a different goal in mind, e.g. low power, higher stability or smaller chip area. The Tab II is given in order to point out that the results obtained within this paper are of the same order like those found in current literature.

TABLE II. Characteristics comparison

Circuits	Technology process [nm]	Supply voltage [V]	Temperature coefficient [ppm/°C]	Temperature range [°C]	Power [μW]
[1]	130	3.3	11.5	-40:125	600
[8]	350	2.5	12.1	5:95C	276
[9]	180	1.8	6.5	-20:90	150
[10]	130	3.3	5.5	-40:105	1300
[11]	180	1.8	2	-20:140	620
This work	130	3.3	1.01	-40:125	3710

VII. Conclusion

A temperature stable bandgap reference voltage circuit is presented in this paper. Concerning both temperature and supply voltage stability, this bandgap voltage reference represents an advancement compared to our previous work. On the other hand, power consumtpion is increased, since a more complex topology employing a separately developed operational amplifier is used. Process variation analysis performed (both corner and Monte Carlo) show that the output is influenced by the fabrication process. This characteristic remains to be worked upon during the layout phase, when more attention will be dedicated to device mismatch.

Acknowledgment

This work is funded from the European Union's Seventh Framework Program for research, technological development and demonstration under grant agreement no. 289481. - project SENSEIVER.

References

[1] A. Pajkanovic and V. Malbasa, "Bandgap voltage reference in 130 nm: Design and schematic level simulation," in *X International Symposium on Industrial Electronics INDEL 2014, Banja Luka, November 06-08, 2014*, 2014, pp. 89–93.

[2] T. Lee, *The RF CMOS Integrated Circuit Design*. Prentice Hall, 1998.

[3] M. C. Schneider and C. Galup-Montoro, *CMOS Analog Design Using All-Region MOSFET Modeling*. Cambridge University Press, 2010.

[4] B. Razavi, *Design of Analog CMOS Integrated Circuit*. McGraw-Hill, 2001.

[5] A. I. A. Cunha, M. C. Schneider, and C. Galup-Montoro, "An mos transistor model for analog circuit design," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 10, pp. 1510–1519, 1998.

[6] A. Pajkanovic, "An operational amplifier implementation: Design, simulation and layout," in *XIV International Symposium INFOTEH-JAHORINA 2015, Jahorina, March 18-20, 2015*, 2015, p. in press.

[7] P. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 6, pp. 1212–1224, 2005.

[8] Y.-H. Lam and W.-H. Ki, "Cmos bandgap references with self-biased symmetrically matched current-voltage mirror and extension of sub-1-v design," *Very Large Scale Integrateion (VLSI) Systems, IEEE Transactions on*, vol. 18, no. 6, pp. 857–865, 2010.

[9] Y. Yusoff, H. C. Lah, N. Razali, S. N. Harun, and T. Yew, "Design and characterization of bandgap voltage reference," in *Semiconductor Electronics (ICSE), 10th IEEE International Conference on*, 2012, pp. 686–689.

[10] M. Lazic, D. Mirkovic, and P. Petkovic, "Integrated voltage reference design in 130 nm cmos technology process (in serbian)," in *Proceedings of 57th ETRAN Conference, Zlatibor, Serbia*, 2013, pp. EL2.1.1–6.

[11] A. Martinez-Nieto, M. Sanz-Pascual, P. Rosales-Quintero, and S. Celma, "A bandgap voltage reference in 0.18μm cmos technology," in *Circuits and Systems (MWSCAS), IEEE 56th International Midwest Symposium on*, 2013, pp. 97–100.

Design of a power supply and voltage references for a wirelessly powered ADC in 0.18-μm CMOS

Niko Bako, Željko Butković and Adrijan Barić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 1000 Zagreb, Croatia

***Abstract*—A low power tag with integrated ADC designed in 0.18-μm CMOS process is presented. The tag extracts energy from electromagnetic waves and stores the charge on external capacitor. The stored charge is used to supply the tag circuitry: supply unit, current and voltage references and voltage regulators and ADC. The obtained referent voltage shows good stability over process, voltage and temperature variations. The mean value of the reference voltage is 485.8 mV and the standard deviation is σ= 7 mV in the temperature range -10° to 70°. The mean temperature coefficient (TC) is 7 ppm/°C and the standard deviation is σ_{TC}= 30 ppm/°C in the same temperature range.**

I. Introduction

In recent years there has been a growing interest in integration of sensing and telemetry to create wireless sensor networks. Integration of radio frequency identification (RFID) systems with sensors has potential usage in infrastructure and object monitoring, identification of harmful agents and implantable biomedical devices [1]. Also, wirelessly powered healthcare sensors enable continuous health monitoring [2], [3]. In such systems available power is limited because the energy is very often extracted from environment and used to supply chip (tag). One way of energy harvesting is harvesting from electromagnetic waves.

The key component in intelligent RFID systems is an analog to digital converter (ADC) because it has to provide high resolution and low power consumption [1]. The designers of a such ADCs are focused on lowering power consumption and increasing the ADC speed. The examples of a low power ADCs are presented in [1]–[5]. Very often generation and stability of reference voltage for the ADCs is challenging task because classic bandgap voltage references are not used in the tags due to low supply voltage of the the tags. The papers [1], [4] presents ADC but without reference and supply voltage while papers [2], [3] presents the whole chip with on-chip voltage references but it is not clear how the references are designed. This paper focuses on design of supply voltage and fully differential reference voltages generation for the fully differential 9-bit ADC presented in [5]. The proposed circuity are integrated with ADC and the proposed circuit (tag) is implemented in 0.18-μm CMOS process.

This paper is organized as follows. Section II presents the proposed circuit, Section III presents simulation results of the proposed circuits with ADC and Section IV brings the conclusion.

II. System description

Fig. 1 shows the block diagram of the proposed tag. The ADC is a 9-bit fully differential cyclic ADC with the following performance characteristics; the supply voltage range is from 1.4 to 1.8 V, the differential reference voltage is 500 mV and the typical common mode voltage is 700 mV. At the sampling rate of 10 kS/s the ADC consumes 7.4 μW from 1.4 V supply (current consumption is 5.3 μA). Input energy, extracted from electromagnetic waves, must be sufficient to ensure normal operation of the ADC under described conditions. The input impedance of the ADC is capacitive.

The tag powering strategy, which reduces required input power, is adopted from [6]. The RF/DC converter converts harvested RF energy to DC voltage V_{DD}. The charge needed to produce the voltage V_{DD} is stored in the external capacitor C_{STORE} and when the tag is active, V_{DD} is used as a supply for the rest of the tag circuitry. When V_{DD} is low, only voltage references are enabled so the tag consumes the minimal power. When V_{DD} exceeds targeted upper voltage limit V_{HIGH} the voltage detection unit will enable the rest of the tag circuitry (the tag is active). When the tag is active, the harvested energy stored in the capacitor is decreasing as the capacitor discharges reducing the supply voltage. At the low targeted voltage limit V_{LOW} the voltage detection circuit will turn off the circuitry so power consumption is minimal and the supply voltage will rise again. The designed tag operates in voltage range $V_{HIGH} - V_{LOW} \approx 0.4$ V so the tag circuitry has to be insensitive to supply variations.

A. Voltage detection and voltage limiter

Fig. 2 shows the voltage reference circuit, the voltage detection and power limiter circuit presented in [6]. A low-power voltage and temperature compensated voltage reference is presented in [7]. The circuit in Fig. 2 generates two 485 mV reference voltages which are insensitive to voltage and temperature variations. One is for the ADC (ADC_{REF}) and the other is for the voltage detector, voltage limiter and for voltage regulators (VDD_{REF}). The reference generation for

This work has been partly supported by the European Union from the European Regional Development Fund under the project Wirelessly powered microelectronic circuits for distributed sensor networks, project code RC.2.2.08-16, contracted through the Ministry of Science, Education and Sports of Croatia

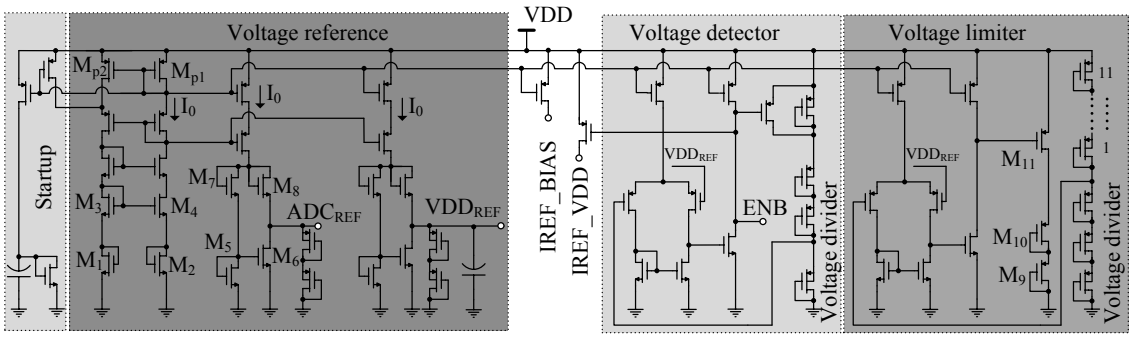


Fig. 2. Schematic of voltage detector circuit and voltage limiter.

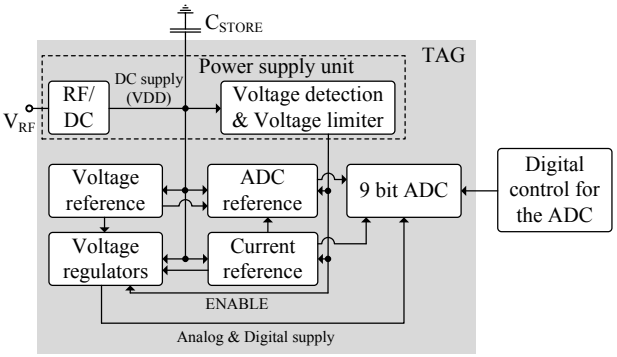


Fig. 1. Block schematic of implemented solution.

the ADC and for the supply is separated in order to avoid possible interferences generated by the voltage regulators.

The stability of the reference voltage depends on process parameters so Monte Carlo mismatch analysis is performed. The mean value of the reference voltage is 485.8 mV and the standard deviation is σ= 7 mV. The mean temperature coefficient (TC) is 7 ppm/°C and the standard deviation is σ_{TC}= 30 ppm/°C. The temperature coefficient is calculated in range -10 °C to 70 °C.

The voltage detector is a low power comparator with the hysteresis. Since the power consumption must be low, four pMOS transistors are used as a voltage divider instead of using large resistors. The transistors bulk and source terminals are shorted and each transistor has the same threshold voltage (i.e. same characteristics) so they are considered to be resistor with equal resistance. When V_{DD} exceeds high level V_{HIGH} = 1.94 V (4·VDD_{REF}) the comparator digital output ENB is set low. This signal enables the rest of the circuitry and the supply voltage $IREF\_VDD$ for current references so C_{STORE} discharges and voltage decreases. When the voltage reaches low level V_{LOW} = 1.455 V (3·VDD_{REF}), ENB is set high and the tag circuitry is turned off so C_{STORE} voltage rises. When V_{DD} is in range 1.455 V – 1.94 V the tag is active so the voltage hysteresis is 485 mV.

The voltage limiter detects if V_{DD} exceeded the breakdown voltage (2.2 V). When the voltage is too high the limiter will very fast turn on the transistor M_{11} and pull large current which will discharge the capacitor C_{STORE} and decrease V_{DD}. In order to prevent a short circuit V_{DD} - ground, two pMOS-based resistors are added (M_9 and M_{10}).

The consumption of the circuit from Fig. 2 must be as low as possible. The overall current consumption in voltage range from 1.4 V to 2 V is 60 nA in a typical case. The voltage reference circuit consumes 41 nA while voltage detection and voltage limiter consumes 12 nA and 7 nA respectively. The worst case consumption is 100 nA (fast corner, 70 °C).

Because very low current flows (15 nA) into the transistors $M_5 - M_8$ and the parasitic capacitances of the transistor gates are large, the reference voltages reaches 485 mV very slowly (the settling time is ≈500 μs). On the other hand ADC_{REF} must be stable before the comparator enables the tag circuitry. This can be achieved if ADC_{REF} is generated at the same time as, VDD_{REF}, so when the tag is enabled (ENB is low) ADC_{REF} has already settled at 485 mV.

B. Current references

Fig. 3 shows the current reference circuit. The circuit generates six current references of ∼100 nA for voltage regulators and for ADC references and one reference current of ∼150 nA for the ADC. The overall current consumption of the circuit is 1.2 μA. The stability of the reference current over temperature and voltage variations is not crucial because generated currents are used as reference currents for the operational amplifiers used in tag. It is only important that these currents are as low as possible due to the power consumption restriction. The current variations over temperature and voltage variations are shown in Fig. 4.

In order to minimize the current consumption the reference supply voltage $IREF\_VDD$ is disabled and the circuit is turned off when ENB is high. After ENB is set low $IREF\_VDD$ is equal to the voltage V_{DD}. The startup circuit is modified to ensure fast turn on. The parasitic capacitance C_p slows down start up and, after enabling the reference circuit, this capacitance charges with current $IREF\_BIAS$ which is generated in the voltage reference circuit. The transistor M_{s1} is always on because it is connected to V_{DD}. After enabling

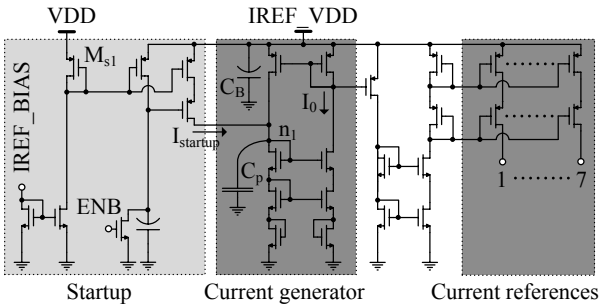


Fig. 3. ADC references.

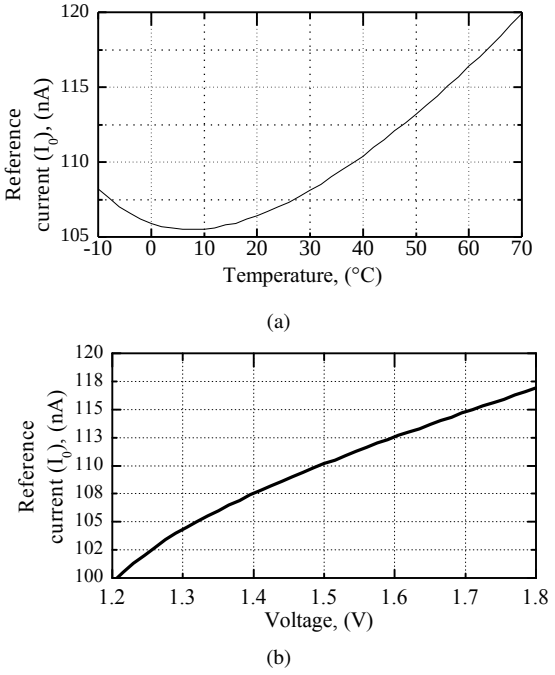


Fig. 4. Reference current variations as a function of temperature and supply voltage obtained by simulations. (a) Reference current variation as a function of temperature (b) Reference current variation as a function of supply voltage.

$IREF\_VDD$ the current $I_{startup}$ will quickly charge C_p and improve startup time.

C. ADC references

The ADC reference circuit is shown in Fig. 5. Input is the referent voltage REF_{ADC} and the outputs are three voltages V_{RHIGH}, V_{RMID} and V_{RLOW}. Three voltages are necessary because the ADC is fully differential. V_{RHIGH} and V_{RLOW} are high and low referent voltages while V_{RMID} is the common mode voltage. The voltage REF_{ADC} is first amplified 2 times so the voltage V_{REF} = 970 mV. V_{REF} is divided and then three reference voltages are generated. To reduce current consumption the feedback resistors should be large. The area is reduced because pMOS transistors M_{R1} and M_{R2} are used instead of resistors. The transistors have the same (W/L) ratio but consists of multiple devices to enable good matching. The overall (W/L) ratio is 1µm/8µm. The source and bulk terminals of the transistors M_{R1} and M_{R2} are shorted so the transistors have the same threshold voltage and resistance R_{DSon}. According to [8] the voltage V_{REF} is

$$V_{REF} = REF_{ADC}\left(1 + \frac{m}{n}\right) \quad (1)$$

where m is the number of transistors between the nodes V_{REF} and V_{FB} and n is the number of transistors between the nodes V_{FB} and ground. Since m=n=1 the voltage V_{REF}=V_{RHIGH} = 970 mV. Low reference voltage is V_{RLOW}= 485 mV and the common mode voltage is $V_{RMID} \approx$ 730 mV. The capacitors C_{R1}, C_{R2} and C_{R3} are blocking capacitors. Their task is to filter any interference which could couple from the power supply V_{DD}. Also, the buffers $OA_2 - OA_4$ prevents switching noise from the ADC during AD conversion.

The power consumption of the circuit is crucial so, in order to reduce power consumption, the operational amplifiers $OA_1 - OA_4$ are directly connected to V_{DD} without voltage regulators. This means that the amplifiers must have good power supply rejection ratio (PSRR). Besides PSRR the reference buffers $OA_2 - OA_4$ also must be capable driving capacitive loads since the ADC inputs are purely capacitive load. Because the ADC has 9-bit resolution the buffers open loop gain is A_{V0} = 60 dB. With 60 dB open loop gain voltage error at the output of buffers is less than 1/2 of differential nonlinearity (DNL) for a 9-bit resolution e.g. $V_{REFH} - V_{REF}$= 0.45 mV. During AD conversion for each bit the buffers must charge or discharge ADC input capacitors (the DAC capacitors) very fast and the reference voltages must settle before the conversion of the next bit starts. The settling time accuracy S is defined as $S = 2^{-b}$ where b is the number of bits [9]. Under assumption that the buffers are modelled as a single-pole systems the gain bandwidth (GBW) is given in [9]

$$GBW = -\frac{\ln S}{\beta T_S} \quad (2)$$

where β is the feedback factor of the closed-loop system (for buffers β=1) and T_S is the settling time. The setting time is defined by the speed of the ADC. The ADC output voltage settles within 2 µs so the buffers must settle much faster and T_S is chosen to be 0.5 µs. According to (2) for 9-bit resolution and T_S = 0.5 µs GBW is $\approx$ 2 MHz.

The classic Miller compensated amplifier has very low gain bandwidth (GBW) which depends on compensation capacitor C_C and it is not suitable for this application. Also the speed is limited by the slew rate. The amplifier presented in Fig. 6 is capable to source and sink large currents so the slew rate is not an issue although quiescent current of the amplifier is only 1 µA. The GBW can be improved if the amplifier is load compensated as the amplifier in Fig. 6.

Because of the small dimensions the amplifier offset is σ= $\pm$2.6 mV. Since the proposed ADC is fully differential circuit the buffers OA_2 and OA_4 from Fig. 5 are layouted close to each other. In this way the differential offset can be reduced.

The Table I presents the performance of the amplifier from Fig. 6 when C_C = 1.8 pF. The transient simulations confirmed that, when phase margin (PM) is 54°, the buffer amplifier

TABLE I
PERFORMANCE OF THE AMPLIFIER

A_{vo} (dB)	61
GBW (MHz)	2.34
Phase margin (°)	54
PSRR@900 MHz (dB)	-70
Quiescent current (μA)	1

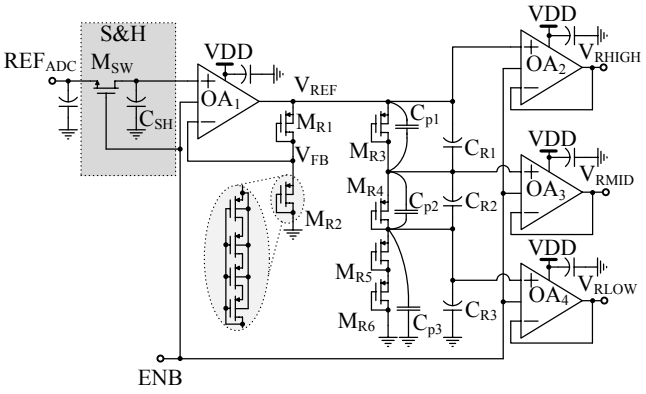


Fig. 5. The ADC reference circuit.

has no overshoots. Fig. 7 shows transient behaviour of differential reference voltage $V_{REF\_DIFF} = V_{RHIGH} - V_{RLOW}$ during AD conversion. During AD conversion the buffers OA_2 and OA_4 charges the ADC input capacitors so voltage V_{REF_DIFF} drops. The voltage settles to its nominal value of 485 mV within 0.5 μs which is enough for the ADC maximal conversion rate of 10 kS/s.

D. Voltage regulators

The ADC is insensitive to power supply variations but the voltage regulators are added to separate the digital and analogue supply. The ADC does not need state of the art regulators so the regulators, used in this design, are actually non-inverting amplifiers shown in Fig. 8. The amplifier gain is 3 so the output voltage voltage ($AVDD$ or $DVDD$) is $AVDD = 3VDD_{REF}$ ≈1.45 V. An operational amplifier used in the voltage regulators is the same amplifier which

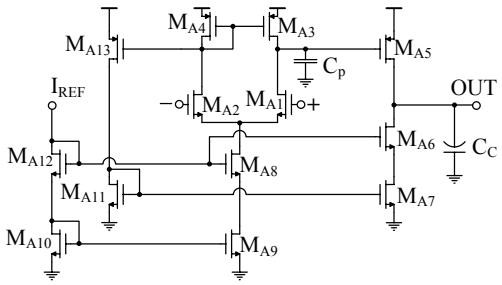


Fig. 6. Operational amplifier.

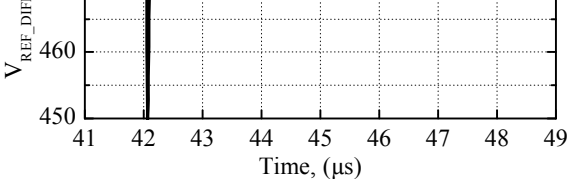


Fig. 7. Transient simulation of the ADC references.

TABLE II
VOLTAGE REGULATOR PERFORMANCE

Line regulation (%)	3
Load regulation (mV/μA)	3
PSRR@900 MHz (dB)	-75
Quiescent current (μA)	1
Load transient response - rising edge (μs)	1
Load transient response - falling edge (μs)	0.7

is used for the ADC reference generation (Fig. 6). The pass element is the transistor M_{A5}. To save the area a pMOS transistors $M_{d1} - M_{d3}$ are used instead of resistors. The transistors have the same (W/L) ratio, but each transistor consists of two parallel devices to enable good matching. In order to track fast transition on $AVDD$ and $DVDD$ pins the feedback loop, which is formed by $M_{d1} - M_{d3}$, must have fast response time. The parasitic capacitance C_{pFB} slows loop performance so it must be reduced. This can be done by reducing (W/L) of $M_{d1} - M_{d3}$ so the transistor ratio is 1μm/2μm. The value of the blocking capacitor C_{B2} is 34 pF and it is implemented as an nMOS capacitor (MOSCAP). When load current varies in range 0 – 5.3 μA (maximal ADC current) the voltage $AVDD$ settles within 1 μs.

Table II presents typical performance of the designed regulator.

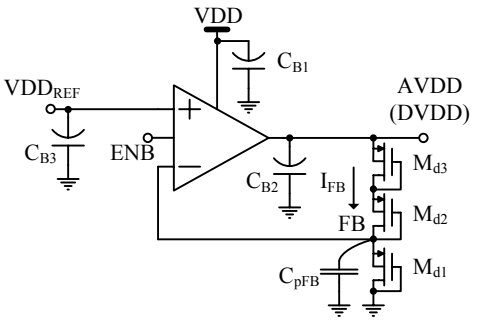


Fig. 8. The schematic of voltage regulators.

E. RF/DC converter

The schematic of RF/DC converter is shown in Fig. 9. The transistors are low threshold voltage nMOS transistors with the typical threshold voltage of 220 mV. The capacitors are metal–insulator–metal (MIM) capacitors. The converter must be able to deliver minimal DC output voltage of 1.94 V for the load current of 100 nA (load is the circuitry for voltage detection and voltage limiter) with minimum input amplitude. The optimal number of half-voltage rectifier stages (*N*), transistors size and capacitors value are obtained through simulation. The simulations have shown that for $W/L = 35\mu m/0.24\mu m$ (all transistors have the same dimensions), $C = 0.6$ pF (all capacitors have the same dimensions) and *N*=24.

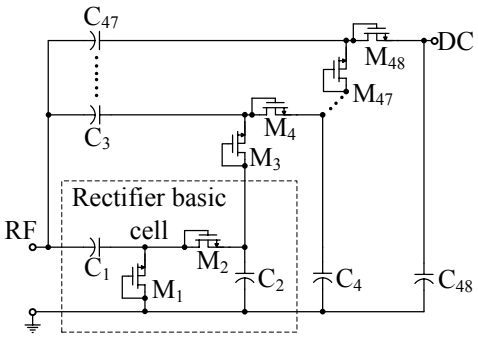


Fig. 9. Block schematic of implemented solution.

F. Tag implementation

The tag is implemented in a 0.18-μm CMOS process and occupies 0.96 mm^2 (1.2×0.8 mm^2) excluding pads. Fig. 10 presents the tag layout. Most of the area is occupied by the ADC and the ADC references. The blocking capacitors $C_{R1} - C_{R3}$, used in the ADC references (Fig. 5), are MiM capacitors. The capacitors consumes most of the reference area.

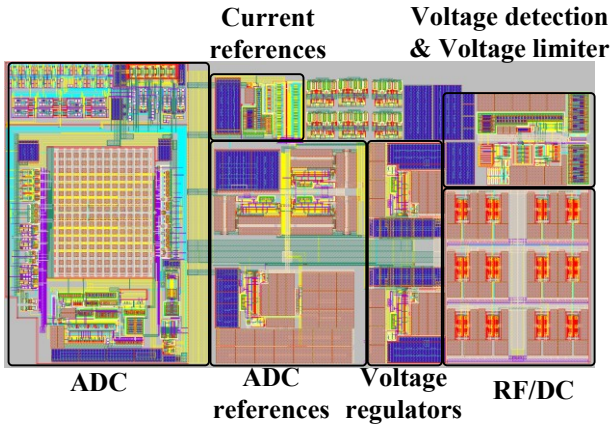


Fig. 10. Layout of the tag.

III. The Tag Performance Results

Fig. 11 shows generated voltage reference V_{REF_DIFF}, analogue and digital supply voltages $AVDD$ and $DVDD$.

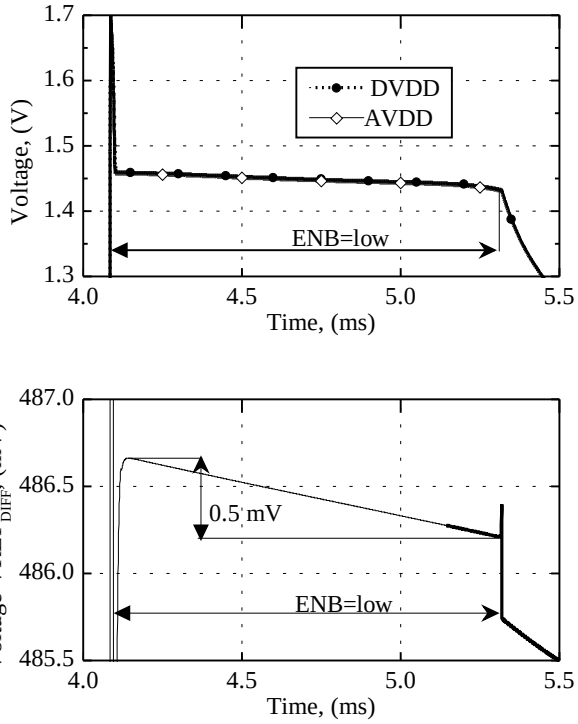


Fig. 11. The generated voltages $AVDD$, $AVDD$ and V_{REF_DIFF}

In order to speed up simulation time, RF signal amplitude is 300 mV while frequency is 900 MHz. The capacitance of the storage capacitor C_{STORE} is 11 nF. With this capacitance the tag is active 1.1 ms which is sufficient to ensure normal operation of the ADC (ADC conversion cycle is 80 μs). The digital signal ENB (active low) enables voltage regulators and voltage reference circuits. After the signal is set low the differential reference voltage V_{REF_DIFF}, analogue and digital supply voltages $AVDD$ and $DVDD$ are settled to final value after 40 μs. The supply voltages varies when the tag is active. This is because VDD_{REF} slightly changes as the supply voltage V_{DD} decreases.

A. The ADC performance results

In order to test performance of the whole tag, the ADC static and dynamic characteristics are simulated. The post layout simulations of the tag would be too complicated so only the ADC extracted parasitic parameters are included in simulations. Because of the same reasons the RF/DC verilog A model, presented in [10], is used instead of the transistor level model. The RF/DC verilog A model models the rectifier in time domain and it is easily implemented in circuit simulator and simulated with the rest of the tag circuitry. Compared with transistor level simulations the model achieves the simulation speed-up factor from 1080 to 3.84E6 times for a root-mean-square error (RMSE) below 4.8 mV.

Static characteristics are shown in Fig. 12. The simulated differential nonlinearity (DNL) is 0.075/-0.18 LSB and the integral nonlinearity (INL) is 0.22/-0.2 LSB which is the same

result as obtained in [5]. Dynamic characteristics are simulated for a full scale 1 kHz sine wave (V= 967 mVpp) and 10 kS/s sampling rate with included transient noise. The external capacitor C_{STORE} is 220 nF. With this capacitor the ADC can take 60 samples before the tag is off. Simulations with included transient noise are repeated 8 time which gives 480 samples. The signal spectrum is shown in Fig. 13. The spurious free dynamic range (SFDR) is 66 dB and signal-to-noise and distortion ratio (SNDR) is 52 dB.

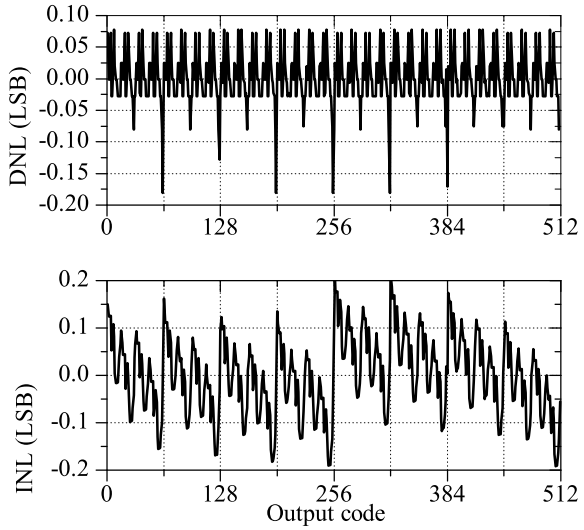


Fig. 12. Integral nonlinearity (INL) and Differential nonlinearity (DNL).

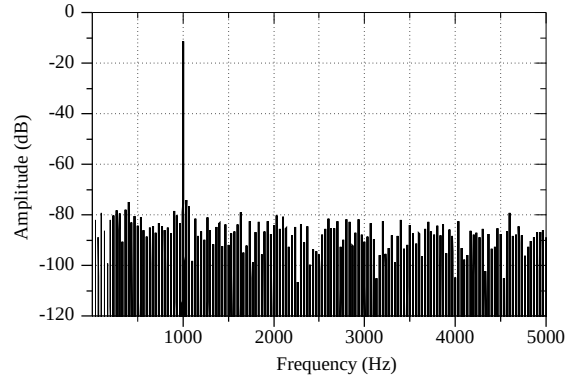


Fig. 13. FFT spectrum of 1 kHz input signal at 10 kS/s.

The maximal overall current consumption, including ADC, is 13 μA. The average power consumption is 22 μW. The most of the power is consumed by the ADC reference circuit (10 μW) and ADC (7.4 μW).

IV. Conclusion

This paper presents the design of power supply and voltage references for a fully differential 9-bit cyclic ADC. The supply, voltage references and ADC are integrated in tag. The tag extracts supply voltage V_{DD} from incoming electromagnetic waves. The supply voltage is stored in the external capacitor C_{STORE}. When the tag is active V_{DD} is used as a supply for the rest of the tag circuitry. The tag is on when V_{DD} exceeds the upper voltage level V_{HIGH} ≈1.94 V. When V_{DD} is lower than the lower voltage level V_{LOW} ≈1.455 V the tag is off. The generated reference voltage for the ADC and for the voltage regulators shows good stability over temperature and supply variations. The mean value of the reference voltage is 485.8 mV and the standard deviation is σ= 7 mV. The mean temperature coefficient (TC) is 7 ppm/°C and the standard deviation is σ_{TC}= 30 ppm/°C.

The tag performance is tested together with the ADC. The simulated static characteristic of the ADC, DNL and INL, are 0.075/-0.18 LSB and 0.22/-0.2 LSB respectively. Dynamic characteristics are following: the spurious free dynamic range (SFDR) is 66 dB and signal-to-noise and distortion ratio (SNDR) is 52 dB. The equivalent number of bits (ENOB) is 8.35. The overall power consumption is 22 μW.

References

[1] D. D. Venuto, D. T. Castro, Y. Ponomarev, and E. Stikvoort, "0.8 μW 12-bit SAR ADC sensors interface for RFID applications," *Microelectronics Journal*, vol. 41, no. 11, pp. 746 – 751, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0026269210001217

[2] Y. Zhang, F. Zhang, Y. Shakhsheer, J. Silver, A. Klinefelter, M. Nagaraju, J. Boley, J. Pandey, A. Shrivastava, E. Carlson, A. Wood, B. Calhoun, and B. Otis, "@a Batteryless 19 μ w MICS/ISM-Band Energy Harvesting Body Sensor Node SoC for ExG Applications," *IEEE Journal of Solid-State Circuits*.

[3] L. Yan, J. Yoo, B. Kim, and H.-J. Yoo, "A 0.5-$\mu\mathrm{V}_{\mathrm{rms}}$ 12-μW Wirelessly Powered Patch-Type Healthcare Sensor for Wearable Body Sensor Network," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 11, pp. 2356–2365, 2010.

[4] D. Zhang, A. Bhide, and A. Alvandpour, "A 53-nW 9.1-ENOB 1-kS/s SAR ADC in 0.13-μm CMOS for Medical Implant Devices," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 7, pp. 1585–1593, 2012.

[5] N. Bako and A. Baric, "A low-power fully differential cyclic 9-bit ADC," in *19th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2012*, 2012, pp. 504–507.

[6] S. Radiom, M. Baghaei-Nejad, K. Aghdam, G. A. E. Vandenbosch, L.-R. Zheng, and G. G. E. Gielen, "Far-Field On-Chip Antennas Monolithically Integrated in a Wireless-Powered 5.8-ghz Downlink/UWB Uplink RFID Tag in 0.18-μm Standard CMOS," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 9, pp. 1746–1758, 2010.

[7] G. De Vita, G. Iannaccone, and P. Andreani, "A 300 nW, 12 ppm/°C voltage reference in a digital 0.35 μm CMOS process," in *Symposium on VLSI Circuits, 2006. Digest of Technical Papers. 2006*, 0-0 2006, pp. 81 –82.

[8] J. Heidrich, D. Brenk, J. Essel, M. Heinrich, M. Jung, G. Hofer, G. Holweg, R. Weigel, and G. Fischer, "Design of a low-power voltage regulator for rfid applications," in *International Conference on Computational Technologies in Electrical and Electronics Engineering (SIBIRCON), 2010 IEEE Region 8*, 2010, pp. 552–557.

[9] J. A. M. Jarvinen, M. Saukoski, and K. A. I. Halonen, "A 12-Bit Ratio-Independent Algorithmic A/D Converter for a Capacitive Sensor Interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 3, pp. 730–740, 2008.

[10] V. Ceperic, N. Bako, and A. Baric, "A symbolic regression-based modelling strategy of ac/dc rectifiers for {RFID} applications," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7061 – 7067, 2014.

RFID UHF Protocol Implementation in Distributed Sensor Networks

Ivo Broz*, Niko Bako**, Željko Butković**, Adrijan Barić **
* Locus Ltd, Zagreb, Croatia
** University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
ivo.broz@lite.hr, niko.bako@fer.hr, zeljko.butkovic@fer.hr, adrijan.baric@fer.hr

Abstract - In today's market significant segment is taken by the applications using sensors to track physical events in the surrounding environment. Market raise is particularly related to the distributed sensor networks based on the wireless communications (like Internet of Things IoT devices). One of the hardest challenges in this field is associated with the power supply requirements. Battery operated electronics is limited by the battery capacity and battery lifetime. This paper proposes solution based on the RFID EPC-Gen2-protocol-based readers supplying power to the sensor nodes, consisting of the microelectronic circuit and the associated sensor. Besides the power, RFID protocol provides means of wireless communication, without the need for the battery supply on the sensor node side. Analogue circuits for receiving and demodulating RF ASK modulated signal from the reader, and transmitting back the signal to the reader using backscattering technique, are presented. Subset of the command protocol is implemented to support data communication including sensor identification through the unique ID code and sensor reading in the form of the digital word provided by the on-chip ADC. RTL description for digital control is developed in VHDL and synthesized. The prototype is processed in UMC Mixed-Mode/RF 180 nm technology.

I. INTRODUCTION

The main challenge in design of low cost, low power sensor nodes in wireless sensor networks is the power supply. RF communication takes the significant part of the power consumption and the time period before the battery requires replacement is often limited. Battery replacement requires human intervention and monitoring that raises the cost of the device and overall system. Sensor nodes without the battery supply, powered by energy harvesting are ideal for the sensor networks.

RFID standard EPC Gen2 [1] operating in the UHF range around 900 MHz (frequency range depends on the region) is present on the market for long time and is still mainly used for the identification of goods or people. The idea presented in this paper is to use standard RFID readers for powering and communication with the sensor nodes in the sensor network, with each node implemented as an integrated circuit (IC) with a sensor without the battery supply. The reader provides power to the IC and communicates according to the RFID protocol. The proposed IC in comparison with a regular RFID tag includes the on-chip ADC. It requires powering in the same way as the digital logic for the RFID communication. The ADC provides sensor reading and this binary data, sent together with the unique ID number of the sensor node, is sent to the reader, as the central place for gathering information from all the sensor nodes in the network.

This work has been partly supported by the European Union from the European Regional Development Fund under the project „Wirelessly powered microelectronic circuits for distributed sensor networks", project code RC.2.2.08-16, contracted through the Ministry of Science, Education and Sports of Croatia.

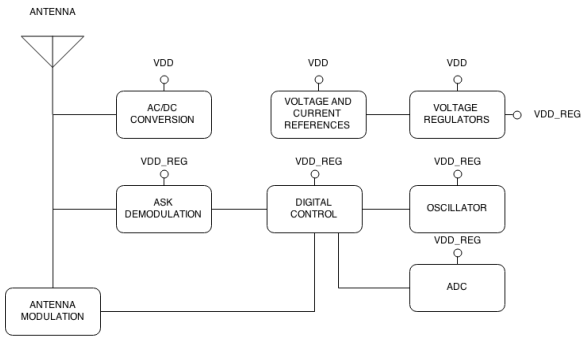


Figure 1. Block Schematic of Proposed IC Tag System.

IC tag system is composed of the energy harvesting module (AC/DC conversion), voltage and current references, voltage regulators, ADC, RF demodulator for the received signal (ASK demodulation), RF modulator for the transmit signal (antenna modulation), oscillator and the digital control logic for the protocol implementation. Fig. 1 shows the block schematic of the proposed system.

In this paper RF modules for modulation and demodulation and the digital logic for the RFID protocol are presented. Section II presents basic principles of the RFID communication and elaborates respective analogue and digital blocks for the protocol implementation. Section III provides the information about the prototype while the section IV gives the conclusion.

II. RF COMMUNICATION

A. *RF communication principle*

RFID system is always consists of the RFID reader and the tag. The tag is an IC chip that carries information that the reader fetches. The basic information is the unique identification number (ID) that directly identifies each tag and indirectly person or goods associated with the tag.

Depending on the power supply, tags can be active (battery supply for powering RF communication and digital logic), semi passive (battery supply for powering digital logic) and passive (no battery supply is provided, all power comes from RF signal harvesting) [2].

Readers and tags operate in different RF frequency bands: LF (125/134 kHz), HF (13.56 MHz) and UHF (880 – 960 MHz, 2.4 GHz) each with its own standard.

This paper presents a passive tag solution operating in UHF range 880 – 960 MHz based on EPC Gen2 protocol [2].

Passive tags don't use battery supply, so the complete power necessary for the RF communication, digital protocol operations and additional functions (in this case ADC conversion) must come from the energy harvesting of the RF signal. Energy must be stored during energy harvesting and the external capacitor is used for this purpose.

Passive tag doesn't incorporate any active RF modules due to limited power supply availability. Communication from the tag to the reader is provided using backscatter technique.

Backscattering is based on the following principle. Antenna on the tag side receives RF signal from the reader (forward link) that induces voltage on it and causes current to flow in antenna load (tag circuits). The principle of reciprocity says that every structure that receives an RF wave also transmits a wave. This transmission is proportional to the current induced in the antenna. This leads to the general principle of backscattering; by the changing antenna load it is possible to change the amplitude of an induced current causing the respective change of an amplitude of the transmitted wave (reverse link). The principle is displayed in Fig 2.

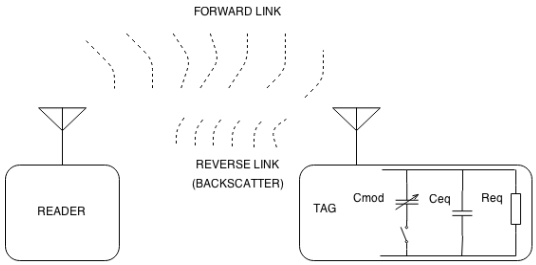


Figure 2. Reader – Tag Communication.

The tag impedance can be approximated by the parallel combination of the capacitor *Ceq* and the resistor *Req*. Resistor *Req* is equivalent to the tag power consumption, while *Ceq* capacitive load comes from the load of the AC/DC converter and the RF demodulation modules. The *Cmod* capacitance is the capacitor used for the modulation of the antenna load.

If the reader transmits RF signal with the constant amplitude, by changing antenna load it is possible to modulate backscatter signal as communication link from the tag to the reader. The largest backscatter signal is obtained when the antenna impedance is matched to the impedance of the tag on the transmitting frequency without connecting *Cmod* (switch in off state). If *Cmod* is connected in parallel to the tag impedance (switch in on state), matching is reduced and the backscatter signal has lower amplitude. This is the general principle which uses low power consumption from the tag side for establishing communication with the reader.

B. Modulation/demodulation

RFID Gen2 protocol defines three modulation types for RF signal communication: double sideband amplitude shift keying (DSB-ASK), single sideband amplitude shift keying (SSB-ASK) and phase reversal amplitude shift keying (PR-ASK) using a pulse interval encoding (PIE) format.

DSB-ASK modulation is chosen for the implementation of a reader to a tag communication due to simple solution for the demodulation of RF signal. RF signal in time domain is shown qualitatively in Fig. 3 (frequency of the RF signal is much higher than the periods of high and low amplitude, so transitions would not be visible in the correct time scale).

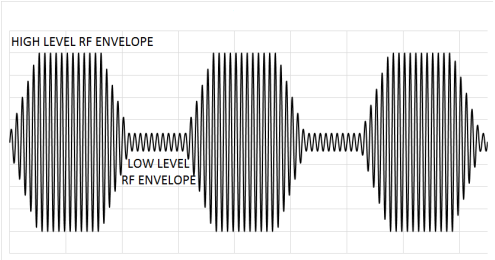


Figure 3. ASK Modulated RF signal.

RF signal is amplitude modulated with two states: the high state and the low state, represented by the high amplitude and low amplitude of the signal envelope. The low amplitude must be in the range from 0 to 20% of the high envelope amplitude according to the protocol.

Demodulation is performed by the envelope detector circuit and the comparator that extracts the RF envelope into pulse train digital signal [3]. This signal has to be decoded into the actual binary information.

Digital signal is encoded with PIE (Pulse Interval Encoding) format. Information about logical "0" and "1" is decoded from the ratio of periods of the high state and the low state according to the Fig. 4. All time intervals are expressed in "*Tari*" values defined by the standard which are in absolute values from 6.25 μs to 25 μs.

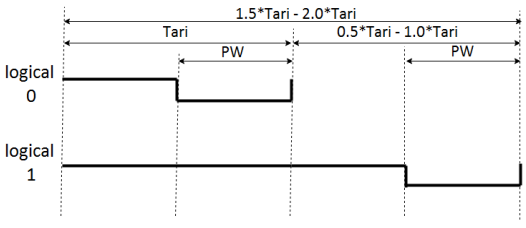


Figure 4. PIE encoding.

Extraction of the digital information is performed in the digital circuitry, by counting the periods of high and low states and calculating the ratio between these two values.

C. Envelope detector

The envelope detector circuit takes as the input the RF signal with the ASK modulation and generates digital signal with PIE encoding (PWM signal) as the output. The circuit consists of two main blocks: the rectifier circuit and the comparator with the circuit for the referent voltage.

The basic rectifier cell architecture is used for providing DC power and demodulation. The circuit of the voltage doubler cell is shown in Fig 5.

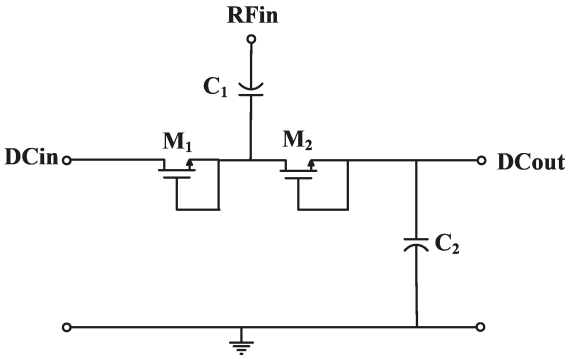


Figure 5. Voltage Doubler Cell.

This architecture is suitable for RFID applications due to its simplicity and re-usability in form of multiplication configuration (connection in series) for achieving larger output voltages.

In an RF signal period when the signal is negative and the *DCin* input is grounded, *M1* NMOS transistor connected as a diode is turned on and the capacitor *C1* is charged to (V_{pk} - V_{on}), where V_{pk} is peak RF amplitude and V_{on} is the drain source voltage of the transistor in on state. During this period the device *M2* is in off state and *C2* in disconnected from the RF signal.

In the positive signal period the device *M1* turns off while *M2* turns on. The capacitor *C2* is charged through the path *RFin*, capacitor *C1* and device *M2* (in on state), to the voltage defined in (1)

$$V_{C2} = V_{pk} + (V_{pk} - V_{on}) - V_{on} = 2*(V_{pk}-V_{on}). \quad (1)$$

Equation (1) is ideal, without taking into count RC time constant of equivalent resistance of diode-connected-devices and capacitors, effecting charging and discharging characteristics on the RF carrier frequency.

The basic voltage doubler cell can be connected in series by connecting the output *DCout* to the input *DCin* of the following stage. In that way the RF signal having the low amplitudes, by multiplication of the stages, produces the output with the voltage levels suitable for further analogue signal processing.

The basic voltage doubler cell is instantiated twice and in this configuration represents the low level cell in the hierarchy of the envelope detector module. The schematic is shown in Fig. 6.

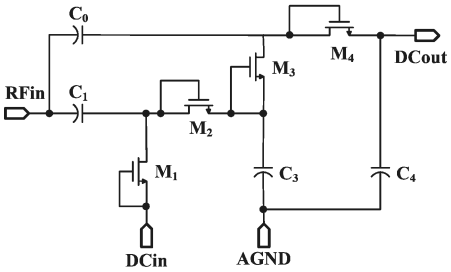


Figure 6. Envelope Detector Cell.

The cell is connected in series three times and the output of the last cell in the series is connected to the diode connected NMOS load, at the pin *ENV_ANA_OUT*, shown in Fig. 7.

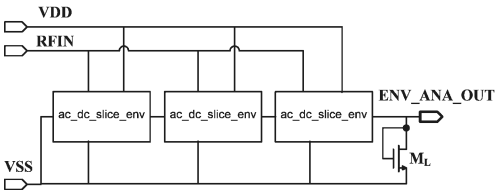


Figure 7. Envelope Detector Module.

The design challenge is to the optimize the design of the voltage basic voltage doubler cell and the NMOS diode connected load M_L to provide acceptable characteristics of these two parameters: timing of the PWM output signal (regarding rise and fall times) and the acceptable voltage level of the signal envelope to be used as the input in the following comparator stage.

The design trade-offs are needed in order to get optimal characteristics of the following requirements. RF signal amplitude at the input of the envelope detector restricted the voltage levels at the output. If higher capacitor values are used with the smaller amplitude of RF signal, higher voltages are achieved at the output. Higher capacitor values demanded NMOS devices with lower equivalent resistance in the voltage doubler cell to keep the timing constant inside acceptable values restricts by the protocol regarding rise and fall times. Communication from the reader to the tag can operate at *"Tari"* values down to 6.25 μs which is a limiting factor for the speed of the envelope detection. The rise and fall times are especially dependent on the dimensions of the NMOS connected load M_L at the output of the envelope detector module. When the NMOS load has lower equivalent resistance (higher width to length (W/L) ratios) to suit rise and fall time requirements, it also lowers the voltage level at the output (in the high envelope state) which limits the useful voltage range of the comparator stage.

Besides all this, the technology corners and temperature range significantly influences the circuit characteristics.

Extensive SPICE simulations are run across technology corners and temperature range from 0ºC to 85ºC and different levels of RF signal amplitudes from 250 mV to 450 mV. The circuit exhibits slowest response on slow corners and low temperatures and this is the worst case for the operation of the envelope detector.

The last stage before the digital control is the comparator stage which produces digital signal from the demodulated signal envelope.

The comparator stage consists of the comparator (Fig. 8) and the circuit for the referent voltage. Since the voltage and current reference block provides regulated power supply for the analogue and the digital circuits, the comparator stage was designed with defined supply voltage range and with the low power consumption as the priority demand. The reference voltage for the comparator is implemented as a simple voltage divider with diode-connected isolated NMOS devices with the bulk connected on the device source for each device separately. This eliminates body effect [4] for the devices and provides more precise voltage division.

The comparator uses a simple architecture of a differential pair and a buffer at the output [5]. It uses bias current of 200 nA. The comparator doesn't need to detect small differences of the input signal, so the offset is not an issue and differential pair transistors are relatively small devices. The input range of the signal is suitable for the NMOS type devices. The output buffer provides additional gain and driving strength for routing to the digital control logic.

The comparator in this application works with the signal transition at its input in the order of microseconds when the signal changes state. It results in the propagation time of the comparator of few hundred nanoseconds. The propagation time is in this case sufficient for the application. Digital logic, which counts the periods of the high state and the low state of the signal, is running on the system clock of 2 MHz, not synchronized with the demodulated signal coming from the envelope detection module. Therefore the transition of the comparator output signal can be detected in the worst case with the delay of almost 500 ns if the comparator changes state just after the clock transition. Tolerances defined by the EPC Gen2 protocol for the *"Tari"* value and the respective intervals of the high and the low signal state are loose, so the inaccuracy in the detection of the high and the low signal periods does not influence functionality.

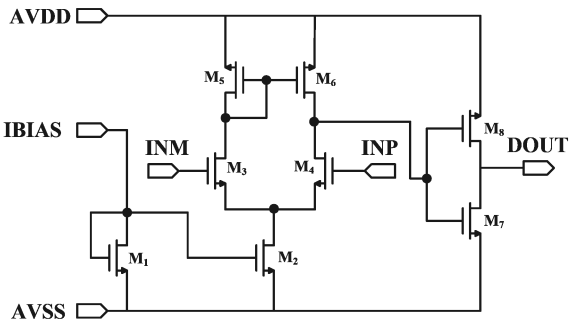


Figure 8. Comparator.

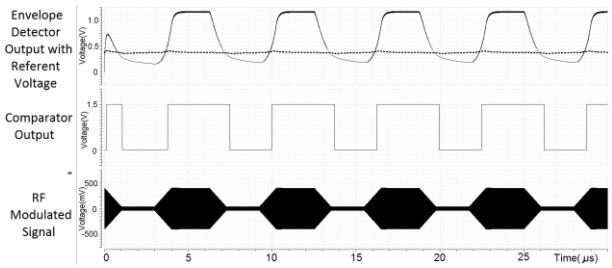


Figure 9. Envelope Detector Signal Waveforms.

The simulation results of the envelope detector circuit are shown in Fig. 9. RF signal is modulated with 400 mV amplitude in high state and 40 mV in the low state with the rise and fall time in accordance to the Gen2 standard. The analogue signal output of the envelope detector is displayed together with the referent voltage in the first waveform row. The referent voltage is around 400 mV, while the analogue signal spans from 100 mV for the low state to 1.2 V for the high state. The comparator output is in the second row. This digital signal is fed directly to the digital control logic for signal processing. The RF modulated signal is shown in the third row.

Simulations are run with the amplitude range of RF signal from 275 mV to 450 mV. The temperature range is from 0º C to 85º C and the technology process corners is used for slow, typical and fast case. The supply voltage is 1.5V. The results obtained by the simulations using these parameters fall into the allowed values defined by the protocol.

D. Layout

The layout was designed with the special care to the interferences that could experience analogue blocks from the RF. Besides that the analogue signals and digital signals both are present on the single chip. Each block is surrounded by N+ and P+ guard rings and layout is floor-planned and blocks are placed keeping in mind interferences that come from RF signal path and digital logic [6].

Fig. 10 shows the layout of the RF and analogue blocks arranged in a rectangular block with the dimensions approx. 550 um x 700 um.

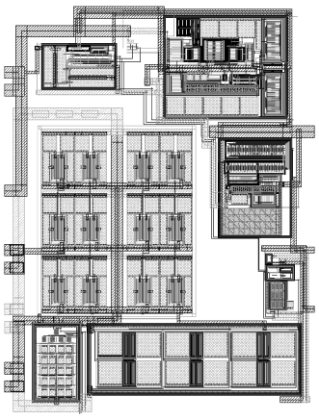


Figure 10. Layout of RF and Analogue Blocks.

E. Digital control

The digital control supports the EPC Gen2 protocol requirements for the communication with standard RFID readers. Protocol implements many features, but for the purpose of this application only a subset of commands is chosen to be included in the digital control. These commands and procedures from the IC side define a set of communication procedures with the reader that includes exchange of data regarding chip identification number and the data reading from the ADC. The minimal protocol subset is implemented to have low power consumption in the limited power budget.

Digital control is designed in VHDL, simulated and synthesized in Cadence tools. Digital control system is designed hierarchically, divided into seven major blocks, shown in the Fig. 11.

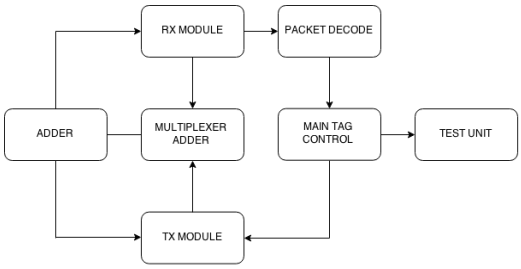


Figure 11. Digital Control Hierarchy.

The digital logic operates on the externally provided system clock of 2 MHz accompanied with the asynchronous reset signal for resetting the digital logic.

The input signal to the digital logic is the PWM signal coming from the envelope detector block.

Output signals are the digital signal for the modulation of the antenna load (backscattering) and some test signals for monitoring the state of the logic and for the debugging purpose.

The main Tag Control block is the block that synchronizes and controls the operation of the complete digital control. It places the tag in the receive mode or the transmit mode and controls the state of the tag (*Ready*, *Arbitrate*, *Reply* and *Acknowledged* states). Response to the protocol specific set of command and procedures (*"Inventory"*) allow sending of the required data over the communication channel.

The RX Module is in charge of receiving data coming in the form of the PWM signal and decoding it to the raw digital information.

The Packet Decode module decodes the information from the RX Module and forwards it to the Main Tag Control. CRC5 cyclic redundancy check and the generation of the pseudo-random digital word is a part of the block functionality.

The Pseudo-random number generation is implemented as 16-bit-LFSR (Linear Feedback Shift Register) [7]. The coefficients for the feedback are chosen to generate long sequences without repetition. Additional randomness is assured by feeding LSB bit in LFSR with the LSB bit of the counter in RX module. Since each chip has its own clock uncorrelated to the other chips and clocks, the LSB bit in each chip's counter is also random.

The TX module implements FM0 baseband modulation for the communication from the chip (tag) to the reader. It also includes FIFO buffer and CRC16 cyclic redundancy check as the part of the protocol requirements.

The Adder block implements 12-bit adder with the command sequences performing adding and subtracting using second complement arithmetic [8].

The Multiplexer Adder block is used for sharing resource of Adder block among other blocks. The architecture consisting of a multiplexer and one adder is chosen to minimize digital logic and have low power consumption. It is feasible because blocks and their corresponding operations are mainly done in sequence.

The Test Unit block is added for testing and debugging purposes. By setting two digital inputs it is possible to monitor 4-bit signals which corresponds to the states of internal blocks' state machines. Two additional test output signals monitor bit by bit information in the receive and the transmit mode.

F. Digital Logic Implementation Details

Logical and physical synthesis are performed on the VHDL code for the digital logic. The statistics on the final synthesized layout is summarized in Table I.

TABLE I. DIGITAL LOGIC STATISTICS

Gate count	Number	Area[a]	Overall Area[a] (with routing and power rails)
Sequential	493	0.032	0.112
Combinational	1520	0.022	

a. In mm^2

III. PROTOTYPE

The designed chip was sent for processing in UMC L180 Mixed-Mode/RF technology as Europractice mini@sic MPW run. The chip will be packaged in PLCC 84 pin package with test pins that will be used for the measurement purposes.

Beside analogue pins for testing, all input and output digital pins will be available for measurement and monitoring purposes. Beside the measurement on the synthesized logic, there is another option envisaged for the measurement and debugging. The external FPGA board can be connected to the input and output pins and used for synthesizing updated version of VHDL for the digital logic. In this way new version of the digital logic could be tested combined with the actual circuits and responses from the analogue blocks.

Measurements will be done with the standard RFID reader/development board with the low level communication data for the logging and the debugging.

IV. CONLUSION

This paper presents concept for intelligent wireless sensor networks based on the UHF EPC Gen2 RFID protocol. Widely used protocol, with many standard RFID readers available on the market gives a good ground for such a concept. The proposed compact chip solution, with a limited subset of communication protocol commands and the simple RF receiver and transmitter architecture, is an interesting solution for the low cost sensor networks without the battery supply. The simulation results of RF communication path showed that the concept is feasible. Measurements of the processed chip using standard RFID readers providing power and communication will show how the simulation results correspond to the chip functionality in the real environment. Built-in test features for both analogue blocks and digital control will allow detailed measurements, post analysis and debugging.

REFERENCES

[1] "EPC™ Radio-Frequency Identify Protocols Generation-2 UHF RFID, Specification for RFID Air Interface," EPCglobal Inc, 2013, version 2.0.0

[2] Daniel M.Dobkin, "The RF in RFID Passive UHF RFID in Practice," Elsevier Inc, USA, 2008.

[3] Ahmed Ashry, Khaled Sharaf, Magdi Ibrahim, "A Compact Low-power UHF RFID tag," Elsevier Microelectronics Journal 40, March 2009, pp. 1504–1513.

[4] Behzad Razavi, "Design of Analog CMOS Integrated Circuits," Mcgraw-Hill, 2001, pp. 23-26.

[5] R. Jacob Baker, "CMOS Circuit Design, Layout and Simulation," IEEE Press, USA 2005, pp. 909–918.

[6] Alan Hastings, "The Art of Analogue Layout," Prentice-Hall, Inc, 2001.

[7] M.Sahithi et al."Implementation of Random Number Generator Using LFSR for High Secured Multi Purpose Applications," International Jornal of Computer Science and Information Technologies, Vol 3(1), 2012.

[8] Volnei A. Pedroni, "Circuit Design with VHDL," MIT Press, 2004, pp. 187-225.

Temperature dependence of injection-locked Fabry-Pérot laser emission in WDM-PON architectures

Marko Šprem, Dubravko Babić, Marko Bosiljevac, Zvonimir Šipuš
Applied Optics Laboratory, Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3 / XII, HR-10000 Zagreb, Croatia
e-mail: marko.sprem@fer.hr

Abstract - We investigate the influence of temperature on multimode Fabry-Pérot laser output in a proposed colorless WDM system with self-seeding via modulation averaging reflector and with BLS-seeding. In particular, we investigate the connection between the information about the optical laser output from the laser integrated monitor photodetector and optical power output of the entire system with temperature change of the laser diode.

I. INTRODUCTION

Self-seeded colorless wavelength-division-multiplexed passive optical networks (WDM-PON) may give rise to access networks with the lowest cost bidirectional link per Gbps per user [1-4]. They have been investigated as a part of New Generation PON phase 2 development efforts on networks that are independent of present PON solutions. However, high transmitter cost still limits the widespread adoption of self-seeded WDM-PON and hence there is an opportunity and a market need for cost-effective colorless WDM transmitters. This work is directed towards creating an adaptable low-cost multi-wavelength transmitter for WDM-PON.

Self-seeding WDM-PON architecture comprises an extended resonator with an optical gain and modulation source at the client end and reflector and a wavelength-selecting filter in the remote node. This cavity can be many kilometers long. The filter in the remote node is a channel in an athermal array-waveguide grating (AAWG) with any number of channels tuned to the dense WDM ITU grid (AAWG with 16, 32, 40, 80, 96 channels are available today). Filtering the emission using many channels of the AAWG is also referred to as spectrum slicing. The number of channels and the corresponding channel separation (200, 100, 50, and 25 GHz available today) are adjusted to efficiently fill the communications window of the modern optical fiber. Presently, the target data rate per wavelength for spectrum-sliced systems is 10 Gbps at channel separations 100 GHz and smaller.

The filtered amplified spontaneous emission light emitted by self-seeded WDM-PON transmitters exhibits a fundamental noise and rate limits owing to the statistical properties of thermal light it emits [5], but up to date transmission rates over 10 Gbps have been reported in self-seeded systems [6], hence target is achievable.

Among the disadvantages of the self-seeded network architecture is intensity variation of the optical wave returning to the gain element (due to the presence of modulation in the seeding light) and fragile link stability. The first problem reduces the link margin but can be reduced by saturating the gain element (laser or reflective semiconductor optical amplifier) [7] or by using modulation averaging technique [8]. The gain elements are generally very polarization dependent, while the birefringence in the fiber and other optical elements degrade, add noise and can destabilize the link. Stabilization of self-seeded systems has been one of its difficult problems. Time reversal of the accrued polarization is one approach [9], while depolarizing the beam and stabilizing modulation using modulation-averaging reflectors (MAR) located in the remote node is another [8][10] (used in this work). The transmitting part of a self-seeded WDM-PON system is illustrated in Fig. 1 when a MAR is connected at the output of the AAWG (1). Light oscillates between the FP-LD and the MAR while it is being filtered with the AAWG and coupled out (as P4) of the cavity using a coupler.

Self-seeding is an alternative approach to the previous generation WDM-PON architectures which employ a broadband light source (BLS) to provide seeding light to identical transceivers connected to the WDM ports on the AAWG in the remote node. The BLS is typically located in the central office. Fig. 1, possibility (2), shows a simplified configuration of a transmitter using BLS to seed the gain elements. The light emitted from the BLS is filtered by the AAWG (also referred to as spectral slicing) and brought to the client end where it is used to injection lock the FP-LD or reflective semiconductor optical amplifier [11].

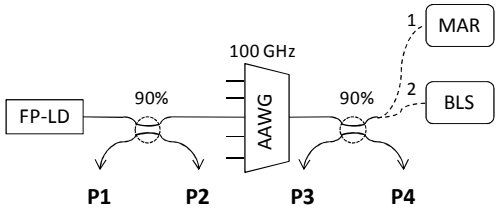


Figure 1. Measurement setup with a MAR or BLS seeding

The longitudinal modes of FP-LDs shift with temperature around 0.1 nm/°C or less. When using uncooled FP-LDs for injection locking, we strive to ensure that the seeding light matches the wavelength of one of the FP-LD resonant modes. If the wavelengths did not match, for example with temperature change of FP-LD, the output power of the FP-LD could not be sufficient for successful transmission. One way for eliminating this risk is to use AWG with channel frequency width wider than the frequency mode spacing of FP-LD, so that at least two mode wavelengths fit within the channel bandwidth. In this way, the seeding light always excites at least one mode regardless of the temperature of the FP-LD. A problem occurs when using AAWGs with channel spacing that is smaller than the longitudinal mode separation in the FP-LD. However, manufacturing of long FP-LDs is problematic because of higher cost. For 100 GHz wavelength separation, the FP-LD length has to approach 1 mm.

In this paper we present the temperature dependence of the light-emission characteristics of a short-cavity FP-LD as groundwork for temperature tuning of FP-LD performance in WDM-PON systems. We use a FP-LD extracted from a 16-channel 200-GHz BLS-seeded WDM-PON system and we are applying it to a denser 100 GHz athermal AWG by self-seeding and the original BLS.

II. Experimental Setup

The measurement setup, shown in Fig. 1, is an extended cavity comprising of an injection-locked Fabry-Perot laser diode (FP-LD), a 32-channel athermal array-waveguide grating (AAWG) with 100-GHz channel spacing, and either (1) a modulation-averaging reflector (MAR) or (2) broad-band light source (BLS) terminating the extended cavity on the other end. The elements are coupled with standard singlemode fiber and optical couplers were introduced between the FP and the AAWG, and between the AAWG and the cavity termination. By connecting the MAR (1) at the end of the cavity, the test setup forms a self-seeded multi-wavelength optical source, while connecting the BLS (2), the test setup forms a BLS-seeded multi-wavelength optical source. The test setup is designed with the intention to mimic a real-life WDM-PON system with the exception of the extended fiber lengths: distribution fiber between the client end and the remote node, and the feeder fiber between the remote node and the central-office. The fiber segments in our experiment have lengths equal to the patch cords of the optical couplers used. The couplers provide sampling of the optical power and spectra incident on the FP-LD (P1), emitted by the FP-LD (P2), reflected by the MAR or emitted by the BLS (P3), and emitted by the multi-wavelength optical source (P4).

A microcontroller was used to set the laser drive current I_{LD}, the power to the heater (10-Ω resistor), to monitor the integrated monitor photo diode (IMPD) current I_{PD}, and read the temperature T using the LM35 thermometer chip (Fig. 2). The optical powers (P1-P4) were measured using a table-top optical power meter with analog output which was also fed to the microcontroller. In this way, all the experiments, data acquisition and export were conducted using the microcontroller. The

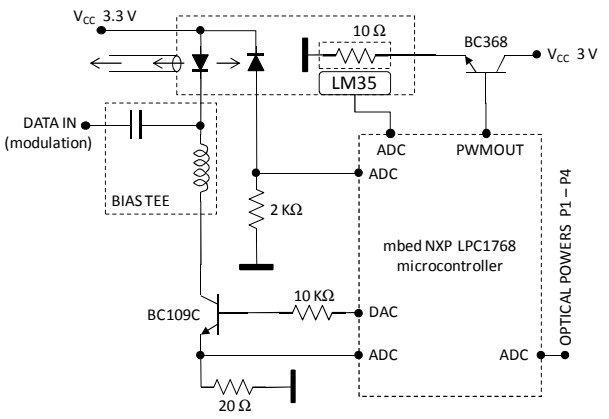


Figure 2. FP-LD driver scheme. Driver current (I_{LD}) was driven by the DAC port and monitored as voltage on a 20 kΩ resistor by the ADC port. The photodetector current (I_{PD}), temperature from LM35 temperature sensor and optical detectors (P1 - P4) were all sensed by the available ADC ports on the microcontroller. The heater (10Ω) was driven by pulse width modulated output. DATA IN port was not used.

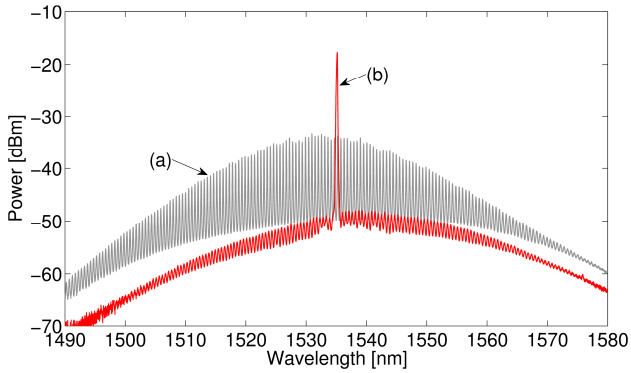


Figure 3. Emission spectrum measured at P2 when (a) neither MAR or BLS are connected on the other end of the cavity ("free-running emission") and (b) the emission spectrum measured when MAR is connected, namely, the FP-LD is injection-locked. The laser drive current in both cases was I_{LD}=35 mA.

advantage of this approach was its simplicity and it was a preliminary test for integrating the required functionality into a single fiber-optic module that would contain the FP-LD and a microcontroller.

The FP-LD was extracted from an optical network terminal (ONT) of a commercial BLS-seeded WDM-PON system (ADC PONy Express 16 DWDM PON). The laser is packaged in a Transmitter Optical Subassembly (TOSA) along with an integrated monitor photodiode (IMPD) and mounted on a temperature-controlled stage. The FP-LD laser came with a ~2% front-facet coating and about 60 GHz mode separation. Fig. 3(a) shows the free-running (no optical feedback) emission spectra of the FP-LD. When no optical feedback is provided, the laser does not reach threshold within the operating current range planned by the control circuit, hence the free-running spectra (Fig. 3(a)) never shows any dominant wavelengths. When the external cavity is terminated with a MAR, the AAWG provides optical feedback peaking at only one wavelength thus injection locking the FP-LD. This is manifested with a strong and narrow emission shown in Fig. 3(b) and reduced laser gain (shallower resonance fringes in the emission spectrum). Channel 3 of

AAWG with wavelength 1535.1 nm was used for all experiments.

For all experiments, FP-LD was operated in continuous-wave (CW) mode, without any modulation signal on data in port (Fig. 2). In all of the experiments a feedback control loop was implemented on the microprocessor to adjust the laser drive current (I_{LD}) in order to maintain either constant photodetector current (I_{PD}) or optical powers (P2 or P4) as we vary the FP-LD temperature. We refer to these measurement modes as *constant I_{PD} mode*, *constant P1, or P2 modes*. The effective coupling of optical energy from the cavity formed by the FP-LD and the AAWG (and possibly MAR) to the outside world (P4) is referred to as output coupling.

III. Results and Discussion

In the proposed injection locking schemes, a temperature change of the FP-LD should cause a change in output laser power (P2), photodetector current (I_{PD}), and the system output power (P4). Temperature increase causes a wavelength shift of longitudinal modes towards longer wavelengths. Since we are using an intra-cavity filter (channel 3 of the AAWG) it is clear that when any one of the FP-LD longitudinal modes coincide with the selected channel of the AAWG, the output power (P4) at given current should increase or conversely, if the output power (for example P4) was maintained constant, the laser current I_{LD} required to deliver the specified output power should reduce. Furthermore, inasmuch as the photodetector current I_{PD} is a measure of the optical power emitted from the laser, the I_{PD} can be used to determine when (as the temperature varies) the FP-LD mode coincides with the externally asserted filter center wavelength. For example, while in self-seeding mode, using the feedback loop to adjust the I_{LD} to maintain I_{PD} constant, should show a minimum in the I_{LD} when the extended cavity has lowest loss, namely, when the FP-LD mode coincides with the AAWG channel. The experiments reported in this paper show that this simple interpretation is generally true, but that the relationship between the photodetector current (I_{PD}), laser drive current (I_{LD}) and the outputs of the optical system (P1, P2, and P4) is a bit more complex and not entirely understood. For all measurements shown, we took data with heating and cooling temperature-scan directions. We observed a small (<1°) shift in temperature between the data for those two scans, while the relationships between currents (I_{LD}, I_{PD}) and optical outputs (P1, P2, and P4) remained unchanged. Hence we show only results for the cooling scans.

Our first group of experiments (sections A, B and C) uses a self-seeded configuration illustrated with Fig. 1 and connection 1 (MAR). In sections D and E, we characterize the same relationship by using a BLS (Fig. 1, connection 2 (BLS)).

A. *MAR self-seeding: Constant I_{PD}*

In this measurement, we varied the laser temperature from 43°C to 24°C while we kept I_{PD} constant and adjusted I_{LD} accordingly. We monitor the laser emission spectrum at P2 and note the output powers from P1, P2, and P4. The spectra shows that wavelengths of multiple modes of the FP-LD all drift towards longer wavelengths (Fig. 4), but that there is substantial mode-pulling when any of the FP-LD modes come close to the externally asserted low loss wavelength. This is visible by the peak wavelength jumping to or away from the constant value of 1535.1 nm.

Fig. 5 shows the I_{PD}, I_{LD}, P1, P2, and P4 normalized and plotted on the same graph. As the temperature increases and longitudinal modes of the laser go in and out of resonance with the externally asserted AAWG mode, the drive current oscillates and grows on the average. The I_{LD} average current increase is a result of laser gain reduction with temperature. When at some temperature the laser mode matches the AAWG channel, this means that the losses of the extended resonator (from FP-LD to MAR) are at minimum. Since lower resonator losses mean lower threshold gain, we expect the laser current to be at a minimum as well. The optical power P2 oscillates synchronously with the laser current I_{LD} and exhibits extrema that coincide with the extrema in the I_{LD} current. This is also expected as more current through the gain forces the laser to emit more spontaneous emission.

The temperature dependence of the emitted power P4 and the power returning to seed the FP-LD also exhibits an oscillation versus temperature. As temperature effectively controls the output coupler of this extended resonator, one would expect the minima in the laser-emitted power P2 to coincide with the maxima of the emitted and seeding powers P4 and P1, respectively. The variation in the seeding power P1 is proportional to P4 since P1 is effectively a filtered and reflected version of P4. However, surprisingly, the P4 extrema do not coincide with the extrema of P2: There is a small, but noticeable temperature offset between the P4 maximum and P2 minimum (shown in Fig. 5) where ideally there should be none. We observe a temperature offset between 1.2°C and 1.6°C. Optical power P3 was observed but not shown in the results, because it was the same as power P1 with small power difference due to the AWG channel attenuation.

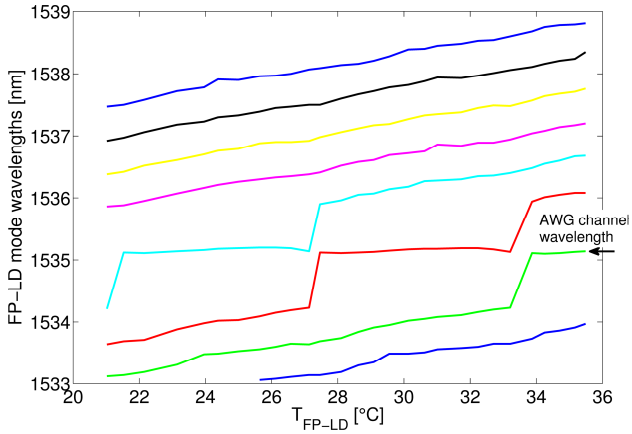


Figure 4. (MAR) Plot of the wavelengths of distinguishable local maxima in the emission spectrum P2 of the injection locked FP-LD as temperature varies.

B. *MAR self-seeding: Constant P2*

In this measurement, the I_{LD} current was driven to maintain constant power P2. The results, shown in Fig. 6 show behavior consistent with what was observed in Constant I_{PD} measurement previously: P2 is proportional to I_{LD}, except for the growth in average I_{LD} with temperature, and a small offset between the temperatures of maximum P4 and I_{PD}. Relative to temperature change, the change of I_{LD} current seems to be almost linear, with very small variations. Compared with results in Fig. 5, it can be confirmed that variations in laser power P2, with temperature change, depend mostly on variations of I_{LD} current.

C. *MAR self-seeding: Constant P1*

In this experiment, the I_{LD} current was driven to maintain constant power P1. The results show different behavior from what was observed in Constant P2 measurement previously: I_{PD}, I_{LD} and P4 all seem to be proportional with one another, except for the growth in average I_{LD} with temperature (Fig. 7). The temperature offset between maxima is very small. Relative to P4, the I_{LD} is in offset by ~ 0.2°C, and the I_{PD} is by ~ - 0.2°C. Compared with results in Fig. 6, the slope of the curves is to the right.

D. *BLS seeding: Constant I_{PD}*

In this method, the BLS was connected to the terminating end of the extended resonator as shown with connection (2) in Fig. 1. The seeding power was kept constant at P1= -21.6 dBm. The results of driving the laser current I_{LD} for constant I_{PD} current show similar variations in observed powers and currents, but without any temperature offset (Fig. 8). Note that the maximum in the optical power P4 coincides with the minimum in optical power P2 because when the laser mode matches the externally asserted mode (AAWG) the output coupling (from laser to P4) is lowest.

E. *BLS seeding: Constant P2*

In this experiment, the I_{LD} current was driven to maintain constant power P2. The results are shown in Fig. 9. It can be seen that the same almost linear variation of I_{LD} current was also observed as in for the self-seeded case shown in Fig. 6. The results are consistent with the constant I_{PD} measurements. Relative to the self-seeding

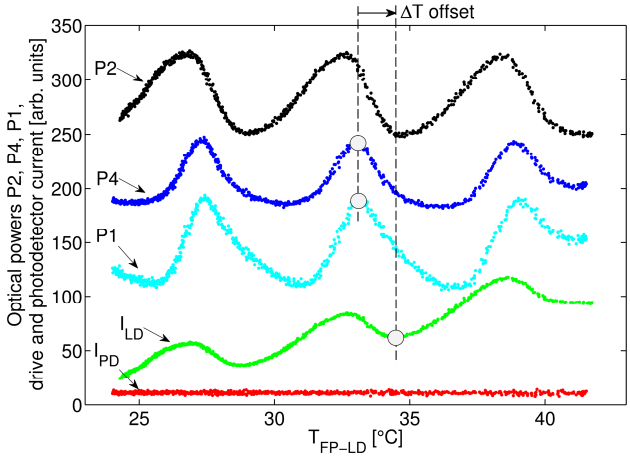


Figure 5. (MAR) Driving I_{LD} current for constant I_{PD} current. All values are plotted with linear units; P2≈[-12.94, -11.66] dBm; P4≈[-15.95, -14.67] dBm; P1≈[-21.73, -24.28] dBm; I_{LD}≈[20, 30] mA; I_{PD}≈250 μA.

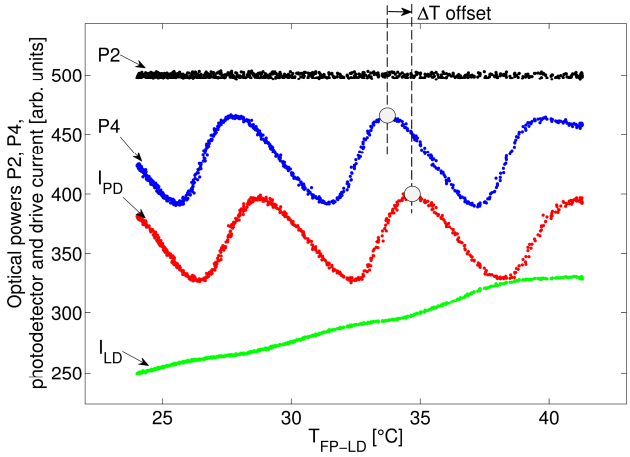


Figure 6. (MAR) Driving I_{LD} current for constant power P2. All values are plotted with linear units; P2≈ -9.5 dBm; P4≈[-14.56, -13.40] dBm; I_{PD}≈[300,370] μA; I_{LD}≈[25, 34] mA.

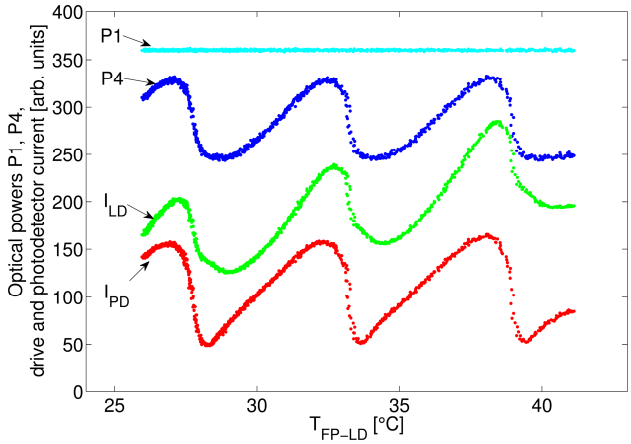


Figure 7. (MAR) Driving I_{LD} current for constant power P1. All values are plotted with linear units; P1≈ -24.40 dBm; P4≈[-15.89, -14.65] dBm; I_{LD}≈[23, 38] mA; I_{PD}≈[150, 265] μA.

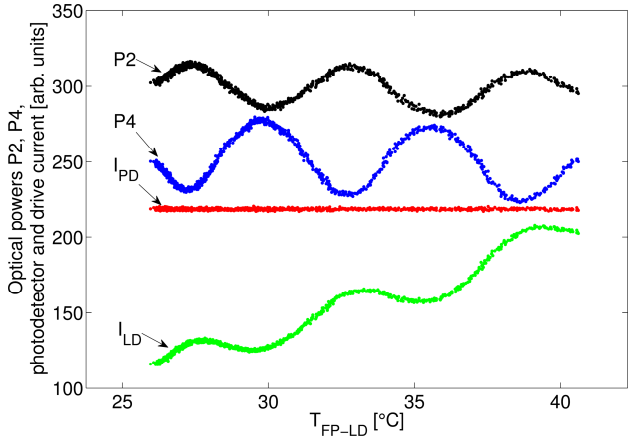


Figure 8. (BLS) Driving I_{LD} current for constant I_{PD} current. All values are plotted with linear units; P2≈[-9.28, -8.77] dBm; P4≈[-15.67, -14.65] dBm; I_{PD}≈250 μA; I_{LD}≈[27, 38] mA.

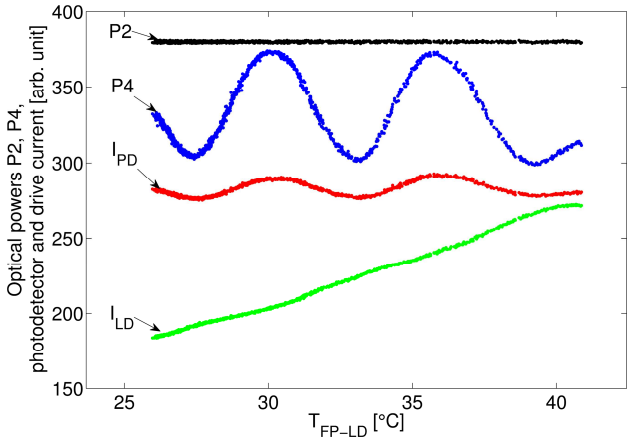


Figure 9. (BLS) Driving I_{LD} current for constant power P2. All values are plotted with linear units; P2≈ -8.99 dBm; P4≈[-15.71, -14.31] dBm; I_{PD}≈[240, 255] µA; I_{LD}≈[28, 37] mA.

results, the variations in I_{PD} current were significantly smaller.

IV. Conclusion

Injection locking of FP-LDs by either self-seeding or using a broadband-light-source, allows one to use the available optical and electrical variables at the client end (I_{PD}, I_{LD}, P1 and P2) to gain information about the interaction between the modes of the FP-LD and the AAWG at the remote node which is hundreds of thousands of meters away. The extended cavity formed by the FP-LD and the AAWG and MAR in the remote node is not coherent, but it exhibits a clear threshold behavior.

In this work, we presented a simple microprocessor-driven control system featuring feedback control loops for stabilizing the laser output depending on either monitor photodiode current, or any other optical power measured in the system. The developed algorithms will be implemented in a complete temperature and output optical power stabilizing system – the goal of this project.

The origin of the temperature offset between temperature-dependence of light emitted from the laser (P2) and the light emitted from the multi-wavelength optical source (P4) is not entirely understood. However, this effect occurs in a self-seeded system, but not in a BLS-seeded system which clearly indicates that the offset is related to the mode pulling and the mutual interaction of the laser with the AAWG mode.

Acknowledgment

This work was funded in part by HRZZ research project “EXCAR - Sustainable Optical Communication and Sensor Systems based on Extended-Cavity Resonators” IP-11-2013-3425.

References

[1] C.-H. Lee, W. V. Sorin, and B. Y. Kim, “Fiber to the home using a PON infrastructure,” J. Lightw. Technol., vol. 24, no. 12, pp. 4568–4583, 2006.

[2] C. F. Lam, “Passive Optical Networks: Principles and Practice“. New York: Academic, 2007.

[3] [B. Kim and B.-W. Kim, “WDM-PON development and deployment as a present optical access solution,” Conf. on Opt. Fiber Comm., San Diego, 2009.

[4] P. Chanclou, A. Cui, F. Geilhardt, H. Nakamura, and D. Nesset, “Network operator requirements for the next generation of optical access networks,” IEEE Netw. 26(2), 8–14, 2012.

[5] J. W. Goodman, “Statistical Optics,” New York: Wiley Classics Library, 2000.

[6] M.L. Deng, Y. Ling, X.F. Chen, R.P. Giddings, Y.H. Hong, X.W. Yi, K. Qiu, and J.M. Tang1, “Self-seeding-based 10Gb/s over 25km optical OFDM transmissions utilizing face-to-face dual-RSOAs at gain saturation,” *Opt. Express*, vol. 22, no. 10, p. 11954, 2014.

[7] A. D. McCoy, P. Horak, B. C. Thomsen, M. Ibsen, and D. J. Richardson, “Noise suppression of incoherent light using a gain-saturated SOA: implications for spectrum-sliced WDM systems,” *J. Lightwave Technol*. 23(8), 2399–2409, 2005.

[8] T. Komljenović, D. Babić, and Z. Šipuš, “Modulation-Averaging Reflectors for Extended-Cavity Optical Sources,” J. Lightwave Technol 29(15), 2249–2258 , 2011.

[9] L. Marazzi, P. Parolari, M. Brunero, A. Gatto, M. Martinelli, R. Brenot, S. Barbet, P. Galli, and G. Gavioli, “Up to 10.7-Gb/s High-PDG RSOA-Based Colorless Transmitter for WDM Networks,” *IEEE Photon. Technol. Lett.*, vol. 25, no. 7, p. 637, 2013.

[10] T. Komljenović, D. Babić, and Z. Šipuš, “C and L band Self-seeded WDM-PON Links using Injection-locked Fabry-Pérot Lasers and Modulation Averaging,” OFC, San Francisco, 2014.

[11] S.-G. Mun, J.-H. Moon, H.-K. Lee, J.-Y. Kim, and C.-H. Lee, “A WDM-PON with 40 Gb/s (32 x 1.25 Gb/s) capacity based on wavelength-locked Fabry-Perot laser diodes,” Opt. Express, 16, 15, 11361-11368, 2008.

Fiber-Optic Vibration Sensor for High-Power Electric Machines

Bojan Igrec[1], Marko Bosiljevac[2], Smiljko Rudan[1], Zvonimir Šipuš[2], Dubravko Babić[2]

[1]Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia.
e-mail: bojan.igrec@gmail.com, marko.bosiljevac@fer.hr

Abstract – We demonstrate a fiber-optic sensor for monitoring mechanical vibrations in high-power electric machines built using 3D-printing. We describe the design, mechanical and optical modeling, and the results of its preliminary characterization. The measurements show linear sensor response in the frequency range 10 to 150 Hz.

I. Introduction

The overwhelming majority of the world's electric power is created by turbine-driven electric generators, while almost half of the energy in the world is consumed by electric motors. Many of these motors and almost all of power generators operate continuously, uninterrupted, and are brought off-line only for planned inspections and maintenance. A catastrophic failure of large generators or motors is generally a serious safety, structural, and financial problem. These machines therefore have to be protected from accidental over-stressing and damage. Presently this protection makes use of remote monitoring of the performance using a variety of sensors (temperature, vibrations, and mechanical stress) along with predictive data analytics to minimize the damage, time, and financial loss. Monitoring of vibration amplitude or acceleration is an essential element of this protection, performance optimization, and prolonging machine's life.

One approach to building vibration sensors is by employing electro-mechanical technology where inertial force is converted to an electric signal via piezoelectric sensing or time-varying capacitive or inductive coupling [1]. Electro-mechanical sensors regularly require electrical power to be delivered to the sensor. In high-power electric machines this is a problem, as sensors are most often needed in places where very strong alternating electric and magnetic fields are present. This, in turn, results in large induced electromagnetic forces in the circuit of the sensor and high danger of voltage breakdown. In such environments, one prefers vibration sensors that have no conductive parts.

Fiber-optic sensor, on the other hand, can be made non-conductive and can convert mechanical vibration into a modulation of an optical signal that is generated and sensed remotely. The common approach employed in fiber-optic vibration sensors is that the optical fibers themselves are allowed to bend and vibrate depending on their own weight and moment of inertia [2] - [6]. Optical fibers are made out of quartz or polymer and are encapsulated within insulating materials, while light propagation through them is not affected by electric or magnetic fields. This makes them suitable for installation in areas where very strong fields are present. Additionally, due to the very low light-attenuation in present-day optical fibers, the location where the light is generated and sensed can be meters or kilometers away from the sensor. Therefore, fiber-optic vibration sensors are an ideal match for this environment as the only connections between the sensors and the outside world are optical fibers that are electrically insulating, flexible, and lightweight.

The objective of this work to develop a highly sensitive, lightweight, and inexpensive fiber-optic vibration sensor for high-power electric machines and similar applications. The target specifications are bandwidth 10 Hz to 150 Hz and acceleration range from 0.1 to 10g. In this paper, we report on the design and demonstrate the operation of a prototype fiber-optic vibration sensor made using 3D printing. The prototype is simple to assemble, adjust, and repair. We discuss the design and show results of frequency response and linearity characterization.

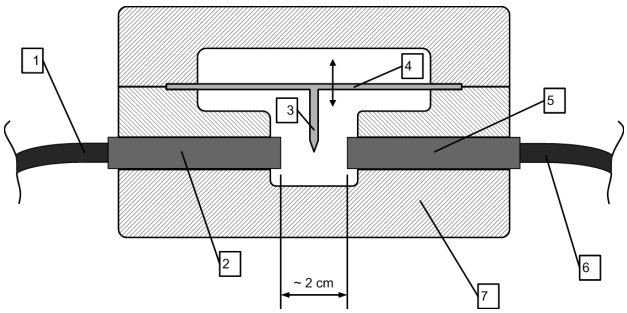


Figure 1 – Schematic of the fiber-optic vibration sensor developed and demonstrated in this work: (1) optical fiber, (2) graded-index collimating lens (source of light), (3) light-modulating blade, (4) bendable beam, (5) light-receiving graded-index collimating lens, (6) optical fiber, and (7) the housing.

II. Sensor Design

The working principle of our fiber-optic sensor is explained with the help of Figure 1. A multimode optical fiber (1) is terminated with a collimating graded-index (GRIN) lens (2) which emits a collimated beam of constant light intensity towards another identical GRIN lens (5) which focuses the beam back into another multimode fiber (6). The light beam is partially blocked using a blade (3) attached to a bendable beam (4) whose motion perpendicular to the optical axis of the beam modulates the intensity of the light reaching the GRIN

lens (5). The housing (7) is attached to a machine body and as the entire sensor vibrates in the *z*-direction, the blade moves relative to the optical beam and its motion is directly translated into intensity modulation of the light leaving with fiber (6).

The weight of the blade (3) and the beam (4) spring constant define the frequency response of the sensor and we desire flexibility in adjusting this response. In common fiber-optic vibration sensors the optical fibers themselves bend and vibrate depending on their own weight and moment of inertia and there one's primary design parameter is the length of the free-standing fiber (and a very limited range of fiber diameters and materials) [2]-[6]. In the structure shown in Figure 1, however, the dimensions and the material used in the bendable beam as well as the weight of the blade can all be adjusted without much restriction. In addition, the beam extending only halfway into the cavity and the blade attached at its end can be realized, while multiple sensors for different axes may be placed into the same housing enabling triaxial vibration measurements. In this work we demonstrate vibration measurement along only one axis.

The design of the sensor prototype was done iteratively considering available components, dimensions, and performance. We first selected GRIN lenses with beam diameter 1.8 mm. This defined the maximum blade excursion (i.e., beam deflection). We wanted the excursion of the blade at maximum acceleration (10g) not to exceed the width of the collimated beam. In this way we ensure that 100% light modulation will coincide with the maximum linear dynamic range of the sensor. In order to achieve the largest dynamic range, we desire that when there is no vibration, exactly one half the light passes through the sensor (the blade is at its central position). Under vibration, the largest excursion of the blade should be one half of the total (peak-to-peak) blade travel. If the blade was to travel outside the linear dynamic range, the light response would saturate introducing nonlinearity and higher harmonics into the response – an undesired feature which would obscure the fundamental intensity harmonic readout.

In order to accurately estimate the effect of the aperture excursion and its effect on the light transmission for our sensor, we modeled light transmission between the lenses for several practical lens-to-lens distances This analysis was performed analytically using ray tracing from an extended optical source (details are given in [7] and [8]). This excursion maximum was determined to be 1 mm. The next task was to design the mechanical part of the sensor that would operate in this range.

The analytic mechanical modeling of the sensor is based on a single beam suspended at two ends with a concentrated mass in its center. The ends of the beam are fixed and the central portion vibrates in direction perpendicular to the optical axis with amplitude $z_p(t)$, the blade attached to the center of the beam moves with amplitude $z(t)$. The solution for this problem is given by the following differential equation [9]-[11]:

$$\frac{d^2z}{dt^2} + 2\zeta\omega_0\frac{dz}{dt} + \omega_0^2 z = 2\zeta\omega_0\frac{dz_p}{dt} + \omega_0^2 z_p \qquad (1)$$

where z and z_p are the blade vertical position and housing vertical position, respectively. The two adjustable parameters in this linear model are the resonant frequency ω_0 and the damping factor ζ. The resonant frequency depends on the spring constant K:

$$\omega_0^2 = K/m \qquad K = 192EI/\ell^3 \qquad (2)$$

where E is the Young's modulus, I moment of the inertia along the beam, ℓ bendable length of the beam, and m the mass of the blade concentrated in the center of the beam. Here K [kg/s²] is the spring constant for vertical deflection of a bendable beam fixed at ends and force applied in the middle [12]. Modulation of light is proportional to the relative position between the blade and the housing, hence the relevant variable is $h = z - z_p$ and its Fourier transform, the transfer function $H(\omega) = h(\omega)/z''_p(\omega)$, which is defined as the optical modulation spectrum divided by the acceleration spectrum and is given by

$$H(\omega) = \frac{\omega_0^2}{\omega_0^2 + 2i\zeta\omega - \omega^2}. \qquad (3)$$

Inasmuch as the system has two poles, it is also clear that the step-response will be monotonic only when $\zeta \geq 1$. Our prototype design did not force this condition, but rather focused on demonstrating the technology.

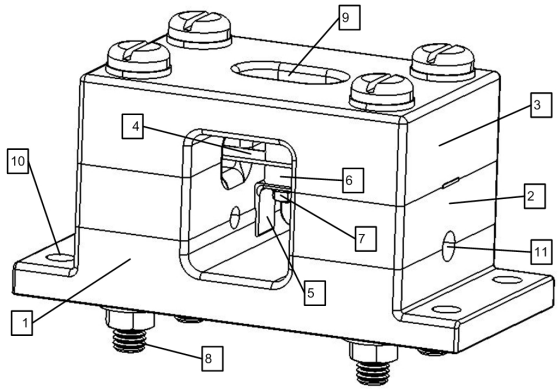


Figure 2 – Isometric view of the sensor

An isometric view of the designed sensor is shown in Figure 2. The housing comprised of three parts: the base plate (1), middle plate (2), and a cover plate (3). The GRIN lenses were clamped within a semi-cylindrical holes (11) and in between the base and the middle plates. The bendable beam (4) is clamped between the middle (2) and cover (3) plates. The plates were designed to allow the reception of beams (or membranes) of different lengths. The blade (5) was attached to a beam holder (6) using a nut/bolt (7). Provisions to adjust the weight (and tighten the nut) was made via hole (9) in the cover plate. In the construction of the housing we made an effort to reduce the number of sharp edges by making them round as much as possible. Four joints with metric thread M5 and related matrices were used to hold the plates together. In our prototype, the bolts were made out of steel, but nylon bolts were ultimately planned.

The dimensions of the beam and the weight of the blade (with the nut/bolt (7) combination) were adjusted to realize $\omega_0 > 150$ Hz. The damping ratio for the polymer to be used in the 3D printing was not known, so we used $\zeta = 0.025$, which is an average of the highest metal value ~0.01 and rubber ~0.05 [13]. The dimensions of the beam were thickness $b = 0.5$ mm, length of bendable section $\ell = 35$ mm, and width $a = 6$ mm. The weight of the beam and the blade were 106 mg and 334 mg, respectively, using the estimated specific weight of the polymer. We used finite-element method (commercial software package FEMAP/NASTRAN) to analyze the mechanical vibration of a weight suspended on a beam where the beam weight and the blade weight were treated as distributed. The computed resonance frequency was ~ 225 Hz.

We also estimated the boundaries of the resonance frequency analytically using equation (2). To this end we took into account that the center section of the beam was tightened with a bolt and two 5-mm wide flat shims, which resulted in an effective bendable length of the beam that was shorter than the physical length ($\ell' = 30$ mm) used in equation (2). Inasmuch as the beam mass is distributed along the beam, it is clear that its contribution to the resonance frequency (2) will be less than if it was concentrated in the center. Assuming that the spring constant was correct (888 kg/s²), using equation (2) we found that the resonance must be larger than 226 Hz, which is obtained with the mass of the blade and the beam assumed to be concentrated in the center. Also it should be smaller than 270 Hz, which is the value obtained if we set the beam mass to zero. These expected values satisfied the target bandwidth and we proceeded to manufacture the sensor.

The GRIN lenses were inserted firmly into the housing and tightened with the two plates. Due to varying tolerance of the housing we adjusted this by inserting a soft wrap around the lenses.

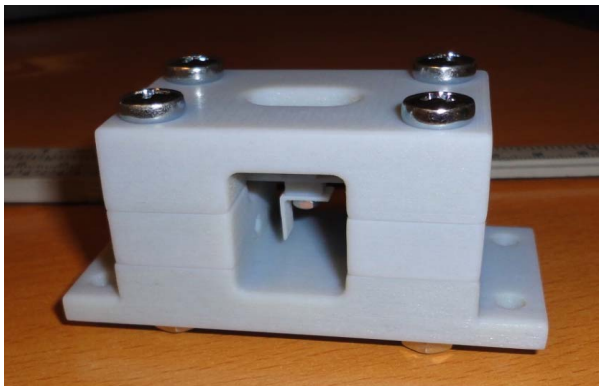


Figure 3 – Photograph of the complete sensor housing

An indentation in the middle and cover plates was made to hold the beam; it has to be held in place firmly with no ability to slip in any direction. The design allows membrane thicknesses between 0.2 and 0.6 mm. The membrane length can be varied ±6 mm. The blade was attached to the beam using an M2 bolt. Four ø5 thru holes were also added to the housing for mounting to the engine block.

The sensor prototype was manufactured using 3D printing using PolyJet Matrix technology on Stratasys CONNEX 350. We used DM 8515 Grey 35 Polymer. The housing and the beam (membrane) were built out of the same material. The density is ~ 1,180 kg/m³ and the Young's module of elasticity was estimated to be ~ 2 GPa, but no information was provided on the damping ratio ζ. The layer thickness was 16 μm.

III. Measurements and Results

The characterization of the sensor was performed by attaching the sensor to a commercially available vibrating table equipped with a powered oscillator and measuring the light transmission through the sensor. The optical fibers were multimode, the optical source a 1550 nm laser, and the receiver is a long-wavelength photodetector with built-in transimpedance amplifier. The acceleration of the vibration stage was monitored using a commercially available acceleration sensor (CAS) with bandwidth greater than 500 Hz [14]. A digital real-time oscilloscope was used to capture the waveforms and read out the data.

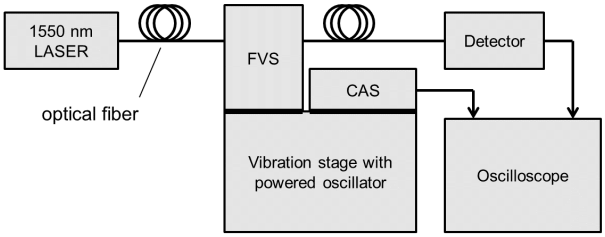


Figure 4 – Block diagram of the characterization setup: The fiber-optic vibration sensor (FVS) and the commercial acceleration sensor (CAS) are mounted on top of the vibration stage with powered oscillator whose amplitude and frequency is manually controlled.

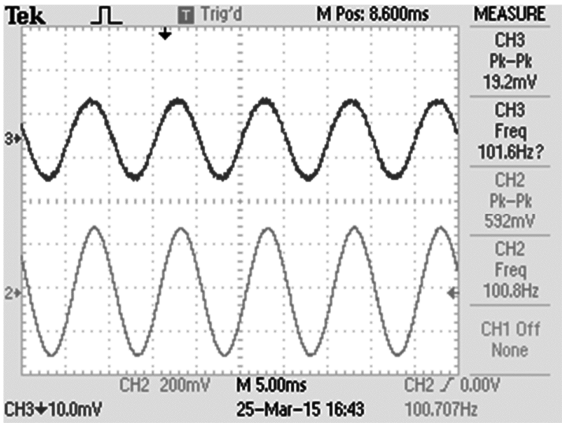


Figure 5 – Typical oscilloscope screen showing the vibrations measured on the fiber-optic vibration sensors (FVS, top curve) and the commercial acceleration sensor (CAS, bottom curve).

The measurement sequence for determining the sensor transfer function were performed as follows: A

frequency was set on the vibrating stage and the vibration magnitude was adjusted so that the commercial acceleration sensor showed 0.1 g for the stage. At this point a reading was made of the peak-to-peak amplitude from the oscilloscope screen (sample shown in Figure 5). The collected data were normalized so that the transfer function would show unity in the low frequency limit. Figure 6 shows the measurement of the transfer function $|H(\omega)|$ for several frequency points along with a fit to the transfer function shown in equation (3).

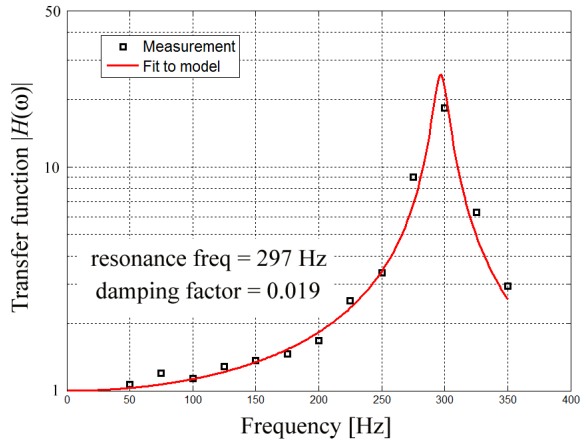


Figure 6 – Fitted transfer function using equation (3) with measured data for acceleration = 0.1 g. The best fitting parameters are listed.

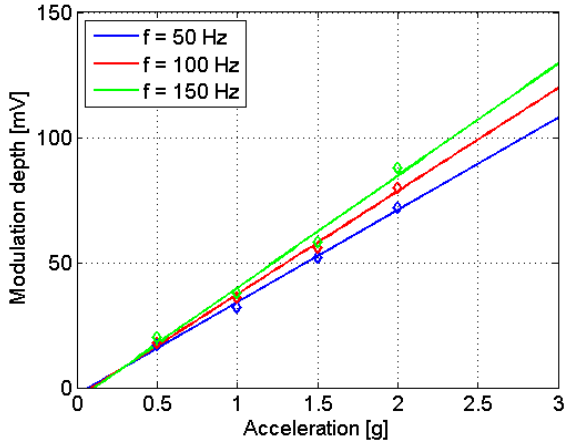


Figure 7 – Modulation depth measured as peak-to-peak voltage swing at the output of the detector versus acceleration for three different vibration frequencies. Maximum acceleration was limited to 3g due to the limit of our reference vibration sensor.

We used the data shown in Figure 6 to determine how close was our design to the sensor we built and to make corrections to the assumed values of the material and/or structure parameters. The fitting parameters for graph in Figure 6 were the resonance frequency (297 Hz) which is close to the values estimated by both numerical and analytical means. We also were able to estimate the damping factor for DM 8515 Grey 35 Polymer: $\zeta \approx 0.019$.

We checked sensor linearity by making measurements at 50, 100, and 150 Hz for different accelerations. The results, shown in Figure 7 show that the sensor is reasonably linear as one would expect.

IV. Conclusion

We demonstrated the design and the characterization of a novel fiber-optic vibration sensor and its straightforward conversion to a prototype by 3D printing.

The sensor shows expected bandwidth (> 150 Hz) and from our measurements, we estimated the damping ratio for DM 8515 Grey 35 Polymer to be $\zeta \approx 0.019$.

The sensor development is still on-going, with effort placed in adjusting the frequency response, acceleration sensitivity, and to minimize the size of the sensor to create a manufacturable prototype. The theoretical analysis shows that the frequency response will continue to exhibit a resonance, unless the damping factor of the system is increased to reach unity. This means some form of damping, such as, a viscous fluid or some other dissipative material, should be added to the structure.

Acknowledgment

The authors acknowledge the help of Center for Additive Technologies at the Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia and the BICRO Proof-of-Concept program (Croatia) for providing partial funding for this work.

References

[1] T. K. Gangopadhyay, “Prospects for Fibre Bragg Gratings and Fabry-Perot Interferometers in fibre-optic vibration sensing”, Fibre Optics Laboratory, Central Glass and Ceramic Research Institute, CSIR, Kolkata 700032, India, 9 January 2004

[2] T. Bosselmann, M. Villnow, M. Willsch, “Fiber-optic Vibration sensor”, United States, Patent Application Publication US 2011/0274386 A1, Nov. 10, 2011

[3] M. Villnow, M. Willsch, T. Bosselmann, "Novel fiber-optical accelerometer for monitoring tasks of power generators", SENSOR+TEST Conference 2011, paper B4.2 – SENSOR Proceedings, p. 277 (2011).

[4] C. Baldwin, J. Niemczuk, J. Kiddy, T. Salter: Review of Fiber Optic Accelerometers, Systems Planning & Analysis, Inc., Advanced Engineering Development Program, 4865 Walden Lane, Lanham, MD 20706.

[5] C. Doyle, G. F. Fernando, *Biaxial Fiber-Optic Accelerometers*, Engineering System Department, RMCS Cranfield University, Shrivenham, Swindon, SN6 8LA, UK

[6] M. Cloutier, J. Pronovost, M. Cloutier: Fiber optics accelerometer, United States Patent Application Publication, US 2007/0247613 A1, Oct. 25, 2007.

[7] B. E. A. Saleh, and M. C. Teich, *Fundamentals of Photonics,* John Wiley & Sons, 2007.

[8] B. Igrec, "Fiber-optic sensor for vibration measurement", Master’s Thesis at Faculty of Mechanical Engineering and Naval Architecture (in Croatian), 2015.

[9] D. J. Inman, *Engineering Vibration*, Virginia Polytechnic Institute and State University, Prentice Hall, Inc. 2001.

[10] S. S. Rao, *Mechanical Vibrations*, Fifth Edition, University of Miami, 2011.

[11] H. Bachmann, *Vibration Problems in Structures: Practical Guidelines*, Birkhauser Verlag, Berlin, 1995.

[12] E. Oberg, F. D. Jones, H. L. Horton, H. H. Ryffel: *27th Edition Machinery’s Handbook*, Industrial Press Inc., 2004., New York

[13] T. Irvine, *Damping properties of materials Revision C*, November 8, 2004

[14] Analog Devices ADXL335 Small, Low Power, 3-Axis ±3g Accelerometer Data Sheet, 2010.

Comparison of Electrical and Electromagnetic Properties of a Planar Inductor

Jurica Kundrata, Adrijan Barić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
E-mail: jurica.kundrata@fer.hr, adrijan.baric@fer.hr

***Abstract*— This paper compares the electromagnetic and electrical properties of a planar inductor used in a low-profile organic light-emitting diode (OLED) driver. A luminary consists of the OLED cell and its driver which uses a planar inductor. The proximity of the OLED cell and the constrained area makes the planar inductor lossy. The planar inductor represents a loop antenna and a main source of radiated emissions of the driver. The properties of the planar inductor are analysed w.r.t. the quality factor and the magnetic moment. The analysis uses a two-level experiment based on EM simulations to explore the geometrical parameters of the inductor. The experiment shows that the track width and the fill-ratio of the inductor have the largest effect on its properties. It is concluded that a planar inductor which introduces minimum power losses and radiated emissions to the OLED driver needs to be optimized w.r.t. the track width and the fill-ratio.**

***Keywords*—thick film inductor, electromagnetic modelling**

I. Introduction

Power converters utilize magnetic devices to store energy and low-profile applications use planar magnetic devices [1]. An example of a planar magnetic device which is commonly used in low-profile devices is the planar inductor [2]-[4].

Organic light-emitting diode (OLED) cells are planar, diffuse lighting devices and they form luminaries with a very low profile. The OLED luminary described in [5] drives the OLED cell using a buck converter. The buck converter utilizes a planar inductor which is considered lossy because of the proximity of the OLED cathode and the planar, distributed design. The planar inductor due to its inductive properties stabilizes the output current, while introducing power losses to the buck converter operation. Furthermore, the ripple of the output current of the buck converter causes the planar inductor to generate radiated emissions. The design of the planar inductor thus affects the operation of the buck converter and its radiated emissions.

The electrical properties of the planar inductor w.r.t. its geometrical design parameters is studied in [6]. A planar inductor on a printed circuit board (PCB) is used in a power converter and it is parametrized in [7]. The design of a planar inductor is optimized w.r.t. the power converter efficiency in [8] and a multiobjective optimization of the planar inductor is done in [9]. The electromagnetic compatibility (EMC) of the buck converters, i.e. its conducted emissions are studied in [10] and [11]. The radiated emissions of a power converter are considered w.r.t. the PCB layout in [12]. A buck converter which uses an air-core solenoid is characterized w.r.t. the magnetic field emissions in [13].

This work was supported by the European Commision under the Seventh Framework Programme (FP7) through project IMOLA „Intelligent light management for OLED on foil applications“ (Grant Agreement No. 288377).

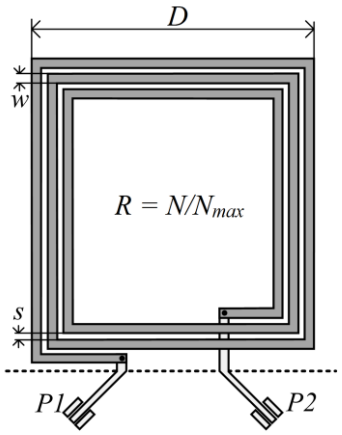


Figure 1. The design of the planar inductor and its parameters.

This paper compares the effects of the design parameters on the electrical properties of the planar inductor related to the buck converter and the radiated emissions of the planar inductor.

The analysis of the planar inductor is based on EM simulations described in Section II. The electrical properties of the planar inductor related to the buck converter are modelled in Section III, while Section IV presents the analysis of the radiated emissions of the planar inductor. The effects of the design parameters of the planar inductor on the electrical and electromagnetic properties are investigated using a screening experiment in Section V. The discussion of the experiment results is presented in Section VI, while Section VII contains the conclusions.

II. Electromagnetic Simulations

The planar inductor is a part of an OLED luminary which consists of the OLED cell and the backplane. The backplane contains the OLED driver, i.e. the buck converter that uses the planar inductor. The dimensions of the OLED cell limit the size of the backplane and thus the dimensions of the planar inductor.

This paper analyses a luminary which uses an OLED cell with dimensions 118x116 mm^2. The luminary is connected to a supply V_{IN} = 24 V and the buck converter supplies the OLED cell with the current I_{OUT} = 400 mA and the voltage V_{OUT} = 13.7 V. The buck converter operates at the switching frequency f_{SW} = 3.6 MHz.

Fig. 1 shows the layout of the planar inductor and its design parameters. The planar inductor has a square shape and its design is determined by the track width *w*, the track spacing *s*, the fill-ratio *R* and the outer dimension *D*. The fill-ratio *R* is the ratio of the number of turns *N* and the maximum number of turns which is determined by the pitch spacing *w+s* and the outer dimension *D*. The inductor structure has two ports *P1* and *P2*. The electrical properties of the planar inductor are simulated using COMSOL Multiphysics which result in two-port *S*-parameters [14].

The simulated substrate has two dielectric layers and three metal layers. The planar inductor is patterned on the 250 micron thick FR4 substrate and it uses two top metal layers which are made from 35 micron thick copper. The bottom metal layer represents the OLED cathode and the dielectric layer between the OLED cathode and the planar inductor represents the glass encapsulation.

The EM simulations analyse the frequency range 300 kHz to 300 MHz, which includes the switching frequency of the converter and the resonant behaviour of the inductor.

III. ELECTRICAL PROPERTIES

The electrical properties of the inductor are commonly represented by the quality factor *Q* which is defined at the frequency *f* as

$$Q(\mathrm{f}) = \frac{2\pi f L_S}{R_S} \tag{1}$$

where L_S is the series inductance and R_S is the series resistance of the inductor. The quality factor gives a ratio of the energy capacity of the inductor and its power losses.

The planar inductor stabilizes the output current of the buck converter, while introducing power losses in the conversion process. The stability of the output current is determined by the inductance, while the power losses are caused by the resistance of the planar inductor. The quality factor *Q* thus represents reasonable choice to represent the electrical characteristics of the planar inductor w.r.t. the buck converter.

The EM simulations of the planar inductor result in two-port *S*-parameters. The *S*-parameters are transformed to *y*-parameters which are used to define the interport, series admittance Y_3 by using

$$Y_3 = -\frac{y_{12} + y_{21}}{2}. \tag{2}$$

The series inductance L_S is extracted at the switching frequency by using

$$L_S = \mathrm{Im}\left[1/Y_3\right]/\left(2\pi f_{SW}\right) \tag{3}$$

and the series resistance R_S is extracted by using

$$R_S = \mathrm{Re}\left[1/Y_3\right]. \tag{4}$$

IV. ELECTROMAGNETIC PROPERTIES

The planar inductor consists of a number of turns and each turn acts as a loop antenna. The radiated emissions of the loop antenna are determined by its magnetic moment, i.e. the product of the loop area and current [15]. The magnetic moment *m* of the planar inductor is defined by the sum of the magnetic moments generated by the individual turns of the inductor which equals

$$m = \sum_{i=1}^{N} I_L A_{L,i} \tag{5}$$

where *N* is the number of turns of the planar inductor, I_L is the inductor current and $A_{L,i}$ is the area of the *i*-th inductor turn.

The current of the planar inductor I_L is determined by the ripple of the output current of the buck converter. The ripple current is related to the design parameters of the buck converter and the inductance of the planar inductor. The peak-to-peak value of the ripple current ΔI_{out} is defined as

$$\Delta I_{out} = \frac{V_{OUT}\left(V_{IN} - V_{OUT}\right)}{2\pi f_{SW} L_S V_{IN}}. \tag{6}$$

The magnetic moment of the planar inductor is derived by taking into account (6) and by determining the area of each inductor turn which leads to

$$m = \Delta I_{out} \sum_{i=1}^{N} \left[D - w - 2\left(i-1\right)\left(w+s\right)\right]^2. \tag{7}$$

V. GEOMETRICAL PARAMETER SCREENING

The geometrical design parameters of the planar inductor are screened by using a two-level full-factorial experiment [16]. Table 1 shows the levels of the design parameters used in the screening experiment. The planar inductor for each combination of the design parameters is simulated and the simulation results are used in determining the quality factor *Q* by using (1)-(4) and the magnetic moment *m* by using (6)-(7).

Fig. 2 shows the quality factor *Q* of the planar inductor w.r.t. the design parameters. Fig. 2 shows the relative value of the quality factor w.r.t. the two levels of the design

TABLE I
THE DESIGN PARAMETERS OF THE SCREENING EXPERIMENT.

Design parameter / Level	-1	+1
Track width *w* [mm]	0.2	0.5
Track spacing *s* [mm]	0.2	0.5
Fill-ratio *R* []	0.05	0.3
Outer dimension *D* [mm]	75	100

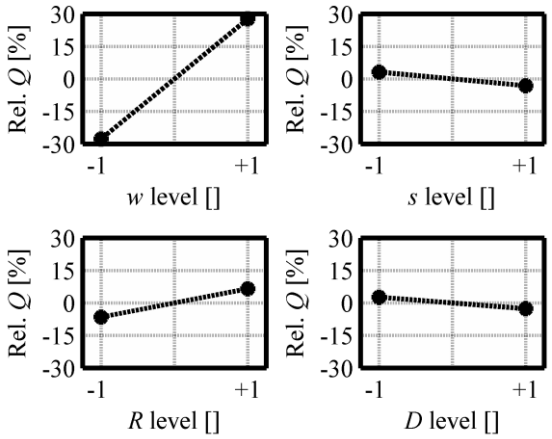


Figure 2. The quality factor Q w.r.t. the design parameters of the planar inductor.

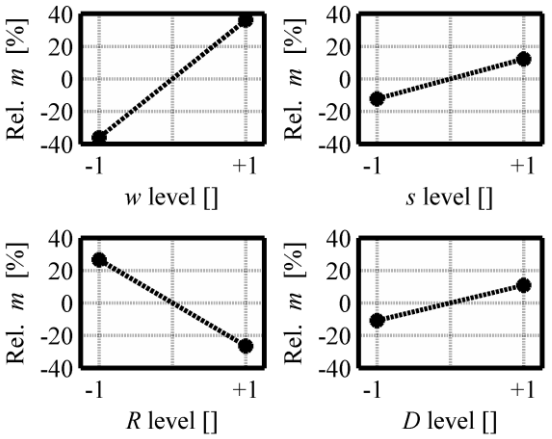


Figure 3. The magnetic moment m w.r.t. the design parameters of the planar inductor.

parameter. The results show that the track width w has the largest effect on the quality factor. It is observed that the effect of the fill-ratio R is relatively small, while the effects of the track spacing s and the outer dimension D on the quality factor are even smaller. The effects of the track width and the fill-ratio on the quality factor are positive, while the effects of the track spacing and the outer dimension are negative.

Fig. 3 shows the magnetic moment m of the planar inductor w.r.t. the design parameters. This figure shows the relative value of the magnetic moment w.r.t. the two levels of the design parameter. The results show that the track width w and the fill-ratio R have the largest effect on the magnetic moment. It is observed that the effects of the track spacing s and the outer dimension D on the magnetic moment are comparably smaller. The effects of the track width, the track spacing and the outer dimension on the magnetic moment are positive, while the effects of the fill-ratio are negative.

Table II shows the results of the analysis of variance (ANOVA) of the simulation results [17]. Table II shows the F-statistics and the p-value for the quality factor Q and the magnetic moment m w.r.t. the design parameters of the planar inductor. The results confirm the observations based on Figs. 2 and 3. The track width has the largest effect on the quality factor. The fill-ratio also has a notable effect, while the track spacing and the outer dimension have the comparably smallest effects on the quality factor. It is also confirmed that the track width and the fill-ratio have the largest effect on the magnetic moment.

VI. Discussion

Increasing the track width w reduces the resistance R_S of the inductor, while minimally influencing the inductance L_S of the inductor and thus the track width has a positive and a large effect on the quality factor Q.

The fill-ratio R determines the number of turns of the inductor. Increasing the fill ratio, i.e. the number of turns increases both the inductance L_S and the resistance R_S of the inductor. The results in Fig. 2 show that the inductance increase is larger than the resistance increase and thus the fill-ratio has a positive, notable effect on the quality factor Q.

The track spacing s influences the number of turns by modifying the track pitch. Increasing the track spacing while maintaining a fixed fill-ratio reduces the number of turns of the inductor. Due to the decrease in the number of turns, the inductance and the resistance of the inductor decrease as well. The inductance decrease is larger than the resistance decrease and thus the quality factor is reduced. The results on Fig. 2 show that this effect is indeed negative, but it is comparably smaller than the effects of the track width and the fill-ratio.

The outer dimension D influences the number of turns and the area of inductor. This influences the length of the inductor turns. These influences affect the inductance and the resistance of the inductor. The results show that these influences for the simulated parameter levels effectively cancel each other resulting in a very small, negative effect.

The magnetic moment m of the inductor is related to the ripple of the converter output current ΔI_{out}, i.e. the inductance L_S of the inductor. It is also related to the number of turns and the area of the inductor. Increasing the track width w while maintaining the fill-ratio R leads to a reduced number of turns. The reduced number of turns decreases the inductance of the inductor and consequently increases the ripple current of the buck converter. The results on Fig. 3 show that increasing the ripple current affects the magnetic moment much more than the decreased number of turns and results in a large, positive effect of the track width on the magnetic moment.

The fill-ratio R determines the number of turns and consequently influences the inductance of the inductor. Increasing the fill-ratio increases the inductance and thus reduces the current ripple. The results in Fig. 3 confirm that the inductance of the planar inductor has the largest effect on the magnetic moment and the fill-ratio has a large, negative effect on the magnetic moment.

Increasing the track spacing s reduces the number of turns and the inductance of the inductor. These changes

TABLE II
THE ANALYSIS OF VARIANCE OF THE SCREENING EXPERIMENT.

	Quality factor Q		Magnetic moment m	
	F	p-value [%]	F	p-value [%]
w	21.3	0.07	37.5	0.01
s	0.27	61.2	7.73	1.79
R	1.20	29.7	20.9	0.08
D	0.20	66.5	1.62	23.0

effectively cancel each other thus resulting in a comparably small, positive effect of the track spacing on the magnetic moment.

The outer dimension *D* directly influences the area of the inductor. Increasing the outer dimension increases the area and the inductance of the inductor. These changes effectively cancel each other thus resulting in a comparably small, positive effect of the outer dimension on the magnetic moment.

In summary, the results show that the track width *w* and the fill-ratio *R* have the largest influences on the quality factor *Q* and the magnetic moment *m* of the planar inductor. It is recommended to include these parameters in a co-optimization of the quality factor and the magnetic moment.

VII. CONCLUSIONS

A luminary consists of the OLED cell and its driver which uses a planar inductor. The planar inductor is the main source of radiated emissions of the driver. The properties of the planar inductor are analysed using EM simulations. The simulations are analysed w.r.t. the quality factor and the magnetic moment. The two-level screening experiment shows that the track width and the fill-ratio of the inductor have the largest effect on its properties. It is concluded that quality factor may be increased by increasing the track width of the inductor, while the magnetic moment is most sensitive to the changes in the inductance of the inductor.

REFERENCES

[1] Z. Ouyang and M. Andersen, "Overview of planar magnetic technology; fundamental properties," *Power Electronics, IEEE Transactions on*, vol. 29, no. 9, pp. 4888–4900, Sept 2014.

[2] I. Kowase, T. Sato, K. Yamasawa, and Y. Miura, "A planar inductor using Mn-Zn ferrite/polyimide composite thick film for low-voltage and large current DC-DC converter," *Magnetics, IEEE Transactions on*, vol. 41, no. 10, pp. 3991–3993, Oct 2005.

[3] J.-J. Lee, Y.-K. Hong, S. Bae, J.-H. Park, J. Jalli, G. Abo, R. Syslo, B.-C. Choi, and G. Donohoe, "High-quality factor Ni-Zn ferrite planar inductor," *Magnetics, IEEE Transactions on*, vol. 46, no. 6, pp. 2417–2420, June 2010.

[4] Y. Yazaki, K. Ishidate, K. Hagita, Y. Kondo, S. Hattori, M. Sonehara, T. Sato, T. Watanabe, Y. Seino, and N. Matsushita, "Embedded planar power inductor in an organic interposer for package-level DC power grid," *Magnetics, IEEE Transactions on*, vol. 50, no. 11, pp. 1–4, Nov 2014.

[5] A. Monte, J. Doutreloigne, J. Windels, and P. Bauwens, "Driving electronics for OLED lighting," in *SID Mid-Europe Spring Meeting 2013*, 2013, pp. 12–15.

[6] J. Kundrata and A. Baric. "Design of a planar inductor for DC-DC converter on flexible foil applications," *MIPRO, 2012 Proceedings of the 35th International Convention*, pp. 54-59, May 2012.

[7] E. Waffenschmidt, B. Ackermann, and J. Ferreira, "Design method and material technologies for passives in printed circuit board embedded circuits," *Power Electronics, IEEE Transactions on*, vol. 20, no. 3, pp. 576–584, May 2005.

[8] C. Marxgut, J. Muhlethaler, F. Krismer, and J. Kolar, "Multiobjective optimization of ultraflat magnetic components with PCB-integrated core," *Power Electronics, IEEE Transactions on*, vol. 28, no. 7, pp. 3591–3602, July 2013.

[9] T. Andersen, C. Zingerli, F. Krismer, J. Kolar, N. Wang, and C. Mathuna, "Modeling and pareto optimization of microfabricated inductors for power supply on chip," *Power Electronics, IEEE Transactions on*, vol. 28, no. 9, pp. 4422–4430, Sept 2013.

[10] E. Rondon-Pinilla, F. Morel, C. Vollaire, and J.-L. Schanen, "Modeling of a buck converter with a SiC JFET to predict EMC conducted emissions," *Power Electronics, IEEE Transactions on*, vol. 29, no. 5, pp. 2246–2260, May 2014.

[11] C. Fuentes, B. Allongue, G. Blanchot, F. Faccio, S. Michelis, S. Orlandi, J. Pontt, J. Rodriguez, and M. Kayal, "Optimization of DC-DC converters for improved electromagnetic compatibility with high energy physics front-end electronics," *Nuclear Science, IEEE Transactions on*, vol. 58, no. 4, pp. 2024–2031, Aug 2011.

[12] A. Bhargava, D. Pommerenke, K. Kam, F. Centola, and C. W. Lam, "DC-DC buck converter EMI reduction using PCB layout modification," *Electromagnetic Compatibility, IEEE Transactions on*, vol. 53, no. 3, pp. 806–813, Aug 2011.

[13] S. Orlandi, B. Allongue, G. Blanchot, S. Buso, F. Faccio, C. Fuentes, M. Kayal, S. Michelis, and G. Spiazzi, "Optimization of shielded PCB air-core toroids for high-efficiency DC-DC converters," *Power Electronics, IEEE Transactions on*, vol. 26, no. 7, pp. 1837–1846, July 2011.

[14] *COMSOL Multiphysics 4.3 Users Guide*, 2012.

[15] D. Cheng, *Field and Wave Electromagnetics*. Addison-Wesley Longman, Inc., 1996.

[16] T. Santner, B. Williams, and W. Notz, *The Design and Analysis of Computer Experiments*. Springer, 2010.

[17] D. C. Montgomery, *Design and Analysis of Experiments*. John Wiley & Sons, 2008.

Effects Influencing Repeatability of Direct Power Injection Measurements

Filip Fajdetic, Mislav Mikic, Tvrtko Mandic, Adrijan Baric
University of Zagreb, Faculty of Electrical Engineering and Computing
Zagreb, Croatia
Email: filip.fajdetic@fer.hr, tvrtko.mandic@fer.hr

***Abstract*—The electromagnetic compatibility (EMC) performance of the voltage regulators is the significant factor in the EMC performance of the complete electronic system. In this paper the characterization of the voltage regulator by direct power injection (DPI) method is presented. Although, the DPI method is defined by various standards, the repeatability of the measurement can be compromised by several factors. The DPI method measurement results are presented for the L4949 voltage regulator. The repeatability of the method is assessed by various factors related to the operating point of the voltage regulator and to the measurement set-up. The obtained results are compared to the results presented by other authors.**

***Index Terms*—DPI method; voltage regulator; measurement reapeatability**

I. Introduction

The electromagnetic compatibility (EMC) characterization of an integrated circuit (IC) is a mandatory step in the design of reliable electronic systems. The level of emission or immunity of an IC can compromise performance of the complete system. The immunity level for an IC is considered as the maximum radio frequency (RF) power level incident to IC through radiation or conduction which does not compromise the predefined normal operating conditions [1].

In this paper electromagnetic (EM) immunity of the low dropout (LDO) voltage regulator is characterized by measurements [2]. The measurements are performed according to direct power injection (DPI) method [3]. This standard defines the testing frequency range 150 kHz - 1 GHz. In this frequency range cable harnesses and/or printed circuit board (PCB) traces can act as efficient antennas. Thus an IC gets the unwanted RF energy through those pins connected to wires of such cables [3].

The same voltage regulator is characterized in [4], [5]. In [4] the DPI measurements are performed using PCB mounted bias tee in the frequency range 1 MHz to 1 GHz. The immunity fail criterion is defined as drift of the output voltage larger than $\pm$200 mV. In [5] the DPI measurements are performed using standalone bias tee in the frequency range 10 MHz to 3 GHz. The fail criterion is defined as drift of the output voltage larger than $\pm$100 mV. In this work the DPI measurements are performed on the same voltage regulator and results are compared to the previously published measurements. The effects influencing the measurement repeatability are discussed.

In Section II the measurement set-up and flowchart of the DPI method are explained. The DUT which is the low dropout voltage regulator L4949 mounted on a PCB are explained in Section III. Section IV brings the measurement results and discussion.

II. DPI Method

A. Measurement Set-up

The measurement set-up is presented in Fig. 1. It consists of the following instruments:

- R&S SMIQ 03B Signal Generator
- AR Amplifier 1W1000
- Tektronix CPS250 Power Supply
- RadiPower USB RF Power Sensors
- AR DC3001A Bi-Directional Coupler
- Bias Tees: Pasternack PE1601 and Mini-Circuits ZX86-12G-S+
- Agilent MSO7034B Oscilloscope
- BNC 6 dB Attenuator
- Desktop Computer (GPIB and USB ports)

The RF signal generated by the signal generator is connected to the 30 dB amplifier. The 6 dB attenuator is connected to the output of the amplifier in order to suppress reflections due to the variations of the output impedance of the amplifier. The bi-directional coupler gives possibility of measuring both forward and reflected power denoted as P_{fwd} and P_{refl}, respectively. Although, only the forward power with respect to the frequency is usually given as the figure of DUT immunity. This is due the fact that many ICs are most susceptible to the disturbances at frequencies where quite high reflections of the incident RF power occur [3].

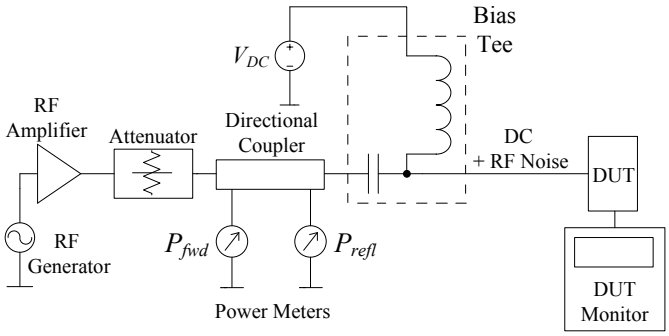


Fig. 1. Direct power injection measurement set-up. RF generator, power meters and DUT monitor are controlled by computer.

The oscilloscope is used to monitor the DUT output voltage level. Together with the RF power meters and the RF signal generator they are connected by desktop computer through GPIB and USB bus and controlled by Virtual Instrument Software Architecture (VISA) standard.

B. Flowchart

The DPI instrumentation is controlled by the graphical user interface designed using Matlab Guide [6]. The developed interface is presented in Fig. 2. It is used to connect to the instrumentation, setting up all relevant parameters related to the DPI method, starting the measurement and finally, saving data and disconnecting.

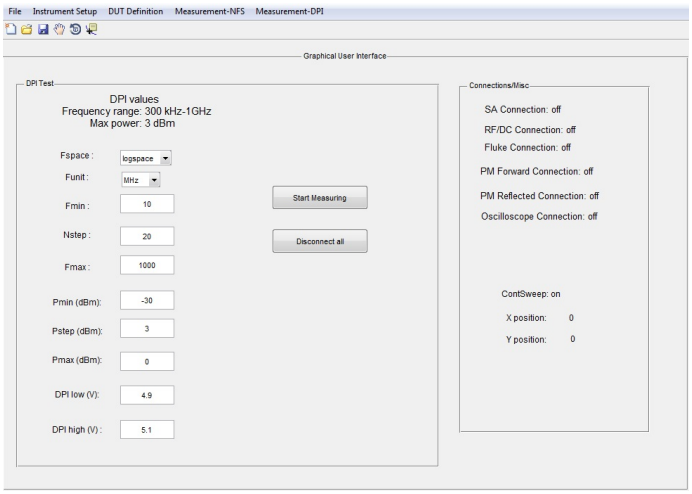


Fig. 2. Graphical user interface for controlling DPI instrumentation and for capturing the measurements.

The DPI method flowchart is shown in Fig. 3. The flowchart is defined according to [3]. The decision for DUT failure is based on the oscilloscope measurement. Every loop iteration the timebase of the oscilloscope is changed according to the frequency of RF disturbance signal. The timebase is set to enable measurement of five periods of the disturbance signal. Furthermore, the RF power meters are provided with the information regarding the frequency of RF disturbance signal in order to increase the measurement accuracy. Total dwell time is equal to 1.1 seconds which is in accordance with the value given by the standard.

III. Test Case: L4949

The DUT used in this paper is the 8-pin low dropout voltage regulator L4949 (ON Semiconductors) in small outline (SO) package. The input voltage can be in range from 5 to 28 V, while the nominal output voltage is 5 V$\pm$1%.

The complete schematic of the DUT is presented in Fig. 4a. The DC input voltage superimposed with RF noise is connected through end-launch SMA connector to the input pin of the voltage regulator. According to L4949 datasheet, two 10 µF capacitors connected to the input (C_S) and output (C_O) of the voltage regulator are recommended.

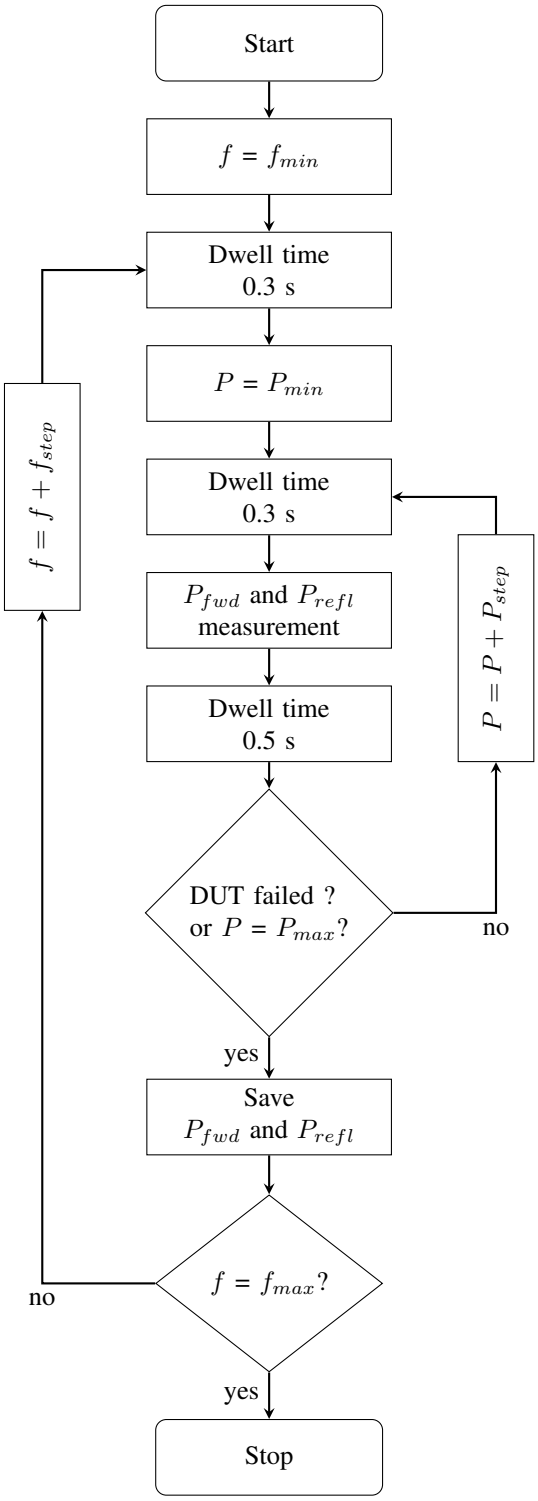
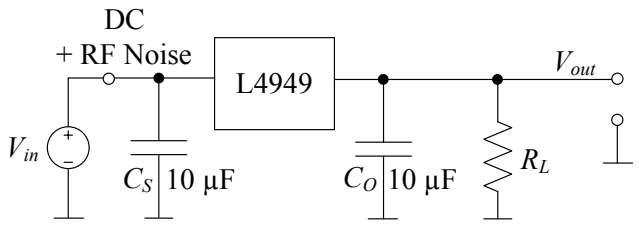


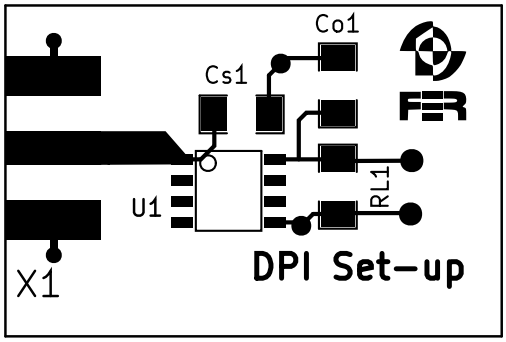
Fig. 3. Flowchart for implemented DPI method.

The top layer of the two-layer PCB is shown in Fig. 4b. All components are mounted on the top layer. The bottom layer of the PCB is a solid ground plane. The PCB is manufactured on the FR-4 substrate having the thickness of 1 mm. The loss tangent and dielectric constant are assumed to be 0.02 and 4.5,

respectively. In order to achieve the characteristic impedance equal to 50 Ω of the microstrip line between the DUT and SMA connector, the designed width of the line is 1.85 mm.



(a)



(b)

Fig. 4. DUT: a) schematic and b) PCB top-layer. Power is injected through end-launch SMA connector X1.

The PCB presented in Fig. 4b has the same dimensions (20 × 30 mm) as the PCB used in [5]. In [5] the stabilizing capacitor C_O with the resistor R_O is used for phase compensation, while in this work only C_O is used. In this work presented layout enables placement of C_S capacitor in order to assess its influence on IC immunity. The test bed presented in [4] is universal DPI fixture for SO packages with on-board bias tee. The transmission line that connects the Vin pin and SMA connector is ≈5 times longer than the line presented in Fig. 4b.

IV. Measurement Results and Discussion

In this section the measurement results are presented and compared to the results presented in [5] and [4]. The effects that influence repeatability of the presented results are discussed.

A. Comparison to [5]

The results presented in [5] are obtained for the fail criterion of the output voltage drift equal to ±2%. The output voltage is measured with the limited bandwidth of the oscilloscope input channel. Therefore, the bandwidth limitation function for the oscilloscope used in this work limits the bandwidth frequency to approximately 25 MHz and aquisition mode is set to 8-samples averaging. The input capacitor C_S is omitted from the measurements and the load resistor is set to 220 Ω. The measurements are performed in the frequency range from 10 MHz to 1 GHz for 20 logarithmically spaced points. The power of the RF generator is changed linearly from -30 dBm to 3 dBm in 1 dB steps.

The comparison to the results presented in [5] for the 6 dB attenuator connected to the output of the amplifier is shown in Fig. 5. The immunity level of the DUT is similar for the frequency range from 30 MHz to 200 MHz. At the frequency equal to 300 MHz the DUT measured in this work exhibits immunity lower approx. 8 dB. The difference of ≈4 dB can be seen up to 1 GHz.

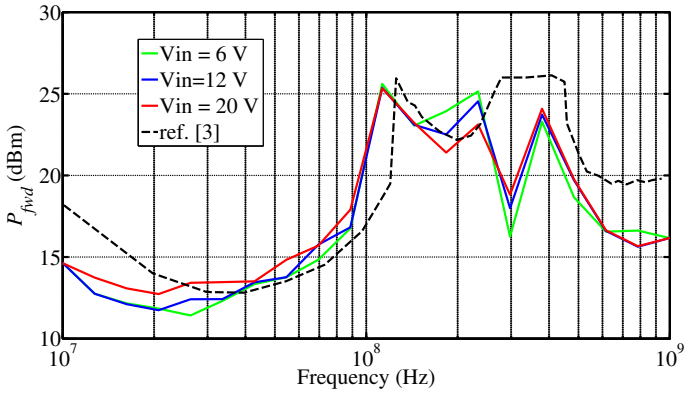


Fig. 5. Immunity levels for three different input voltages of voltage regulator. The results are compared to the results presented in [5]. The regulator input capacitor C_S=0 µF.

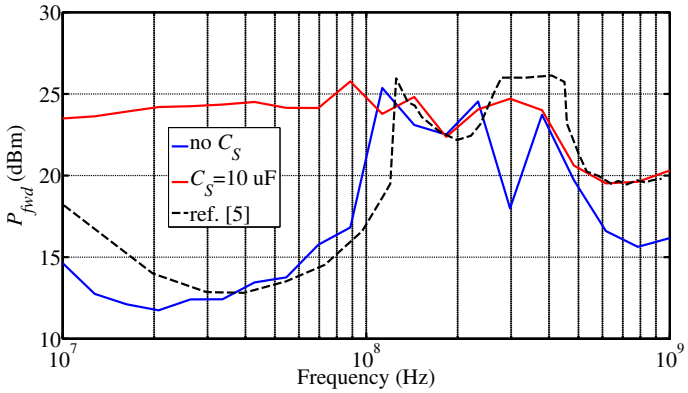


Fig. 6. Comparison of the immunity levels with respect to the input capacitor C_S=10 µF of the voltage regulator. The regulator input voltage V_{in}=12 V.

The measurement results are presented for three different values of the regulator input voltage. Although, the input voltage has an impact on immunity levels at lower frequencies, there is still difference of ≈4 dB at 10 MHz. This difference remains the same for various bias tees used in measurement.

In Fig. 6 the results are presented for the voltage regulator with the C_S capacitor connected at the input. The capacitor value is 10 µF and the input voltage of the regulator is 12 V. The result shows that the immunity level can be substantially improved by using the shunt capacitor at the input of the

regulator. Furthermore, the high frequency behaviour is similar to the results presented in [5].

B. Comparison to [4]

In [4] the results are presented for the fail criterion of the output voltage drift equal to ±4%. Since there is no information about the bandwidth limitation of the oscilloscope, the bandwidth limitation function is not used. In this work the measurements are performed with 350 MHz oscilloscope and therefore the comparison is provided up to the cut-off frequency. There is no load connected at the output of the voltage regulator, as well as no input capacitor C_S. The resistor of 10 kΩ is connected between the voltage regulator pins Vout and Reset. The input voltage of the regulator is equal to 12 V.

The measurements are performed in the frequency range from 1 MHz to 350 MHz for 20 logarithmically spaced points. The power of the RF generator is changed linearly from -30 dBm to -3 dBm in 1 dB steps. The 6 dB attenuator is connected to the output of the amplifier.

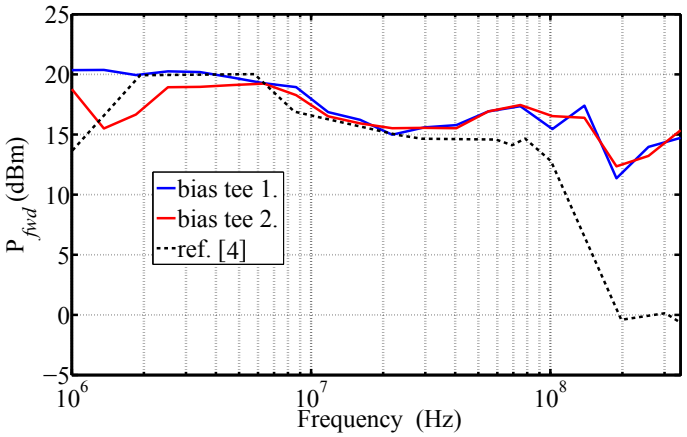


Fig. 7. Measurement results of the immunity levels compared to the results presented in [4]. The measurements are performed using two different types of bias tees.

The comparison of the measurement results to the results presented in [4] are shown in Fig. 7. Two bias tees are used in measurement set-up: Pasternack PE1601 (bias tee 1.) and Mini-Circuits ZX86-12G-S+ (bias tee 2.). In [4] the PCB bias tee is used, while in this paper the measurements are performed using the standalone bias tees. As can be seen on Fig. 7, the bias tee influence the immunity figure up to several MHz. At higher frequencies, the RF and DC path can be accurately approximated by a short circuit and open circuit, respectively, and the difference is negligible. In the frequency range above 100 MHz there is a difference of ≈15 dB. This difference can be partially attributed to the limited bandwidth of the oscilloscope (350 MHz).

V. Conclusion

This paper presents the direct power injection measurement results for the low dropout voltage regulator L4949. The measurements results are compared to the results already published by other authors. The repeatability of direct power injection measurements is strongly dependent on the measurement set-up and the set-up of the DUT. The same voltage regulator can exhibit qualitatively and quantitatively different immunity figures due to the single component change in the DUT schematic or in the measurement set-up of the DUT monitor.

References

[1] S. Ben Dhia, M. Ramdani, and E. Sicard, *Electromagnetic Compatibility of Integrated Circuits*, 1st ed. New York: Springer, 2006.

[2] F. Fiori and P. Crovetti, "Linear voltage regulator susceptibility to conducted EMI," vol. 4, pp. 1398–1403 vol.4, 2002.

[3] *Integrated Circuits - Measurement of Electromagnetic Immunity, 150 kHz to 1 GHz - Part 4: Direct RF Power Injection Method*, IEC 62132-4, First Edition, 2006.

[4] W. Jian-fei, E. Sicard, A. Ndoye, F. Lafon, L. Jian-cheng, and S. Rong-jun, "Investigation on DPI effects in a low dropout voltage regulator," pp. 153–158, Nov 2011.

[5] T. Matsushima, N. Ikehara, T. Hisakado, and O. Wada, "Improvement of reproducibility of DPI method to quantify RF conducted immunity of LDO regulator," pp. 59–62, Dec 2013.

[6] *Matlab Documentation, GUI Buillding*, The MathWorks, Inc., http://www.mathworks.com/help/matlab/gui-development.html, 2015.

SPICE Analysis of RL and RC Snubber Circuits for Synchronous Buck DC-DC Converters

Raul Blecic*†, Renaud Gillon‡, Bart Nauwelaers†, Adrijan Baric*
*University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
Tel: +385 (0)1 6129547, fax: +385 (0)1 6129653, e-mail: raul.blecic@fer.hr
†KU Leuven, ESAT-TELEMIC, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
‡ON Semiconductor, Westerring 15, 9700 Oudenaarde, Belgium

***Abstract*—Operation of a 1.3-MHz, 12-to-1.2 V, 20-A synchronous buck converter is analyzed by SPICE simulations. Special attention is given to two resonances formed by the capacitance of FETs in the off-state and the total inductance of the input decoupling network. The resonances are the main source of the broadband electromagnetic (EM) interference. Two most common circuit level methods for reducing the EM interference, an RL snubber circuit and an RC snubber circuit, are analyzed and compared in terms of their impact on the radiation characteristics, efficiency and reliability of the analyzed synchronous buck converter.**

***Keywords*—conversion efficiency, electromagnetic (EM) interference, radiated emissions, reliability, switching-mode power supply (SMPS).**

I. Introduction

Switching DC-DC converters outperform linear converters in terms of efficiency and size, which makes them a preferred choice in various applications. However, because of their switching operation, they are sources of EM interference [1]. The resonance formed by the capacitance of the low-side FET in the off-state C_{oss} and the total inductance of the decoupling network is the main cause of ringing of the switching node and of an increased EM interference [2]. The resonant frequency is typically in the range of 50-300 MHz [3]. The resonant current is mostly confined in the input decoupling network because of the output inductor [3].

The methods for reducing the EM interference of switching DC-DC converters can be roughly divided into two categories - printed circuit board (PCB) methods and circuit level methods. The PCB methods minimize the radiated fields by carefully designing the input decoupling network (dimensions of metal plates and traces, stackup of the PCB, positioning of components and vias) [3], [4]. They present an efficient and cheap solution for reducing EM interference of switching DC-DC converters. However, the PCB methods are often limited by thermal considerations and geometrical parameters of the components.

The circuit level methods minimize the radiated fields by adding a resistance into the resonant loop which dampens the resonance. This can be achieved by adding an RL snubber circuit into the resonant loop [4], an RC snubber circuit to the switching node [5], [6], by using decoupling capacitors with high equivalent series resistance (ESR) [7], or by using FETs with high resistance in the off-state R_{oss} [8], [9]. The drawbacks of the circuit level modifications are an increased complexity of the circuit and an increased cost of the design. Additionally, increasing the resistance of the input decoupling network inevitably increases the losses and decreases the conversion efficiency of the DC-DC converter, while adding reactive components increases switching losses and affects the magnitudes of currents and voltages in the resonances which can compromise the reliability of the converter.

In this paper, the operation of a synchronous buck converter is analyzed by SPICE simulations with a special attention given to two resonances formed by the capacitance of FETs in the off-state and the inductance of the input decoupling network. Two most common circuit level methods, an RL snubber circuit and an RC snubber circuit, are analyzed and compared (ideal snubber circuits are considered, their parasitics not being taken into account). An extensive parametric SPICE analysis is performed for a wide range of values of the parameters of the snubber circuits. Impact of the snubber circuits on the radiation characteristics, efficiency and reliability of the synchronous buck converter is analyzed.

This paper is structured as follows. Section II presents the analysis of the operation of a synchronous buck converter. Section III and IV analyze the impact of RL and RC snubber circuits on the characteristics of the converter. Section V concludes the paper.

II. Operation of Synchronous Buck Converter

Synchronous buck converter analyzed in this paper is a 1.3-MHz, 12-to-1.2 V, 20-A output current converter. The PWM, FET drivers, high-side FET and low-side FET are co-packaged in a DrMOS module. More details on the chip can be found in [10]. The FETs are 30-V trench power FETs with a shielded gate [11].

The SPICE model consists of the PWM, FET drivers, high-side and low-side FETs (which include the model of a body diode [10]), package parasitics, output inductor and the equivalent model of the input decoupling network (which consists of the metal plates on the PCB and the decoupling capacitors). The parameters of the SPICE models of active components are extracted from the process simulations, while the parasitic elements of the system are estimated by experiments and finite element simulations. A simplified schematic of the analyzed converter is shown in Fig. 1. An RL snubber circuit and an RC snubber circuit indicated in the figure are analyzed in the following sections.

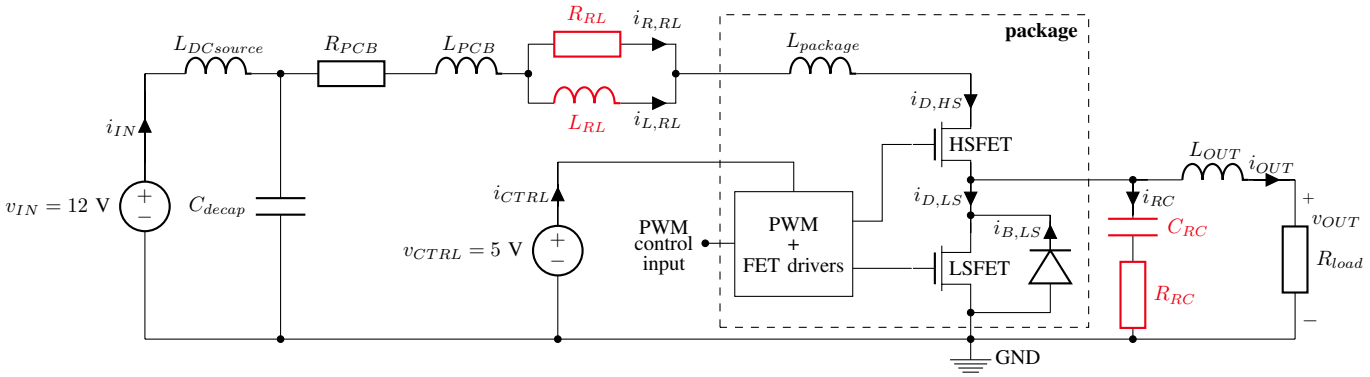


Fig. 1. Simplified schematic of the analyzed synchronous buck converter with some parasitic elements. RL snubber circuit is placed between the package and the decoupling network, while RC snubber circuit is added to the switching node.

The steady-state time-domain waveforms of the nominal simulation (without the snubber circuits) are shown in Fig. 2. The output voltage is set to 1.2 V by the duty cycle of the PWM. The output current of 20 A is set by the load resistance of 60 mΩ. Two resonances can be identified, the first one at the moment when the high-side FET turns on, and the second one, at the moment when the high-side FET turns off. The waveforms at each resonance are analyzed separately in the following subsections.

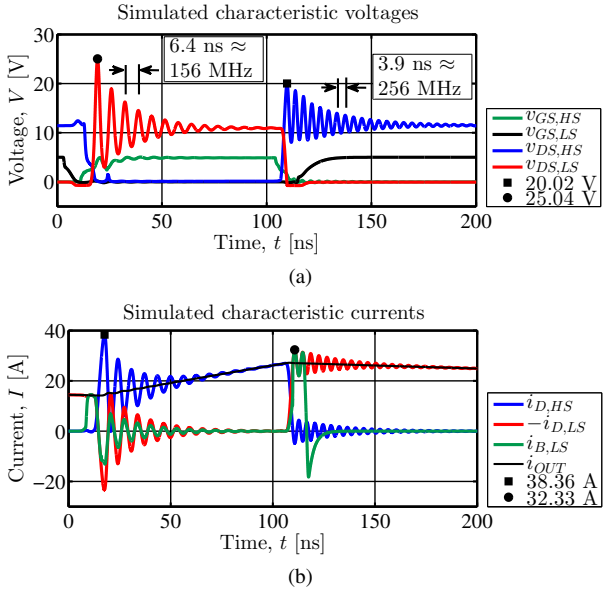


Fig. 2. Steady-state time-domain waveforms of the nominal simulation: (a) characteristic voltages and (b) characteristic currents. 200 ns of one period (774 ns) is plotted to show the waveforms in the resonances in more details.

A. Resonance at the leading edge

At the beginning of the simulation, the low-side FET is on, while the high-side FET is off. The low-side FET is turned off before the high-side FET is turned on in order to prevent a shoot-through (a low-resistive connection between the input voltage and the ground). The time during which both FETs are off is commonly referred to as the dead-time. During the dead-time, the current to the load is provided through the body diode of the low-side FET. The voltage of the switching node is tied to the negative value determined by the on-state voltage drop of the body diode (approximately $V_D = -0.7$ V).

Then, the high-side FET turns on. The current through the body diode $i_{B,LS}$ decreases while the current through the high-side FET increases by the rate determined by the input voltage and the inductance of the input network $di/dt = (V_{IN} - V_D)/L_{input}$ (or by the turn-on time of the high-side FET, depending on which of the two is slower). The diode does not switch off immediately, but undergoes a reverse recovery - the current through the diode falls to zero and continues to flow in the reverse direction, until the depletion region is sufficiently depleted so that the body diode is able to support the blocking voltage [12], [13], [14].

Then, the charge from the input decoupling capacitors charges the depletion layers of the transistor and the voltage of the switching node starts to increase. This charge, and the remaining charge in the depletion region of the body diode excite the resonant loop. The resonant loop is formed by the capacitances of the low-side FET in the off-state (which consists of a drain-to-source capacitance and the capacitance of the body diode, $C_{oss} = C_{ds} + C_j$) and the total inductance of the input decoupling network.

B. Resonance at the trailing edge

Similar process happens at the trailing edge. The high-side FET turns off and the current to the load is provided through the body diode of the low-side FET (the low-side FET is still off to prevent a shoot-through). The current through the high-side FET decreases, while the current through the body diode of the low-side FET increases by the rate determined by the inductance of the low-side FET (or by the turn-off time of the high-side FET, depending on which of the two is slower). This change triggers the resonance in the resonant loop formed by the capacitance of the high-side FET in the off-state and the total inductance of the input decoupling network.

Finally, the low-side FET turns on and the body diode of the low-side FET undergoes a reverse recovery. The charge accumulated in the depletion region of the body diode discharges through the on-resistance of the low-side FET and does not excite the resonant loop.

C. Comparison of two resonances

The two resonances have different resonant frequencies and different rate of decay, which is a consequence of different sizes of the high-side FET and the low-side FET. The high-side FET is smaller than the low-side FET, so the C_{oss} that corresponds to the high-side FET is smaller than that of the low-side FET. This results in a lower resonant frequency of the resonance related to the leading edge of $v_{DS,LS}$ compared to that related to the trailing edge. The on-resistance of the high-side FET $R_{dson,HS}$ is larger than the on-resistance of the low-side FET $R_{dson,LS}$ for the same reason (the high-side FET is smaller than the low-side FET). Therefore, the resonance related to the leading edge is more dampened and decays faster than the one related to the trailing edge.

The inductance forming the resonances is the same for both resonances and consists of the total inductance of the input decoupling network - the parasitic inductance of the package, the inductance of the PCB interconnects and the parasitic inductance of the decoupling capacitors.

D. Radiation characteristics

The energy trapped in the resonant loop at each resonance is either dissipated in its lossy elements or radiated into the space. The radiation characteristics of a DC-DC converter can be described by those of a small loop antenna [3], which is given by (assuming a constant spatial current distribution) [15]:

$$E = \eta \frac{\pi S I \exp(-\jmath k r)}{\lambda^2 r} \sin(\theta) \propto S I f^2, \qquad (1)$$

where η is the impedance of the free space ($\approx$ 377 Ω), λ is the wavelength ($\lambda = c_0/f$), c_0 is the speed of light in the free space, f is the frequency, k is the wavenumber, S is the effective area of the resonant loop, I is the spectrum of the resonant current (approximated by the current through the high-side FET), and r is the distance to the observation point.

The equation (1) suggests two ways of reducing the EM interference - reducing the effective area S is the basis for the PCB methods, while reducing the high-frequency spectrum of the resonant current I by increasing the resistance in the loop is the basis for the circuit level methods.

The maximum radiated fields as a function of frequency calculated by (1) for the nominal simulation (without the snubber circuits) and assuming $S = 1$ mm^2 are shown in Fig. 3. Two peaks can be identified. The first one corresponds to the resonance related to the leading edge, while the second one corresponds to the resonance related to the trailing edge.

The impact of the RL and RC snubber circuits on the radiation characteristics of the converter is analyzed by evaluating the frequencies of the two peaks and the radiated fields at those frequencies for each set of the parameters of the snubber circuits.

E. Conversion efficiency

Losses in switching DC-DC converters are a sum of conductive losses in the resistive elements and the switching losses in the reactive elements of the circuit. To assess the impact of RL and RC snubber circuits on the conversion efficiency of the analyzed converter, the conversion efficiency is calculated

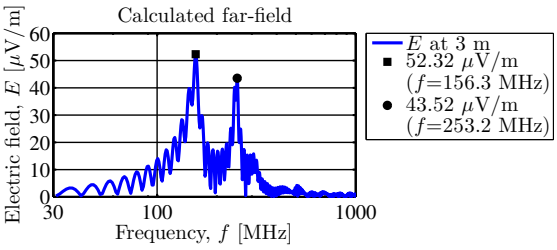


Fig. 3. Maximum radiated fields as a function of frequency calculated by (1) (at $r = 3$ m). The first peak corresponds to the resonance related to the leading edge, while the second one corresponds to that related to the trailing edge.

from the time-domain waveforms simulated for each set of the parameters of the snubber circuits as follows:

$$P_{OUT} = \text{average}(i_{OUT} \cdot v_{OUT}) \qquad (2)$$

$$P_{IN} = \text{average}(i_{IN} \cdot v_{IN}) + \text{average}(i_{CTRL} \cdot v_{CTRL}) \qquad (3)$$

$$\eta = P_{OUT}/P_{IN} \qquad (4)$$

Additionally, to get an insight into the dissipation of each element of the converter (low-side FET, body diode, high-side FET, input decoupling network, output inductor and snubber circuits), their dissipation is calculated as:

$$P_{ELEMENT} = \text{average}(i_{ELEMENT} \cdot v_{ELEMENT}) \qquad (5)$$

F. Reliability considerations

The peaks of the currents and voltages of the low-side FET and the high-side FET in the resonances can be several times larger than the nominal values of the converter. Adding reactive components into the resonant loop impacts those peaks which can compromise the reliability of the converter if the peaks exceed the maximum rated values of the devices.

The FET in the off-state has to sustain the input voltage and the voltage overshoot in the resonance, while the FET in the on-state has to sustain the output current and the resonant current. At the resonance related to the leading edge, $v_{DS,LS}$ and $i_{D,HS}$ can exceed their maximum rated values, while at the resonance related to the trailing edge, $v_{DS,HS}$ and $i_{D,LS}$ can exceed their maximum rated values.

Reliability of the analyzed converter is addressed by evaluating the maximum values of the time-domain waveforms of the currents and voltages simulated for each set of the parameters of the snubber circuits.

III. RL Snubber Circuit

An RL snubber circuit is added to the circuit model in order to analyze its impact on the radiation characteristics, conversion efficiency and reliability of the analyzed converter. The RL snubber circuit provides a high-resistance path for the resonant current, while providing a low-impedance path for the output current. The location of the snubber is shown in Fig. 1. The parametric simulations of the RL snubber circuit are performed for R in the range of 0.2 to 2 Ω with the 0.2-Ω step, and for L in the range of 0.2 to 2 nH with the 0.2-nH step. Altogether, 100 simulations are performed. The duty cycle is adjusted in each simulation to maintain the same output current

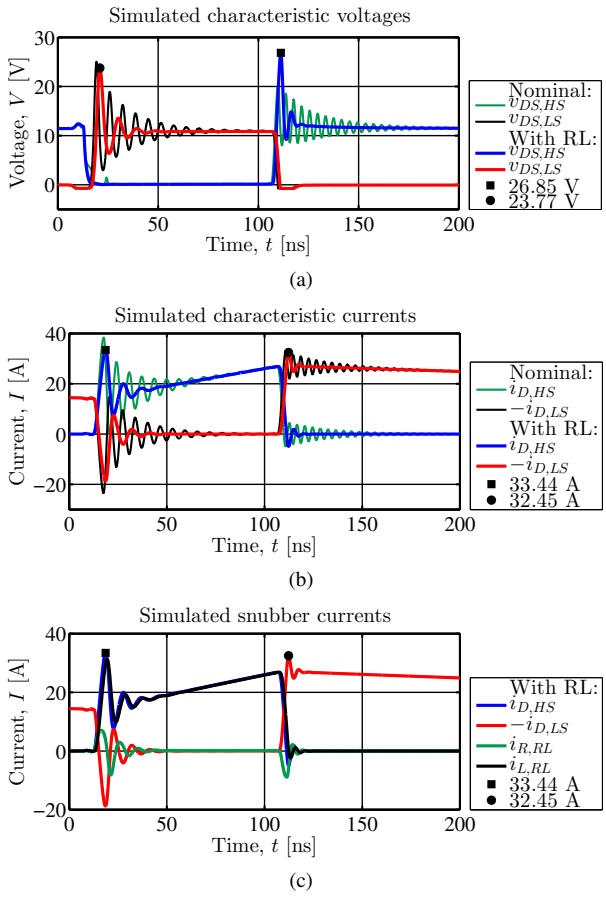


Fig. 4. Steady-state time-domain waveforms of the analyzed converter with RL snubber circuit ($R = 0.8\ \Omega$, $L = 0.8$ nH): (a) characteristic voltages, (b) characteristic currents, and (c) currents through the RL snubber circuit.

and voltage. The steady-state time-domain waveforms of the characteristic voltages and currents for $R = 0.8\ \Omega$ and $L = 0.8$ nH are shown in Fig. 4.

A. Radiation characteristics

The impact of an RL snubber circuit on the radiated fields is addressed as follows. The radiation characteristics of the converter are calculated by (1) under the assumption that adding the snubber circuit does not significantly alter the effective area of the radiating loop. The values of the two resonant frequencies and the corresponding maximum values of the radiated field are extracted for each combination of R and L. The results are given relative to the nominal simulation (without the snubber circuit) and are shown in Fig. 5. The reduction up to 18 dB is achieved. The results for $R = 0.8\ \Omega$ and $L = 0.8$ nH are compared to the nominal simulation (without the snubber circuit) in Table I.

B. Conversion efficiency

Adding an RL snubber circuit increases the losses of the circuit and decreases the conversion efficiency. The losses of the snubber circuit are calculated from the time-domain waveforms by (5). The efficiency is calculated by (4). The results are shown in Fig. 6. The results for $R = 0.8\ \Omega$ and

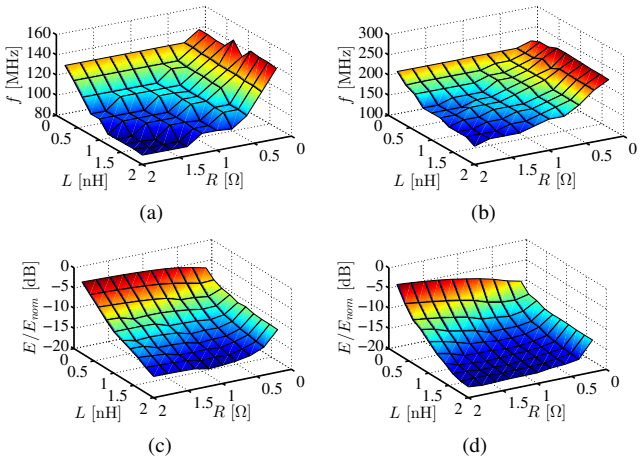


Fig. 5. Impact of RL snubber circuit on the radiated fields of the analyzed converter: resonant frequency of the resonance related to the (a) leading edge and (b) trailing edge, and maximum radiated fields in the resonance related to the (c) leading edge and (d) trailing edge.

TABLE I. RESONANT FREQUENCIES AND MAXIMUM VALUES OF RADIATED FIELDS OF THE ANALYZED CONVERTER WITH THE RL SNUBBER CIRCUIT ($R = 0.8\ \Omega$, $L = 0.8$ NH).

Parameter	Maximum value nominal simulation	Maximum value with the snubber circuit	Difference
$f_{res,1}$	156 MHz	112 MHz	-44 MHz
$f_{res,2}$	253 MHz	200 MHz	-53 MHz
E_1	52.32 µV/m	16.94 µV/m	-9.80 dB
E_2	43.52 µV/m	9.80 µV/m	-12.95 dB

$L = 0.8$ nH are compared to the nominal simulation (without the snubber circuit) in Table II.

The losses of the RL snubber circuit are dominantly a function of the inductance of the RL snubber circuit. When the high-side FET turns off, the current path between the input network and the load is disrupted. The energy accumulated in the inductance of the RL snubber is dissipated in the resistance of the RL snubber circuit. The energy accumulated in the inductance is a function of the output current. Therefore, the losses will be high for high-current converters.

Adding an RL snubber circuit adds additional source of losses. However, the reduction of the conversion efficiency is partially compensated by the reduction of the switching losses in the low-side FET, body diode, high-side FET, input decoupling network and output inductor as a consequence of the more dampened resonance compared to the nominal simulation (without the snubber circuit, see Table II).

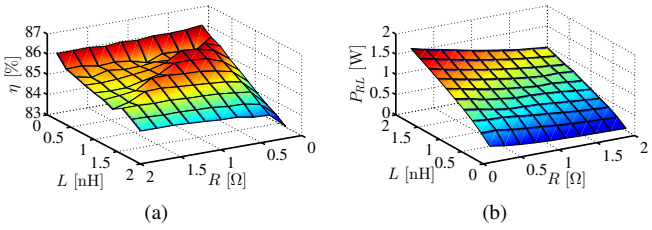


Fig. 6. Impact of RL snubber circuit on the efficiency of the analyzed converter: (a) conversion efficiency and (b) losses of the RL snubber circuit.

TABLE II. LOSSES IN THE ANALYZED CONVERTER WITH THE RL SNUBBER CIRCUIT ($R = 0.8\ \Omega$, $L = 0.8$ nH).

Element	Loss (P, W) nominal simulation	Loss (P, W) with the snubber circuit	Relative difference
Low-side FET	1.31	1.28	-2.29 %
(of which body diode)	(0.46)	(0.43)	-6.52 %
High-side FET	1.19	0.96	-19.33 %
Input network	0.98	0.90	-8.16 %
Output inductor	0.26	0.13	-50.00 %
Other	0.22	0.21	-4.55 %
RL snubber	-	0.61	-
η, %	86.69	86.32	-0.43 %

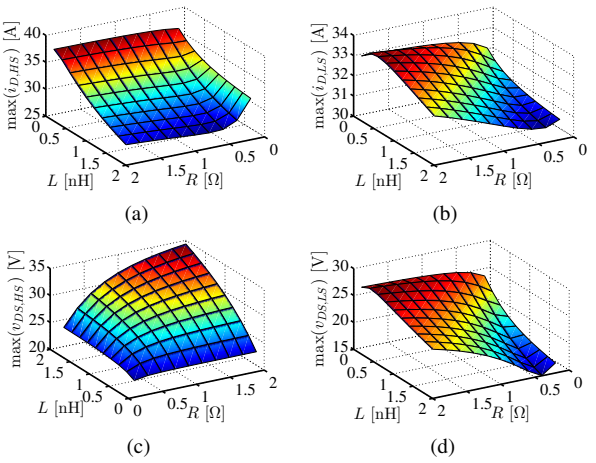


Fig. 7. Impact of RL snubber circuit on the reliability of the analyzed converter: maximum values of (a) $i_{D,HS}$, (b) $i_{D,LS}$, (c) $v_{DS,HS}$ and (d) $v_{DS,LS}$.

C. Reliability considerations

The maximum values of the currents and voltages of the high-side FET and the low-side FET are evaluated for each combination of R and L. The results are shown in Fig. 7. The results for $R = 0.8\ \Omega$ and $L = 0.8$ nH are compared to the nominal simulation (without the snubber circuit) in Table III.

The maximum value of $i_{D,HS}$ decreases when the RL snubber circuit is added to the circuit as a consequence of the more dampened resonance compared to the nominal simulation (without the snubber circuit).

The maximum values of $i_{D,LS}$ and $v_{DS,LS}$ can decrease as a consequence of the more dampened resonance compared to the nominal simulation (without the snubber circuit), or slightly increase as a consequence of the interaction between the body diode and the inductance of the input decoupling network, depending on which of the two effects is dominant.

The critical impact of adding an RL snubber circuit is an increased voltage overshoot of $v_{DS,HS}$ compared to the nominal simulation (without the snubber circuit). When the high-side FET turns off, the current path between the input network and the load is disrupted. The energy accumulated in the inductance of the RL snubber is dissipated in the resistance of the RL snubber circuit creating a voltage drop over the resistance and increasing the overshoot of $v_{DS,HS}$. The voltage drop is a function of the output current, so the overshoot will be high for high-current converters.

TABLE III. MAXIMUM VALUES OF CURRENTS AND VOLTAGES OF THE ANALYZED CONVERTER WITH THE RL SNUBBER CIRCUIT ($R = 0.8\ \Omega$, $L = 0.8$ nH).

Waveform	Maximum value nominal simulation	Maximum value with the snubber circuit	Relative difference
$i_{D,HS}$	38.36 A	33.44 A	-12.83 %
$i_{D,LS}$	32.33 A	32.45 A	+0.37 %
$v_{DS,HS}$	20.02 V	26.85 V	+34.12 %
$v_{DS,LS}$	25.04 V	23.77 V	-5.07 %

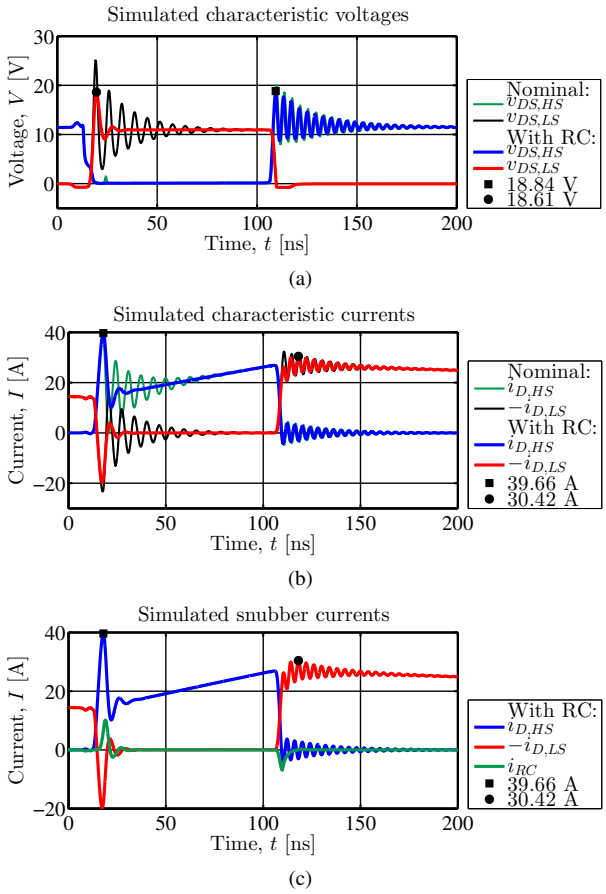


Fig. 8. Time-domain waveforms of the analyzed converter with RC snubber circuit ($R = 1\ \Omega$, $C = 2$ nF): (a) characteristic voltages, (b) characteristic currents, and (c) current through the RC snubber circuit.

IV. RC SNUBBER CIRCUIT

Equivalently to an RL snubber circuit presented in the previous section, an RC snubber circuit is added to the circuit model to analyze its impact on the radiation characteristics, conversion efficiency and reliability of the analyzed converter. The RC snubber circuit provides a high-resistance path for the resonant current, while providing a high-impedance path for the output current. The location of the snubber circuit is shown in Fig. 1. The parametric simulations of the RC snubber circuit are performed for R in the range of 0.2 to 2 Ω with the 0.2-Ω step, and for C in the range of 0.2 to 2 nF with the 0.2-nF step. Altogether, 100 simulations are performed. The duty cycle is adjusted in each simulation to maintain the same output current and voltage. The steady-state time-domain waveforms of the characteristic voltages and currents for $R = 1\ \Omega$ and $C = 2$ nF are shown in Fig. 8.

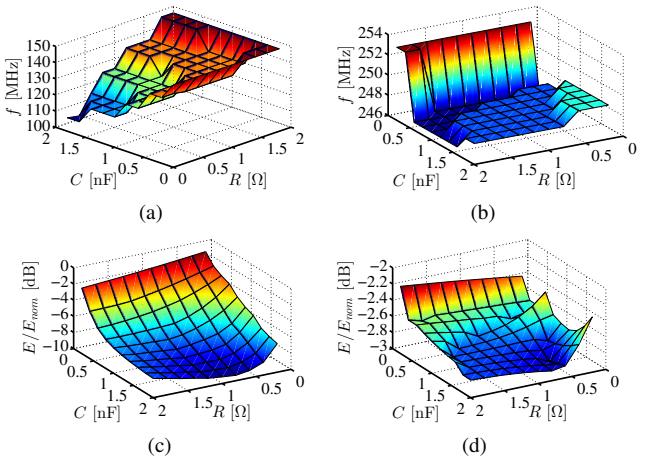


Fig. 9. Impact of RC snubber circuit on the radiated fields of the analyzed converter: resonant frequency of the resonance related to the (a) leading edge and (b) trailing edge, and maximum radiated fields in the resonance related to the (c) leading edge and (d) trailing edge.

TABLE IV. RESONANT FREQUENCIES AND MAXIMUM VALUES OF RADIATED FIELDS OF THE ANALYZED CONVERTER WITH THE RC SNUBBER CIRCUIT ($R = 1\ \Omega$, $C = 2$ nF).

Parameter	Maximum value nominal simulation	Maximum value with the snubber circuit	Difference
$f_{res,1}$	153 MHz	146 MHz	-7 MHz
$f_{res,2}$	253 MHz	248 MHz	-5 MHz
E_1	52.30 µV/m	16.71 µV/m	-9.91 dB
E_2	43.54 µV/m	37.67 µV/m	-1.26 dB

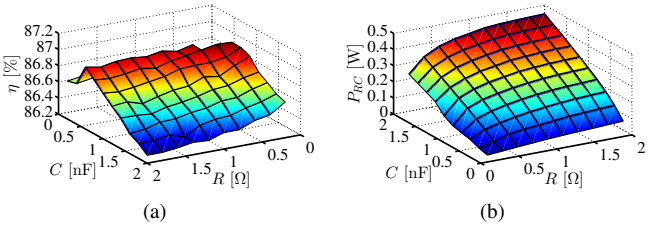


Fig. 10. Impact of RC snubber circuit on the efficiency of the analyzed converter: (a) conversion efficiency and (b) losses of the RC snubber circuit.

A. Radiation characteristics

The radiated fields of the converter are calculated by (1) under the assumption that adding the snubber does not significantly alter the effective area of the radiating loop. The results are given relative to the nominal simulation (without the snubber circuit) and are shown in Fig. 9. The reduction up to 10 dB is achieved. The results for $R = 1\ \Omega$ and $C = 2$ nF are compared to the nominal simulation (without the snubber circuit) in Table IV.

The RL snubber circuit shows better effectiveness in reducing the radiated fields than the RC snubber circuit. This is a consequence of the placement of snubber circuits: the RL snubber circuit is placed into the resonant loop, while the RC snubber circuit is placed parallel to the FET. Therefore, the resonant current has to flow through the RL snubber circuit, while it is divided between the RC snubber circuit and the capacitance of the FET in the off-state C_{oss} which limits its effectiveness.

Moreover, while the RL snubber circuit affects both resonances equally, the RC snubber circuit is more effective in reducing the resonance related to the leading edge. The RC snubber circuit is short-circuited by the low-side FET in the on-state, so it is not effective in dampening the resonance related to the trailing edge. However, the capacitance of the RC snubber circuit provides a charge tank for the load at the moment when the high-side FET turns off. Therefore, the peak of the current from the body diode of the low-side FET is reduced and the resonance related to the trailing edge is less excited. Another RC snubber circuit can be placed parallel to the high-side FET in order to dampen the resonance related to the trailing edge. Alternatively, it can be added to the input decoupling network to be effective in dampening both resonances.

B. Conversion efficiency

Adding an RC snubber circuit increases the losses of the circuit and decreases the conversion efficiency. The losses of the snubber circuit are calculated from the time-domain waveforms by (5). The efficiency is calculated by (4). The results are shown in Fig. 10. The results for $R = 1\ \Omega$ and $C = 2$ nF are compared to the nominal simulation (without the snubber circuit) in Table V.

The losses of the RC snubber circuit are dominantly a function of the capacitance of the RC snubber circuit. When the high-side FET turns off, the energy accumulated in the capacitance of the RC snubber is dissipated in the resistance of the snubber circuit. The losses are a function of the input voltage, so they will be high for high-voltage converters.

Adding an RC snubber circuit adds additional sources of losses. However, the reduction of the conversion efficiency is partially compensated by the reduction of the switching losses in the low-side FET, body diode, high-side FET, input decoupling network and output inductor as a consequence of the more dampened resonance compared to the nominal simulation (without the snubber circuit, see Table V).

C. Reliability considerations

The maximum values of currents and voltages of the high-side FET and the low-side FET are evaluated for each combination of R and C. The results are shown in Fig. 11. The results for $R = 1\ \Omega$ and $C = 2$ nF are compared to the nominal simulation (without the snubber circuit) in Table VI.

The maximum value of $i_{D,LS}$ decreases when the RC snubber circuit is added to the converter. The decrease is

TABLE V. LOSSES IN THE ANALYZED CONVERTER WITH THE RC SNUBBER CIRCUIT ($R = 1\ \Omega$, $C = 2$ NF).

Element	Loss (P, W) nominal simulation	Loss (P, W) with the snubber circuit	Relative difference
Low-side FET	1.31	1.27	-3.05 %
(of which body diode)	(0.46)	(0.42)	(-8.70 %)
High-side FET	1.19	1.13	-5.04 %
Input network	0.98	0.94	-4.08 %
Output inductor	0.26	0.11	-57.69 %
Other	0.22	0.21	-4.55 %
RC snubber	-	0.41	-
η, %	86.69	86.40	-0.33 %

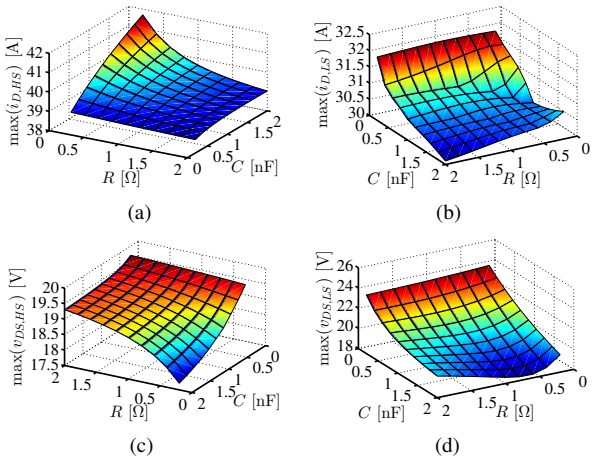


Fig. 11. Impact of RC snubber circuit on the reliability of the analyzed converter: maximum values of (a) $i_{D,HS}$, (b) $i_{D,LS}$, (c) $v_{DS,HS}$ and (d) $v_{DS,LS}$.

TABLE VI. MAXIMUM VALUES OF CURRENTS AND VOLTAGES OF THE ANALYZED CONVERTER WITH THE RC SNUBBER CIRCUIT ($R = 1\ \Omega$, $C = 2$ nF).

Waveform	Maximum value nominal simulation	Maximum value with the snubber circuit	Relative difference
$i_{D,HS}$	38.36 A	39.66 A	+3.39 %
$i_{D,LS}$	32.33 A	30.42 A	-5.91 %
$v_{DS,HS}$	20.02 V	18.84 V	-5.89 %
$v_{DS,LS}$	25.04 V	18.61 V	-25.68 %

dominantly a function of the capacitance of the RC snubber circuit. The capacitance provides a charge tank for the load at the moment when the high-side FET turns off. Therefore, the peak of the current from the body diode of the low-side FET decreases. The overshoot of $v_{DS,HS}$ decreases for the same reason.

The maximum value of $v_{DS,LS}$ decreases as a consequence of the more dampened resonance compared to the nominal simulation (without the snubber circuit).

The critical impact of adding an RC snubber circuit is an increased current overshoot of $i_{D,HS}$ compared to the nominal simulation (without the snubber circuit). When the high-side FET turns on, an additional current is needed to charge the capacitance of the RC snubber circuit. The additional current is a function of the input voltage, so the overshoot will be high for high-voltage converters.

V. CONCLUSION

The impact of RL and RC snubber circuits on the operation of synchronous buck converters is analyzed by SPICE simulations. The RL snubber circuit shows better effectiveness in dampening the resonances and reducing the radiated fields than the RC snubber circuit. The reductions of the calculated maximum radiated fields up to 18 dB are obtained for the RL snubber circuit, and up to 10 dB for the RC snubber circuit. The downside of the snubber circuits is the degradation of the conversion efficiency and the increase of the voltage and current stresses on the transistors. The losses of the RL snubber circuit are a function of the output current, while the losses of the RC snubber circuit are a function of the input voltage. The RL snubber circuit increases the overshoot of the voltage of the high-side FET as a function of the output current, while the RC snubber circuit increases the overshoot of the current through the high-side FET as a function of the input voltage. Therefore, the RL snubber circuit is more suitable for high-voltage, low-current converters, while the opposite is valid for the RC snubber circuit.

ACKNOWLEDGEMENT

This work is carried out in the frame of the INVENT project and financed by the IWT (Flanders, Belgium).

REFERENCES

[1] F. L. Luo and H. Ye, "Investigation of EMI, EMS and EMC in power DC/DC converters," in *Proc. IEEE Int. Conf. on Power Electron. and Drive Systems*, vol. 1, Nov. 2003, pp. 572–577.

[2] K. Kam, D. Pommerenke, C.-W. Lam, and R. Steinfeld, "EMI analysis methods for synchronous buck converter EMI root cause analysis," in *Proc. IEEE Int. Symp. Electromagn. Compat.*, Aug. 2008, pp. 1–7.

[3] A. Bhargava, D. Pommerenke, K. Kam, F. Centola, and C. W. Lam, "DC-DC Buck Converter EMI Reduction Using PCB Layout Modification," *IEEE Trans. Electromagn. Compat.*, vol. 53, no. 3, pp. 806–813, Aug. 2011.

[4] K. Kam, D. Pommerenke, F. Centola, C. wei Lam, and R. Steinfeld, "EMC guideline for synchronous buck converter design," in *Proc. IEEE Int. Symp. Electromagn. Compat.*, Aug. 2009, pp. 47–52.

[5] K. Harada and T. Ninomiya, "Optimum Design of RC Snubbers for Switching Regulators," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-15, no. 2, pp. 209–218, Mar. 1979.

[6] C. U-Yaisom, W. Khanngern, and S. Nitta, "The study and analysis of the conducted EMI suppression on power MOSFET using passive snubber circuits," in *Proc. Int. Symp. on Electromagn. Compat.*, May 2002, pp. 561–564.

[7] K. Zeineddine, A. Uan-Zo-li, C. hao Chen, J. Hahn, and D. ho Han, "A power-efficient method to mitigate the EMI of Switched-Mode Power Supplies," in *Proc. Int. Conf. on Power Electron.*, June 2011, pp. 453–459.

[8] K. Kam, D. Pommerenke, A. Bhargava, B. Steinfeld, C.-W. Lam, and F. Centola, "Quantification of Self-Damping of Power MOSFET in a Synchronous Buck Converter," *IEEE Trans. Electromagn. Compat.*, vol. 53, no. 4, pp. 1091–1093, Nov. 2011.

[9] J. Roig, C.-F. Tong, F. Bauwens, R. Gillon, H. Massie, and C. Hoggatt, "Internal self-damping optimization in trench power FETs for high-frequency conversion," in *Proc. IEEE Applied Power Electron. Conf. and Exposition (APEC)*, Mar. 2014, pp. 137–142.

[10] J. Roig, S. Mouhoubi, F. De Pestel, N. Martens, F. Bauwens, H. Massie, L. Golonka, and G. Loechelt, "Body-diode related losses in Shield-Plate FETs for SiP 12V-input DC/DC buck converters operating at high-frequency (4MHz)," in *24th Int. Symp. on Power Semiconductor Devices and ICs (ISPSD)*, June 2012, pp. 291–294.

[11] G. Loechelt, G. Grivna, L. Golonka, C. Hoggatt, H. Massie, F. De Pestel, N. Martens, S. Mouhoubi, J. Roig, T. Colpaert, P. Coppens, F. Bauwens, and E. De Backer, "A high-speed silicon FET for efficient DC-DC power conversion," in *24th Int. Symp. on Power Semiconductor Devices and ICs (ISPSD)*, June 2012, pp. 85–88.

[12] P. O. Lauritzen and C. Ma, "A simple diode model with reverse recovery," *IEEE Trans. on Power Electron.*, vol. 6, no. 2, pp. 188–191, Apr. 1991.

[13] B. J. Baliga, *Fundamentals of Power Semiconductor Devices*, 1st ed. Springer, 2008.

[14] R. Erickson and D. Maksimovic, *Fundamentals of Power Electronics*, ser. Power electronics. Springer, 2001.

[15] C. Balanis, *Antenna Theory: Analysis and Design, 3rd Edition*. Wiley, 2005.

Dual-channel Programmable Precise Time-delay Pulse Signal Generator for Synchronous DC-DC Converter Applications

Filip Hormot, Josip Bačmaga, Adrijan Barić
University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
Tel: +385 (0)1 6129547, Fax: +385 (0)1 6129653, e-mail: filip.hormot@fer.hr

***Abstract*—A precise two-channel programmable time-delay pulse signal generator for synchronous DC-DC converter applications is presented. The developed system is required to generate high-side (HS) and low-side (LS) signal in frequency range from 1 MHz to 15 MHz and duty cycle values from 5% to 50%. While both generated signals have the same switching frequency, the pulse width and time-delay of each signal is adjusted independently of the other signal. The most notable characteristic of the developed generator is the ability to adjust dead-times between the generated signals as precise as possible in order to use the signal generator for testing the semiconductor devices for synchronous DC-DC converter applications. The developed signal generator has the ability to set the pulse width and time-delay with the precision of 250 ps. The system is designed to transmit generated signals using either the BNC or optical fiber cable. The functionality of the system is verified by measurements performed under various operating conditions.**

***Index Terms*—driving signals, pulse-width modulation, dead-time adjustment, fiber optic transmitter**

I. Introduction

The main advantage of synchronous over asynchronous rectifiers used in DC-DC converters is significant reduction in power losses which leads to the higher overall converter efficiency. If the MOSFET drain to source resistance in the on-state R_{DS-on} is sufficiently small, then its conduction power losses are lower than those obtained using a diode. The diode junction potential limits the lowest forward voltage drop of the diode and its conduction losses cannot be reduced further. This adverse effect in asynchronous rectifiers is even more expressed in common low-voltage (1.2 V, 3.3 V or 5 V) power supplies where diode conduction losses make the largest contribution in the overall losses of the DC-DC converter [1]. However, since the synchronous topology uses two MOSFETs, it has to be ensured that both of them are never turned on at the same time, so that shoot-through does not occur. The prevention of the shoot-through effect requires a more complex circuit for generation of the driving signals.

The typical model of a buck converter is shown in Fig. 1(a). The high-side (HS) and low-side (LS) MOSFET switches are driven using the two-channel pulse signal generator with adjustable signal properties. The example of the appropriate driving signals for the HS and LS MOSFET generated by a signal generator in 3.3V CMOS technology is shown in Fig. 1(b). Because of the non-zero values of rise and fall times of driving signals, the so-called dead-times t_{d1} and t_{d2} ensure

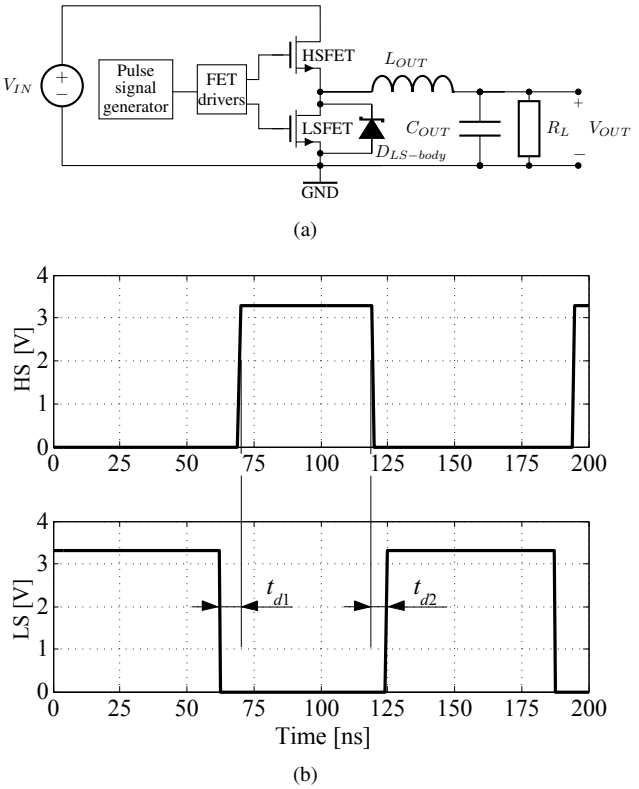


Fig. 1. Synchronous buck converter: (a) the model of a buck converter, and (b) ideal driving signals for proper work of the buck converter.

that one MOSFET is fully turned off before the other one enters the conduction mode. It can be seen that dead-times are not equal so, apart from the signal frequency, the driving signal for the HS MOSFET is controlled separately of the driving signal for the LS MOSFET and conversely.

If the influence of the dead-times t_{d1} and t_{d2} on the operation of the MOSFET switches is tested in realistic operating conditions, such as in a synchronous DC-DC converter, the pulse generator has to be able to adjust both dead times as precise as possible. Depending on the parameters of the DC-DC converter used for testing the operation of MOSFET switches, such as load resistance R_L, inductance and capacitance of the LC filter L_{OUT} and C_{OUT}, shown in Fig. 1(a), as well as the switching frequency; the dead-times

t_{d1} and t_{d2} have a different impact on the converter efficiency. Therefore, if the dead-times are adjusted separately, the impact of each dead-time on the DC-DC converter efficiency can be investigated. Furthermore, since the DC-DC converters aim at high switching frequencies in order to use smaller passive components and increase the density of the design, the pulse generator for testing purposes should be able to generate driving signals in the low megahertz range.

Since during the t_{d1} and t_{d2} neither one of MOSFET switches conducts, the current will flow through the body diode of the LS MOSFET and cause higher conduction losses than during the MOSFET conduction. Therefore, it is important for every model of the tested MOSFET switch under different operating conditions to identify the optimal dead-time values for the highest efficiency of the MOSFET switch, as well as the DC-DC converter itself.

Testing the influence of the dead-time values on the behaviour of the DC-DC converter implies the implementation of open-loop systems which are able to regulate the parameters of the driving signals regardless of any other system parameters. Therefore, the adaptive dead-time control techniques using the closed-loop systems presented in [2], [3], [4], [5] do not investigate the impact of different dead-time values on the operation of the DC-DC converters. The influence of the dead-time on the reverse-recovery effect in DC-DC converters has been tested in [6]. However, the achieved precision of setting the dead-time is only 10 ns and much more precise adjustment could be obtained.

Section II presents the developed precise pulse signal generator for synchronous DC-DC converter applications and describes its working principle. The functionality of the developed signal generator is verified by the measurement results given in Section III. The conclusion is given in Section IV.

II. The Developed Dual-Channel Driving Signal Generator

The developed system is required to generate two driving signals for HS and LS MOSFET that are controlled separately and have the same switching frequency of up to 15 MHz. Both generated signals should have duty cycle value of up to 50%, while one of them should be inverted relative to the other one in order to ensure the proper operation of the DC-DC converter test circuit. Therefore, if the HS signal would be used to drive the LS MOSFET and conversely, the generator would be able to achieve duty cycle values higher than 50% as it can be seen in Fig. 1. The important requirement of the generator is the ability to adjust the pulse width of each generated signal and relative time-delay between them with a sub-nanosecond precision for all switching frequencies and duty cycle values of interest. The dead-time adjustment is performed directly by adjusting the pulse width and time-delay of each signal.

The developed signal generator has the ability to adjust pulse width and time-delay of each generated signal, i.e. to adjust each dead-time, with the resolution of 0.25 ns for the complete range of switching frequencies and duty cycle values. This is the most important characteristic of the developed

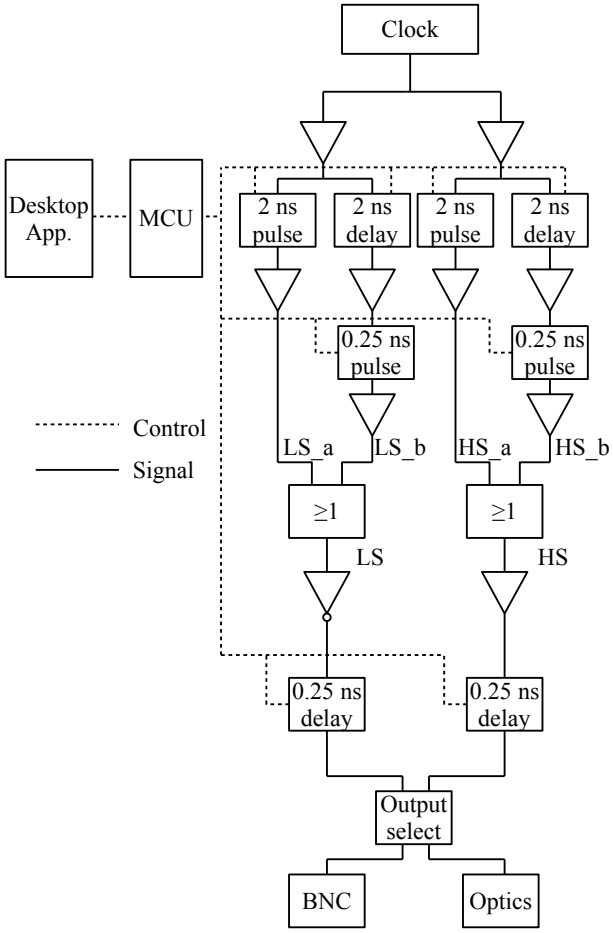


Fig. 2. Block diagram of the developed driving signal generator.

generator that overcomes the drawbacks of the similar driving signal generator developed earlier and presented in [7]. The block diagram of the developed driving signal generator is shown in Fig. 2.

All pulse width and time-delay elements are realised using the timing element DS1123L produced by Maxim Integrated that has both monostable and time-delay functions [8]. The used timing elements consist of an 8-bit step register (255 time-steps) and have a fixed time-step of either 2 ns or 0.25 ns. The 2-ns timing elements are used to obtain duty cycle values up to 50% while 0.25-ns blocks ensure the highest precision of the signal generator. In other words, the 2-ns blocks are used for coarse regulation of the generated signals while 0.25-ns blocks are used for fine adjustment of signal properties.

As it is shown in Fig. 2, each signal is constructed using one of these two methods: construction by 0.25-ns time-step division (signal path HS_b and/or LS_b) or construction by the pulse combination (HS_a+HS_b and/or LS_a+LS_b). Construction by the 0.25-ns time-step division is performed simply by setting the zero value into 2-ns-delay block and writing the wanted number of 0.25-ns time-steps into the 0.25-ns-pulse block. It can only occur if the wanted pulse width is less than 63.75 ns (0.25 ns $\times$ 255 time-steps). In

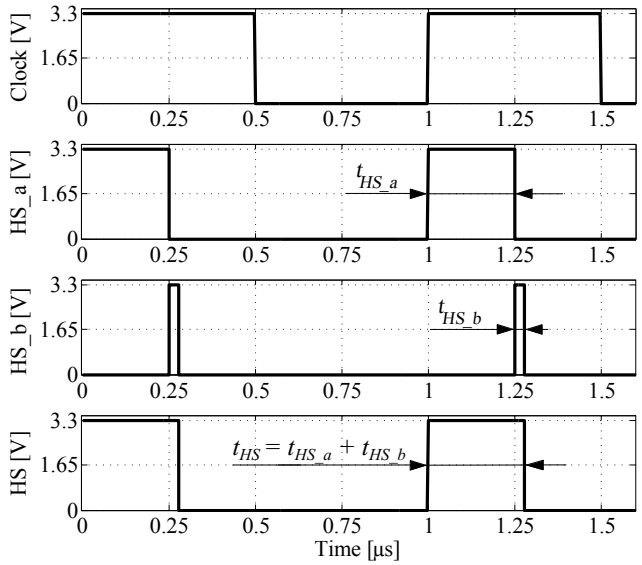


Fig. 3. Working principle of the signal construction by pulse combination (pulse width smaller than 63.75 ns) in the HS channel. (HS_a – wide pulse, HS_b – short pulse, HS – constructed PWM pulse).

this case, the signal is generated completely using the 0.25-ns block, i.e. using only the HS_b and/or LS_b signal path.

Otherwise, if the wanted pulse width is greater than 63.75 ns, the signals are constructed using the combination of both 2-ns and 0.25-ns timing elements (signal paths HS_a+HS_b and/or LS_a+LS_b). Construction by the pulse combination consists of two stages and the principle of the signal construction for one (HS) channel is shown in Fig. 3. First, a wide pulse HS_a is generated with a time-step of 2 ns (2-ns-pulse block). Second, a delayed short pulse HS_b is generated with a time-step of 0.25 ns (0.25-ns-pulse block). The short pulse HS_b is delayed (using the 2-ns-delay block) at the end of the wide pulse and, using an OR logic gate (noted with $\geq$1 in Fig. 2), those two pulses are added together resulting in a wide pulse with a precise time step adjustment as shown in Fig. 3. The same principle of the signal construction is applied for the LS driving signal.

The buffer stages, drawn as the blocks with point-down triangle in Fig. 2, are placed in between every stage of the pulse generator in order to prevent the input impedance of the following stage from affecting the preceeding stage. In order to generate appropriate driving signals for a synchronous buck converter test circuit, the output of the LS signal is inverted relative to the HS signal. Both buffer and inverter are from the same family of the integrated circuits so there is no difference in the inserted time-delay between the HS and LS signal path. The PCB traces carrying the HS and LS driving signals are routed to have the same physical dimensions so the insertion of any unwanted additional time-delay between generated signals is prevented. The 0.25-ns-delay block at both channels are used to insert a relative time-delay between generated signals. Therefore, by controlling the pulse widths and time-delay of both signals, the full control of dead-time adjustment is achieved and any time-delay mismatch can be calibrated.

As shown in Fig. 2, the developed generator has the ability to select either BNC (50 Ω) output or optical transmitter. The driving of the test circuit using the optical fiber is useful in the high-voltage DC-DC applications where, because of both functionality and safety reasons, the galvanic isolation between the pulse generator and high-voltage test circuit has to be implemented.

The operation of every timing element is controlled by the microcontroller unit (MCU in Fig. 2) that is controlled using the PC application shown in Fig. 4. Using the application, it is possible to set all relevant characteristics of generated signals. Additionaly, using the set values, the application calculates the duty cycle values of both HS and LS signal or number of time-steps written into the register of the corresponding timing-element.

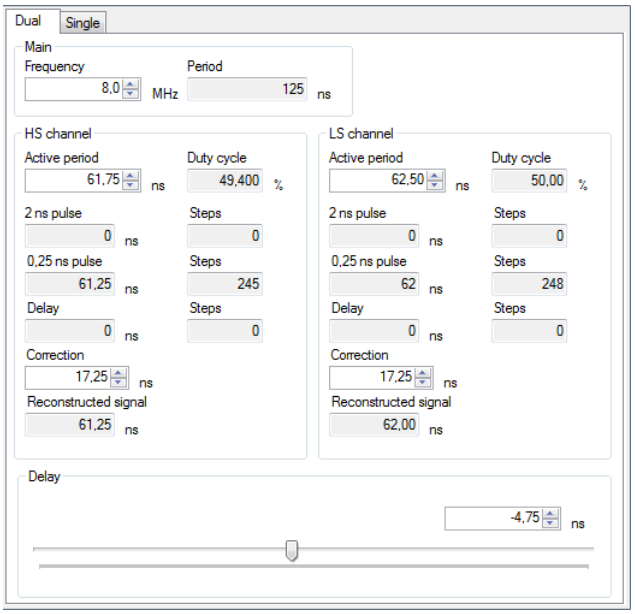


Fig. 4. The PC application used to control the operation of the developed signal generator.

III. Evaluation of the Developed Signal Generator

Before the developed signal generator is verified by measurements, the calibration of the system has to be performed. Besides the avoidance of the typical systematic errors (voltage offset, measurement errors, etc.), the calibration is also used to eliminate the mismatch of the time-delays due to the different number of elements between the signal paths LS_a (HS_a) and LS_b (HS_b).

The functionality of the developed signal generator is evaluated by measuring the HS and LS signal at both BNC and fiber-optic output. The focus of the evaluation of the system is set on testing the system precision (resolution) and system repeatability. Additionally, the functionality of the system at different ambient temperatures is also tested. All the tests are performed under different test cases as presented in Table I.

TABLE I
TEST TYPES AND TEST CASES FOR THE EVALUATION OF THE DEVELOPED SIGNAL GENERATOR.

Test type	Test case
Resolution	@f = 8 MHz (BNC and optics)
Repeatability (pulse width)	HS – 100 ns and LS – 150 ns @f = 1 MHz HS – 50.75 ns and LS – 55.75 ns @f = 5 MHz HS – 22.25 ns and LS – 26.75 ns @f = 10 MHz
Thermal stress	35 °C – 65 °C (5 °C step) @f = 8 MHz and Δt = 250 ps

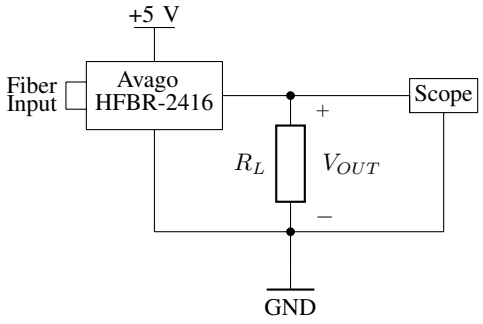


Fig. 5. The optical receiver circuit for testing the operation of the generated optical driving signals (R_L = 2.6 kΩ).

All waveforms are recorded using the Agilent MSO7034B oscilloscope and Agilent 10073D passive probe configuration as follows: 1:1 ratio, DC coupling, bandwidth limit: OFF, input impedance: 50 Ω, averaging: 256 samples. The functionality of the optical transmitter is verified using the circuit with Avago HFBR-2416 optical receiver as shown in Fig. 5.

The system resolution is verified by measuring the time-delay of one-step increment that gives the minimum difference between two pulse widths. The frequency of the generated driving signals was set to 8 MHz as noted in Table I. The recorded waveform of the rise edge for BNC and optical output are shown in Fig. 6. The voltage scales for recorded waveforms differ because the output of the optical receiver used only for testing purposes has to be buffered before it could be used as a driving stage for a MOSFET switch. The time-delay after a one-step increment for a BNC output has a value of exactly 250 ps, which was expected. Due to the slightly distorted signal observed at the output of the optical receiver, the 'bandwidth' function of the oscilloscope was turned on for this test so rise time of recorded signal is somewhat larger than for the BNC output. In order to get a clearer view of the resolution of the optical signal, four time-step increments were applied. The measured time-delay equals 1 ns which is also 250 ps for a single time-step increment.

The ability to achieve the wanted pulse width, i.e. the repeatability of the system, is tested at three different switching frequencies as shown in Table I. The repeatability test consisted of a constant reprogramming of the system for the given parameters in a repeating time period of 3 seconds for 50 test cases. Both HS and LS signals are monitored and the results of the repeatability test are shown in Fig. 7 and Fig. 8

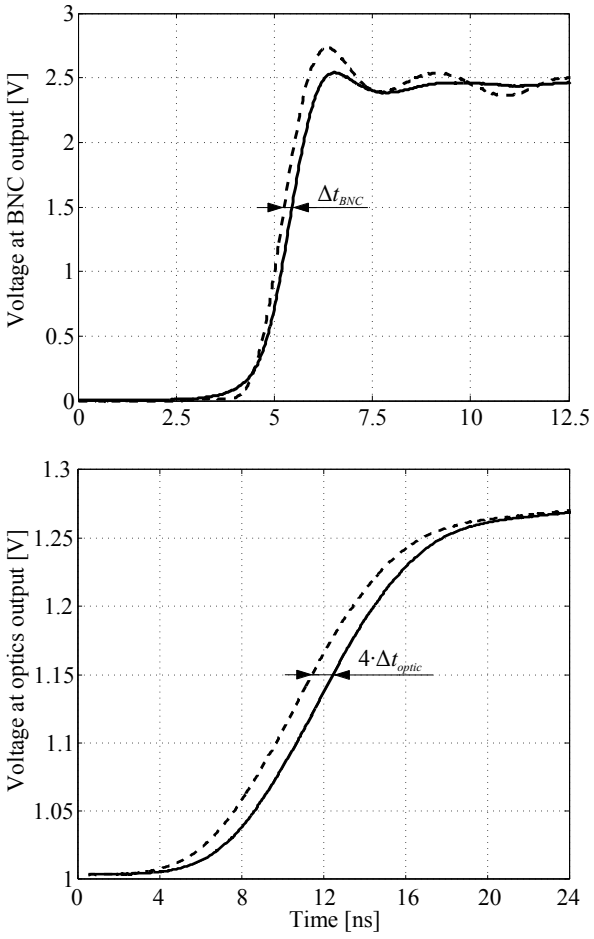


Fig. 6. Measured resolution of the generated signal before (dashed line) and after (solid line) a one time-step increment for BNC (top figure, Δt_{BNC} = 250 ps) and optical output (bottom figure, $4{\cdot}\Delta t_{optic}$ = 1 ns.

for BNC and optical output, respectively. For the HS and LS signal at BNC output, the difference of the measured pulse width and set value is less than a minimum time-step value of 250 ps. Somewhat larger deviation compared to higher switching frequencies is observed at 1 MHz. This could be mainly attributed to the less precise measurement due to the larger time-scale set on the oscilloscope in order to measure the wanted pulse width. However, as long as the deviation is lower than the minimum time-step value, it can be considered negligible. The obtained results for the optical output also show the ability of the system to achieve the deviation of the pulse width lower than 250 ps. However, the deviation is somewhat larger than 250 ps for signal frequency of 1 MHz. Besides the same cause as for the signals at the BNC output, the already mentioned distortion of the optical signal could also affect the measurement results.

In order to test the functionality of the system at higher ambient temperatures, thermal stress was applied to the system. The ability of the system to achieve its resolution of 250 ps at signal frequency of 8 MHz, as well as the stability of the set frequency value, is tested at different ambient temperatures as

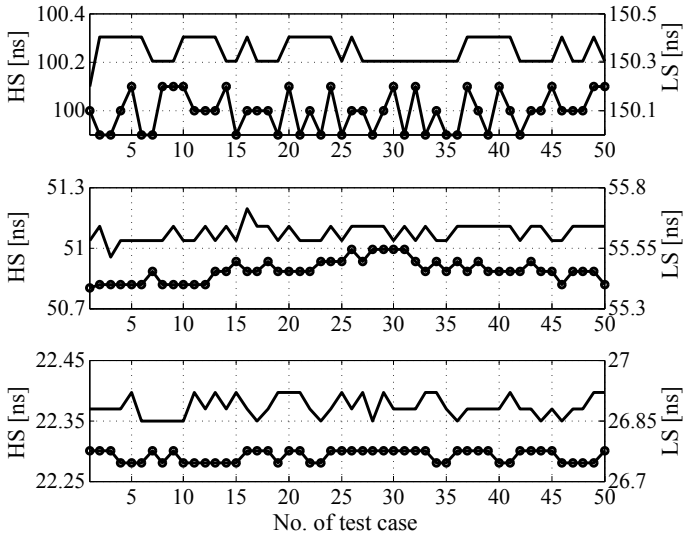


Fig. 7. Repeatability of set pulse width for HS (without marker) and LS (with marker) channel for BNC output.

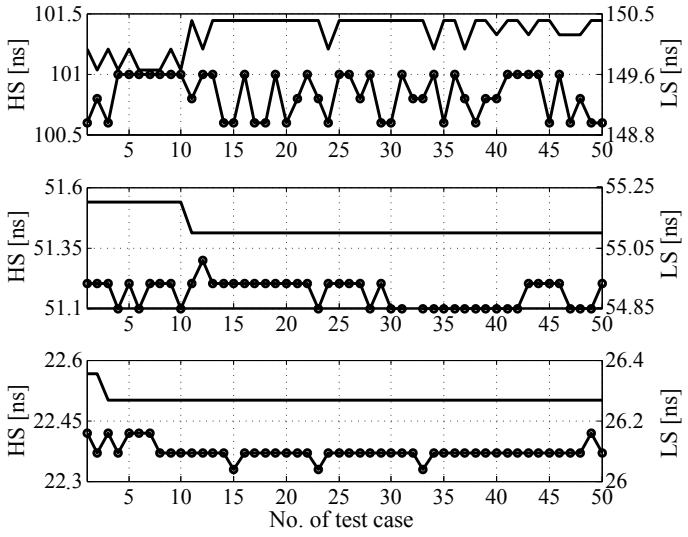


Fig. 8. Repeatability of set pulse width for HS (without marker) and LS (with marker) channel for optical output.

shown in Table I. The highest ambient temperature was chosen to be 65 °C so the temperature limit of 70 °C for the used timing element is not reached [8]. The results of the performed thermal stress are shown in Fig. 9. The displayed (normalised) quantities are the values of the signal frequency and pulse width divided by set values (8 MHz and 250 ps, respectively) written in Table I. It can be observed that signal frequency slightly increases as ambient temperature rises while the value of a single time-step increment is constant up to the highest ambient temperature. This frequency increase, that could be assigned to the impact of the temperature on the integrated oscillator circuit (clock generator), is still less than 0.3% for all ambient temperature values.

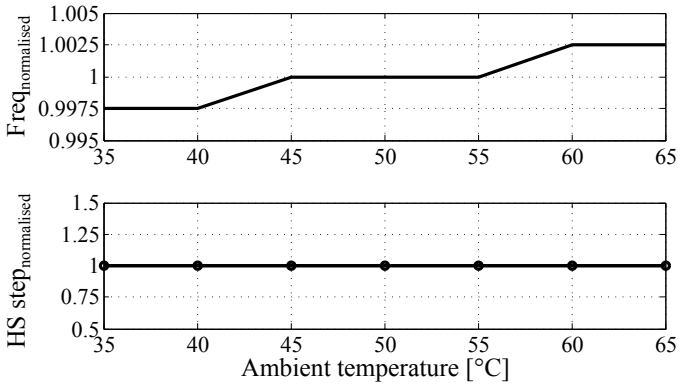


Fig. 9. Impact of ambient temperature on the signal frequency and pulse width for one (HS) channel for BNC (without marker) and optical output (with marker).

IV. Conclusion

In order to characterise the influence of switching frequency, duty cycle and dead-time values on the operation of different models of MOSFET switches used in synchronous DC-DC converters, it is necessary to implement a driving signal generator with precisely adjustable pulse width and relative time-delay between generated signals. Using the digital timing element with fixed time-step value as a basic block for pulse generation and time-delay achievement, it is possible to build a signal generator for synchronous rectifier with the sub-nanosecond precision. The measurement results showed that developed generator keeps its precision under various operating conditions. The optical output of the system, used for driving switches in high-voltage applications, also showed excellent stability during the performed tests.

References

[1] R. W. Erickson and D. Maksimović, *Fundamentals of Power Electronics*, 2nd ed. Kluwer Academic Publishers, 2004.

[2] L. Mei, D. Williams, and W. Eberle, "A predictive analog dead-time control circuit for a buck converter," in *IEEE Canadian Conf. on Electrical and Computer Engineering (CCECE)*, May 2013, pp. 1–5.

[3] S. Lee, S. Jung, C. Park, C.-T. Rim, and G.-H. Cho, "Accurate Dead-Time Control for Synchronous Buck Converter With Fast Error Sensing Circuits," *IEEE Trans. on Circuits and Systems*, vol. 60, no. 11, pp. 3080–3089, Nov 2013.

[4] S. Zhen, B. Zhang, P. Luo, K. Yang, X. Zhu, and J. Li, "A high efficiency synchronous buck converter with adaptive dead time control for dynamic voltage scaling applications," in *International Conf. on VLSI and System-on-Chip (VLSI-SoC)*, Oct 2011, pp. 43–48.

[5] T. Reiter, D. Polenov, H. Proebstle, and H.-G. Herzog, "PWM Dead Time Optimization Method for Automotive Multiphase DC/DC-Converters," *IEEE Trans. on Power Electronics*, vol. 25, no. 6, pp. 1604–1614, June 2010.

[6] D. Polenov, T. Reiter, R. Baburske, H. Pröbstle, and J. Lutz, "The influence of turn-off dead time on the reverse-recovery behaviour of synchronous rectifiers in automotive DC/DC-converters," in *European Conf. on Power Electronics and Applications*, Sept 2009, pp. 1–8.

[7] J. Bačmaga, K. Bene, B. Pejčinović, and A. Barić, "Evaluation of the operation of Depletion-mode SiC Power JFET in DC-DC converter applications," in *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2014, pp. 130–135.

[8] Maxim Integrated. (2015) DS1123L Product Details. [Online]. Available: http://www.maximintegrated.com/en/products/digital/clock-generation-distribution/DS1123L.html

Effects of dead time and semiconductor device voltage drops of output voltage of multilevel converters

Ivan Mrčela Viktor Šunde Marinko Kovačić
University of Zagreb
Faculty of Electrical Engineering and Computing
Department of Electrical Machines, Drives and Automation
Croatia

***Abstract*— This paper presents the results of a study of the effect of dead time and finite semiconductor device voltage drops on the output voltage of three-level neutral point clamped power converters. Results showed that finite voltage drops and dead time caused a DC voltage shift, and lowered the base harmonic amplitude of output voltage. These effects were analysed and the results confirmed via simulation models. Voltage drops effect was confirmed using *Ansys Simplorer*, while the dead time effect was confirmed using *Matlab*. The results showed that the output voltage of the three-level neutral point clamped converters was dependant on the listed parameters. A method is proposed to suppress the negative effects these parameters have on converter voltage.**

I. Introduction

Multilevel topologies research has been intensifying over the past two decades. Due to their inherent advantages over classical two-level converters, primarily in lower switching losses, they are used in a growing number of applications, primarily in the mid-voltage range. In these uses, three topologies stand out among the many multilevel topologies: neutral point clamped, flying capacitor and cascaded. These topologies can be found with renewable energy sources, HVDC transmission, auxiliary power supplies in trains and trams, etc, [1].

Due to an increased number of input voltage levels and an increased number of switches, multilevel converter switches switch at a fraction of the input voltage, depending on the converter topology. This enables the use of semiconductor modules with lower voltage ratings. Despite the increased number of switches, multilevel converter switching losses are lower than those for two-level converters. An increase in the number of output voltage levels reduces the Total Harmonic Distortion (THD) factor. Lower THD factor enables the use of a smaller output filter. This, in return, decreases the converter's dimensions and mass. On the other hand, an increase in the number of switches means a more complex control algorithm. Multiple output voltage levels are achieved by dividing the input voltage into several smaller voltages using a voltage divider (often a series of capacitors), [2], [3], or using several galvanically isolated voltage sources (i.e. fuel cells), [4], [5], [6]. Furthermore, due to the series-connected input capacitors in some converter topologies, the switching states that connect an output phase leg to the neutral point cause voltage ripple.

A common issue with converter topologies of any number of voltage levels are converter nonlinearities, especially in open-loop modulation strategies. The most important nonlinearity is introduced by the necessary blanking time to avoid the shoot-through of either the DC link with two-level converters, or the input capacitors with multilevel converters, caused by the devices' finite turn-on and turn-off times. To ensure that two semiconductor switches operating in counterstroke never conduct at the same time, an initial delay in each switching period, called dead time, is introduced to the control signals.

Another common issue are the semiconductors' on-state voltage drops. If voltage drops vary within the semiconductors of one phase leg, a DC voltage component of both phase and line voltages can occur [7], [8]. These DC voltage components are fairly small, but if the load has a small impedance, the DC current components may be significant. One example of a low-impedance load is a transformer. Since transformers are designed to work near saturation, even a small DC current can lead to transformer saturation [9].

This paper investigates the effects of semiconductor voltage drop and dead time on converter output voltage. These effects are analyzed, and then simulated via *Matlab* and *Simplorer*. Finally, a simple compensation method is proposed.

II. Three-level neutral point clamped inverter

The term "multilevel" was first used by Nabae et al. in 1980, [10]. This paper also presented the neutral point clamped three-level inverter. The converter topology is shown in Fig. 1.

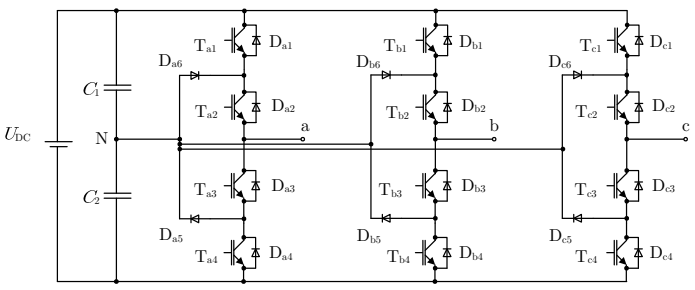


Fig. 1. A neutral point clamped three-level converter topology

The converter consists of 12 transistors and freewheeling diodes, along with six stand-alone diodes, called clamping diodes, all divided equally between three output legs.

The transistors T_{k1} and T_{k3}, as well as T_{k2} and T_{k4}, $k = \mathrm{a}, \mathrm{b}, \mathrm{c}$, work in counterstroke. There are three possible switch states, depending on the transistor on-state combination. The phase leg a switch states, along with their matching output voltages, are presented in Table I. The switch states and output voltages of the remaining two legs are generated in the same manner.

TABLE I

NEUTRAL POINT CLAMPED THREE-LEVEL INVERTER SWITCH STATES AND MATCHING PHASE LEG A OUTPUT VOLTAGE LEVELS

T_{a1}	T_{a2}	T_{a3}	T_{a4}	$U_{a\mathrm{N}}$
1	1	0	0	$+\frac{U_{DC}}{2}$
0	1	1	0	0
0	0	1	1	$-\frac{U_{DC}}{2}$

Control signals for the transistors of one output phase leg are generated by comparing one or more low-frequency modulation signals and one or more high-frequency carrier signals. The number and waveform of modulation signals and carrier signals depend on the control system. Figure 2 shows the assumption of one modulation signal per leg, and two carrier signals, along with the corresponding phase and line voltages.

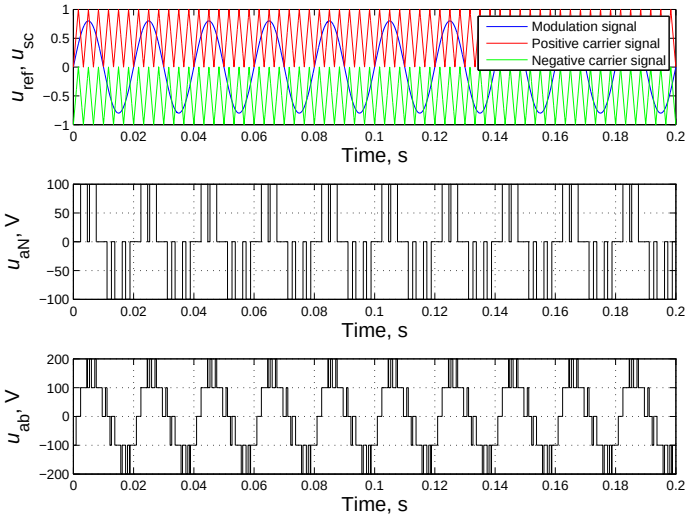


Fig. 2. Phase a modulation signal and carrier signals waveforms and phase and line voltage waveforms

Transistors T_{k1} and T_{k2} are in the on-state in those intervals in which the modulation signal is positive and greater than the positive carrier signal. Transistors T_{k2} and T_{k3} are in the on-state in those intervals in which the modulation signal is either positive and lesser than the positive carrier signal, or negative and greater than the negative carrier signal. Transistors T_{k3} and T_{k4} are in the on-state in those intervals in which the modulation signal is negative and lesser than the negative carrier signal.

Transistors conduct when in the on-state, and the output current direction of the matching phase leg matches the direction of the collector-emitter current. If the output current direction is the opposite of the transistor collector-emitter current, the according freewheeling diode conducts in that time interval.

Clamping diodes connect the phase voltages to the neutral point. In case a phase current is positive, diodes D_{k6} conduct, whilst if a phase current is negative, diodes D_{k5} conduct.

Fig. 3 depicts the phase a current paths. For instance, if phase a to neutral voltage is positive, and phase a current is negative, then the transistors T_{a1} and T_{a2} are in the on-state, as circled in Fig. 3, and diodes D_{a1} and D_{a2} are conducting. Other switching states and current paths are derived in the same manner.

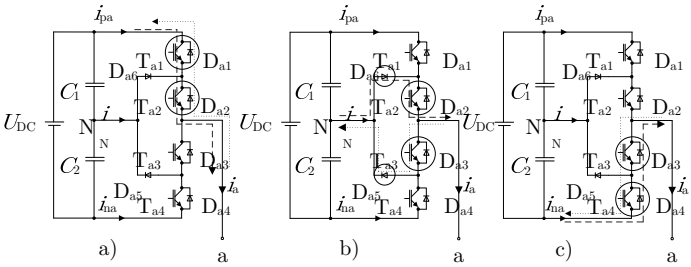


Fig. 3. Phase a current paths

III. VOLTAGE DROP EFFECT

It is generally assumed that all semiconductor voltage drops are either negligible, or equal. This need not necessarily be true. Semiconductor voltage drops in the on-state are listed in manufacturer datasheets, and vary with temperature and different production series. Datasheets usually list the minimum, typical and maximum voltage drop values for two working temperatures. Voltage drop values for the in-between temperatures are usually given graphically. For instance, the on-state voltage drops of IGBTs within one semiconductor module may vary as much as 1 V for 1200 V, 300 A semiconductor modules.

Figs. 4 a) and b) show the voltage drop effects on the phase-to-neutral voltage of an arbitrary phase leg, k, in the case of a positive modulation signal $u_{\mathrm{ref},\,k}\,(t)$ and positive and negative phase leg k load currents, respectively. u_{g}^{+} is the control signal for the phase leg switches that switch the load to the positive DC bus, u_{g}^{0} the control signal for the phase leg switches that switch the load to the neutral point, $u_{k,\,\mathrm{id}}$ the ideal output voltage waveform, and $u_{k,\,\mathrm{st}}$ the actual output voltage waveform.

Figs. 5 a) and b) show voltage drop effects on the phase-to-neutral point voltage of an arbitrary phase leg, k, in the case of a negative modulation signal $u_{\mathrm{ref}}\,(t)$ and positive and negative phase leg k load current, respectively. u_{g}^{+} is the control signal for phase leg switches that switch the load to the negative DC bus.

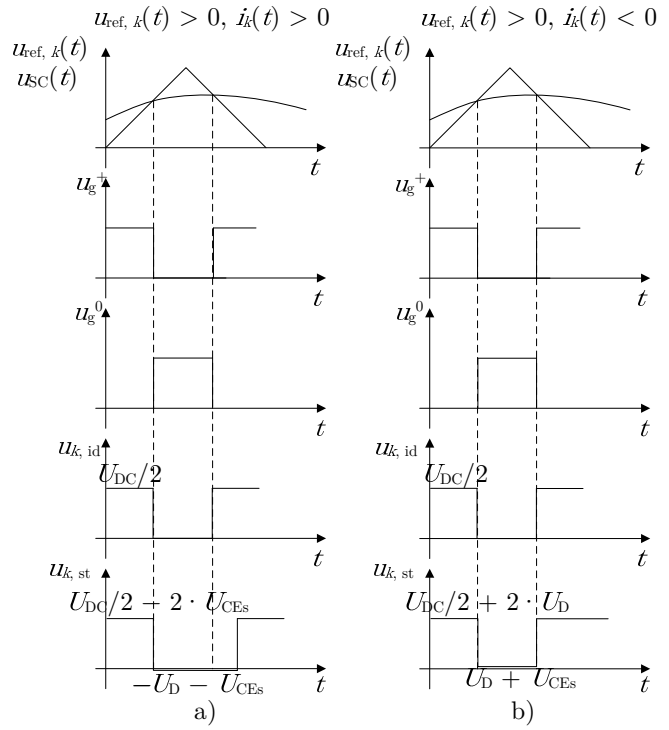


Fig. 4. Effect of semiconductor voltage drops on phase k phase-to-neutral-point voltage in case of a positive modulation signal and positive and negative load currents

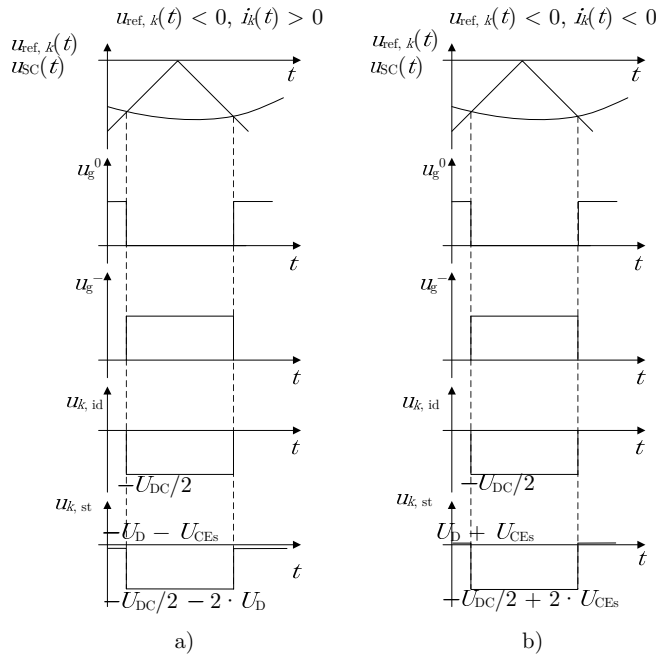


Fig. 5. Effect of semiconductor voltage drops on phase k phase-to-neutral-point voltage in case of a negative modulation signal and positive and negative load currents

During intervals where the modulation signal is greater than the carrier signal, and the phase current is positive, transistors T_{k1} and T_{k2} are in on-state and conducting, as depicted in Fig. 3.

During intervals where the modulation signal is greater than the carrier signal, and the phase current is positive, transistor T_{k2} and diode D_{k6} are conducting. In that case, transistor and diode voltage drops increase the one-cycle average. The same principle may be applied in the case of a negative modulation signal value. Therefore, if semiconductor voltage drops are equal for all transistors and diodes, no DC voltage component is present in the output voltage. In the case the semiconductor voltage drops vary from one semiconductor to another, the output voltage contains a DC component. The largest DC component can be predicted from the actual phase leg voltages shown in Figs. 4 and 5.

If transistor T_{k1} and T_{k2} voltage drops are maximum, and transistor T_{k3} and T_{k4} minimum, a maximum negative output voltage DC component can be expected. Let this be the negative worst-case scenario. Similarly, if transistor T_{k1} and T_{k2} voltage drops are minimum, and transistor T_{k3} and T_{k4} maximum, a maximum positive output voltage DC component can be expected. Let this be the positive worst-case scenario. DC components of the phase leg voltages U_{aN}, U_{bN} and U_{cN} were simulated for both worst-case scenarios.

The DC voltage components of phase leg a were simulated via *Simplorer*. The simulation parameters are listed in Table II.

TABLE II

THREE LEVEL NPC SIMULATION MODEL PARAMETERS

Parameter	Symbol	Value
Load resistance	R_{d}	$1.2\,\Omega$
Load inductance	L_{d}	$1920\,\mu\mathrm{H}$
Source voltage	U_{DC}	$750\,\mathrm{V}$
Modulation signal frequency	f_{ref}	$50\,\mathrm{Hz}$
Carrier and modulation signal frequency ratio	m_f	50
Modulation and carrier signal amplitude ratio	m_a	0.7

Semiconductor voltage drops were taken from the *Mitsubishi CM300DY-24A* power module datasheet. No minimum voltage drop was listed, instead, it was estimated at 2 V; the typical voltage drop is listed as 2.1 V, and the maximum voltage drop as 3 V for 25 °C. Since no listing was provided for the freewheeling diode voltage drops, a value of 2.17 V is assumed for the minimum, typical and maximum values. This value matches the value listed in the *IXYS MEE75-12DA-ND* datasheet, which is the clamping diode module.

The simulation results are presented in Table III. The middle column presents the results for the positive worst-case scenario, whilst the right column presents the results for the negative worst-case scenario.

TABLE III

PHASE LEG DC VOLTAGE COMPONENTS FOR WORST-CASE SCENARIOS FOR 25 °C

Voltage	Value	Value
U_{aN}	$969.81\,\mathrm{mV}$	$-835.42\,\mathrm{mV}$
U_{bN}	$843.42\,\mathrm{mV}$	$-811.72\,\mathrm{mV}$
U_{cN}	$747.75\,\mathrm{mV}$	$-801.74\,\mathrm{mV}$

The results show that phase leg voltages can contain a significant DC component, which can cause significant DC current components through a low-impedance load.

Line voltage u_{ab}, u_{ac} and u_{bc} DC components were simulated for various transistor T_{a1} and transistor T_{a2} voltage drops. This allowed for an investigation of the relationships between different voltage drops of one phase semiconductor components and line voltages. The results are shown in Fig. 6.

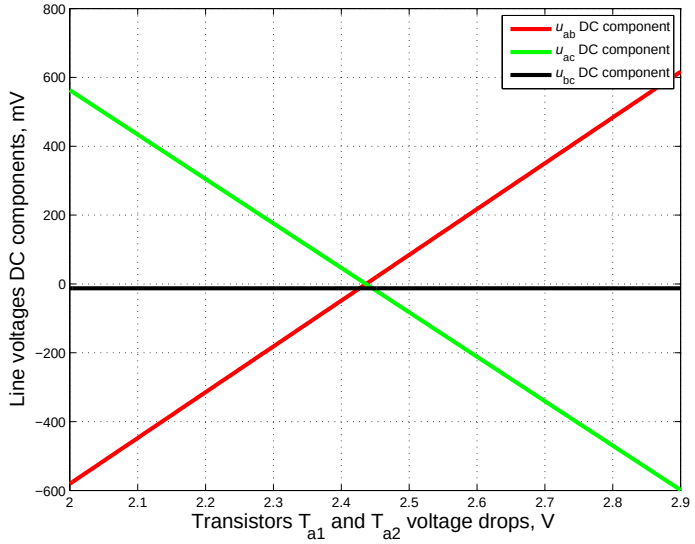


Fig. 6. Line voltages u_{ab}, u_{ac} and u_{bc} DC components

Fig. 6 shows that different voltage drops of one phase leg's semiconductor components can cause significant line voltages, which can cause significant DC current components through a low-impedance load.

A hypothetical situation was simulated, in which the converter temperature reached 125 °C. It was assumed that, before the converter temperature rose, all voltage drop values had been typical for 25 °C. After the temperature had risen, phase leg a semiconductor voltage drop values were assumed to cause the negative worst-case scenario DC component for 125 °C. Semiconductor voltage drop values for the remaining two phase legs were assumed typical for 125 °C. Voltage drop values were estimated the same way they had been estimated for 25 °C. Transistor $\mathrm{T_{a1}}$ and $\mathrm{T_{a2}}$ voltage drops are 3.3 V, transistor $\mathrm{T_{a3}}$ and $\mathrm{T_{a4}}$ voltage drops are 2.4 V, the remaining transistors voltage drops are 2.4 V, all freewheeling diodes voltage drops are 2.3 V, and all clamping diodes voltage drops are 1.85 V.

TABLE IV

PHASE LEG A AND LINE DC VOLTAGE COMPONENTS FOR 125 °C

Voltage	Value
U_{aN}	−511.65 mV
U_{ab}	−536.81 mV
U_{ac}	−521.5 mV
U_{bc}	20 mV

IV. DEAD TIME EFFECT

It is generally assumed that the output voltage is fully defined by the converter switching states. However, if the control signal contains dead time, this assumption becomes invalid and during dead time the output voltage is defined by the load phase current direction.

The influence of dead time on phase leg k voltage for the positive modulation signal is shown in Figs. 7 a) and b) for positive and negative load current, respectively. t_{d} denotes dead time duration, t_{on} turn-on time and t_{off} turn-off time.

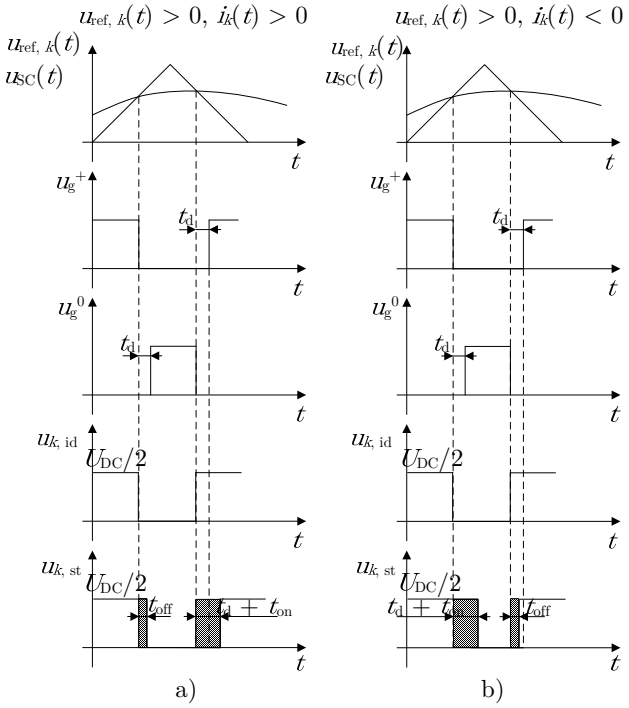


Fig. 7. Dead time influence on phase k to the neutral voltage - positive modulation signal

Dead time influence on phase leg k voltage for the negative modulation signal is shown in Figs. 8 a) and b) for positive and negative load current, respectively.

For instance, in the case that the modulation signal is positive and greater than the carrier signal, and the matching phase leg load current is positive, as shown in Fig. 7 a), transistors T_{k1} and T_{k2} are in the on-state and the phase leg output voltage is positive. When the modulation signal intersects the carrier signal, the transistor T_{k1} control signal shifts from one to zero (logical values). After t_{off}, the transistor current becomes zero and the transistor is now in the off-state. The output current commutes from transistor T_{k1} to diode D_{k6} and the phase leg voltage becomes zero. t_{d} after the modulation signal and the carrier signal intersect, transistor T_{k3} control signal shifts from zero to one, and after t_{on} transistor T_{k3} is in the on-state and the phase leg k output voltage remains zero. When the modulation signal and the carrier signal intersect for the second time, transistor T_{k3} is in the off-state after t_{off} and the phase current commutes back

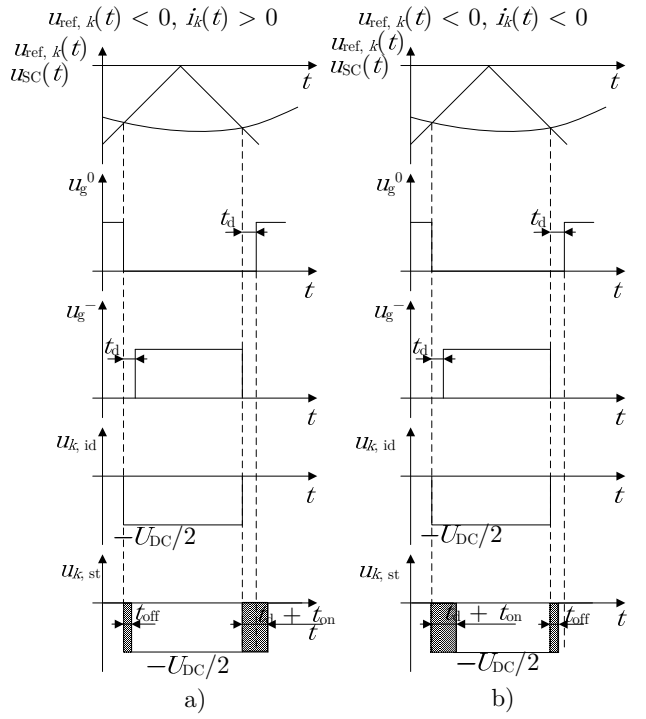


Fig. 8. Dead time influence on phase k to the neutral voltage - negative modulation signal

to diode D_{k6} and the output voltage remains zero. $t_{\mathrm{d}} + t_{\mathrm{on}}$ after the modulation signal and the carrier signal intersect for the second time transistor T_{k1} is in on-state and the output voltage is positive again.

Therefore, it is clear that during dead time, the converter's output voltage is not controlled by the modulation signal but, instead, is determined by the load current direction. This is equivalent to the discontinuation of the modulation signal. Hence, dead time has an effect on the converter's harmonic spectrum. This effect is prominent at lower frequencies. One-cycle average voltage deviation is constant for one current direction due to the dead time effect. This deviation is shown in Fig. 9, as Δu. Δu_1 is fundamental harmonic of Δu, u_{ref} is the ideal output voltage fundamental harmonic, u_1 is the actual output voltage fundamental harmonic, i_1 is the actual output current fundamental harmonic, and ϕ is the phase shift between the output voltage and output current fundamental harmonics.

Let M be the total number of switchings within one modulation signal period T_{r}. One-cycle average of the voltage deviation (over the course of one modulation signal period) is:

$$\Delta\overline{u} = \frac{M\left(t_{\mathrm{d}} + t_{\mathrm{on}} - t_{\mathrm{off}}\right)U_{\mathrm{DC}}}{2T_{\mathrm{r}}}. \tag{1}$$

It's fundamental harmonic amplitude is:

$$\Delta\hat{\overline{u}}_1 = \frac{4}{\pi}\Delta\overline{u}. \tag{2}$$

The effect dead time has on the output voltage can be calculated from the phasor diagram, Fig. 10. Using the law of

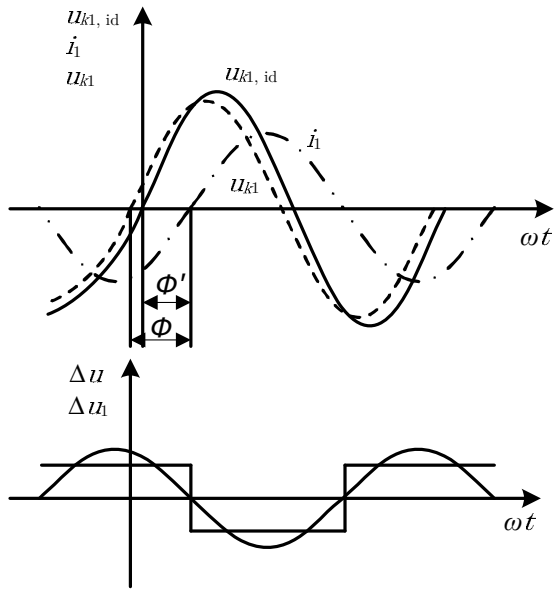


Fig. 9. Output voltage fundamental harmonic distortion due to dead time

cosines, the effect of dead time on the output voltage, normed to the output voltage fundamental harmonic, can be calculated as:

$$\frac{\hat{U}_1}{\hat{U}_{\mathrm{ref}}} = -\eta\cos\left(\phi\right) + \sqrt{1 - \eta\sin^2\left(\phi\right)}, \tag{3}$$

where η is the normed value of the voltage deviation fundamental harmonic:

$$\eta = \frac{8}{\pi}\frac{M\left(t_{\mathrm{d}} + t_{\mathrm{on}} - t_{\mathrm{off}}\right)}{T_{\mathrm{r}}m_{\mathrm{a}}}. \tag{4}$$

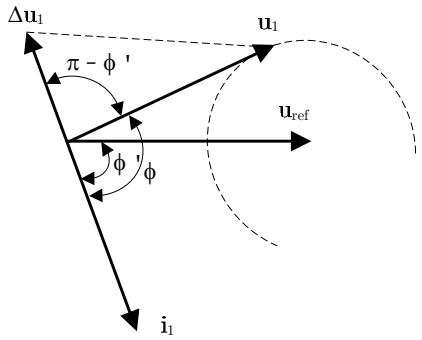


Fig. 10. Effect of the output voltage dead time

The effect dead time has on the converter output current's amplitude and phase shift was simulated via *Matlab* for various dead time and amplitude modulation ratio m_{a}. Model parameters are listed in Table V.

The simulation results are listed in Table VI and they show that the output current amplitude and phase deviations are larger for larger dead time values. This is contributed by the fact that the controllable voltage percentage drops when

TABLE V

DEADTIME EFFECT SIMULATION PARAMETERS

Parameter	Symbol	Value
DC source voltage	U_{DC}	750 V
Capacitance	C_{DC}	50 µF
Load resistance	R_{d}	1.2 Ω
Load inductance	L_{d}	1920 µH
Transistor voltage drops	U_{CEs}	2.9 V
Diode voltage drops	U_{D}	2.3 V
Simulation stop time	T_{end}	0.2 s
Step size	H	0.25 µs
Solver	-	ode1 (Euler)

TABLE VI

OUTPUT CURRENT FUNDAMENTAL HARMONIC AMPLITUDE AND PHASE DEVIATIONS DUE TO DEAD TIME EFFECTS

$\mathrm{m_a}$	0.2				
t_{D}	2 µs	4 µs	6 µs	8 µs	10 µs
ΔI_{a}	2.6 A	5.99 A	9.73 A	12.38 A	15.57 A
$\Delta\phi$	0.5°	3.5°	6.5°	7°	7.5°
$\mathrm{m_a}$	0.4				
t_{D}	2 µs	4 µs	6 µs	8 µs	10 µs
ΔI_{a}	2.35 A	4.66 A	6.86 A	9.9 A	12.94 A
$\Delta\phi$	0.5°	0.96°	1.46°	1.92°	2.4°
$\mathrm{m_a}$	0.6				
t_{D}	2 µs	4 µs	6 µs	8 µs	10 µs
ΔI_{a}	2.37 A	4.6 A	6.94 A	9.18 A	11.39 A
$\Delta\phi$	0.47°	0.9°	1.34°	1.82°	2.3°
$\mathrm{m_a}$	0.8				
t_{D}	2 µs	4 µs	6 µs	8 µs	10 µs
ΔI_{a}	2.85 A	5.67 A	8.54 A	11.2 A	14 A
$\Delta\phi$	0.3°	0.6°	0.9°	1.2°	1.5°
$\mathrm{m_a}$	1				
t_{D}	2 µs	4 µs	6 µs	8 µs	10 µs
ΔI_{a}	2.77 A	5.66 A	8.47 A	11.22 A	13.87 A
$\Delta\phi$	0.66°	0.95°	1.24°	1.53°	1.84°

dead time duration increases. They are also dependent on the amplitude modulation ratio. This fact was not predicted in the previous analysis, as the narrow impulse effect was neglected. Narrow impulses are present in areas where the modulation signal is near its greatest or lowest value, and around zero. When a narrow pulse is present, dead time lasts longer than the switch state, a situation present near the carrier signals' amplitudes. This influence is prominent for small or large amplitude modulation ratios, because of the higher narrow impulse content for such amplitude modulation ratios.

V. PROPOSED COMPENSATION METHOD

Since both semiconductor voltage drops and dead time influences are constant within one-half of the modulation signal and load current period, a simple modification to the basic modulation signal can be generated to compensate for these influences. This method is based on the so-called volt-second method, where the modulation signal is modified in such a way that the modified output voltage one-cycle average is equal to the ideal output voltage one-cycle average. This method is known from two-level converters, both for carrier-based modulation [11], [12], and for space-vector control [13]–[15].

Semiconductor voltage drops and dead time influences can be combined into one equation, depending on the modulation signal and current signs and on the modulation signal waveform. That equation can be simplified by neglecting the modulation signal waveform, due to its negligible contribution to the voltage deviation. These simplified influences are listed in Table VII.

TABLE VII

SIMPLIFIED OUTPUT VOLTAGE DEVIATIONS DUE TO SEMICONDUCTOR VOLTAGE DROPS AND DEAD TIME

u_{ref}	i_k	$\Delta\overline{u}_k$
+	+	$\frac{\mathrm{t_d}+\mathrm{t_{on}}-\mathrm{t_{off}}}{T_{\mathrm{s}}} \cdot \left(\frac{U_{\mathrm{DC}}}{2} - U_{\mathrm{CEs}} + U_{\mathrm{D}}\right) + \frac{U_{\mathrm{D}}+U_{\mathrm{CEs}}}{2}$
+	−	$-\frac{\mathrm{t_d}+\mathrm{t_{on}}-\mathrm{t_{off}}}{T_{\mathrm{s}}} \cdot \left(\frac{U_{\mathrm{DC}}}{2} - U_{\mathrm{CEs}} + U_{\mathrm{D}}\right) - \frac{U_{\mathrm{D}}+U_{\mathrm{CEs}}}{2}$
−	+	$\frac{\mathrm{t_d}+\mathrm{t_{on}}-\mathrm{t_{off}}}{T_{\mathrm{s}}} \cdot \left(\frac{U_{\mathrm{DC}}}{2} - U_{\mathrm{CEs}} + U_{\mathrm{D}}\right) + \frac{U_{\mathrm{D}}+U_{\mathrm{CEs}}}{2}$
−	−	$-\frac{\mathrm{t_d}+\mathrm{t_{on}}-\mathrm{t_{off}}}{T_{\mathrm{s}}} \cdot \left(\frac{U_{\mathrm{DC}}}{2} - U_{\mathrm{CEs}} + U_{\mathrm{D}}\right) - \frac{U_{\mathrm{D}}+U_{\mathrm{CEs}}}{2}$

If a compensation signal is superpositioned onto the original modulation signal, with regard to the modulation signal and output current signs, the influence listed in Table VII can be cancelled. This method requires very little computational power and a handful of measurements, and is therefore suitable for implementation in industrial systems.

VI. CONCLUSION

This paper presented the effects dead time and semiconductor voltage drops have on the three-level neutral point clamped converter output voltage. The influence on the one-cycle average output voltage value was analyzed analyticaly and the results were confirmed via simulations. Simulation results showed that different semiconductor voltage drops cause significant phase leg and line voltage DC components. Semiconductor voltage drop temperature dependence was analyzed and the results were also confirmed via simulation. Simulation results show that various semiconductor voltage drops cause a temperature-dependent output voltage DC component. The influence dead time has on the output current fundamental harmonic amplitude and phase shift was also simulated. Simulation results show that dead time lowers the amplitude of the output current fundamental harmonic, and that it introduces an additional phase shift. A compensation method was proposed to diminish the effects of dead time and semiconductor voltage drops, based on the conducted analysis.

REFERENCES

[1] J. Rodriguez, J.-S. Lai, and F. Z. Peng, "Multilevel inverters: a survey of topologies, controls, and applications," *Industrial Electronics, IEEE Transactions on*, vol. 49, no. 4, pp. 724–738, 2002.

[2] J. K. Steinke, "Switching frequency optimal pwm control of a three-level inverter," *Power Electronics, IEEE Transactions on*, vol. 7, no. 3, pp. 487–496, 1992.

[3] J. Pou, J. Zaragoza, P. Rodriguez, S. Ceballos, V. M. Sala, R. P. Burgos, and D. Boroyevich, "Fast-processing modulation strategy for the neutral-point-clamped converter with total elimination of low-frequency voltage oscillations in the neutral point," *Industrial Electronics, IEEE Transactions on*, vol. 54, no. 4, pp. 2288–2294, 2007.

[4] S. Kouro, K. Asfaw, R. Goldman, R. Snow, B. Wu, and J. Rodriguez, "Npc multilevel multistring topology for large scale grid connected photovoltaic systems," in *Power Electronics for Distributed Generation Systems (PEDG), 2010 2nd IEEE International Symposium on*, pp. 400–405, IEEE, 2010.

[5] S. Rivera, S. Kouro, B. Wu, J. Leon, J. Rodriguez, and L. Franquelo, "Cascaded h-bridge multilevel converter multistring topology for large scale photovoltaic systems," in *Industrial Electronics (ISIE), 2011 IEEE International Symposium on*, pp. 1837–1844, IEEE, 2011.

[6] S. Kouro, C. Fuentes, M. Perez, and J. Rodriguez, "Single dc-link cascaded h-bridge multilevel multistring photovoltaic energy conversion system with inherent balanced operation," in *IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society*, pp. 4998–5005, IEEE, 2012.

[7] X.-c. C. Jun Honda, *Application Note AN-1144: IRS20957S Functional Description*. International Rectifier, 2008.

[8] Y. Nakamura, H. Funato, and S. Ogasawara, "Compensation of output voltage distortion analysis of pwm inverter with lc filter caused by device voltage drop," in *Power Electronics and Applications, 2007 European Conference on*, pp. 1–9, IEEE, 2007.

[9] M. Macan, I. Bahun, and Ž. Jakopović, "Output dc voltage elimination in pwm converters for railway applications," in *Conference on Electrical Drives and Power Electronics, EDPE2011 The High Tatras*, 2011.

[10] A. Nabae, I. Takahashi, and H. Akagi, "A new neutral-point-clamped PWM inverter," *Industry Applications, IEEE Transactions on*, no. 5, pp. 518–523, 1981.

[11] D. Leggate and R. J. Kerkman, "Pulse based dead time compensator for pwm voltage inverters," in *Industrial Electronics, Control, and Instrumentation, 1995., Proceedings of the 1995 IEEE IECON 21st International Conference on*, vol. 1, pp. 474–481, IEEE, 1995.

[12] A. R. Munoz and T. A. Lipo, "On-line dead-time compensation technique for open-loop pwm-vsi drives," *Power Electronics, IEEE Transactions on*, vol. 14, no. 4, pp. 683–689, 1999.

[13] T. Sukegawa, K. Kamiyama, T. Matsui, and T. Okuyama, "Fully digital, vector-controlled pwm vsi-fed ac drives with an inverter dead-time compensation strategy," in *Industry Applications Society Annual Meeting, 1988., Conference Record of the 1988 IEEE*, pp. 463–469, IEEE, 1988.

[14] W. S. Oh, "Dead-time compensation of a current controlled inverter using the space vector modulation method," *International journal of electronics*, vol. 80, no. 2, pp. 277–289, 1996.

[15] J.-W. Choi and S.-K. Sul, "Inverter output voltage synthesis using novel dead time compensation," *Power Electronics, IEEE Transactions on*, vol. 11, no. 2, pp. 221–227, 1996.

Model of three level neutral point clamped converter (NPC) for grid connected photovoltaic systems

Mateja Novak, Viktor Šunde and Željko Jakopović
Faculty of Electrical Engineering and Computing/Department of Electrical Machines, Drives and Automation, Zagreb, Croatia

Abstract - Photovoltaic modules as direct converters of solar energy into electrical energy are typically connected to the power grid trough standard two level converters. A photovoltaics system can be connected to the power grid if it can satisfy the IEEE standardized conditions set by the power grid operator. The conditions include limitations of THD factor and current harmonics. Multilevel converters which have a lower THD factor than two level converters can satisfy these conditions and in the same time work at higher switching frequencies lowering the losses and the size of the output filter. The model implemented in Simulink® MATLAB consists of a photovoltaic array, boost dc-dc converter with MPPT control and three level neutral point clamped converter connected to the power grid. The simulation results are presented, analyzed and compared to the standard model with a two level converter to validate the benefits of the usage of multilevel converters for grid connected photovoltaic systems.

I. INTRODUCTION

The world energy consumption by the year 2050 is projected to reach 30 terawatts (TW). In order to stabilize the concentration of CO_2 in the atmosphere, about 20 TW would have to be generated by the non-CO_2 energy plants [1]. According to renewables 2014 global status report the capacity of installed solar power has already reached 139 GW [2]. As a distributed energy source it presents a challenge for the grid operator in terms of the availability as well as energy quality.

Typically the solar plants are connected to the grid by standard 2-level converters. The drawback of connecting the solar plants trough high frequency converters is injection of low level current harmonics to the grid and increasing the total harmonic distortion (THD) factor. To limit the grid pollution, standardized conditions defined by the IEEE and IEC which include THDV factor limitation to maximum of 5% and current harmonics limitations have to be met in order to obtain a permit from the grid operator to connect the solar power plant to the public grid. [3] Multilevel converters which have a lower THD factor then standard two level converters can satisfy these conditions and in the same time provide many benefits that can help in raising the efficiency of the energy conversion lowering the switching losses by working at higher switching frequencies and allowing the usage of smaller output filters [5][7][8]. Semiconductor manufacturers have recognized these values and are already offering complete phase-leg modules for various multilevel converter topologies [4][6].

Multilevel converter topology chosen for this simulation model is a standard three level neutral point clamped (NPC) topology, Fig. 1. NPC converter topology is currently mostly used for large variable speed drives and wind turbines in medium and high voltage range to improve the conversion efficiency.

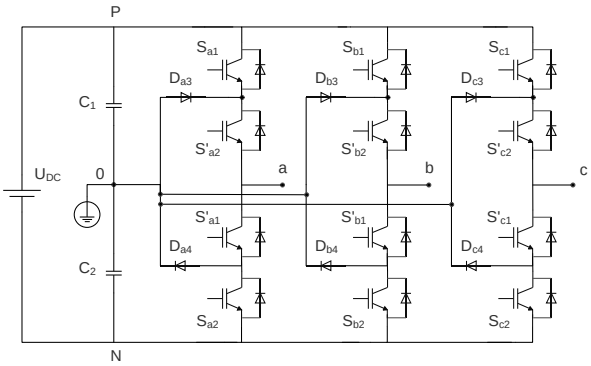


Figure 1. Three level NPC topology

As the losses in energy transport are proportional to square value of the current it is logical that the transport by higher voltage will provide the reduction of energy losses. For large scale photovoltaic power plants the dc link voltage reached the maximum voltage class of the semiconductor switches. In standard 2-level voltage source converters the voltage class of semiconductor switches has to be slightly higher than the dc-link voltage. Semiconductor switches in NPC topology bear half the voltage of 2-level, therefore the rated voltage of devices used is half of that in 2-level permitting the usage of NPC with higher dc-link voltage. A higher dc-link voltage also means that the output voltage of the converter will be higher.

To study the behavior of multilevel converters for grid connected PV systems in terms of energy quality and transmission efficiency a simulation model was implemented in Simulink ® MATLAB. The model consists of a photovoltaic array, boost dc-dc converter with MPPT control unit and grid connected 3-level NPC converter, Fig 2.

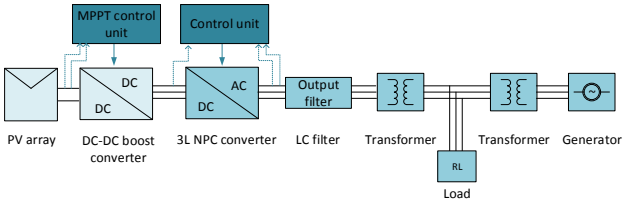


Figure 2. Model layout

II. SIMULATION MODEL

In subsections of this chapter a brief guidance is given for choosing the model parameters along with selected parameter values used in the simulation model presented in Fig.3.

A. *PV array*

Photovoltaic array consists out of 520 photovoltaic modules with rated nominal power of 305 W_p connected in 40 strings. Module characteristics are extracted from NREL System Advisor Model [9].To simplify the model temperature dependence of the PV array is not evaluated in the model. PV Module/Array parameters are given in Table I. Sampling times of current and voltage measurements of the PV array are T_S=1μs.

TABLE I. PV MODULE/ARRAY PARAMETERS

Parameter	Module	Array
V_{OC}	64.2 V	834.6 V
I_{SC}	5.96 A	238.4 A
V_{MPP}	54.7 V	711.1 V
I_{MPP}	5.58 A	223.2 A

B. *DC-DC boost converter with MPPT control*

In order to draw the maximum out of available solar power the duty cycle of the IGBT switch in the dc-dc boost converter is calculated by the MPPT (Maximum Power Point Tracker) control unit. Input parameters of the unit: PV array current, voltage and power measurements are calculated as average values over 1000 samples making the sample time of the MPPT unit T_S=1ms.

The implemented MPPT algorithm is based on the Incremental Conductance (IC) algorithm with a variable duty cycle increment as presented in Fig. 4. [10].
Variable increment (δ_V) is calculated as (1) if $\delta_V < \delta_{max}$:

$$\delta_v = N \cdot \left| \frac{dP}{dV} \right|_{\delta max} \quad (1)$$

Parameter N is chosen to satisfy (2) for a fixed step IC algorithm with maximum duty cycle increment.

$$N = \frac{\delta_{max}}{\left| \frac{dP}{dV} \right|_{\delta max}} \quad (2)$$

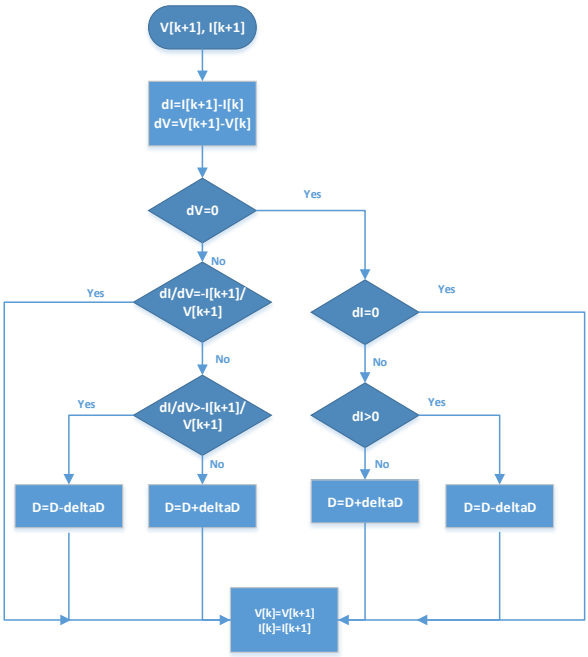


Figure 4. Incremental conductance algorithm

For maximum irradiation of 1kW/m^2 average value of duty cycle is 0.29. Switching frequency of the converter is set to 10 kHz. Parameters of the DC-DC boost converter are given in Table II.

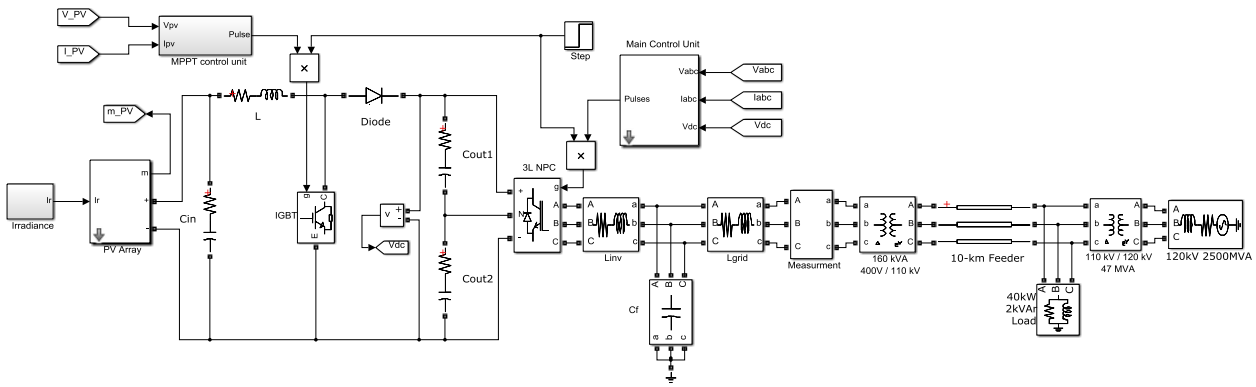


Figure 3. Simulation model in Simulink® MATLAB

TABLE II. DC-DC BOOST CONVERTER PARAMETERS

Parameter	Specification
C_{in}, R_c	100 uF, 1 mΩ
L, R_L	0.5 mH, 5 mΩ
δ_{max}	0.01
f_{PWM}	10 kHz
C_{out1}, C_{out2}	10 mF

C. NPC converter and control unit

Initial values of the dc-link capacitors were chosen as proposed in [11] and later adjusted to get the desired behavior of dc-link voltage during simulation. For the 160 kVA NPC converter nominal dc-link voltage V_{dc} is set to 1000 V, line rms output voltage is 400 V with nominal frequency of 50 Hz.

Control unit of the NPC converter includes a phase locked loop (PPL) block, current and DC voltage PI regulators, voltage reference generating block and PWM pulse generator. The d-axis current component reference is defined as output value of DC voltage regulator while the q-axis component reference is set to zero in order to maintain unity power factor. PWM pulse generator uses two triangular amplitude shifted 8 kHz (f_{sw}) carrier signals.

D. Grid connection

Output filter of the NPC converter is a wye connected LCL filter, Fig 5. designed as described in [12].

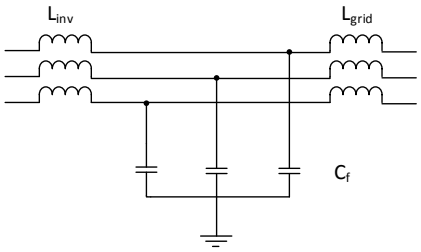


Figure 5. Wye connected LCL filter topology

To find the suitable values of filter parameters it is necessary to calculate the base impedance (Z_B) (3) and capacitance (C_B) (4). The voltage value used in equation is line rms value; frequency ω_N is nominal output frequency of the NPC converter and S nominal power of the converter in VA.

$$Z_B = \frac{3 \cdot U_{Lrms}}{S} \tag{3}$$

$$C_B = \frac{1}{\omega_N \cdot Z_B} \tag{4}$$

Maximum value of filter capacitance is 5% of the base capacitance because it is considered that the maximum power factor variation see by the grid is 5%.

$$C_f = 0.05 \cdot C_B \tag{5}$$

Inverter inductance is calculated for the worst case scenario when the duty cycle of the dc-dc boost converter is 0.5 (6). The inductor should be able to limit the ripple of output current to 10% of the rated amplitude value (7).

$$L_{inv} = \frac{V_{dc}}{16 \cdot f_{sw} \cdot \Delta I_{out,max}} \tag{6}$$

$$\Delta I_{out,max} = 0.1 \frac{P_n \sqrt{2}}{3 \cdot V_{PHrms}} \tag{7}$$

Relation factor *r* between the grid inductance and inverter (7) is defining the ripple attenuation and is usually set to 0.5.

$$L_{grid} = r \cdot L_{inv} \tag{8}$$

Calculated parameters for the LCL filter of the simulation model are presented in Table III.

TABLE III. LCL FILTER PARAMETERS

Parameter	Specification
L_{inv}, R_L	0.24 mH, 1 mΩ
L_{grid}, R_L	0.12 mH, 1 mΩ
C_f	50uF

The converter is connected through transformer to the 110 kV grid. A 10 km feeder represents the transmission grid connecting the generator.

III. SIMULATION RESULTS AND ANALYSIS

Simulation is performed for two irradiance curves representing different weather conditions. The half sine curve represents an ideal sunny day with maximum 1kW/m^2 irradiance peak, Fig. 6. To simplify and accelerate the simulation the average 12h of available daylight is scaled to fit the simulation time of 1s. Second curve represents rapidly changing irradiance as a result of cloud shading, Fig 7.

On both Figures can be seen that with the help of implemented MPPT algorithm the power output curve follows the irradiance change. The output voltage remains constant with peak phase to ground value of 320 V.

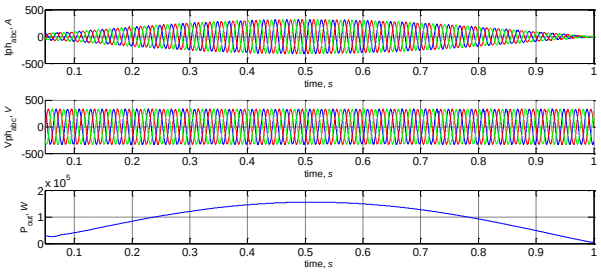


Figure 6. Output current, voltage and power for ideal irradiance curve

DC-link voltage for both irradiance curves is similar, Fig 8. During the first 1ms the capacitors are being charged without regulation of the main control unit.

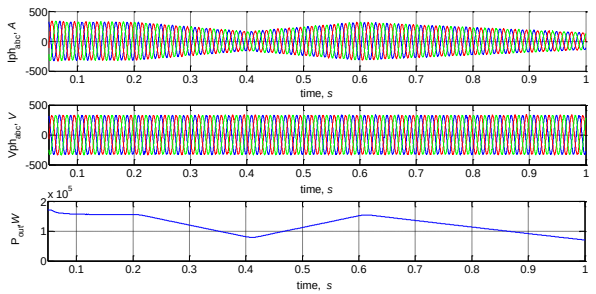


Figure 7. Output current, voltage and power for changing irradiance curve

After the main and MPPT control unit are enabled the dc-link voltage drops to the value of reference voltage. The voltage ripple remains within ±1% of the reference link voltage.

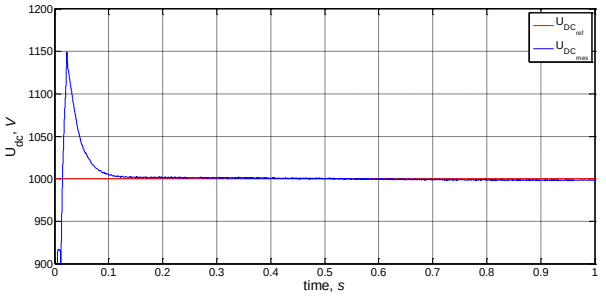


Figure 8. Simulated DC link voltage

Created model also copes very well with more extreme irradiance changes. Daily irradiance data uploaded on Sunny portal for a public available plant in Croatia was fitted on a 1 second scale and set as input irradiance variable for the simulation model [13].

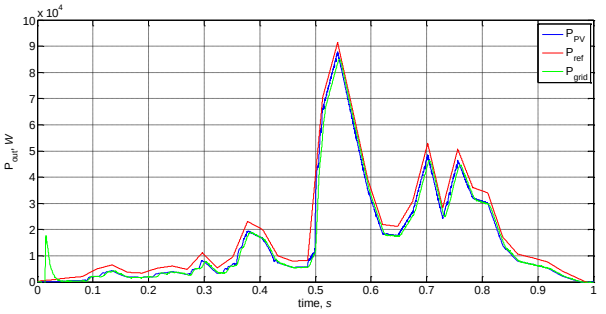


Figure 9. Simulated output power for extreme irradiance change

The results presented in Fig. 9 show the reference power obtained from the irradiance data (P_{ref}), PV power measured before the NPC converter (P_{pv}) and the power transmitted to the grid (P_{grid}). Even under difficult conditions the MPPT algorithm managed to draw out nearly the maximum of the available power.

IV. TWO LEVEL VSC AND THREE LEVEL NPC COMPARISON

In order to show the benefits of using 3-level NPC converter on higher switching frequencies for connecting the PV systems a comparison with an equally specified 2-level voltage source converter (VSC) model has been made. The primary focus of the comparison as stated in the introduction is the current distortion expressed through TDD (Total Demand Distortion) factor of the current waveform. TDD represents the ratio of the root-sum-square value of the harmonic current to the maximum demand load current hence at rated load, in this case maximum constant irradiation level, current TDD and THD are the same value. [3] For following switching frequencies 2, 4 and 8 kHz results of the simulation are presented in Table IV. During the simulation, irradiance level is set at constant value of 1kW/m^2. While performing FFT analysis over 10 cycle sample of the current signals, the sampling time of model was reduced to 0.1µs.

TABLE IV. CONVERTER OUTPUT CURRENT THD FACTOR

Carrier frequency	2-level VSC before filter	2-level VSC after filter	3-level NPC before filter	3-level NPC after filter
2 kHz	7.03%	4.97%	3.54%	2.6%
4 kHz	5.51%	2.34%	3.03%	1.79%
8 kHz	4.93%	1.35%	2.47%	0.72%

Filters for each frequency were designed as described in section II.D. It also has to be taken into account that a certain amount of current ripple has been produced through the boost dc-dc MPPT controlled converter.

When raising the switching frequency it can be seen that the usage of 3-level NPC converter produces less current distortion and can be used on higher switching frequencies and stay under 5% THD factor limit.

Additionally to prove that a reduction of filter size without a huge impact on the current distortion is possible for NPC converter, the simulation was also run for 4 kHz switching frequency but with a filter half the size as used for simulation in Table IV. Fig. 10 shows the result of FFT analysis of converter output current for 4 kHz switching frequency with smaller filter.

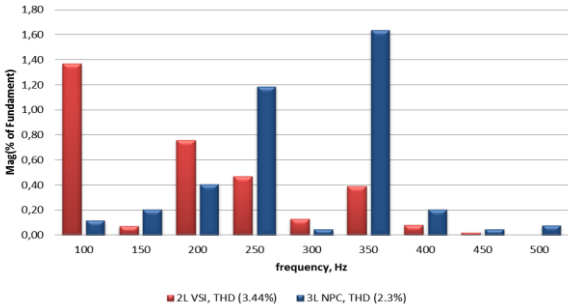


Figure 10. FFT analysis of I_{out}

The THD factor for the NPC converter output current was 2.3%, whereas for the 2-level VSC it was 3.44% making the difference between the two converters with smaller filters two times higher than when using larger filters in favor of the NPC converter.

V. CONLUSION

Developed model of NPC converter for grid connected PV systems showed that NPC converter performs well on high switching frequencies. The NPC converter output

current distortion was almost two times lower than when using standard 2-level VSC converter. It was shown that even with the reduced filter size the current distortion of the NPC is below the 5% THD factor limit.

In order to expand the analysis and comparison with 2-level converters the next step of this simulation model development will be to include switching losses and verify trough simulation results the thesis of reducing the switching losses and raising the conversion efficiency by usage of multilevel converters.

ACKNOWLEDGMENT

This work has been partly supported by the European Regional Development Fund under the project „Advanced technologies in power plants and rail vehicles“.

REFERENCES

[1] Zweibel K., “The Terawatt challenge for thin film PV”, National Renewable Energy Laboratory, August 2005, pp. 18–22.

[2] REN21. 2014. “Renewables 2014 Global Status Report”, pp 47 – 49

[3] IEEE Power and Energy Society, IEEE Std 519™ – 2014 Recommended practice and Requirements for Harmonic Control in Electric Power Systems

[4] Spenger M., Alvarez R., Staudt I., Bernet S., “Characterization of a New 1.2 kV IGBT 3L-NPC Phase-leg Module for Low voltage applications”, Proceedings of the 14th European Conference on Power Electronics and Applications (EPE 2011), Birmingham, August, 2011,pp. 1–10

[5] Chen Z., Yuan L., Zhao Z., Sun X., “Power losses in two- and three-level three phase photovoltaic inverters equipped with IGBTs”, 15th International Conference on Electrical Machines and Systems (ICEMS), Sapporo, October 2012., pp 1–6

[6] Zanin A., SKiM® IGBT Modules – Technical Explanations, July 2014

[7] Schöner C., Derix D., Hensel A., “Comparison and evalution of different three-level inverter topologies for PV systems”, 16th European Conference on Power Electronics and Applications, Lappeenranta, August 2014, pp 1-10

[8] Fujii K., Kikuchi T., Koubayashi H., Yoda K., “1-MW advanced T-type NPC Converters for Solar Power Generation System”, 15-th European Conference on Power Electronics and Applications (EPE), Lille, Sepetember 2013, pp 1-10

[9] http://www.mathworks.com/examples/simpower/156-detailed-model-of-a-100-kw-grid-connected-pv-array, Sybille P.G.G., Hydro-Quebec Research Institute (IREQ), Osorio C., Chandrachood S., The Mathworks Copyright (c) 2012

[10] Liu F., Duan S., Liu F., Liu B., Kang Y., “A Variable Step Size INC MPPT Method for PV Systems”, IEEE Transactions on industrial electronics, vol. 55, no. 7, July 2008, pp 2622-28

[11] Bueno Pena E. J., “Optimization of the behaviour of a NPC Three-level converter connected to the grid”, PhD Thesis, University of Alcala, 2005, pp 23-26

[12] Reznik A. , Simões M.Godoy , Al-Durra A., Muyeen S. M., “LCL Filter Design and Performance Analysis for Grid Interconnected Systems”, IEEE Transactions on Industry Applications, vol.50, no.2, July 2013, pp 1225-32

[13] Sunny Portal, Meteorological data for SE SOLVIS Varazdin, Croatia,https://www.sunnyportal.com/Templates/PublicPageOverview.asp

Fast development of a DSP-based control of three-phase four-leg inverter for auxiliary power supply converters using hardware-in-the-loop

Igor Musulin*, Josip Babić* and Željko Jakopović**
* Končar–Electrical Engineering Institute Inc., Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
igor.musulin@koncar-institut.hr

Abstract - This paper presents an easy and relatively fast way to design and implement control system of three-phase four-leg inverter in DSP (digital signal processor). The inverter model is emulated using hardware-in-the-loop simulator and control system is coded in TI DSP identical to the target system using graphically based block-diagram programming environment. A multi-loop cascade control is applied and two control strategies are implemented: i) the PI (proportional-integral) controllers with symmetrical sequence current and voltage decomposition and ii) the PIR (proportional-integral-resonant) controllers. Experimental results with symmetrical and single-phase loads are presented. The control algorithms are evaluated based on the execution time. Results show that PIR based control executes roughly three times faster that PI control.

I. INTRODUCTION

When designing new power electronics converter two different models are often used: time discrete model and average model. In time discrete model current and voltage waveforms are mathematically represented for each interval. This model is mainly used for detailed converter analysis. However, mathematical expressions can sometimes become complicated and, since the model is nonlinear, it is not suitable for the development of the control algorithm. For this reason, a linear frequency domain model is used and an averaging is done to get an average model. A control design engineer will usually spend a lot of time and work hours on the development, testing and optimization of a new design and/or algorithm while working with one of the previously mentioned off-line models. Furthermore, extensive experiments and tests on target converter, which can often be time and money consuming, can lead to new problems.

One promising alternative is using a digital real-time simulator that is capable of emulating power electronics hardware. Hardware-in-the-loop (HIL) simulation enables testing of control software and hardware for power electronics by replacing the real electrical system with a real-time emulator. Analog and digital signals are replaced and emulated using a virtual model. This accelerates the development of control algorithms. Some of the advantages of using hardware-in-the-loop simulators include safe testing environment, ability to test the full operating range, including failures test, repeatable testing, flexible reconfiguration and automated testing [1]. Hardware-in-the-loop simulators are already used in industries like automotive, aerospace, power systems, power quality (FACTS) and energy transfer (HVDC), but have not yet been extensively used for the development of the power electronic systems. This is mainly due to the inability to accurately simulate fast voltage-source converters with high-frequency PWM controls and large number of switches [2]. However, with the rapid development of fast field-programmable gate array (FPGA) chips, a wide range of commercially available HIL devices with low latency are available.

This paper is organized as follows. Chapter II introduces the problem of non-symmetrical loads in auxiliary power supply converters in railway vehicles. Three inverter topologies are presented, two of which are today considered as standard solutions in auxiliary power supply converters in railway vehicles. The last one, a three-phase four-leg inverter, with its modulation technique and control strategy is described in more details in Chapter III and Chapter IV. DSP and HIL implementation are described in Chapter IV. Simulation and experimental results, including the measurement of execution time, (Chapter V) and conclusion (Chapter VI) are given at the end.

II. NON-SYMMETRICAL LOADS IN POWER SYSTEM

A recent survey has shown that today non-linear loads, e.g. rectifiers, frequency converters, arc furnaces, controlled motor drives, etc. are approaching between 75 ÷ 90 % of the loading in electric utility distribution systems [3]. Non-symmetrical loads, e.g. single- or two-phase loads in a three-phase system lead to current and voltage unbalance. A large number of non-linear and non-symmetrical loads generate harmonics, cause increase of neutral currents, overheating of drives and equipment, reduction in the effective torque, etc.

While the problem is not so evident in low voltage electric grid due to somewhat evenly distributed single-phase loads, the influence of non-symmetrical loads is more apparent in isolated systems like uninterruptible power supply, distributed renewable energy sources operating in island mode and auxiliary power supply for ships, airplanes and railway coaches. For example, three-phase output inverter in an auxiliary power supply for trains

must generate three-phase and single-phase output with grounded neutral wire. The three-phase output is used for supplying air-conditioning units and auxiliary inverter, main inverter and transformer fans. Single-phase output supplies evaporators and condenser blower fans, lighting and 230 V general purpose power sockets for passenger use.

Single-phase output can be obtained from a three-phase inverter by:

- an additional single-phase 400/230 V or delta/wye three-phase transformer,
- a split dc-linked capacitors and tying the neutral grounded point to the midpoint of the dc-linked capacitors or
- a four-leg inverter topology and tying the neutral grounded point to the mid-point of the fourth leg.

Solution with a transformer is the most widespread, mainly due to its simplicity and minimal changes in design and control algorithm. Galvanic isolation and elimination of zero currents are considered advantages, while transformer price and weight are sometimes seen as disadvantages, mostly in applications like train or tram converters where the total mass of converter plays a significant role. Second solution, if rearranged actually consists of three single-phase half bridge inverters. This enables the independent control of the each leg, i.e. phase. In case of unbalanced loads, neutral current flows through the dc capacitors. Voltage oscillations across capacitors are directly influenced by neutral current [4]. Therefore, high value capacitors or additional balancing circuits must be used for voltage balancing.

III. THREE-PHASE FOUR-LEG INVERTER TOPOLOGY AND MODULATION

A three-phase four-leg inverter topology show in Fig. 1, although first mentioned in the literature in the early 90is, has not been extensively used in the commercial applications. The neutral wire is connected to the neutral point of the added fourth leg. Among main advantages are higher utilization of dc voltage (more than 15 % compared to conventional three-phase three-leg inverter) and neutral current flows through the fourth leg and not through the dc capacitors, thus minimizing dc input voltage oscillations. Complex modulation scheme and large number of semiconductor devices can be regarded as a disadvantage. Nevertheless, this type of inverter presents an ideal solution for four-wire systems for supplying non-symmetrical and single-phase loads since it offers independent control of currents of each phase.

Two modulation strategies for three-phase four-leg inverters operating at constant switching frequency are often used:

- carrier-based pulse-width modulation (PWM), and
- three dimensional (3D) space vector (SV) PWM

A 3D SVPWM is preferable for digital implementation, but requires faster processors since sector determination and switch ON time calculation can take up a lot of processor calculation time and lead to digital delay. On the other hand, carrier-based PWM algorithm [5–10] is easy to implement since it uses only the comparison of reference waveform with the triangular waveform. This modulation technique is similar to sinusoidal PWM where switching states are determined by comparing triangular carrier and three sinus reference waveforms. The only difference is in the reference waveform generation where additional, so-called, offset voltage is added to sinusoidal waveform. Best voltage synthesis, e.g. utilization of dc bus voltage and low harmonic generation can be achieved by operating in the linear modulation region and selecting offset voltage as shown in (1).

$$V_{n0}{}^{*} = \begin{cases} -\dfrac{V_{max}}{2}, & V_{min} > 0 \\ -\dfrac{V_{min}}{2}, & V_{max} < 0 \\ -\dfrac{V_{max} + V_{min}}{2}, & \text{otherwise} \end{cases} \quad (1)$$

where V_{min} = min (V_{an}, V_{bn}, V_{cn}) and V_{max} = max (V_{an}, V_{bn}, V_{cn}).

IV. DESIGN OF CONTROL SYSTEM

A multi-loop cascade control strategy is adopted in this paper. An outer voltage control loop is used for voltage reference tracking. Its output acts as a reference for the current control loop. Current control loop is used for protection of the inverter against overload and for the improvement of the control system response. Three-phase four-leg inverter control is based on dq0 coordinate transformation. However, with unbalanced load d and q channel voltage and current oscillate with frequency 2ω, while 0 channel voltage and current oscillates with ω. Therefore the control bandwidth should be high enough to attenuate disturbance caused by the unbalanced load. On the other

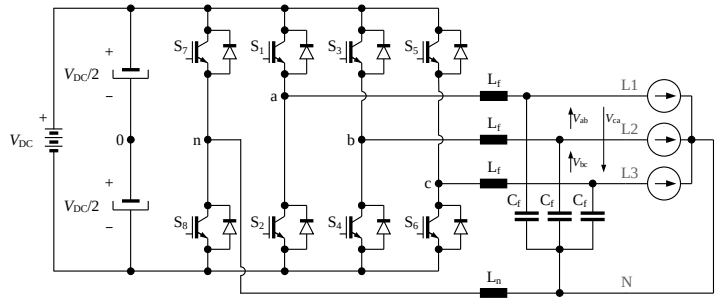


Figure 1. Three-phase four-leg inverter topology

hand, high bandwidth can affect system stability. Furthermore, since control values are oscillating classical PI controller cannot achieve zero steady-state error.

As a solution a transformation of current and voltage waveforms to the symmetrical components, i.e. sequence decomposition is proposed in [11–14]. From there, using abc/dq coordinate transformation and standard PI controllers, one can easily control positive, negative and zero component of a (now) symmetrical system. Instantaneous positive, negative and zero sequence are calculated using (2). Time domain implementation of an operator a ($= -1/2 + j\ \sqrt{3}/2$) is shown in Fig. 2. It can be seen that a 90 ° phase-shift associated with j is needed. This is done using simple second-order low-pass filter tuned at the input frequency ω_{in} with a damping factor $\xi = 1$ as shown in (3) and further explained in [15]. Time domain implementation of a second-order low-pass filter (*LPF*) is shown in Fig. 3. The operator a^2 can be implemented in a similar manner. Subsequently, sequence composition (4) and dq/abc transformation are used to get V_a, V_b and V_c for PWM modulation, Fig 4

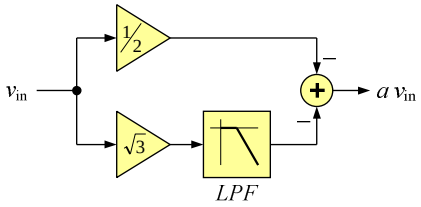


Figure 2. Implementation of operator a in time domain

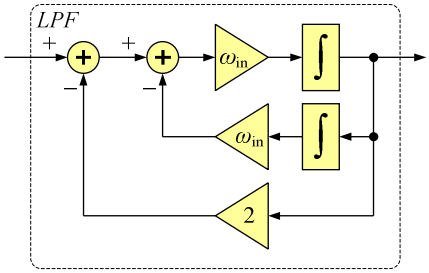


Figure 3. Implementation of *LPF* in time domain

$$v_{\text{abc positive}} = \begin{pmatrix} 1 & a & a^2 \\ a^2 & 1 & a \\ a & a^2 & 1 \end{pmatrix} \begin{pmatrix} v_a & v_b & v_c \end{pmatrix}^T$$

$$v_{\text{abc negative}} = \begin{pmatrix} 1 & a^2 & a \\ a & 1 & a^2 \\ a^2 & a & 1 \end{pmatrix} \begin{pmatrix} v_a & v_b & v_c \end{pmatrix}^T \quad (2)$$

$$v_{\text{abc zero}} = \begin{pmatrix} 1 & 1 & 1 \\ a & a & a \\ a^2 & a^2 & a^2 \end{pmatrix} \begin{pmatrix} v_a & v_b & v_c \end{pmatrix}^T$$

$$LPF = \frac{\omega_{in}^2}{(s+\omega_{in})^2} \quad (3)$$

$$v_{abc} = \begin{pmatrix} 1 & 1 & 1 \\ a & a & a \\ a^2 & a^2 & a^2 \end{pmatrix} \begin{pmatrix} v_{a\,positive} & v_{a\,negative} & v_{a\,zero} \end{pmatrix}^T \quad (4)$$

Another solution is using PR (Proportional-Resonant) or PIR (Proportional-Integral-Resonant) controllers for control [9], [13], [16–19]. PIR controller is represented by (5). Resonant based controllers have an infinitive gain at resonant frequency ω_0, the controller can successfully eliminate steady-state error when control values are oscil-

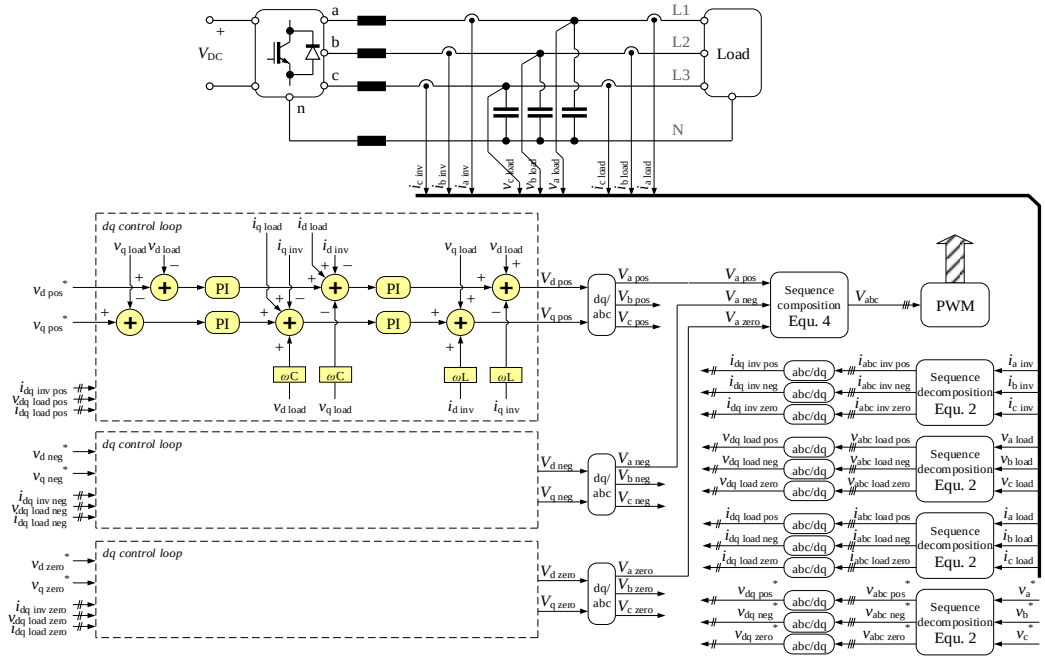


Figure 4. Three-phase four-leg inverter topology

lating with ω_0. Another advantage of using resonant based controllers is the simple control strategy, no additional coordinate transformations and fastest time of execution. The last point will be further explored in Chapter VI. It is important to note that abc/dq0 coordinate transformation must be used here.

$$G_c(s) = K_P + K_I \frac{1}{s} + K_R \frac{s}{s^2 + \omega_0^2} \quad (5)$$

where $\omega_0 = 2\omega$ for d and q frame and $\omega_0 = \omega$ for 0 frame.

V. DSP AND HARDWARE-IN-THE-LOOP IMPLEMENTATION

The control algorithm is coded using GRAP, graphical block-diagram based programming environment [20], so there is no need for textual coding and the programming is done relatively fast. Previously developed control algorithm in Matlab/Simulink environment is easily transferred using corresponding optimized block elements.

The code itself is implemented in Texas Instruments TMS320F28335 controlCARD with operating core frequency of 150 MHz. Three different control tasks are used, namely CT1, CT2 and CT3. The fastest control task is set to 20 kHz and it is used for acquisition, A/D conversion and PWM waveform generation. The second task (at 5 kHz) is used for sequence (de) composition, control sequence, and communication while the slowest task (at 500 Hz) is used for monitoring and diagnostics.

Using the HIL simulator one can easily implement any desired topology, as it was done with the three-phase four-leg inverter topology. A total of four voltage measurements (input dc voltage and three output ac phase voltages) and six current measurements (three inverter line currents and three load line currents) are used for control. Additional current and voltage measurements can also be used. HIL simulation time step is set to 1 μs. Experimental setup is shown in Fig. 5.

VI. HARDWARE-IN-THE-LOOP SIMULATION AND EXPERIMENTAL RESULTS

Two control strategies from Chapter IV are used: PI control with coordinate transformation and PIR control. An input dc voltage is set to 700 V while output reference ac phase voltage is set to 230 V. The inverter switching frequency is 3 kHz. Maximum three-phase output power P_{max} is set to 24 kW with cos φ = 1 while minimum power is set to P_{min} 500 W with cos φ = 1 and corresponds to the power of the main and auxiliary converter fans.

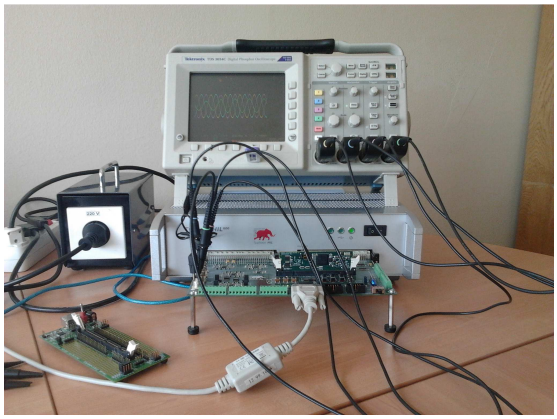


Figure 5. HIL experimental setup with controller and interface board

In Fig. 6 and 8 an output voltage and current response in the case of switching a 12 kW ideal resistive single-phase load into already symmetrically loaded three-phase output ($P_{three\text{-}phase} = P_{max}$) is shown. Dashed vertical line represents the start of transition. Both algorithms exhibit the similar performance regarding the settling time. For the first controls strategy, before switching a single-phase loads negative and zero sequence are equal to zero or approximately zero due to non-ideal acquisition and noise in measured channels. In the second algorithm, resonant part is equal to zero and PIR controller behaves as PI. Zero (0) part of abc/dq0 transformation is also equal or approximately zero.

Output voltage and current response for a switching of a 24 kW ideal resistive single-phase load from a minimum load ($P_{three\text{-}phase} = P_{min}$) is shown for two control strategies in Fig. 7 and Fig 9. It can be seen that both algorithms

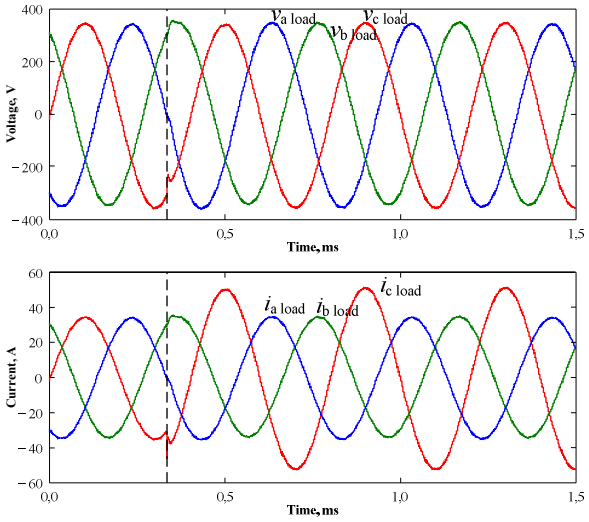


Figure 6. PI control; $P_{three\text{-}phase} = P_{max}$

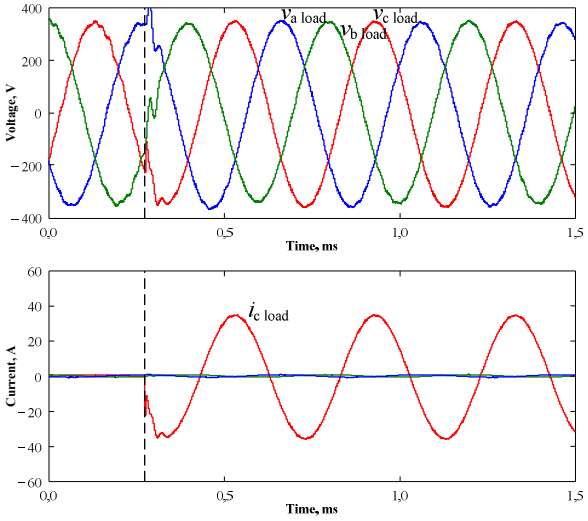


Figure 7. PI control; $P_{three\text{-}phase} = P_{min}$

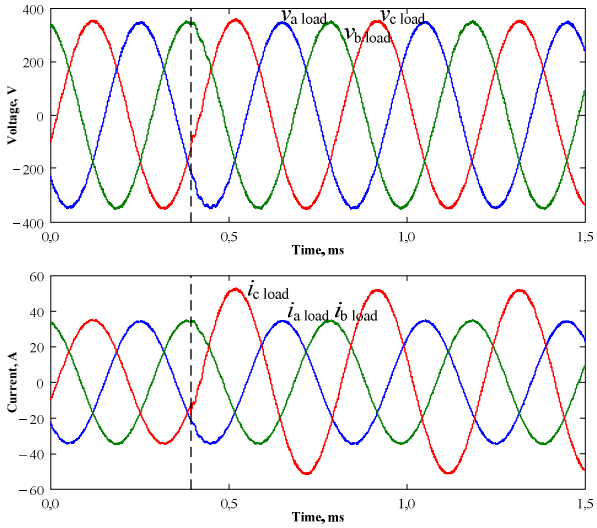


Figure 8. PIR control; $P_{three\text{-}phase} = P_{max}$

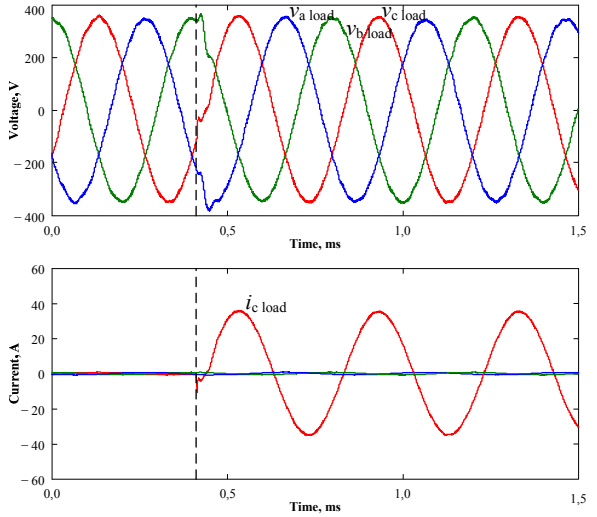


Figure 9. PIR control; $P_{three\text{-}phase} = P_{min}$

behave similar response and produce symmetrical output voltage in case of unsymmetrical loads. The response time depends on the controller parameters. While there is certainly a room for the further improvement, three-phase output response in an auxiliary power supply for trams and trains generally does not need to be a fast one. Loads are quite robust and the settling time of up to two periods is acceptable.

The identical waveforms can be seen on the HIL analog outputs, Fig. 10 and Fig 11. HIL input/output is ± 5 V so appropriate scaling must be done. In this case, voltage and current scaling factor is set to 100 V = = 1 V and 100 A = = 1 V, respectively.

Many papers dealing with the control of a three-phase four-leg inverter using coordinate transformation refer to complex control structure as an obvious disadvantage over PR or PIR based control. In this paper, apart from testing different control strategies, the execution time is measured for each one. Five distinctive parts of control algorithm are selected: sequence decomposition, sequence composition, coordinate transformation (abc/dq(0)and dq(0)/abc) and control. Other parts of the application program like system definition, A/D conversion, scaling, PWM and recorder are the same for each algorithm and are not measured. Execution times are given in Table I.

PI based control strategy (Fig. 5) has four sequence decomposition calculations (12,72 μs × 4 = 50,88 μs) and one sequence composition calculation. In Table I it can be seen that the sequence (de) composition takes roughly 65 % of total execution time. There are also twelve abc/dq and three dq/abc coordinate transformations per one task. Although single abc/dq transformation last only 0,78 μs compared to 1,21 μs for abc/dq0, a total coordinate transformation time is two times larger due to number of calculations. Also, control is executed three times per task, once for each system, i.e. positive, negative and zero. It is interesting to note that the control itself is executed faster (18,36 μs compared to 19,37 μs). This leads to conclusion that digital implementation of resonant part of PIR controller used in this example takes a lot of computing time. For PIR control strategy there is no sequence (de) composition and coordinate transformation is executed just five times (4 × abc/dq0 + 1 × dq0/abc) per task. The control is done only once per task, but there are three frames to control (d, q and 0), opposite to previous case where there are only two (d and q frame).

Comparing the total execution time for PI and PIR based control strategies, from Table I it can be seen that PIR executes roughly 60 % faster than PI. Since it exe-

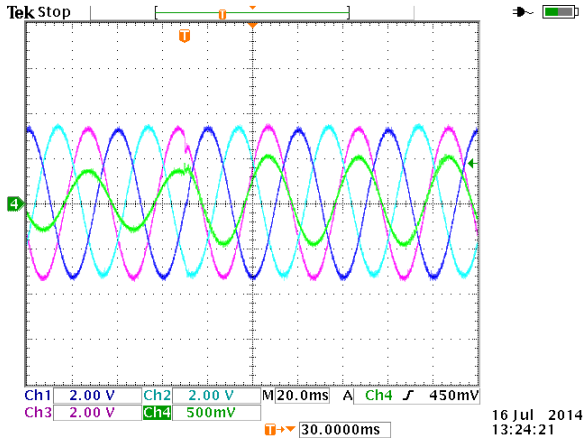


Figure 10. Voltage and current (Ch4: $i_{c\ load}$) waveforms corresponding to test case in Fig. 6; 1 V = = 100 V (A)

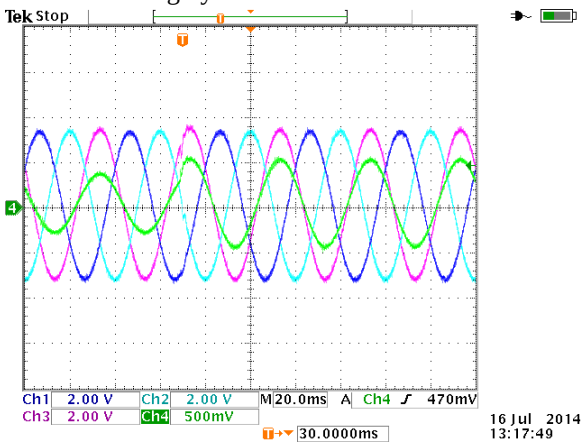


Figure 11. Voltage and current (Ch4: $i_{c\ load}$) waveforms corresponding to test case in Fig. 8; 1 V = = 100 V (A)

TABLE I. EXECUTION TIME

No.	Process	PI control	PIR control
1	Sequence decomposition	50,88 μs	–
2	Coordinate transformation	9,36 μs	4,84 μs
3	Control	18,36 μs	19,37 μs
4	Coordinate transformation	2,88 μs	1,35 μs
5	Sequence composition	7,41 μs	–
	Total (1+2+3+4+5)	88,89 μs	25,56 μs

cutes faster, meaning less code is needed for the same result, overall processor load is lower. This is desirable because usually single processor is used for other functions apart from inverter control including monitoring, communication, sequencing, error detection, etc. Further optimization regarding the stability of an inverter can be achieved be setting the control task CT2 in PIR control strategy to a higher sampling rate. But this depends on the processor load. On the other hand, optimization regarding the PI control strategy can be done by transferring sequence (de) composition calculations to dedicated chips, e.g. FPGA or implementing own digital integrators, e.g. First-Order IIR Digital Integrator based on backward rectangle method with lower computation time than the built-in elements.

VII. CONCLUSION

In this paper a new way of developing control algorithm using hardware-in-the-loop has been investigated. Two different control strategies have been implemented and tested. Special emphasis is given to measuring the execution time. PIR controller has the advantage compared to PI controller mainly due to lack of sequence (de) composition calculation and less coordinate transformation calculations.

Using this way of design, the process of prototyping and testing of complex control systems is simplified. Control is coded on the same TI controlCARD used in the target converter. This enables user to easy implement desired control strategy, test various control algorithms, compare overall exection time and load and, if necessary, further optimize algorithm, all before actually beginning to work on a target model.

REFERENCES

[1] J. Poon, P. Haessig, J. G. Hwang and I. Celanovic, "High-Speed Hardware-in-the-loop Platform for Rapid Prototyping of Power Electronics Systems", 2010 IEEE Conference on Innovative Technologies for an Efficient and Reliable Electricity Supply (CITRES), Sept. 27–29, 2010, Waltham, MA, pp. 420–424.

[2] L. Gregoire, K. Al-Haddad and G. Nanjundaiah, "Hardware-in-the-Loop (HIL) to reduce the development cost of power electronic converters", 2010 International Conference on Power Electronics (IICPE), Jan. 28–30, 2011, New Delhi, pp. 1–6.

[3] N. Ciurro, "Harmonics in Industrial Power Systems", [Online] http://ecmweb.com/design_engineering/harmonics-industrial-power-systems-20090601/, accessed Sep. 2014.

[4] S. El-Barbari and W. Hofmann, "Digital control of a four leg inverter for standalone photovoltaic systems with unbalanced load", 26th Annual Conference of the IEEE Industrial Electronics Society (IECON 2000), Oct. 2000, Nagoya, vol. 1, pp. 729–734.

[5] A. Bellini and S. Bifaretti, "Modulation Techniques for Three-Phase Four-Leg Inverters", 6th WSEAS International Conference on Power Systems, Lisbon, Portugal, Sept. 22–24, 2006, pp. 398–403.

[6] N. V. Nho and M. J. Youn, "Carrier PWM algorithm with optimized switching loss for three-phase four-leg multilevel inverters", IEE Electronics Letters, UK, Jan. 2005, Vol. 41, No. 1, pp. 43–44.

[7] W. Muangjai and S. Premrudeepreechacharn, "Implementation of a Carrier-based Three-dimensional Space Vector PWM Technique for Three-phase Four-leg Voltage Source Converter with Microcontroller", 4th IEEE Conference on Industrial Electronics and Applications (ICIEA 2009), 2009, May, 25–27, 2009, pp. 837–841.

[8] J.-H. Kim and S.-K. Sul, "A Carrier-based PWM Method for Three-Phase Four-Leg Voltage Source Converters", IEEE Transactions on Power Electronics, Jan. 2004, vol. 19, issue 1, pp. 66–75.

[9] E. Demirkutlu and A. M. Hava, "A Scalar Resonant-Filter-Bank-Based Output-Voltage Control Method and a Scalar Minimum-Switching-Loss Discontinuous PWM Method for the Four-Leg-Inverter-based Three-Phase Four-Wire Power Supply", IEEE Transactions on Industry Applications, May–June 2009, vol. 45, issue 3, pp. 982–991.

[10] A. M. Hava and E. Demirkutlu, "Output Voltage Control of a Four-leg Inverter Based Three-phase UPS Utilizing Stationary Frame Resonant Filter Banks", 2007 European Conference on Power Electronics and Applications, Sept. 2–5, 2007, pp. 1–10.

[11] I. Vechiu, H. Camblong, G. Tapia, B. Dakyo O. Curea, "Control of four leg inverter for hybrid power system applications with unbalanced load", Energy Conversion and Management, vol. 48, issue 7, July 2007, pp. 2119–2128.

[12] F. Li, X. Wang, Z. Chen and X. Zhang, "A Laboratory Grid Simulator Based on Three-phase Four-leg Inverter: Design and Implementation", 2011 International Conference on Electrical Machines and Systems (ICEMS), Aug. 20–23, 2011, Beijing, pp. 1–5.

[13] E. Ortjohann, A. Mohd, N. Hamsic, D. Morton and O. Omari, "Advanced Control Strategies for Three-Phase Grid Inverters with Unbalanced Loads for PV/Hybrid Power Systems", 21st European Photovoltaic Solar Energy Conference, Dresden, Germany, Sept. 2006.

[14] H. Camblong, I. Vechiu and O. Curea, "An Innovative VSI Controller for the Generation of Balanced Voltage in Spite of the Presence of Unbalanced Loads", ACC '07 American Control Conference, July 9–13, 2007, New York, NY, pp. 4756–4761.

[15] R. Teodorescu, M. Liserre and P. Rodriguez, "Grid Converters for Photovoltaic and Wind Power Systems", Wiley, Jan. 2011.

[16] W. T. Franke, C. Kürtz and F. W. Fuchs, "Analysis of Control Strategies for a 3 Phase 4 Wire Topology for Transformerless Solar Inverters", 2010 IEEE International Symposium on Industrial Electronics (ISIE), July 4–7, 2010, Bari, pp. 655–663.

[17] R. Teodorescu, F. Blaabjerg, M. Liserre P. C. Loh, "'Proportional-resonant controllers and filters for grid-connected voltage-source converters", IEE Proceedings Electric Power Applications, Vol. 153, Issue 5, Sept. 2006, pp. 750–762.

[18] W. Hu, W. Ma and C. Liu, "Research on a New Control Strategy of Three Phase Inverter for Unbalanced Loads", Journal of Electromagnetic Analysis and Applications, vol. 2, no. 1, 2010, pp. 39–44.

[19] W. Lenwari, "Review of a High Performance Control of AC Signals Based on a Proportional Plus Resonant Compensator", Suranaree Journal of Science and Technology, vol. 15, no. 4, Oct –Dec. 2008, pp. 333–344.

[20] J. Babic, S. Marijan and I. Petrović, "Introducing Model-Based Techniques Into Development of Real-Time Embedded Applications", Automatika: Journal for Control, Measurement, Electronics, Computing and Communications, vol. 52, no. 4, March, 2012, pp. 329–338.

Noise Improvement using SC Filters

I. Volarić*, N. Stojković* and S. Vlahinić*
* University of Rijeka - Faculty of Engineering / Department of automatics and electronics, Rijeka, Croatia
ivolaric@riteh.hr, nino.stojkovic@riteh.hr, sasa.vlahinic@riteh.hr

Abstract – In this paper noise analysis of three different filter designs is performed. Voltage noise spectral density and RMS noise voltage in simply designed active RC, noise optimized active RC and switched capacitor (SC) filter design are calculated for fifth order Chebyshev low pass filter, with a cut-off frequency f_c=3.4kHz, and pass-band ripple α_{max}=-0.5dB. All three designs are realized as a cascade of two second order sections and one first order section. All filter analysis has been performed using MATLAB and SPICE program tools. Through the analysis it is shown that SC based filter has significantly lower noise when compared with active RC based filter.

I. INTRODUCTION

There are several quality assessment parameters of the electrical circuits and systems, and noise is one of them. Demand on noise as lower as possible is also a demand on higher dynamic range. Noise is very important factor for small signal processing in order to extract proper information from the signal. There are several noise origins which produce several types of noise, but in this paper only thermal noise has been considered [1]. One approach utilizes OTA-C filter structure in order to obtain a low noise design, however, main disadvantage of this realization is low dynamic range [2]. Another possible design uses active RC filter sections where noise optimization can be done to obtain low noise [3]. The third design uses switched capacitor (SC) filter structure, [4]. It is shown that this approach will reduce noise level more than other realizations. In this paper, noise properties are analyzed on simply designed active RC filter, noise optimized active RC filter and corresponding SC filter.

SC based circuits replace a resistor with a capacitor and two electric switches (usually MOS transistors) driven by the two phase non-overlapping clock [4, 5]. The time constant τ of such circuit is:

$$\tau = RC = \frac{C}{C_R}\frac{1}{f_{CL}}, \tag{1}$$

where f_{CL} is the clock frequency. Even though SC circuits have increased number of components, the overall size of the circuit is reduced since small valued capacitors can be used for acquiring desired C/C_R ratio, thus allowing the integration of the circuit into a single monolithic integrated circuit. Furthermore, the C/C_R ratio can be realized with great accuracy, even up to 0.1%, while RC error can range from 20% up to 50%. In addition, since τ is dependent of f_{CL} (1), the cut-off frequency of SC based filters can be regulated simply by varying the clock frequency.

In this paper, fifth order Chebyshev low pass filter with a cut-off frequency f_c=3.4kHz and a pass-band ripple α_{max}=-0.5dB is realized. All previously mentioned designs are realized as a cascade of two second order sections and one first order section. Sequence of the sections is such that their Q-factors are in descending order (first one being the greatest, and last one the smallest), and sections are based on the Sallen-Key topology. Through the paper significant improvement of the SC filter over the active RC filter in terms of noise reduction is shown.

II. CONTINUOUS ACTIVE RC FILTER

The continuous time transfer function of the active RC low pass Chebyshev fifth order filter with previously noted parameters is given by:

$$H(s) = H_1(s)\cdot H_2(s)\cdot H_3(s), \tag{2a}$$

$$H_1(s) = \frac{4.727\cdot 10^8}{s^2 + 4784s + 4.727\cdot 10^8}, \tag{2b}$$

$$H_2(s) = \frac{2.176\cdot 10^8}{s^2 + 1.252\cdot 10^4 s + 2.176\cdot 10^8}, \tag{2c}$$

$$H_3(s) = \frac{7740}{s + 7740}. \tag{2d}$$

The Sallen-Key topology of low pass filter is given in Fig. 1. The filter is realized as a cascade of two second order sections and one first order section. The transfer function of the given circuit is:

$$H(s) = \prod_{i=1,2} H_i(s)\cdot H_3(s), \tag{3a}$$

$$H_i(s) = \frac{\left(1+\frac{G_{3i}}{G_{4i}}\right)\frac{G_{1i}G_{2i}}{C_{1i}C_{2i}}}{s^2 + s\left(\frac{G_{1i}+G_{2i}}{C_{1i}} + \frac{G_{2i}G_{3i}}{C_{2i}G_{4i}}\right) + \frac{G_{1i}G_{2i}}{C_{1i}C_{2i}}}, \tag{3b}$$

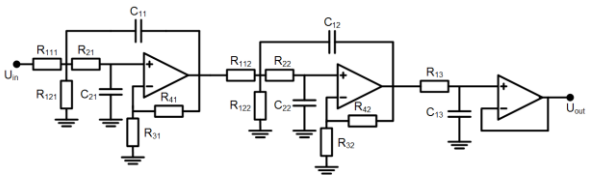


Figure 1. The Fifth order low pass active RC filter.

$$H_3(s) = \frac{\frac{1}{R_{13}C_{13}}}{s + \frac{1}{R_{13}C_{13}}}. \qquad (3c)$$

Element values have been calculated by comparing equations (2a) – (2d) with (3a) - (3d) with $C_{1i} = C_{2i}$ and $R_{1i}=R_{2i}$, and are given in Table 1.

Frequency responses of the filter are shown in Fig. 2. Calculated values are obtained from the filter transfer function (2a) - (2d), while simulated values are obtained by simulating the circuit shown in Fig. 1 with element values from Table 1. The difference between the characteristics is minimal, as can be seen in zoomed inset of Fig. 2.

The equivalent circuit of the active RC filter for the calculation of the voltage noise spectral density is shown in Fig. 3. Current noise sources: $I_{nji} = \sqrt{4kT}/R_{ji}$ were added parallel to the resistors, while voltage noise sources (E_n = 15 nV/√Hz for LT1055) were added at input of the operational amplifiers. The operational amplifier input noise currents were neglected since their values are significantly smaller (E_n = 1.8 fA/√Hz for LT1055). The total voltage noise spectral density is calculated as [5]:

$$V_n^2(\omega) = \sum^{m} |T_{I,k}(j\omega)|^2 \, (I_{n,k})^2 + \sum^{n} |T_{U,l}(j\omega)|^2 \, (E_{n,l})^2 \,, \qquad (4)$$

where $T_{I,k}$ and $T_{U,l}$ are the transfer functions of the current noise sources $I_{n,k}$ and the voltage noise sources $E_{n,l}$, respectively. Table 2 shows section specific transfer function of each noise source, and in order to get a complete transfer function, the section specific transfer function must be multiplied with transfer functions of all following sections. Figure 4(a) shows the total voltage noise spectral density. Voltage noise spectral density functions of all noise components based on their origin are also presented in Fig. 4(b, c, d). The small difference between calculated and simulated values can be explained with differences in noise source models. Real element noise models used in SPICE are slightly different than here used theoretical values made only from thermal noise sources. The noise RMS can be calculated as:

$$(E_n^2)_{ef} = \int_{\omega_1}^{\omega_2} V_n^2(\omega) d\omega \,. \qquad (5)$$

In order to decrease noise, optimization algorithm described in [3] has been applied, and operational amplifier LT1055 has been replaced with low noise

TABLE 1. ELEMENT VALUES OF THE CONTINUOUS TIME ACTIVE RC FILTER SHOWN IN FIG. 1.

	1st section	**2nd section**	**3rd section**
R_{11i}	12.8 kΩ	14.6 kΩ	-
R_{12i}	7.2 kΩ	12.7 kΩ	-
R_{1i}	-	-	12.9 kΩ
R_{2i}	4.6 kΩ	6.8 kΩ	-
R_{3i}	2.6 kΩ	4.1 kΩ	-
R_{4i}	4.7 kΩ	4.7 kΩ	-
C_{1i}	10 nF	10 nF	10 nF
C_{2i}	10 nF	10 nF	-

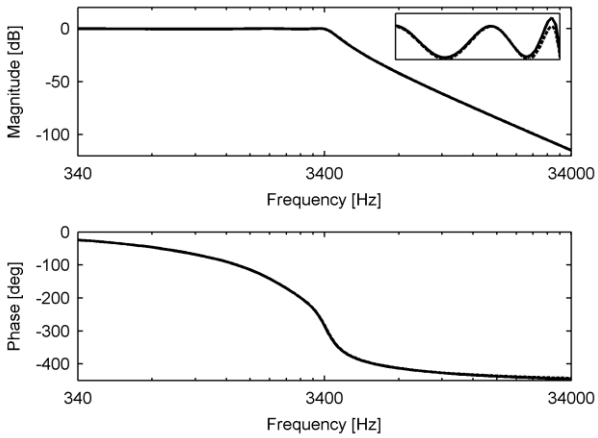


Figure 2. Transfer function of the continuous time active RC filter: simulated (full) and calculated (dashed). Magnitude values between f_c/100 and f_c are zoomed in to a range from -0.5 dB to 0.2 dB.

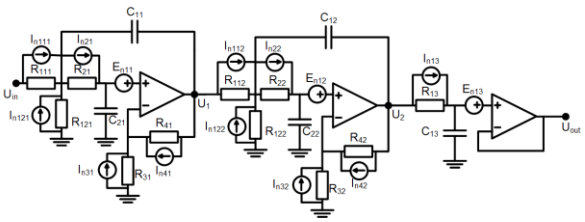


Figure 3. Equivalent noise circuit of the continuous time active RC filter.

LT1007 operational amplifier (E_n = 2.5 nV/√Hz). The newly calculated filter element values are shown in Table 3, while voltage noise spectral density is shown in Fig. 5. The noise RMS value is 140% smaller when compared to non-optimized filter values, shown in Fig. 4. Note that significant decrease of noise is obtained in filter pass-band.

TABLE 2. SECTION SPECIFIC VOLTAGE NOISE TRANSFER FUNCTIONS OF THE CONTINUOUS TIME ACTIVE RC FILTER.

Noise source	Transfer function
$\frac{U_1}{I_{n1\,1i}}, i=1,2$	$\frac{R_{1\,1i}R_{12i}(R_{3i}+R_{4i})}{s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}R_{3i}+s(C_{2i}R_{2i}R_{3i}(R_{1\,1i}+R_{12i})+R_{1\,1i}R_{12i}(C_{2i}R_{3i}-C_{1i}R_{4i}))+R_{3i}(R_{1\,1i}+R_{12i})}$
$\frac{U_1}{I_{n12i}}, i=1,2$	$\frac{R_{1\,1i}R_{12i}(R_{3i}+R_{4i})}{s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}R_{3i}+s(C_{2i}R_{2i}R_{3i}(R_{1\,1i}+R_{12i})+R_{1\,1i}R_{12i}(C_{2i}R_{3i}-C_{1i}R_{4i}))+R_{3i}(R_{1\,1i}+R_{12i})}$
$\frac{U_1}{I_{n2i}}, i=1,2$	$\frac{R_{2i}(R_{3i}+R_{4i})(sC_{1i}R_{1\,1i}R_{12i}+R_{1\,1i}+R_{12i})}{s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}R_{3i}+s(C_{2i}R_{2i}R_{3i}(R_{1\,1i}+R_{12i})+R_{1\,1i}R_{12i}(C_{2i}R_{3i}-C_{1i}R_{4i}))+R_{3i}(R_{1\,1i}+R_{12i})}$
$\frac{U_1}{I_{n3i}}, i=1,2$	$\frac{R_{3i}R_{4i}(s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}+s(C_{2i}R_{12i}(R_{1\,1i}+R_{2i})+R_{1\,1i}(C_{1i}R_{12i}+C_{2i}R_{2i}))+R_{1\,1i}+R_{12i}}{s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}R_{3i}+s(C_{2i}R_{2i}R_{3i}(R_{1\,1i}+R_{12i})+R_{1\,1i}R_{12i}(C_{2i}R_{3i}-C_{1i}R_{4i}))+R_{3i}(R_{1\,1i}+R_{12i})}$
$\frac{U_1}{I_{n4i}}, i=1,2$	$\frac{R_{3i}R_{4i}(s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}+s(C_{2i}R_{12i}(R_{1\,1i}+R_{2i})+R_{1\,1i}(C_{1i}R_{12i}+C_{2i}R_{2i}))+R_{1\,1i}+R_{12i}}{s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}R_{3i}+s(C_{2i}R_{2i}R_{3i}(R_{1\,1i}+R_{12i})+R_{1\,1i}R_{12i}(C_{2i}R_{3i}-C_{1i}R_{4i}))+R_{3i}(R_{1\,1i}+R_{12i})}$
$\frac{U_1}{E_{n1i}}, i=1,2$	$\frac{(R_{3i}+R_{4i})(s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}+s(C_{2i}R_{12i}(R_{1\,1i}+R_{2i})+R_{1\,1i}(C_{1i}R_{12i}+C_{2i}R_{2i}))+R_{1\,1i}+R_{12i}}{s^2C_{1i}C_{2i}R_{1\,1i}R_{12i}R_{2i}R_{3i}+s(C_{2i}R_{2i}R_{3i}(R_{1\,1i}+R_{12i})+R_{1\,1i}R_{12i}(C_{2i}R_{3i}-C_{1i}R_{4i}))+R_{3i}(R_{1\,1i}+R_{12i})}$
$\frac{U_{out}}{I_{n13}}$	$\frac{R_{13}}{1+C_{13}R_{13}s}$
$\frac{U_{out}}{E_{n13}}$	1

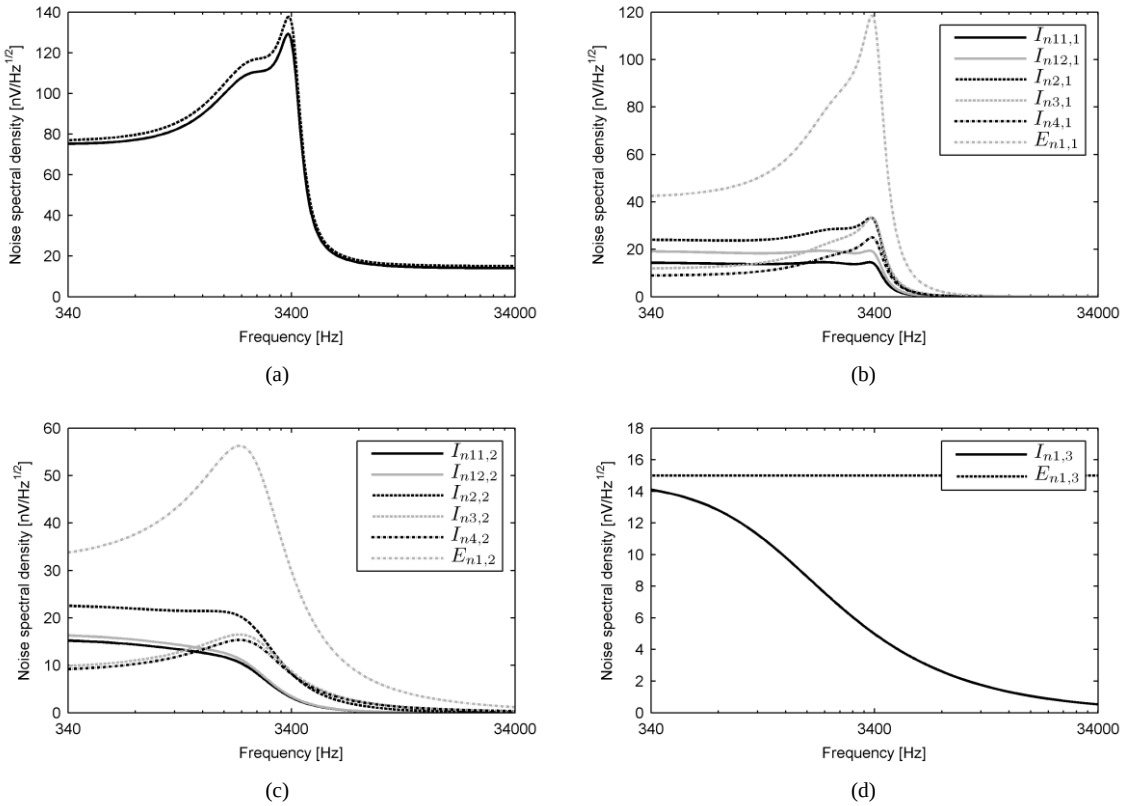


Figure 4. Voltage noise spectral density of the continuous time active RC filter and its components. (a) Total voltage noise spectral density: simulated (full) and calculated (dashed). Noise RMS value is $(E_n)_{ef}$ = 17.595 μV ($f_1 = 0.1\ f_c$, $f_2 = 10\ f_c$). (b) Section 1 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 13.867 μV ($f_1 = 0.1\ f_c$, $f_2 = 10\ f_c$). (c) Section 2 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 8.218 μV ($f_1 = 0.1\ f_c$, $f_2 = 10\ f_c$). (d) Section 3 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 7.054 μV ($f_1 = 0.1\ f_c$, $f_2 = 10\ f_c$).

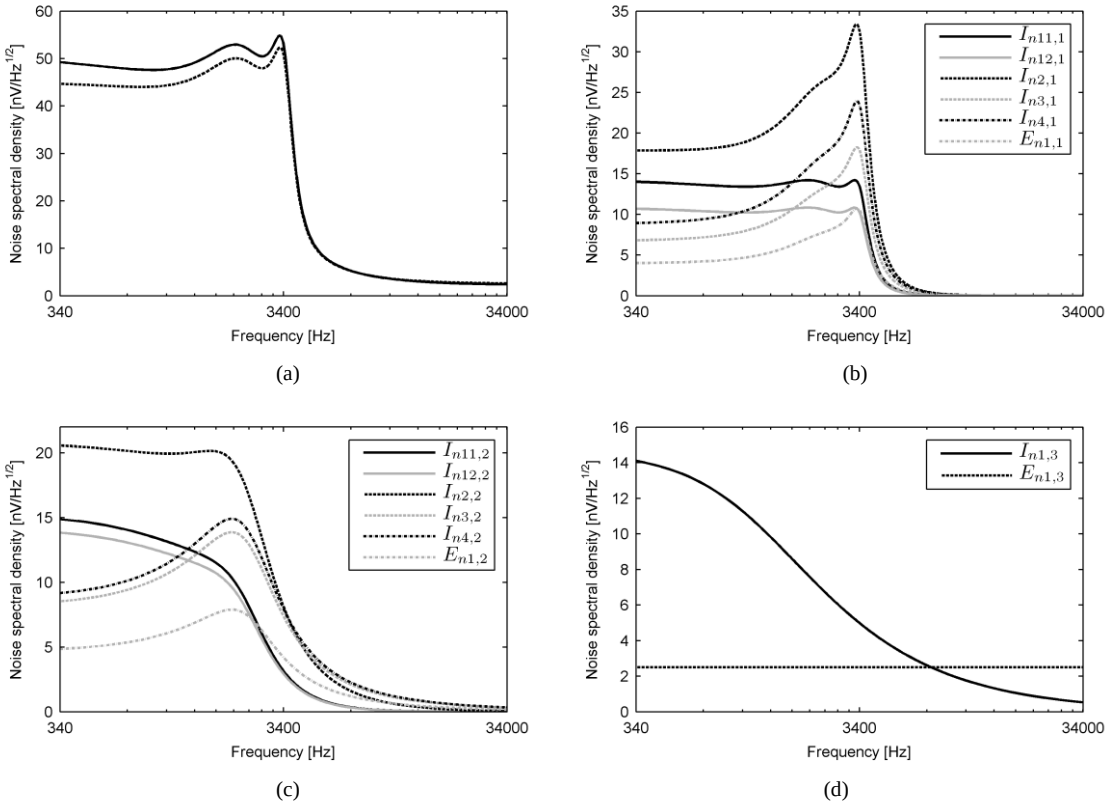


Figure 5. Voltage noise spectral density of the optimized continuous time active RC filter and its components. (a) Total voltage noise spectral density: simulated (full) and calculated (dashed). Noise RMS value is $(E_n)_{ef}$ = 7.212 µV (f_1 = 0.1 f_c, f_2 = 10 f_c). (b) Section 1 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 5.552 µV (f_1 = 0.1 f_c, f_2 = 10 f_c). (c) Section 2 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 4.212 µV (f_1 = 0.1 f_c, f_2 = 10 f_c). (d) Section 3 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 1.855 µV (f_1 = 0.1 f_c, f_2 = 10 f_c).

TABLE 3. ELEMENT VALUES OF THE OPTIMIZED CONTINUOUS TIME ACTIVE RC FILTER SHOWN IN FIG. 1.

	1st section	**2nd section**	**3rd section**
R_{11i}	12.2 kΩ	14.0 kΩ	-
R_{12i}	20.9 kΩ	16.2 kΩ	-
R_{1i}	-	-	12.9 kΩ
R_{2i}	7.7 kΩ	7.5 kΩ	-
R_{3i}	8.0 kΩ	5.4 kΩ	-
R_{4i}	4.7 kΩ	4.7 kΩ	-
C_{1i}	10 nF	10 nF	10 nF
C_{2i}	3.6 nF	8.2 nF	-

III. DISCRETE TIME SC FILTER

The discrete time transfer function is obtained by applying forward Euler transformation ($s \Rightarrow (z-1)/T_S$, where T_S is the sampling time) to the continuous time transfer function, defined by (2a) - (2d), with sampling frequency f_s = 900 f_c:

$$H(z) = H_1(z) \cdot H_2(z) \cdot H_3(z), \quad (6a)$$

$$H_1(z) = \frac{5.048 \cdot 10^{-5}}{z^2 - 1.998z + 0.999}, \quad (6b)$$

$$H_2(z) = \frac{2.324 \cdot 10^{-5}}{z^2 - 1.996z + 0.996}, \quad (6c)$$

$$H_3(z) = \frac{0.003}{z - 0.998}. \quad (6d)$$

Figure 6 shows the electric circuit of the discrete time SC filter based on Sallen-Key topology [5]. The transfer function of the given circuit is:

$$H(z) = \prod_{i=1,2} \frac{n_{1i}}{d_{1i}z^2 + d_{2i}z + d_{3i}} \cdot \frac{C_{R13}}{(C_{R13} + C_{13})z - C_{13}}, \quad (7a)$$

$$n_{1i} = C_{R1i}C_{R2i}(C_{1i} + C_{gi}), \quad (7b)$$

$$d_{1i} = (C_{1i} + C_{gi} + C_{R1i})(C_{1i} + C_{gi} + C_{R2i})(C_{2i} + C_{R2i}), \quad (7c)$$

$$d_{2i} = (C_{1i} + C_{gi})[C_{2i}(2C_{1i} + 2C_{gi} + C_{R1i}) + C_{R2i}(C_{gi} + 2C_{1i} + C_{R2i} + C_{2i})] + C_{R1i}C_{R2i}(C_{R2i} + C_{2i}), \quad (7d)$$

$$d_{3i} = n_{1i} - d_{1i} + d_{2i} = (C_{1i} + C_{gi})(C_{1i}C_{2i} + C_{1i}C_{R2i} + C_{2i}C_{gi}). \quad (7e)$$

By comparing (6a) - (6d) with (7a) - (7e) the element values of the circuit have been calculated and are shown in Table 4.

Transfer function of the filter is shown in Fig. 7. The values were calculated from a discrete time transfer function of the SC filter (6a) - (6d). The Z-transformation hasn't affected desired continues time designed filter specifications, as can be observed by comparing them with Fig. 2.

TABLE 4. ELEMENT VALUES OF THE DISCRETE TIME SC FILTER SHOWN IN FIG. 6.

	1st section	2nd section	3rd section
C_{R1i}	121.1 pF	1.0 nF	560 pF
C_{R2i}	100 pF	69 pF	-
C_{gi}	100 pF	10 nF	-
C_{1i}	150 nF	270 nF	220.8 nF
C_{2i}	1.5 nF	10.5 nF	-

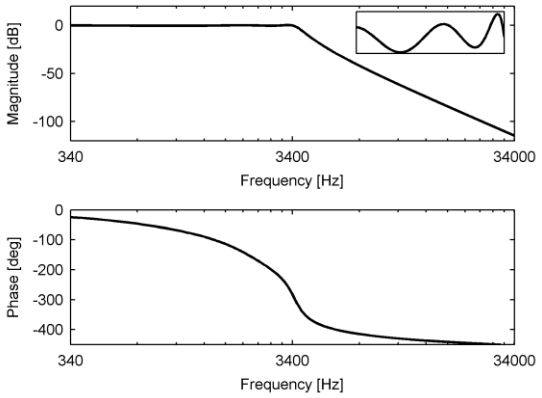


Figure 7. Transfer function of the discrete time SC filter. Magnitude values between f_c/100 and f_c are zoomed in to a range from -0.5 dB to 0.3 dB.

Equivalent circuit for the calculation of the voltage noise spectral density is based on [6], and calculated voltage noise spectral density is shown in Fig. 8. The noise RMS value, defined by (5), is over 5 times lower, when compared with continues time active RC filter, shown in Fig. 4 and over 2 times lower, when compared with optimized active RC filter, shown in Fig. 5. In the end Fig. 9 shows noise for different filter designs discussed in this paper. The noise reduction using SC design is significant.

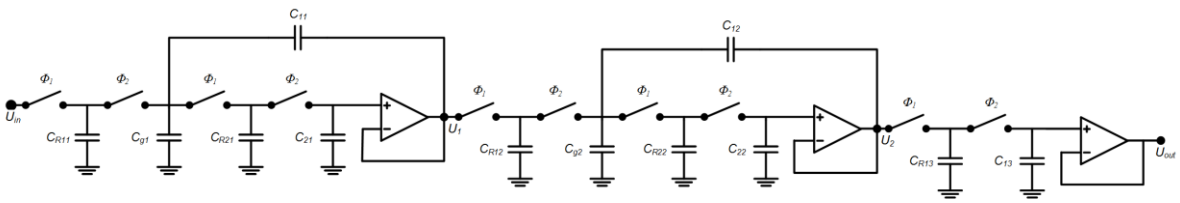


Figure 6. The fifth order low pass SC filter.

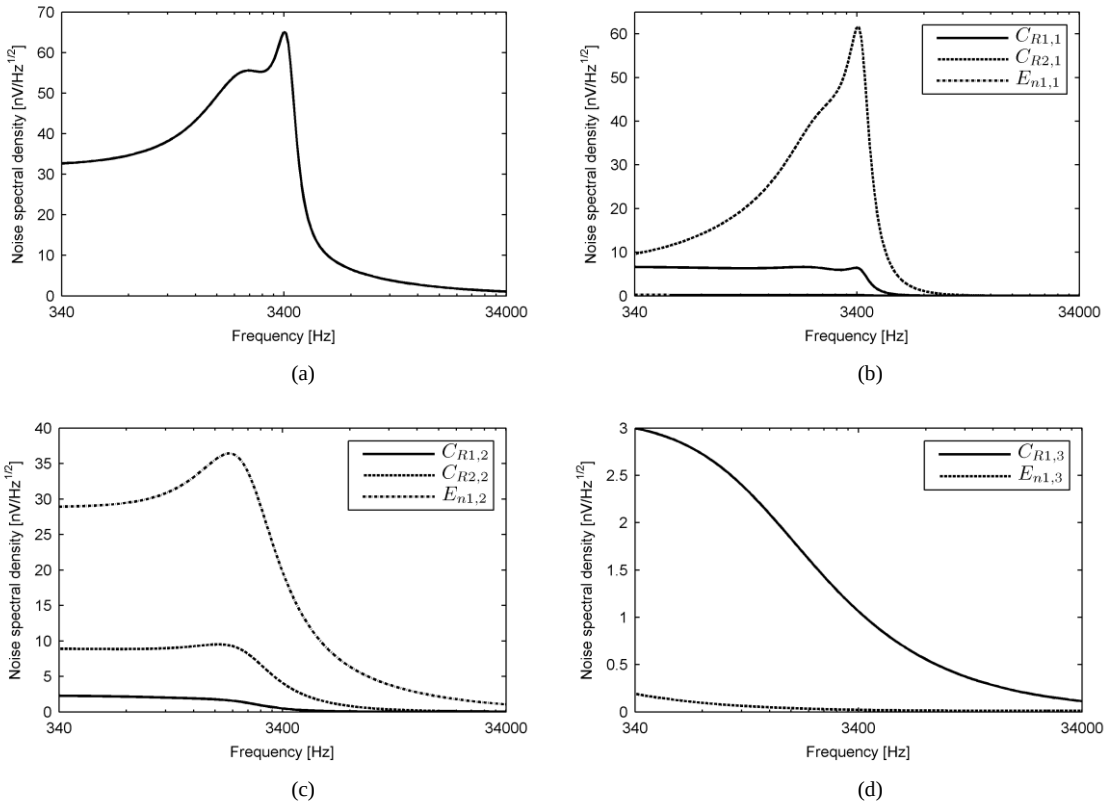


Figure 8. Voltage noise spectral density of the discrete time SC filter and its components. (a) Total voltage noise spectral density. Noise RMS value is $(E_n)_{ef}$ = 3.129 µV (f_1 = 0.1 f_c, f_2 = 10 f_c). (b) Section 1 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 2.448 µV (f_1 = 0.1 f_c, f_2 = 10 f_c). (c) Section 2 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 1.945 µV (f_1 = 0.1 f_c, f_2 = 10 f_c). (d) Section 3 voltage noise spectral density components. Noise RMS value is $(E_n)_{ef}$ = 0.123 µV (f_1 = 0.1 f_c, f_2 = 10 f_c).

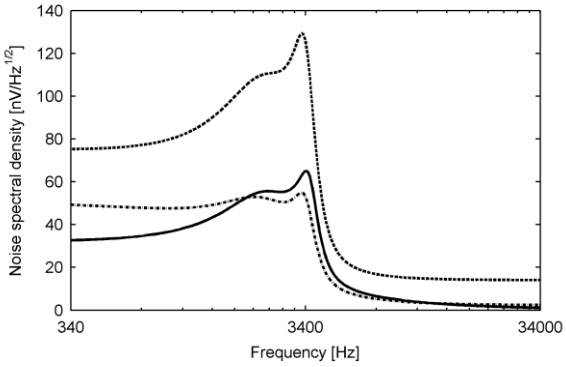


Figure 9. Simulated voltage noise spectral density of: active RC filter (dashed), optimized active RC filter (dot-dashed), and SC filter (full).

IV. CONCLUSION

In this paper noise analysis of three different filter designs are done. It is shown that discrete time SC based filter have better noise figure over the continuous time active RC based filter. Noise reduction measured with RMS noise voltage is more than twice in comparison with noise optimized active RC design and more than five times in comparison with simply designed active RC filter design.

Further improvement of the presented filter design can be implemented using more complex filter structure with feedbacks instead of used cascade filter structure.

REFERENCES

[1] T. H. Lee, The Design of CMOS Radio-Frequency Integrated Circuits, 2nd ed., Cambridge University Press, New York, USA, 2004.

[2] N. Stojković, N. Mijat, "Noise and Dynamic Range of Second Order OTA-C BP Filter Sections", Proceedings of the European Conference on Circuit Theory and Design, vol. 2, pp. 795-798, August 1999, Stresa, Italy.

[3] N. Stojković, E. Kamenar, M. Šverko, "Optimized Second- and Fourth- Order LP and BP Filters", Automatika, no 52., vol. 2, pp. 152-168, 2011.

[4] H. A. de Andrade Serra, "Design of Switched-Capacitor Filters using Low Gain Amplifiers", PhD thesis, Faculdade de Ciencias e Techologia, Universidade Nova de Lisboa, 2012.

[5] W. K. Chen, The Circuits and Filters Handbook, 2nd ed. CRC Press, 2003.

[6] J. H. Fischer, "Noise sources and calculation techniques for switched capacitor filters", IEEE Journal of Solid-State Circuits, vol. 17., pp. 742-752, August 1982.

Phase Demodulation of Magnetic GMR Signal for Virtual Sensing Applications by using Hilbert Transform

Muhammad Adnan*, Dirk Hammerschmidt** and Mario Huemer ***

*CTR Carinthian Tech. Reserach AG, Villach, Austria
**Infineon Technologies AG, Villach, Austria
***Institute of Signal Processing, JKU, Linz, Austria
e-mail: adnan.muhammad@ctr.at

Abstract - GMR (Giant MagnetoResistance) sensors, which are used in conjunction with magnetic encoders (magnetic pole wheels), have a huge number of applications in automotive and machine industry, where they are used to compute the rotational speed of rotating parts like wheels and shafts. In this paper, we present an analytical model of magnetic pole wheel signal as it is sensed by a GMR sensor. Experimental measurements are performed to check the validity of the mathematical model. Furthermore, we present a novel digital signal processing technique, which incorporates the Hilbert transformer to phase demodulate this GMR signals to compute high resolution instantaneous position and speed information. A frequency domain analysis is then deployed to analyze the signal for any low frequency vibrations present in the rotating part.

I. Introduction

MAGNETIC encoders used with magnetic speed sensors have wide applications in industry where rotational speed or position is needed to be known. The conventional method to compute the speed signal from any magnetic sensor, whether it is a GMR (Giant MagnetoResistance) or Hall sensor, involves an analog comparator which generates a square pulse, whose frequency is proportional to the speed of rotation. This method has an inherent discrepancy that very low resolution speed and position information is available. This limitation is a bottleneck in computing high resolution speed and position information precisely. The maximum samples per revolution is limited to $2N$, where N is the total number of pole pairs in the encoder. Fig. 1 illustrates the speed signal generated by a GMR sensor and the corresponding square pulse. The dotted line shows an ideal sinusoid which represents the ideal magnetic field signal while the solid line depicts the actual signal out of the GMR sensor. Conventional speed signal computation methods do not take into account the fact that the pole wheel suffers from manufacturing intolerances. All pole pairs are expected to be of equal length in $radians$. In reality, each pole pair has a different pole length and the pole periods vary from pole to pole in a single pole wheel, and are given by

$$\Delta\theta_i = 2\pi\left(\frac{\Delta t_i}{\sum_{i=0}^{N-1}\Delta t_i}\right) - \frac{2\pi}{N} \qquad (1)$$

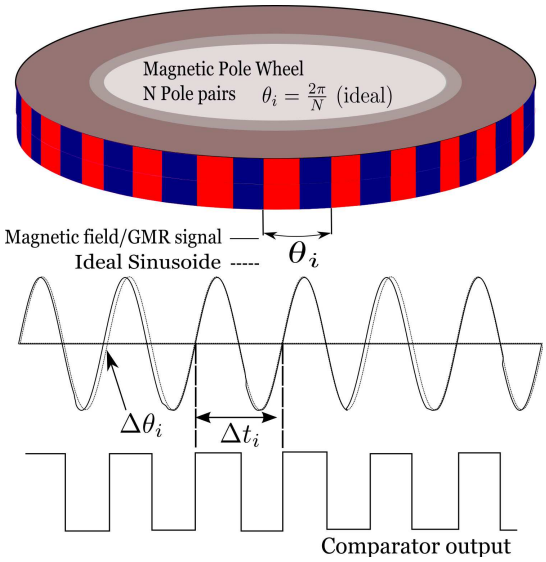


Fig. 1: Speed signal generated by GMR sensor and the resultant square pulse

where N is the total number of pole pairs.

The magnetic field signal of a magnetic encoder can be modeled as a phase modulated sinusoid, as illustrated in Fig. 1, where the phase signal is the pattern of the pole wheel. This is indicated after a discrete interval as $\Delta\theta_i$. Modern digital signal processing techniques can be applied to the analog magnetic field signal to extract the instantaneous values of the phase angle. The time derivative of the instantaneous phase then gives the instantaneous angular speed.

In the following, we present a mathematical treatment of the magnetic pole wheel signal as a phase modulated sinusoid, followed by a demodulation algorithm description using a Hilbert transformer in the frequency domain. Later, we present an experimental setup for measurements, we elaborate a frequency domain analysis of the results, which is followed by concluding remarks.

II. Magnetic pole wheel signal as phase modulated sinusoid

Due to the manufacturing intolerances, the magnetic field, strength from two poles differs and the width of two poles is not equal. This gives rise to sinusoid with slight offsets in phase from each other. The signal from the magnetic pole wheel can be modeled as a sinusoid, which is phase modulated by a pattern generated by the pole wheel intolerances. A general phase modulated signal can be written as [1],

$$x(t) = A_c \cos\left(\omega_c t + \theta_m(t)\right) \tag{2}$$

where $\theta_m(t)$ is a function of the pole wheel pattern, dependent on the geometry of the pole wheel. For all real scenarios $\theta_m(t)$, which is the variation in the phase angle from the absolute value, is very small. It has been observed for a pole wheel of 32 pole pairs to be in the order of 10^{-3} radians [2]. $\theta_m(t)$ is also inversely proportional to the number of pole pairs in a pole wheel. Expanding Eq. 2 and utilizing small signal approximation for $\theta_m(t)$ we get

$$x(t) = A_c \cos\left(\omega_c t\right) - A_c \theta_m(t) \sin\left(\omega_c t\right) \tag{3}$$

where we have assumed the approximations $\cos\left(\theta_m(t)\right) = 1$ and $\sin\left(\theta_m(t)\right) = \theta_m(t)$.

As mentioned earlier, $\theta_m(t)$ is dependent on the pole wheel geometry, which repeats itself each revolution. This can be modeled as a series of functions with periods ranging from T_o to $\frac{T_o}{N}$, where N is the total number of pole pairs and T_o is the period of one complete revolution of the pole wheel. Hence $\frac{T_o}{N}$ is the period of one pole pair, which corresponds to the highest frequency harmonic.

Let Ω_o be the rotational frequency of the pole wheel and ω_o be the rotational frequency of one pole pair then

$$\omega_o = \Omega_o N \tag{4}$$

which is the frequency of the basic pattern repeating itself in each pole pair. In this case, this is also the carrier frequency mentioned above as ω_c. Next is the pattern that is a combination of two pole pairs and it shall repeat at a frequency $\Omega_o(N-1)$. Similarly another pattern which is a combination of three pole pairs shall have frequency $\Omega_o(N-2)$and so on. We can generalize this as

$$\omega_p = \Omega_o(N-p) \qquad p = 0, 1, 2....N-1 \tag{5}$$

This indicates that the lowest frequency is of the pattern, which repeats once each revolution and is the pattern of the whole pole wheel.

To summarize above discussion, we can conclude that $\theta_m(t)$ consists of a linear combinations of functions of pole patterns with fundamental frequencies given by Eq. 5. So $\theta_m(t)$ can be generalized as

$$\theta_m(t) = \sum_{p=0}^{N-1} f_p(t, \psi) \tag{6}$$

where ψ is a space variable added for completeness. Plugging this value of $\theta_m(t)$ into Eq. 3, we get

$$x(t) = A_c \cos\left(\omega_c t\right) - A_c\left(\sum_{p=0}^{N-1} f_p(t, \psi)\right) \sin\left(\omega_c t\right) \tag{7}$$

Since $f_p(t, \psi)$ is periodic with fundamental frequency ω_p, multiplication by a $\sin(\omega_c t)$ will shift this frequency to the carrier frequency which is given by $\Omega_o N$. So the spectrum of the pole wheel signal will have harmonics at $\omega_c \pm \Omega_o(N-p)$ where $p = 0, 1, 2...N$. This pattern of harmonics will repeat itself in the Fourier transform of the signal where all frequencies will be centered around the carrier frequency.

III. Demodulation using Hilbert Transform

Since the magnetic pole wheel signal has been modeled as a phase modulated sinusoid, it can be demodulated by using the Hilbert transform. The Hilbert transform of a signal $x(t)$ is denoted by $x_h(t)$ and is given by [3]

$$x_h(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\alpha)}{t-\alpha} d\alpha \tag{8}$$

and in frequency domain

$$X_h(w) = -j\mathrm{X}(w)\mathrm{sgn}(w) \tag{9}$$

where $\mathrm{X}(w)$ is the frequency spectrum of $x(t)$. Hence a Hilbert transformer has a transfer function given by $-j\mathrm{sgn}(w)$. In essence, it shifts each spectral component of an input signal by $\frac{\pi}{2}$ while keeping the amplitude constant.

Once $x_h(t)$ is known, we can combine it with $x(t)$ to get an analytical signal $\hat{x}(t)$.

$$\hat{x}(t) = x(t) + j x_h(t) \tag{10}$$

The instantaneous phase of this analytical signal shall give us the value of $\phi_m(t)$, which is given by

$$\phi_m(t) = \arctan\left[\frac{x_h(t)}{x(t)}\right] \tag{11}$$

where $\phi_m(t) = \omega_c t + \theta_m(t)$ is the instantaneous phase of the analytical signal.

Hilbert transformer can be implemented either in time domain [4], [5] or in frequency domain [6], [7]. Time domain transformation is ideally a filter whose output is the Hilbert transform of the input signal. This method has inherent drawback that it suffers from bandwidth limitations. The Frequency domain approach is more elegant and utilizes the FFT (Fast Fourier Transform) to produce the analytical signal irrespective of the bandwidth. The Pole wheels' rotational frequency may vary from a couple of rotations per second to tens of rotations per second. The resultant signal bandwidth can be well over a KHz.

A. Implementation strategy

Fig. 2 shows the implementation algorithm for the demodulation of the magnetic pole wheel signal. The Hilbert transform of a sequence of length L is computed by the frequency domain technique where we take the Fourier transform of the input signal, set $-\pi < \omega < 0 = 0$ and take the inverse Fourier transform. For practical purpose, the DFT (Discrete Fourier Transform) of a long sequence is not implementable. This can be circumvented by a sliding window FFT. Each iteration of the loop computes the FFT, then sets $-\pi < \omega < 0 = 0$ and takes the inverse FFT. In the next iteration, the window is

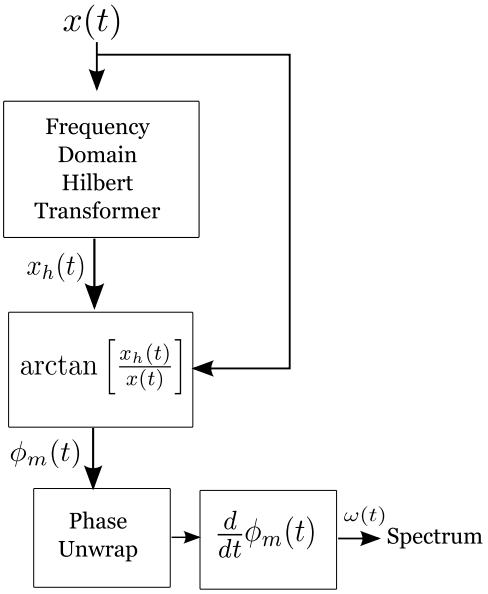


Fig. 2: Block diagram of demodulation algorithm

moved ahead half of its length and the same computation sequence is repeated. For optimal processing L must be power of 2 and must contain at least two periods of the input sinusoid. In essence, we compute the Hilbert transform of a certain length L of the input sequence, shift the window by $L/2$ and continue. After computing the phase angle in each iteration, the sequence from $L/4+1$ to $3L/4$ is retained and edges are discarded. Then this signal is concatenated with the output from the next iteration in the same way (after discarding its edges as well) to get continuous phase values. This prevents the effects of circular convolution from appearing in the phase signal. The technique is illustrated in Fig. 3.

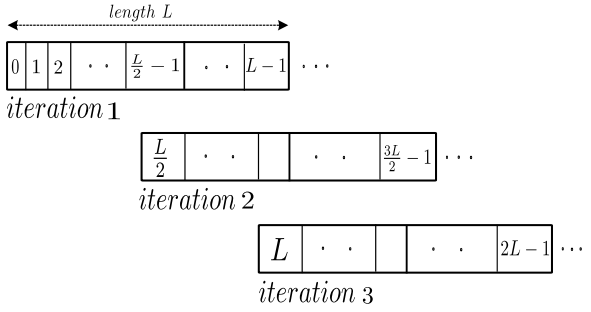


Fig. 3: Sliding window Hilbert transform computation

The Instantaneous phase of the analytical signal is computed using the arctan function. The CORIDC algorithm can be used to perform this task [8] . The computed phase is unwrapped and its time derivative gives the instantaneous rotational speed.

IV. Measurement setup and experimental results

Measurement data for analysis has been obtained from an experimental setup where a pole wheel was mounted on a shaft. A GMR (Giant MagnetoResistance) sensor has been placed at an optimal distance from the pole wheel so as to get a maximum undistorted magnetic field signal. The Magnetic field signal from the GMR has been amplified and then digitized at 100 kHz with a 16 bit ADC (Analog to Digital Converter). The pole wheel was rotated at 14 rev/sec.

Fig. 4 shows the instantaneous phase angle value computed with the above mentioned algorithm. Phase continuity has been ensured by use of the sliding window technique while computing Hilbert transform in frequency domain. Fig. 5

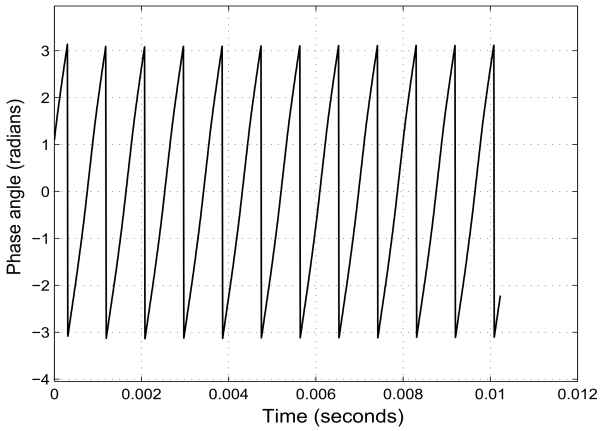


Fig. 4: Instaneous phase computed with Hilbert transform

shows the PSD (Power Spectral Density) of the modulated GMR signal. Fig. 6 shows the power spectral density estimate of the demodulated GMR signal $\frac{d}{dt}(\phi_m(t)$. It shows the lowest pattern frequency at about 14 Hz, which belongs to $p = N - 1$ in Eq. (5) . Below 10 Hz are the vibrations which have been modulated by the carrier frequency and now are visible in the low frequency band after demodulation.

A. Pattern calculation and removal

Pattern calculation and removal is a key aspect to clean the spectrum from regular harmonics and to isolate the vibrations

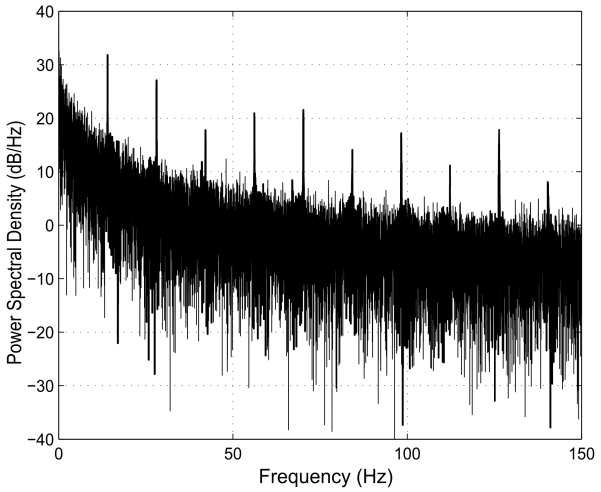


Fig. 5: Spectrum of modulated GMR signal

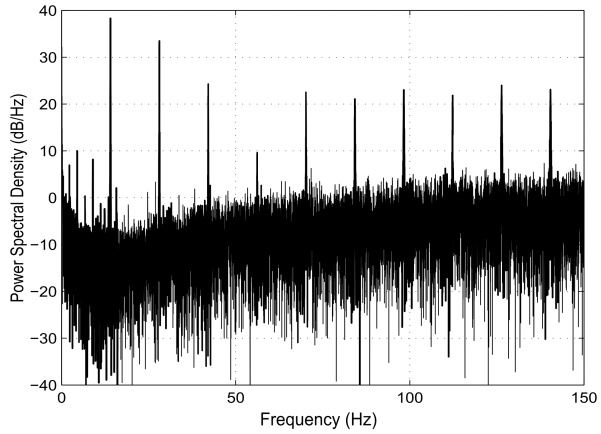


Fig. 6: Demodulated GMR speed signal spectrum

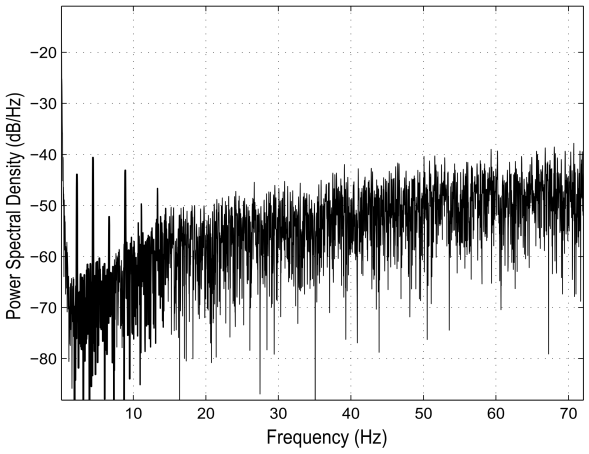


Fig. 7: Spectrum of demodulated GMR speed signal after pattern removal

frequency from the pattern frequency. Two algorithms for pattern calculations are explained in [2]. A Pattern algorithm can be used on the wrapped phase by computing the time difference between two phase shifts in the instantaneous phase. This shall give us an unambiguous pole pair width in time units. The resultant signal is a decimated version of the instantaneous phase with a frequency limited to $\Omega_o N$.

It is important to note that these results are only valid if $\Omega_o N$ satisfies the Nyquist criteria. For most practical purposes where the target is to identify low frequency vibrations which are below $\frac{\Omega_o N}{2}$, this method can be used. For high frequency vibration analysis, the computed phase signal can be re-sampled at regular intervals of angle and desired sampling frequency can be achieved.

Fig. 7 shows the final result after pattern removal, where low frequency vibrations induced from attached rotating parts are clearly distinguishable below 10 Hz. In the actual signal as shown in Fig. 5, these frequencies were modulated by the carrier frequency and are now demodulated which makes them distinguishable in the spectrum.

V. Conclusion

An analytical model of the magnetic pole wheel, as sensed by a magnetic GMR sensor,has been presented. A demodulation method based on Hilbert Transformer has been devised for the demodulation of the rotational speed signal. Frequency analysis has been presented to identify and isolate low frequency vibrations present in the rotating parts. A sliding window, frequency domain Hilbert Transform is used to compute the instantaneous phase of rotating magnetic encoder. This technique can be used in virtual sensing applications where direct measurement of vibrations is either costly or not possible by using existing sensor. In these scenarios, off-line signal processing can be used to identify vibrations present in the rotating parts of a machine or a vehicle.

Acknowledgment

This project has been supported within the COMET - Competence Centers for Excellent technologies Programs by BMVIT, BMWFW and the federal provinces of Carinthia and Styria.

References

[1] S. Haykin, *Communication Systems*, 4th ed. McMaster University: John Wiley & Sons, Inc. 2001.

[2] M. Adnan, D. Hammerschmidt, *Comparative analysis of methods for computing pole angle offsets in magnetic pole wheels using ABS sensor*. PRIME, 9th Conference on. pp. 301-304, June 2013.

[3] B.P. Lathi, *Modern digital and analog communication systems*, 3th ed. Oxford University Press, 1998.

[4] A.V. Oppenheim, R. W. Schafer, *Discrete time signal processing*, 2nd ed. Prentice Hall, 1999.

[5] A. Reilly, G. Frazer, B. Boashash, *Analytic signal generation-tips and traps*. IEEE Trans. Signal Process., vol. 42, no. 11, pp. 324-3245, Nov. 1994.

[6] S.L. Hahn, *Hilbert Transforms in Signal Processing*. Norwood, MA: Artech House, Dec. 1996.

[7] S.L. Marple Jr., *Computing the discrete-time analytic signal via the FFT*. IEEE Trans. Signal Process., vol. 47, no. 9, pp. 2600-2603, Sep. 1999.

[8] J.E. Volder, *The CORDIC Trigonometric computing technique*. The institute of Radio engineers, Inc. 1959.

Thin-film CIGS Solar Modules for Design Driven Applications in the Frame of the FP7 NMP Project SolarDesign

W. Brenner*, N. Adamovic*
* Vienna University of Technology/Institute of Sensor and Actuator Systems, Vienna, Austria
werner.brenner@tuwien.ac.at

Abstract - The FP7 NMP project SolarDesign addresses the demand for aesthetically integrated photovoltaic materials by the development of novel solar cell materials, manufacturing processes and supportive actions (design toolbox). One key technology is a laser based monolithic interconnection process for thin-film solar cells. The material in focus is CIGS (Copper, Indium, Gallium, Diselenide).

I. INTRODUCTION

Photovoltaics (PV) is widely recognized as one of the key technologies for the future energy supply. The cost reduction within the last 15 years is a result of the utilization of learning effects, expansion of production capacities, extensive automation and standardization efforts [1]. For design driven PV integrated product solutions customizable shapes, sizes, colours, transparencies or specific electrical properties are required, which have a decisive influence on the acceptance on the market. SolarDesign thus concentrates on aesthetically integrated photovoltaic materials.

Designers, architects and industrial manufacturers across the world share a common interest in using Photovoltaics (PV) as a decentralized and sustainable source of energy in their designs. Promising markets such as sustainable or temporary building structures, outdoor activities, and mobile computing make necessary decentralized, attractive energy solutions. Indeed photovoltaic modules are the only viable renewable energy solution that can be integrated directly into objects such as devices, textiles and surfaces of buildings [5].

CIGS based photovoltaic (PV) devices show the highest efficiency among the different thin film solar technologies. They obtained a value of 20.3% at cell level (0.5 cm^2) that is already comparable to the 25% record in single crystal Si solar cells, while record efficiencies of 12.8% and 17.6% have been reported for CIGS cells on flexible substrates. However the main problem is to bring the high efficiencies form laboratory scale into industrial production. It has been already achieved to some extent on glass substrates (up to 15% efficiency) [10], [11] but in the realm of flexible CIGS modules the efficiencies range between 8 and 13% depending on the chosen substrate and production process. Besides the high attainable efficiencies CIGS (Copper Indium, Gallium, Diselenide) is a compound semiconductor which allows modifying its band gap and spectral response. Since the intensity of solar radiation on earth is not homogeneous at different wavelengths it is possible to adjust the band gap/absorption edge of CIGS at theoretical maximum efficiency by its composition/mixture of copper-indium-gallium–selenium, which is unique in PV.

Current generations of CIGS thin film modules are optimized for Standard Test Conditions (25°C, irradiance 1000W/m^2, spectrum of Air mass 1.5). By adjusting the compositions and spectral responses of the functional layers of Thin Film Photovoltaics not only the overall efficiency can be improved but also energy yields under diffuse light conditions or at higher operating temperatures are changeable. This uniqueness of CIGS has potential for exploitation.

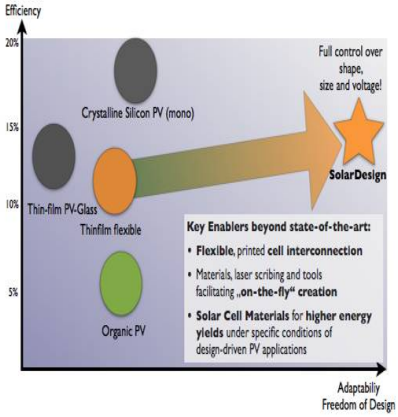


Figure 1. "Technical constraints" versus "design freedom"

Furthermore the designer who wants to integrate solar electricity into his/her work needs a service environment to be assisted in the creative process. Tools should offer support in conceiving, planning and producing the solar design products. The scientific and technical objectives (STOs) of the SolarDesign project are:

STO1: A flexible scribing and printing technology that allows producing a given photovoltaic module according to specific design requirements "on-the-fly". This flexible interconnection is applied on the solar foil (i.e. an endless solar cell) with a minimum width of 300 mm and allows curved solar cells and interconnection patterns with a minimum radius of 10 mm.

STO2: Novel materials for the underlying flexible solar cell technology to extend the design related degrees of freedom.

STO3: Novel materials for satisfying design related requirements on solar module level. Focus is laid on materials for the electrical conducting front grid to allow a high design freedom of patterns.

STO4: A methodological toolbox to provide design rules for the best solar cell super-structure and module design layout for a given application by using numerical modelling and simulation.

STO5: The developed technologies are demonstrated in the following applications: solar charging (cover for tablet PC and solar powered radio), solar powered light, solar powered sensor networks for detection of fire in forest, urban solar lighting (compact solar street lighting system) for Product Integrated Photovoltaics (PIPV), building skins with geodesic curvatures or other related forms for Building Integrated Photovoltaics (BIPV) and integration of PV in textile support.

II. SolarDesign Approach

The basic material is an endless thin-film solar cell based on the CIGS thin film technology as given below.

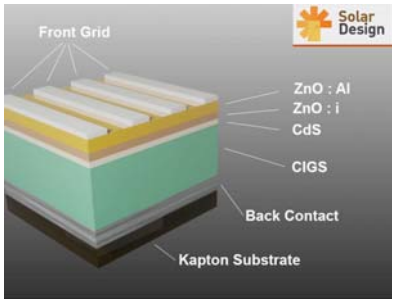


Figure 2. Cross section CIGS solar cell, SolarDesign aims at replacing CdS as buffer layer material

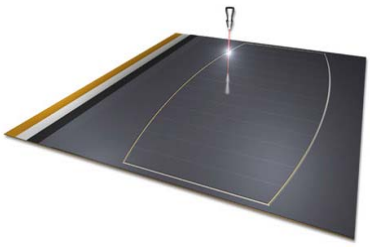


Figure 3. SolarDesign's endless thin-film solar cell

The basic fabric in SolarDesign is an endless produced solar cell. 140 µm to 180 µm wide micro-scale grooves are applied by an ultra short-pulse laser process into the solar foil. With the help of this laser process the edges of a given PV module are isolated and the thin-film layers can be ablated selectively so that the rear contact (which is initially covered by several thin functional layers) and the transparent front contact of the solar foil can be electrically interconnected. The electrical interconnection itself then can be printed in the same production step.

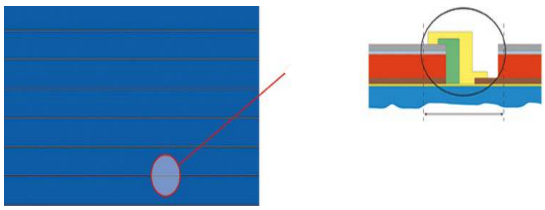


Figure 4. SolarDesign's approach: a flexible scribing and printing technology that allows producing a given photovoltaic module

Laser structuring processes for CIGS Thin-Film edge deletion: a laser structuring process is developed for the edge deletion of CIGS thin-film solar cells on glass and polyimide substrates by Munich University of Applied Sciences [13]. The transparent conducting oxides and the CIGS layers in the edge area of the solar cell are removed in this step. Then the solar cell is covered with a laminate layer which protects it against environmental influences. The edge deletion and the lamination avoid humidity diffusion through the CIGS layer into the cell and the corresponding derogation of the cell efficiency. For the edge deletion an electro-optically Q-switched 880 picosecond IR laser is used. In order to determine suitable process parameters, the pulses per position (PPP) and the fluence are varied. A process window between fluences of 0.8 and 1.5 J/cm² and a PPP of 1.6 to 6 is found. The occurring ablation mechanism is identified as lift-off process (see Fig. 5) allowing extremely efficient material removal. Removal rates down to 4 J/mm³ are determined for a fluence of 2 J/cm² and a PPP of 0.8.

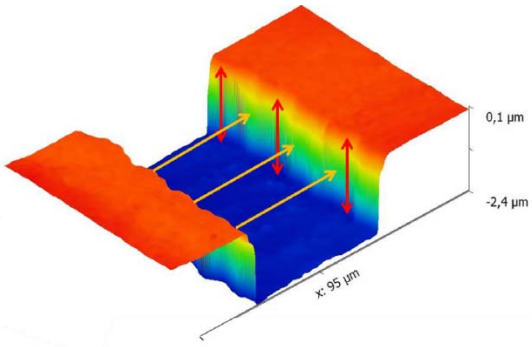


Figure 5. Confocal microscopy image showing the lift-off ablation of the TCO and CIGS layer for the edge deletion process of a CIGS thin film solar cell. Laser parameters: λ = 1064 nm, τ = 880 ps; PPP 4; F0 = 0.8 J/cm². (source: Munich University of Applied Sciences).

To understand the underlying physical mechanisms, a 3D axis symmetric finite element model was created and numerically solved. The model is verified by a direct comparison of experimental and numerical results. It includes volume absorption of the laser pulse, heat diffusion in the electron gas, and the lattice, thermal expansion of the solid phase and further volume expansion from phase transition to fluid and gas, and finally the mechanical motion of the layer caused by the resulting stress wave and the interaction with the substrate. The simulation revealed that irradiation of the Molybdenum layer with an ultra-short pulse causes a rapid acceleration in the direction of the surface normal within a time frame of a hundred picoseconds to a peak velocity about 100 m/s. The molybdenum layer continues to move as an oscillating membrane, and finally forms a dome after about 100 ns. The calculated strain at the edges of the dome exceeds the tensile stress limit at

fluencies that initiate the "lift-off" in experimental investigations. In addition, the simulation reveals that the driving mechanism of the "lift-off" is the ultrafast expansion of the interface layer and not the generated gas pressure [4], [7].

In the frame of SolarDesign M. Acciarri and S. Binetti [8] focus on the second generation thin film devices: the alloy Cu(In,Ga)Se2 (CIGS) which is the most promising material due to the high efficiencies. Nowadays there are two main approaches to produce CIGS solar cells: sputtering followed by selenization evaporation and coevaporation process. High efficiency solar cells are obtained by the sequential three stage co-evaporation process. On the other hand sputtering system is more suitable for an industrial application but the efficiencies are systematically lower. In SolarDesign a new hybrid sputtering and evaporation system was developed to obtain a scalable process combining the advantages of both techniques. This technique has been already applied to deposition on glass and stainless steel substrates. Good results were obtained reaching an efficiency of 15% on glass and 13% on stainless steel substrates respectively. A new research line on the alternative buffer layer is developed by UNIMIB. Thereby, one main thrust in the CIGS solar cells is to replace the cadmium sulfide CdS buffer layer with other material. Currently, a CdS thin film deposited by Chemical Solution Deposition (CSD) technique is utilized as a buffer layer in CIGS solar cells. The further process is not manufacturing-friendly due to the use of cadmium and the liquid waste. Thereby, the cell production is limited by the window materials, and there is need to develop wider band gap buffer layers alternative to CdS. The UNIMIB material candidates as alternative buffer layers in CIGS solar cells include zinc tin oxides (ZnxSnyOz), indium sulfide (InxSy) and zinc sulfide (ZnS) deposited by radio frequency sputtering technique which can be easily integrated into a roll to roll system.

Recently, a special attention is on the deposition of the CIGS absorber layer on flexible substrates [4], [8].

Sunplugged successfully installed a hybrid sputtering and evaporation prototype which is based on the CIGS formation process initially developed by University Milano-Bicocca [13].

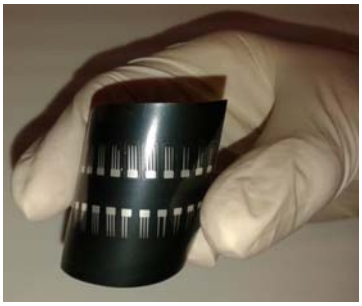


Figure 6. Flexible substrate (source: UNIMIB)

The full-scale prototype is able to produce the absorbtion layer endlessly on flexible substrates up to a maximum of 320 mm. Running in full production mode the prototype is able to produce about 10 running meters per hour.

Figure 7. Hybrid sputtering and evaporation prototype based on the CIGS formation process process initially developed by University Milano-Bicocca (source: Sunplugged)

A typical problem of PV systems is the power loss due to temperature increase, because modules often operate close to the building resp. product envelope with low ventilation. EURAC therefore evaluated and compared the PV temperature conditions of different PV module categories (in terms of PV technology and material type). A simple linear expression for the evaluation of the PV module temperature is $T_{mod}=T_{amb}+kG$ which links T_{mod} with the ambient temperature T_{amb} and the incident solar radiation flux G. Within this expression the value of the dimensional parameter k, known as the Ross coefficient, depends on several aspects (i.e. module type, wind velocity and integration characteristics). However, dispersed values for this parameter can be found in literature (in the range of 0.02-0.06 K m^2/W) according to different module types [12].

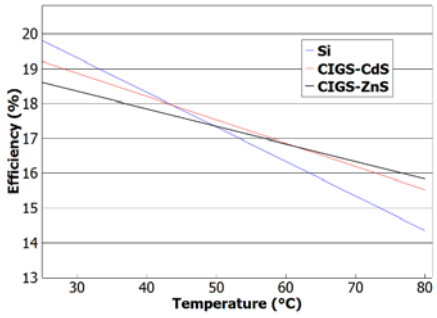


Figure 8. Overview of temperature coefficients of different thin film photovoltaic technologies [9]

As can be seen in Figure 8, the efficiency decrease of ZnS-buffered CIGS (Copper, Indium, Gallium, Diselenide) is smaller than for CdS-buffered CIGS. With further improvement of alternative buffers (like ZnS, ZnMgO, Zn(O,S)), the efficiency values of CdS-buffered CIGS at 25°C can be reached (demonstrated by Helmholtz Zentrum Berlin [3]. Due to lower temperature coefficient even higher efficiencies at operating temperatures (typically ~60°C for midsummer day) can be expected for alternative buffers – without using toxic Cd. For deposition of different buffer layer materials both sputtering and wet chemical process are available.

Vienna University of Technology (TUW) has finished development of the numerical model that describes solar modules based on CIGS technology. The model has two aspects. Firstly, it is used to optimize the fabrication technology by exploring the impact of physical

parameters of the solar cells. The simulation of the technology process is indispensable, as it saves both time and recourses by reducing the number of tests.

Figure 9. Application street lamp, an example for the need for reduced temperature coefficient to guarantee minimum values of luminescence which are standardized for street lighting, source: SIARQ)

Secondly, the developed model is used to meet the needs of the designers/architects. Specific user defined shapes of modules can be easily achieved, while respecting the electrical requirements. The design of metallization lines with varying thickness or shape, as well as arbitrary banding profiles is enabled. The simulations of the solar modules under different illumination levels has been taken into account, thus enabling the calculation of the daily or yearly energy performance of the considered modules [4].

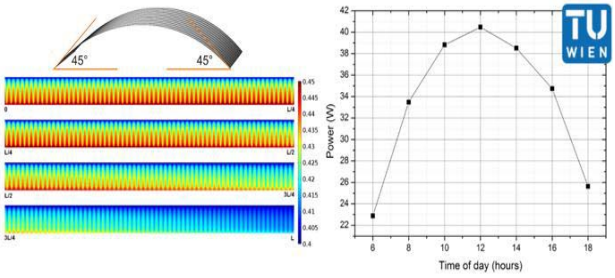


Figure 10. Example of the simulation of an arbitrary curved module: a) module geometry, b) voltage distribution along one cell, b) calculated power yield during a day [source: TU Wien]

The technologies for the printing step are screen printing and dispensing. For very small scale production (lot size below 200) this process is unsuitable because every pattern requires its own screen. Also after the printing of the isolation paste thermal curing needs to be applied. The main advantage of the dispensing technology is that it can be fully digitized and a "lot size one" is possible. It is an addition to the screen printing to fulfill customer demands regarding design for predefined cost limits.

Series interconnection of thin film solar cells is done by monolithic interconnection. It connects the transparent conductive oxide (TCO) of cell 1 with the rear contact of cell 2 and so on. While the rear contact which is normally a high conductive metal has at least a conductance of 35% compared to copper the commonly used TCOs like ITO (Indium Tin Oxide) or AZO (Aluminium-doped Zinc Oxide) have a conductance of less than 2%. This value depends of course of the thickness of the TCO and the layer beneath but a thicker TCO does also mean less transmittance thus reducing the efficiency. One improvement is the use of a metallic front grid, a standard solution for thick film solar cells but still in its infancy stage for thin film applications. There has to be found equilibrium between high conductivity and reduced shading of active areas. One objective in question is to find a solution to this problem coupled with an optimized pattern for front grids.

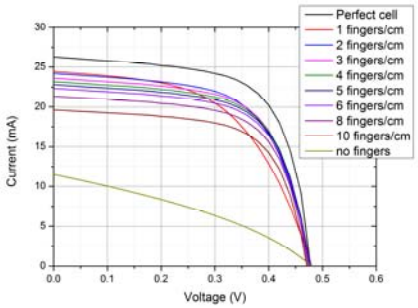


Figure 11. Metallization grid optimization, d_{Ag} = 20 µm, this provides good conductivity, therefore the width of the fingers can be a minimum: w_{finger} = 200 µm

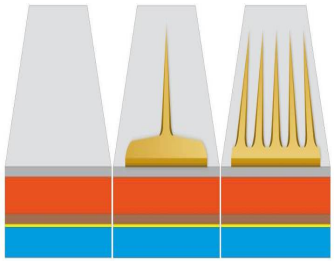


Figure 12. Different layouts of the metallic front contacts. The challenge for the design of the front contact lies in optimization of contradicting parameters like high conductivity and minimized losses by shading

In principle such a front contacted photovoltaic cell body is structured in three parts and two connecting interface regions in between with very different electrical and optical properties. The semiconductor layer (e. g. CIGS/n-layer), the transparent conductive oxide layer (TCO) and the metal grid (metal): the bottom semiconducting absorber layer has optimized high optical absorption and low reflectivity, the intermediated TCO-layer has a low optical reflectivity and absorption, and low electrical resistance, the top metallic grid layer has very high optical reflectivity and almost no optical transmission but very low electrical resistivity.

Encapsulants for CIGS solar modules require attention. The potential solution is a compromise between several technical and economical constraints: it has to be thin, flexible and transparent. It needs to protect the solar cell against external shocks and mainly against moisture. Suitable low-cost high performance encapsulation strategies are a key challenge to achieve sufficient lifetime. Indeed, most of these promising thin film technologies (CIS, CIGS, organic etc.) have to be protected with a gas barrier material to reach sufficient lifetime for the dedicated applications. In order to increase the barrier properties, Atomic layer deposition

techniques are mainly proposed in order to make deposition of several atomic layers of Al_2O_3. The encapsulation process needs the lamination of a barrier foil with gas barrier protection only reachable by depositing alternatively organic and inorganic layers on a polymeric substrate or by lamination of several single barrier films. In addition, photochemical stability of barrier films is an important aspect in the case of PV devices. Nowadays, commercial solutions appear at a pre-industrial level but some questions remain not completely resolved as the cost, stability, lamination process (adhesive), connection etc.

Figure 13. Color variations providing more freedom of PV design, using different novel encapsulants allows custom designed optical appearance (source: CEA)

SolarDesign is working on the development of a user oriented software tool taking into consideration all the parameters for the establishment of a numerical modelling and simulation tool that allows gathering the necessary information during the design process.

The actual innovation of this tool is that enables the user to calculate solar radiation received on any shape, allowing simulating any scenario on a complex shape. The user oriented software tool provides design rules for the best solar cell super-structure and module design layout, and addresses the efficiency in the incorporation of thin film to the different products or structures designed. The tool will help users to calculate a performance simulation to subsequently design the electronic and pattern for the production process. To develop and test the software tool, a building structure with a characteristic and complex shape called "domo" has been chosen on a first approach.

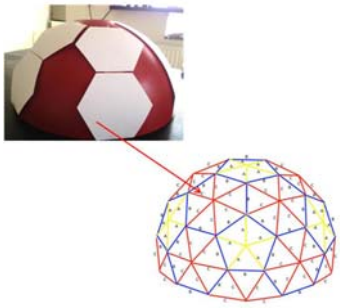


Figure 14. Domo Structure (Source: Julien Bollaro, GAIA, Application simulant la radiation solaire sur une installation afin d'assister le développement de la technologie Thin-Film dans le cadre du projet européen: SolarDesign. 2013)

III. Multi Tools Marketing Methodology

SolarDesign is an SME-targeted collaborative project within the work program topic "Novel materials and design-based solutions for the creative industry". Following this philosophy CEA has defined and applied a multi tools marketing methodology which can be used for projects with several potential key results. For SolarDesign the methodology aims at prioritizing 4 markets (outdoor lighting, BIPV, consumer products and textile) by considering their attractiveness vs. SolarDesign's technology assets. Four different actions were lead: a documentary research based on patents and economic data's analysis, a "field analysis" with qualitative interviews of well-chosen actors, a "quantitative analysis" inspired by Kano approach and a long-term news watch disseminated to the members through a share point dedicated to the project [4].

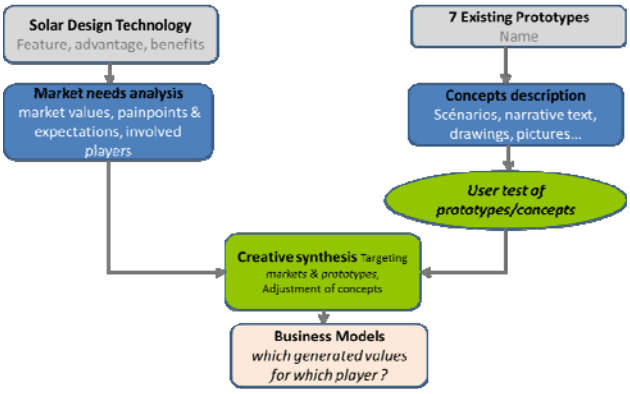


Figure 15. Market analysis (source: CEA)

IV. SolarDesign Consortium

The consortium formed by eleven partners belonging to six different countries, covers the needed expertise, namely, PV design, materials research, thin-film deposition and characterization, laser processing, interconnection manufacturing, and integration in products (PIPV) and buildings (BIPV). E.g.: company Sunplugged is developing own CIGS thin-film solar cells on flexible substrates production line. Vienna University of Technology (TUW) provides the numerical modeling and simulation tool that allows prediction of energy yields, performance, electrical-, optical-, thermal- and mechanical characteristics during the design process. Munich University of Applied Sciences (MUAS) has been involved in research of laser scribing processes of CIGS thin-film solar cell. Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) develops encapsulates for modules (new materials, new pigments, up scaling up to large surfaces). CEA has experience in BIPV (integration of modules into tile) and via CEA-INES has the module manufacturing capabilities. Consortium:

- Technische Universität Wien (TUW) Austria www.isas.tuwien.ac.at
- Sunplugged GmbH (SUN) Austria www.sunplugged.at
- Faktor 3 ApS (F3) Denmark www.faktor-3.dk
- Innovatec Sensorisatión y Comunicación S.L (INN) Spain http://www.innovatecsc.com

- Studio Itinerante Arquitectura S.L. (SIARQ) Spain www.siarq.net
- RHP Technology GmbH & Co KG (RHP) Austria www.rhp-technology.com
- Asociación de Industrias de las Technologias Electrónicas y de la Información del País Vasco (GAIA) Spain www.gaia.es
- Munich University of Applied Sciences (MUAS) Germany www.fb06.fh-muenchen.de/lhm
- Accademia Europea Bolzano (EURAC) Italy www.eurac.edu/en/eurac/welcome
- Università degli Studi di Milano-Bicocca (UNIMIB) Italy www.unimib.it/go/102/Home/English
- Commissariat à l'energie atomique et aux energies alternatives (CEA) France www.cea.fr

V. STANDARDS

The FP7 NMP SolarDesign Project is focused on new PV integrated product solutions. To achieve this, the project has the goal to develop new materials, flexible production - and business processes in PV powered product design and architecture. SolarDesign has decided to integrate standardization into the project's efforts [2]. Standards can enhance the economic value of research and innovation projects. Main questions according to [6] are:

- Are SolarDesign's R&D results expected to be compatible and interoperable with other technologies?
- Is there a need to ensure reliability, safety and comparability of R&D activity or result?
- Will results have a potential European or international market application?

Integrating photovoltaics in architecture and industrial design is in its infancy and taking into account different European perspectives is crucial for its success. This is especially true for Building Integrated PV where legal conditions like construction codes or feed-in regimes are differing from one European country to another. By seeing design from a stylish or sensory perspective regional characteristics exist. Taking product designers and PV experts from different European countries can ensure the incorporation of varying design manifestations. Dissemination, implementation and standardization cannot be seen as separate activities. They need coordination because they represent two sides of the same plan for market access. In the next step CEN/CENELEC was identified as relevant standardization partner. By this it is ensured that results are widely accepted and available.

VI. CONCLUSION

A growing number of designers, architects and industrial manufacturers across the world share a common interest in aesthetically integrated photovoltaic materials as a decentralized and sustainable source of energy in their product designs. For solar powered products customizable shapes, sizes, colours, transparencies or specific electrical properties are required. In SolarDesign a new hybrid sputtering and evaporation system was developed to obtain a scalable process combining the advantages of both techniques. This technique has been already applied to deposition on glass and stainless steel substrates. Good results were obtained reaching an efficiency of 15% on glass and 13% on stainless steel substrates respectively. Research on alternative buffer layers focuses on replacing CdS.

ACKNOWLEDGMENT

SolarDesign http://www.solar-design.eu has received funding from the EC 7th Framework Program FP7-NMP-2012-SME-6, Grant Agreement No 310220.

REFERENCES

[1] B. Parida et al.: A review of solar photovoltaic technologies, Renewable and Sustainable Energy Reviews 15 (2011) 1625–1636

[2] W. Brenner, N. Adamovic: Benefits of Standardization in the Research and Innovation Process of the FP7 NMP Project SolarDesign. in: P. Biljanović (ed.): Proc. of mipro2014, the 37th Int. Convention, May 26-30, 2014 Opatija, Croatia, ISSN 1847-3946, http://mipro-proceedings.com/en/inova

[3] 3rd International Workshop on CIGS Solar Cell Technology, Berlin, 17.-19.04.2012

[4] J. Garatea, I. Muñoz (eds.): SolarDesign newsletter 3,. http://www.solar-design.eu/component/phocadownload/category/2-public-downloads

[5] S. Ishizuka et al.: Monolithically integrated flexible Cu(In,Ga)Se2 solar cell submodules, Solar Energy Materials and Solar Cells 94 (2010) p.2052

[6] 2056http://www.cencenelec.eu/research/WhyStandards/Pages/default.aspx

[7] J. Sotrop, A. Kersch, M. Domke, G. Heise, H.P. Huber: Numerical simulation of ultrafast expansion as the driving mechanism for confined laser ablation with ultra-short laser pulses, Applied Physics A: Materials Science and Processing, Vol. 113/Issue 2 (2013) 397-411

[8] M. Acciarri, A. Le Donne, M. Morgano, L. Caccamo, L. Miglio, S.Marchionna, R. Moneta, M. Meschia and S. Binetti,"Hybrid sputtering/evaporation deposition of Cu(In,Ga)Se_2 thin film solar cells", Energy Procedia 10 (2011) 138 – 143.

[9] A. Murawski: Final report of initiative project *Photovoltaikfolie für die Integration in solarthermische Kollektoren*, Land Tirol (2011)

[10] S. Seike et al.: Development of high-efficiency CIGS integrated submodules using in-line deposition technology, Solar Energy Materials & Solar Cells 95 (2011) 254–256

[11] M. Powalla et al.: Highly efficient CIS solar cells and modules made by the co-evaporation process, Thin Solid Films 517 (2009) 2111–2114

[12] L. Maturi, G. Belluardo, D. Moser, M. Dell Buono: BiPV System Performance and Efficiency Drops: Overview on PV Module Temperature Conditions of Different Module Types. Energy Procedia 48 (2014) 1311 - 1319

[13] J. Garatea, I. Muñoz (eds.): SolarDesign newsletter 4,. http://www.solar-design.eu/component/phocadownload/category/2-public-downloads

Measurement System for Object Detection Based on Multielectrode Capacitive Sensor

D. Ćika*, T. Martinović** and H. Džapo**
*Research and Engineering Center d.d., Zagreb, Croatia
**Faculty of Electrical Engineering and Computing, Zagreb, Croatia
drazen.cika@rec-global.com

Abstract - The problem of detecting objects, people and events is encountered in different forms in many application areas. In this paper we investigate the applicability of multielectrode capacitive sensors for monitoring the volume of interest by analyzing the disturbances in electrical field due to the influence of external objects. Capacitive sensors are particularly interesting as a low cost solution with minimum privacy issue concerns. The goal of this research was to characterize the performance of modern capacitance-to-digital converter (CDC) PicoCap PCap02 integrated circuit for detecting various types of objects in a multielectrode system configuration, with human detection as an example application. We present an analysis of an appropriate electrode geometry, propose a custom-designed data acquisition system which complies to the required capacitance measurement range and resolution, discuss sources of measurement errors and methods for minimization of their influence, and elaborate the applicability of the method beyond the simple object detection, by investigating the possibility of object class discrimination from acquired data.

Keywords: capacitive sensors, object detection, data acquisition, electronic instrumentation

I. INTRODUCTION

Object detection is the problem encountered in different forms in many application areas, such as industrial quality control, production process monitoring, security systems, traffic monitoring, inventory tracking etc. Intensive research led to the development of methods and techniques to solve particular practical problems for data acquisition and processing based on information acquired by different types of sensors, including cameras, motion detection sensors (ultrasound, infrared, radars), capacitive and inductive sensors, planar force detection sensors, acoustic sensors etc.[1]. Most advanced methods and achievements have been developed in the area of computer vision, using camera as data acquisition sensor, and employing image processing and machine learning algorithms to solve many practical problems in the field of detection and classification of objects and people. However, those methods are usually computationally demanding, and not suitable for low power embedded systems that are commonly encountered in practice. There are also privacy concerns which might exclude camera sensors in some applications.

Capacitive sensors are interesting due to possibilities for simple and low cost production (both electrode manufacturing process and data acquisition circuitry), and capability for detection of non-magnetic objects, without a need for direct galvanic contact. These characteristics make capacitive sensors an attractive choice also for non-metallic and non-magnetic object detection and discrimination, and also for applications including detection of people and animals. Capacitive sensors are used in many different applications [2], [3], such as precise displacement measurement (micrometers), proximity detection, pressure and force measurement, liquid level measurement, human-machine interfaces (touch screen displays, keyboards), inertial sensors (accelerometers), environmental conditions monitoring etc. [4]

The majority of advances in the area of multielectrode capacitive sensors applications has been achieved for human-machine interfaces, namely touch screen that must precisely determine a finger position in a real-time [5], [6]. Prior work in the area of person detection by means of capacitive sensors proposed using large electrodes in horizontal or vertical plane for detecting persons on the move [7], [8], detecting hand movement in the proximity of the metal processing machinery for increased operator safety [9], or detecting vehicle seat occupancy by electrodes embedded in seats [10], [11], [12]. Previous research also confirmed the applicability of multielectrode capacitive sensors systems for detection and classification of events, human activity, body position and movement [13], [14], [15], by using machine learning approach for sensor data processing.

In this article we investigate the possibility of object detection and discrimination by means of a custom-designed multielectrode capacitive sensor, with a signal conditioning part based on a single chip capacitance-to-digital converter (CDC). The integrated circuit PCap02 (manufactured by acam-messelectronic gmbh) was chosen due to its exceptional measurement characteristics, comparing with other available CDC solutions. The goal of the research was to assess whether the approach to people and object detection by using low-cost compact CDC-based integrated circuits is a viable solution for practical use cases.

II. PCAP02 CAPACITANCE-TO-DIGITAL-CONVERTER

A. PCap02 Integrated Circuit Characteristics

Typical PCap02 application schematic diagram when I^2C interface is used is shown in Fig. 1.

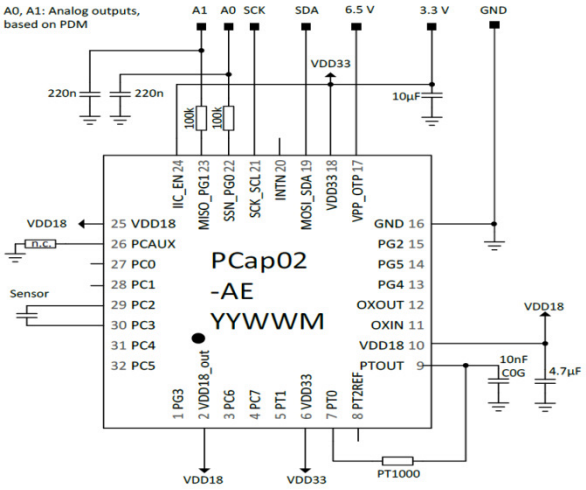


Fig. 1. PCap02A typical schematics, I2C interface

Detailed characteristics and explanation of principle of operation for this component can be found in [16], [17].

PCap02 has the digital signal processor (DSP) integrated in order to achieve on-chip data post-processing in addition to the basic converter functionality. The PCAp02 can handle larger offset capacitance on its inputs than some other comparable CDC chips, therefore it is inherently less susceptible to stray capacitance caused by connecting wires. Table 1 shows dependency between the output data rate and capacitive noise and resolution for 10 pF base and 1 pF span. Only part of the table for floating sensor configuration is shown here, more information can be found in [16].

The PCap02 is based on the discharge time measurement principle, which means that the capacitance is determined from the time needed to discharge the RC-network. The time is measured by high resolution time-to-digital converter (TDC), and the ratio of discharge times is directly proportional to the ratio of capacitor under test and the reference capacitor. The discharge time depends on the capacitor under test, and discharge resistor which can be selected under software control.

Output Data Rate [Hz]	FLOATING Fully compensated		
	RMS Noise [aF]	Eff. Resolution 10 pF base [Bits]	Eff. Resolution 1 pF span [Bits]
2.5	15	19.3	16.0
5	23	18.7	15.4
10	35	18.1	14.8
25	48	17.7	14.4
100	134	16.2	12.9
250	172	15.8	12.5
1,000	330	14.9	11.6
2,000	438	14.5	11.2
4,000	603	14.0	10.7
10,000	838	13.5	10.2

Table 1. PCap02 RMS noise and resolution

B. Connecting Sensors to CDC

Various configurations are possible when capacitance sensors are used, as shown in Fig. 2 and Fig. 3. Sensor plates can be connected in a single or differential configuration, and each can have external ground reference, or it can be floating. Depending on the chosen configuration, different number of electrodes can be used since the total number of measuring inputs on the PCap02 CDC unit is limited to eight, including the input for the reference capacitance measuring. If the single floating configuration is used, four capacitive sensors can be connected to the single CDC unit. Shielded cables are recommended to compensate the external parasitic capacitances against the ground for longer cables, as shown in Fig. 4. This measure alone may not be effective in some cases, so active shield driving might be necessary.

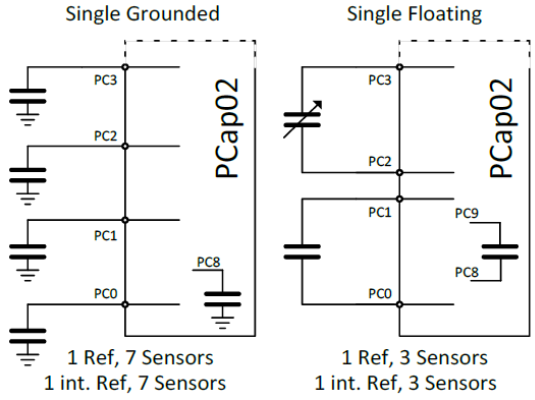


Fig. 2. Options for connecting sensors - single inputs

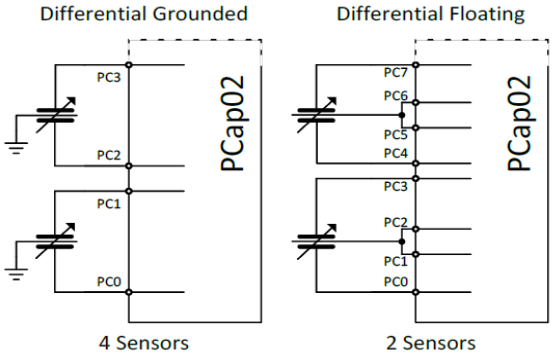


Fig. 3. Options for connecting sensors - differential inputs

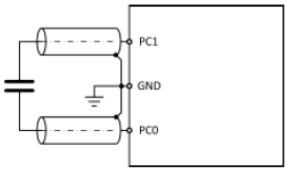


Fig. 4. Connection of shielded cables

III. DATA ACQUISITION SYSTEM PROTOTYPE

A. PCap02 Breakout Board

Although there is readily available PCap02 evaluation board (PCap02-EVA-Kit [18]), we decided to fabricate our own small breakout board, that is more suitable for cases where portability and small dimensions are needed. The breakout board contains only minimum set of components necessary for the PCap02 operation and communication with the host microcontroller. Breakout board dimensions are 20 x 20 mm.

The hand soldered PCap02 ready for test is shown in Fig. 5. The breakout board schematic is shown in Fig. 6. PCap02 pins necessary for functions not implemented on the breakout board are available at header pins distributed along the breakout board edges with the standard 100 mils pin spacing.

The power supply is not part of the breakout board, and user should provide the necessary 3.3 V DC voltage to the board. The 1.8 V DC is generated internally in the PCap02, and available at external pins for powering other components if necessary. The PCap02 breakout board power consumption measured on the prototype is 15 µA with I^2C interface used for communication with the host microcontroller. Additional power consumption should be added to the system power requirements, depending on the microcontroller type and clock speed selection, ratio of active versus sleep time operation, and other microcontroller operation specific parameters.

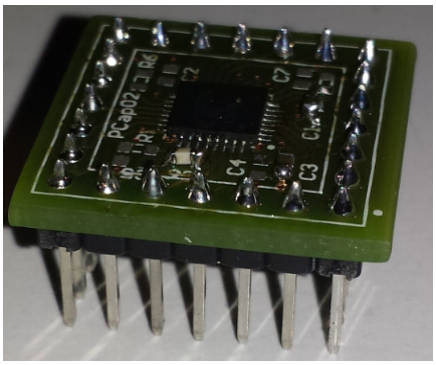


Fig. 5. PCap02 breakout board ready for test

The serial bus communication mode with the host microcontroller (MCU) is preselected by soldering the R3 solder bridge (0 Ω resistor) for the SPI, or the R4 solder bridge for the I^2C interface. The I^2C is more convenient when several boards needs to be connected with a minimum wiring, while the SPI offers higher data transfer rates, which was not a requirement for this project.

B. CDC Module and Firmware Support

The ARM Cortex M4 core family microcontroller was chosen for implementation of the experimental capacitive sensor data acquisition system, and the STM32F4 Discovery evaluation board is used [19]. Firmware was developed in the ARM Keil© uVision integrated development environment (IDE) [20] by defining and implementing hardware abstraction layer (HAL) C functions for the PCap02 sensor integration and support. In case of the MCU platform change, even to some other architecture, changes in the code would be minimized as there are no hardware dependencies.

The I^2C communication and other hardware dependent functions are implemented using the STM32F4 DSP and standard peripherals library [21], and some freely available open source wrapper functions on top of this library [22]. During the experiments, communication with the host PC was necessary in order to transfer data and display graphs, and the USB port was used for that purpose as shown in Fig. 7. The complete environment including the bench type laboratory power supply and a person standing between two electrodes while doing the measurement is shown in Fig. 8.

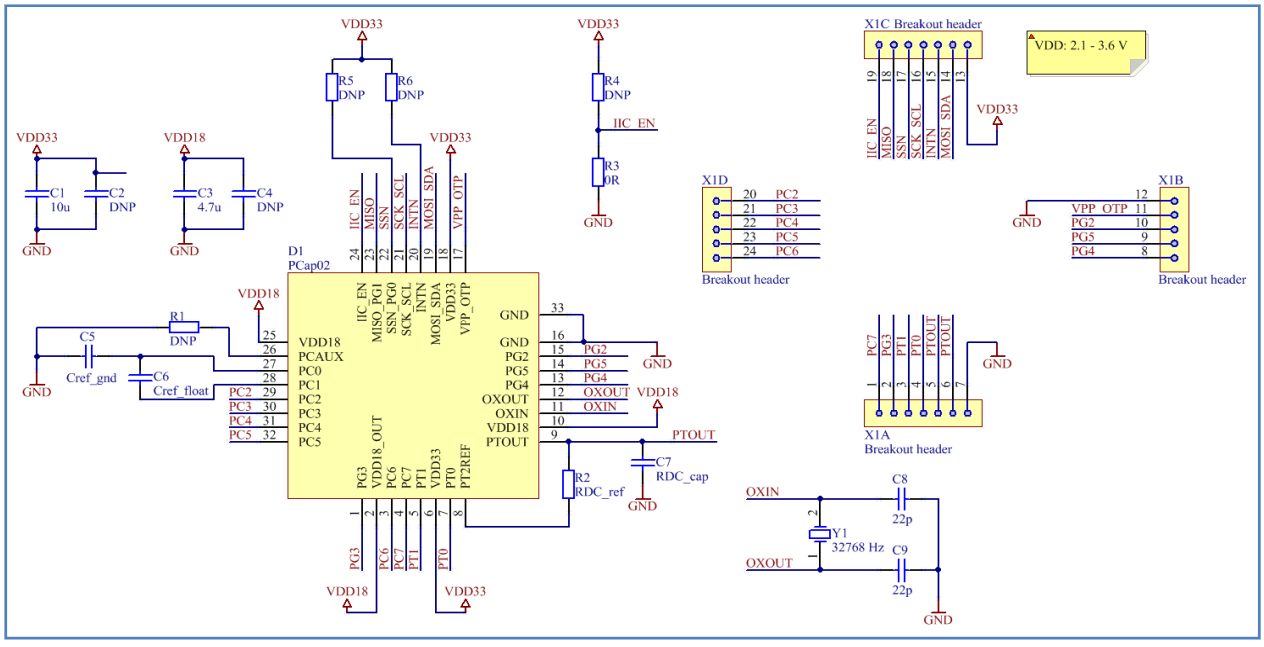


Fig. 6. Schematic diagram for the custom designed PCap02 breakout board

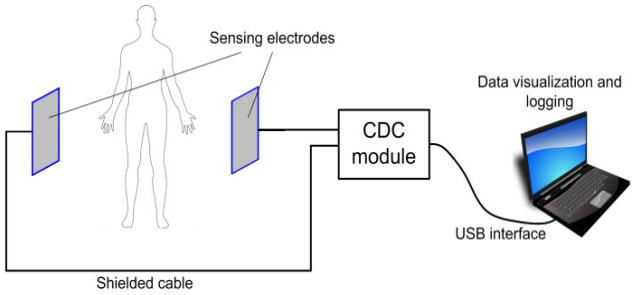


Fig. 7. Measurement setup block diagram

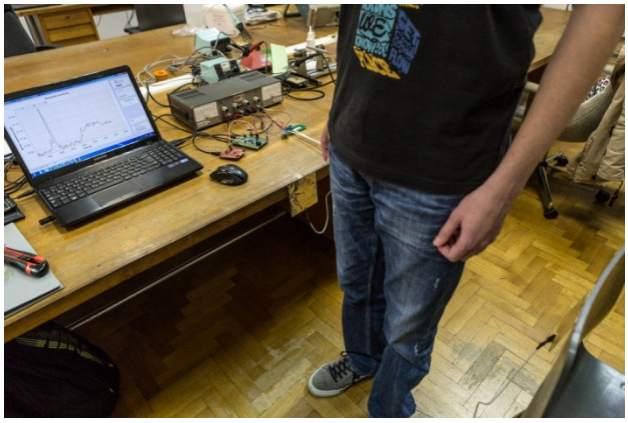


Fig. 8. The measurement setup including single pair of electrodes

IV. SIMULATION AND MEASUREMENT

A. Measurement Setup

The block diagram of the basic measurement setup already shown in Fig. 7 corresponds to the photograph shown in Fig. 8. For different electrode configurations the measuring setup would only be altered regarding the number of inputs. As long as up to three channels are needed, single PCap02 can be used, otherwise additional breakout board should be used.

The multielectrode measurement system can be arranged so that each pair of electrodes captures different area of interest. In the experimental setup described here only one pair of electrodes was used at the time, but the electrode placement and object orientation was varied to observe the resulting capacitance change.

B. Electrostatic Simulation

The expected experimental results can be estimated by means of electrostatic simulation, with different level of accuracy, depending on the tools available and quality of models that can be used with these simulation tools. In this research the simulations were performed by using The Robin Hood Solver [23] because it is fast, it has low memory requirements, scripting in the Lua programming language is supported, and it is easy to use with licensing policy suitable for academic use.

The simulation setup used two different electrode pairs (10 x 10 cm and 20 x 10 cm), and each electrode was divided into 1800 triangle elements. Electrodes were separated by 80 cm thus simulating the standard door.

The simulations were performed by using the metallic cylinder as very simplified model of the human body with satisfactory approximation results. The cylinder diameter was varied in the range of 30 cm, 35 cm, and 40 cm with 180 cm height, and it was divided into 5108 triangle elements for simulation purposes. By using the Robin Hood Solver scripting language, cylinder (i.e. person) is programmed to pass in the middle of the volume between the plates. Simulation results are shown in Fig. 9.

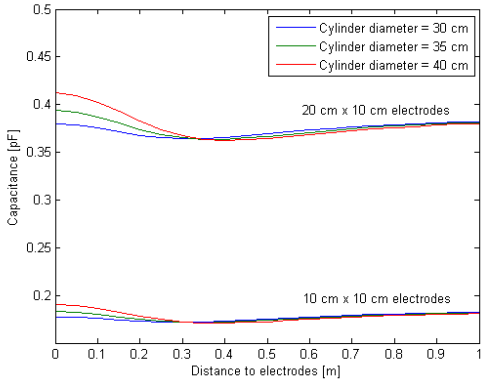


Fig. 9. The Robin Hood Solver simulation results

In a zero position, the cylinder is placed exactly between the sensing plates, and the distance is increased to 1 m in 5 cm steps. Two different sets of electrode dimensions were simulated, 20 x 10 cm, and 10 x 10 cm. The resulting capacitance change is between 15 fF and 50 fF, depending on a sensing plate size and cylinder diameter. While even smaller sensing plates could be used, resulting capacitance changes would be barely noticeable due to parasitic capacitance changes representing noise to the measuring system. These are very demanding and challenging requirements for the measurement system.

C. Measurement Results

The cable connection to the sensor plates is critical, and if the wiring is not done properly it can introduce noise, or result in much smaller measuring span. Measurements were performed by using:

1. unshielded cable,
2. shielded cables with grounded shield,
3. shielded cables with active shield driving.

First experimental setup was similar to those described in [10], [15]. The sensing plates are positioned on the chair backrest and seating area, person was standing, seating, and seating away from the backrest. Plate size is 20 x 30 cm. The results are shown in Fig. 10.

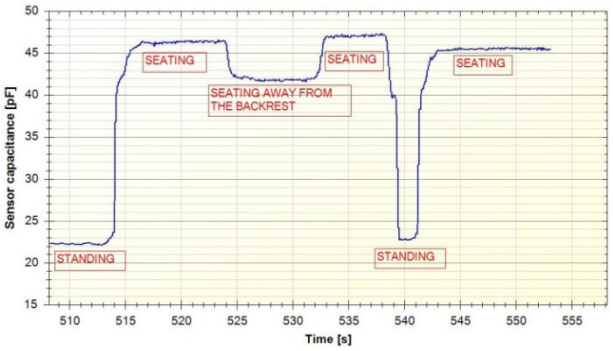


Fig. 10. Sensor electrodes positioned on the chair

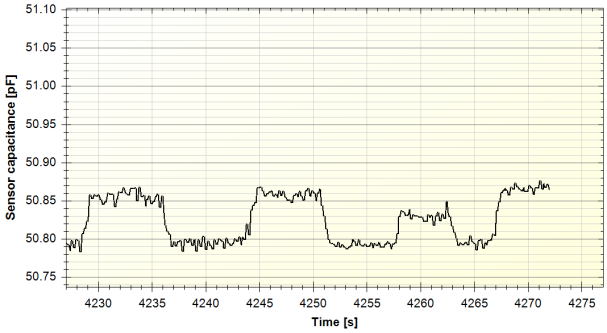


Fig. 11. Active shield, 20 x 15 cm electrodes, 80 cm distance

The distance of only few centimeters from seating person to the sensors could explain much higher capacitance values than simulated and found in following experiments. The unshielded cables also act as a part of electrode and contribute to the total measured capacitance. The noise visible in the results looks insignificant, but it is in the tenths of picofarad range, which is well above the capacity change predicted by the simulation and found in the following 'doorway' measurement setups where object was more distant from the sensor plates.

The second experimental setup was using the shielded cable grounded at the PCB side, but the longer cable still contributes to the input capacitance, and shielded cable of one meter length and more could saturate the inputs of the PCap02.

Best results were obtained in the third experimental setup, when the cable shield is driven with the same signal used for capacitive sensor excitation. Measurement results are shown in Fig. 11. The higher capacitance values correspond to the person passing through the doorway, and standing in the space between the sensor plates for about five seconds. First two peaks were obtained for 'normal' front facing orientation, and third peak which is a bit smaller in capacitance was measured for sideway walking and standing. This is result of a position where the body is a bit more away from the sensor.

The active shield measurement option was realized by using the same signal for the shield drive as the one driving the sensor plates. The TL082 operational amplifier in voltage follower configuration was used for each plate (two amplifiers per measuring channel), as shown in the schematic diagram in Fig. 12.

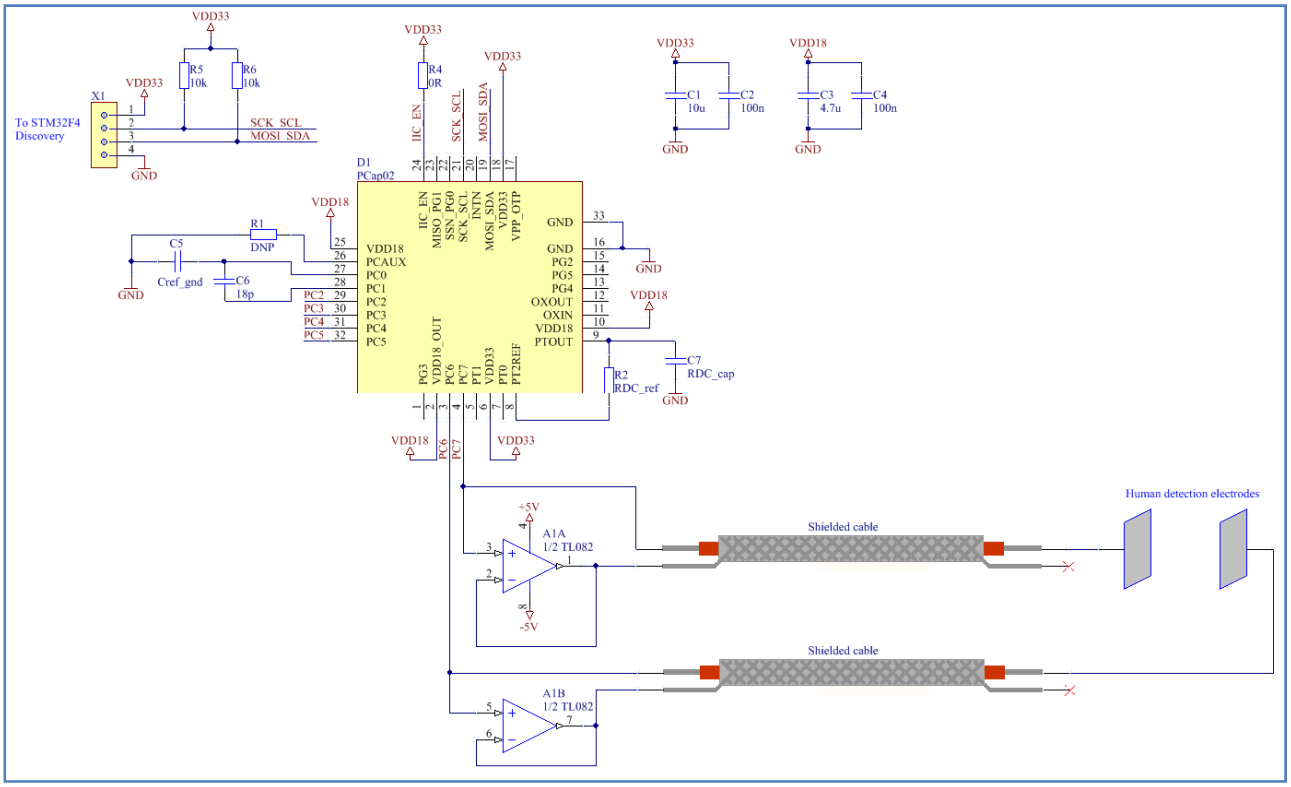


Fig. 12. Complete capacitive sensing circuit (without MCU board) with active shield driving

With the electrodes disconnected, the noise was around 1 fF peak to peak. With electrodes connected using the overall cable length of 3 meters, the noise was around 20 fF, less than half of approximately 50 fF capacitance change detected when person passed through the area observed. Measuring span is 144 pF for the 18 pF reference capacitance selected (span = 8 x C_{ref}).

V. CONCLUSION

The PCap02 CDC capacitive sensor can be used for object detection and discrimination applications based on the multielectrode capacitive sensors. Cabling is critical, and after active shield drive technique was used, the noise was reduced below the detection threshold using the reasonably sized electrodes, while in the same time cables longer than one meter are used. The firmware support developed can be easily applied to other low power MCU architectures, and portable battery operated device can be designed which would autonomously operate for extended period of time. The sensitivity and speed of the measurement setup is high enough to detect if the person was passing slowly or running, or is it standing facing front or sideways, if supported by appropriate detection and discrimination algorithms.

ACKNOWLEDGMENT

The research leading to these results has received funding from the project "Information and communication technology for generic and energy-efficient communication solutions with application in e-/-/m-health (ICTGEN)" co-financed by the European Union from the European Regional Development Fund.

REFERENCES

[1] T. Texiera, G. Dublon, A. Savvides, "Survey of Human-Sensing-Methods for Detecting Presence, Count, Location, Track, and Identity", ENALAB Technical Report 09-2010, vol. 1, no. 1, pp. 1 - 41, September 2010

[2] L. K. Baxter, "Capacitive Sensors - Design and Application", IEEE Press, NY, 1997.

[3] X. Hu, W. Yang, "Planar capacitive sensors - designs and applications", Sensor Review, vol. 30, no. 1, pp. 24-39, Emerald Group Publishing, Bingley, 2010

[4] A. Troiano, E. Pasero. L. Mesin, "New System for Detecting Road Ice Formation", IEEE Transactions on Instrumentation and Measurement, vol. 60, no. 3, pp. 1091 - 1101, March 2011

[5] T. G. Zimmerman, J. R. Smith et al. "Applying Electric Field Sensing to Human-Computer Interfaces", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 280-287, 1995

[6] J. R. Smith, "Electric Field Imaging", PhD thesis, MIT Media Laboratory, Cambridge, MA, 1999

[7] R. N. Aguilar, M. Roelofsz, H. Kerkvliet, G. C. M. Meijer, "Capacitive Human-detection Systems with Auto-calibration", The 9th World Multi-Conference on Systemics, Cybernetics and Informatics, vol. 9, pp. 290-296, Orlando, FL, 2005

[8] F.N. Toth, J.A. Zapico, G.C.M Meijer, "A Low-Cost Capacitive Personnel Detector with Fuzzy Logic Position Estimation",

[9] M. Neumayer, B. George, T. Bretterklieber, H. Zangl, G. Brasseur, "Robust Sensing of Human Proximity for Safety Applications", Instrumentation and Measurement Technology Conference, pp. 458 - 463, Austin, TX, 2010

[10] B. George, H. Zangl, T. Bretterklieber, G. Brasseur, "Seat Occupancy Detection Based on Capacitive Sensing", IEEE Transactions on Instrumentation and Measurement, vol. 58, no. 5, pp. 1487 - 1494, 2009

[11] B. George, H. Zangl, T. Bretterklieber, G. Brasseur, "A Combined Inductive-Capacitive Proximity Sensor for Seat Occupancy Detection", IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 5, pp. 1463 - 1470, 2010

[12] A.S. Zeeman, M.J. Booysen, G. Ruggeri, B. Lagana, "Capacitive seat sensors for multiple occupancy detection using a low-cost setup", IEEE International Conference on Industrial Technology (ICIT), pp. 1228 - 1233, 2013

[13] R. Wimmer, M. Kranz, S. Boring, A. Schmidt, "CapTable and CapShelf - Unobtrusive Activity Recognition Using Networked Capacitive Sensors", Fourth International Conference on Networked Sensing Systems, pp. 85-88, 2007

[14] T. Grosse-Puppendahl, Y. Berghoefer, A. Braun, R. Wimmer, A. Kuijper, "OpenCapSense - A Rapid Prototyping Toolkit for Pervasive Interaction Using Capacitive Sensing", IEEE International Conference on Pervasive Computing and Communications, pp. 152 - 159, 2013

[15] S. Kivikunnas, E. Strommer, M. Korkalainen, T. Heikkila, M. Haverinen, "Sensing Sofa and Its Ubiquitous Use", International Conference on Information and Communication Technology Convergence, pp. 559 - 562, 2010

[16] PCap02A data sheet, "Single-chip Solution for Capacitance Measurement - Volume 1: General Data and Front-end Description", Version 1.6, acam mess-electronic gmbh, 2014

[17] PCap02Ax DSP data sheet, "Single-chip Solution for Capacitance Measurement - Volume 2: Digital Signal Processor ", Version 0.2, acam mess-electronic gmbh, 2013

[18] PCap02-EVA-Kit, "Evaluation System for PCap02 ", 2013

[19] UM1472 User manual, "Discovery kit for STM32F407/417 lines", STMicroelectronics, 2014

[20] Keil uVision IDE, http://www.keil.com/uvision/

[21] STSW-STM32065, "STM32F4 DSP and standard peripherals library", http://www.st.com/web/en/catalog/tools/PF257901#

[22] T. Majerle, Libraries and tutorials for STM32F4 series MCUs, http://stm32f4-discovery.com

[23] Artes Calculi Ltd, Robin Hood Solver, http://www.irb.hr/Gospodarstvo/Spin-off-drustva/Artes-Calculi

IBMS - Intelligent Building Management System Framework

S. Tepić*, P. Pejić*, J. Domšić*, H. Mihaldinec*, H. Džapo*
* Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

***Abstract*—The ubiquitous availability of Internet connection in residential and industrial environments enabled limitless possibilities of connecting local sensor networks using various cloud technologies to provide value-added services for intelligent management of building facilities. One of the challenges in implementing a distributed smart building management systems is to enable interoperability of many different computational platforms by providing a common protocol across the wide range of devices, with different capabilities and resources. In this paper, we present an infrastructure architecture and software framework for building large interconnected sensor networks, with primary application for building management. We establish the terminology, system elements specification, functional description of different hardware and software building blocks, define a thin common interoperability layer with established message semantic and event-driven framework, providing a modular architecture that is easy to adapt and extend. Sensor nodes are connected through the data aggregators, which communicate with the remote scalable cloud service, designed for large sensor data collecting, syncing, storage, and event processing. For testing purposes, we implemented a demo system based on open source technologies that are integrated with our IBMS framework.**

***Keywords*—smart sensors and actuators, sensor networks, intelligent building management, Internet-of-Things**

I. INTRODUCTION

Internet of Things (IoT) is a paradigm that enables interconnection of sensors, actuators, and other types of small devices through the Internet infrastructure in order to enhance the possibilities of user interaction with physical environment [1], [2]. "Things" or devices can range from a simple sensor nodes to very complex embedded systems, depending on their purpose and role [3]. The IoT paradigm is believed to be a next big step in information technology development, after Internet and mobile communication network, which will provide more comprehensive interoperability and intelligence in management of Things.

The concept of Internet of Things was first proposed by MIT Auto-ID Labs, in a context of research of object localization and state recognition using wireless sensor networks [4], [5]. The report released six years later by the International Telecommunication Union (ITU) formally proposed the concept of the Internet of Things. This report shaped an idea in which all objects in the world can exchange information via the networks actively. It can be said that since then the era of the ubiquitous IoT communication has started [4], [6]. This domain has been identified by numerous companies as a source of new business opportunities, since it is expected that in a near future billions of small devices will be connected to the Internet [7].The possible areas of application include various fields, such as infrastructure management, security applications, environment protection, smart agriculture, industry, traffic, business services, urban management etc. [4], [8], [9].

The IoT paradigm demands a huge global distributed information system consisting of numerous devices that need to be identified, tracked, sensed, and processed. Therefore, there is a need to use standardized concepts and interoperable Machine-to-Machine (M2M) communication protocols, and to develop new ones to fit all IoT systems. With the support of new widely present technologies, such as broadband mobile communication and cloud computing, the IoT systems are able to process the objects' state and provide intelligent user interface for management to ease the necessary decisions based on the real-time notifications. Integration with machine learning and artificial intelligence approaches enables the development of cooperative IoT systems that do not require human intervention, thus making them completely autonomous. In implementation of such systems, it is crucial to provide an efficient way of collecting raw data for further processing and analysis. The IoT paradigm introduces heterogeneity, which inevitably induces the system complexity. Therefore, there is a great need for use of middleware that can overcome the complexity by providing the transparency for the end user. The main goal of this paper is to provide a middleware concept and layer implementation, which is simple, extensible, and based on readily available and widely used technologies and solutions.

The paper is structured as follows. The section II discusses the state-of-the-art solutions and the related work. The section III describes proposed system architecture. The section IV explains the inner system composition and implementation details more thoroughly. Paper is concluded with Section V.

II. RELATED WORK

One of important goals of the IoT paradigm is to automate the communication among devices, without a need for human intervention. Hence, the M2M concept takes a substantial segment of research and development in this field. There are several comprehensive IoT platforms in a process of active development, such as GSN [10] and OpenIoT [11]. These platforms strive to provide an overall and generic solution that would enable the interconnection of sensor and device networks with complex software data managements systems, by providing necessary functionality layers. These systems are designed to have a mediator role between sensors and sensor data consumers.

The main focus of our research is concerned with the system integration on a hardware side, and its implementation in small, low-power embedded systems. In general, the use of standard open technologies is preferred to proprietary closed solutions, since the latter are more difficult to maintain, manage and further develop, also with

a questionable long-term support. The common data exchange layer is typically built upon the Internet Protocol (IP), since the Internet infrastructure is omnipresent and readily available in residential, business, industrial and other types of environments. Traditionally, wireless sensor networks employ ZigBee or some other IEEE 802.15.4 based protocol for communication, since the main requirement is to provide a low cost and energy-efficient design, with ultra-low power capabilities of the underlying hardware solution [1]. In order to connect small devices with constrained hardware resources to the Internet, the 6LoWPAN protocol was developed. The main purpose of this protocol is to build TCP/IP stack compatibility for small networked devices, in order to encapsulate the IPv6 packets within ZigBee frames. TinyOS [12] is an instance of a real-time operating system for wireless sensor networks that has a lightweight implementation of 6LoWPAN called *BLIP*. Other alternative solution is *Sensor Networks for an All-IP World*(SNAIL), which integrates four technologies into its IP adaptation architecture, such as mobility, web enablement, time synchronization, and security [13].

Recent efforts in the field of power efficient and space constrained electronic system design lead to the development of low-cost Wi-Fi components and modules that are specifically designed for embedding in wireless sensor network nodes. They can achieve competitive technical characteristics and price when compared to IEEE 802.15.4 and similar technologies. Such solutions enable easier system field deployment and scaling with constantly changing demands. Modern sensor network hardware designs are continuously shifting towards the solutions implementing the native Wi-Fi support and such approach is becoming a dominant choice for the IoT domain. With a careful system design, in certain usage scenarios a battery life of several years can be expected, while providing an easy integration with the existing network infrastructure. Re-use of existing infrastructure is key point for cost savings, faster development, and scalable deployment. Moreover, Wi-Fi enabled devices offer native IP network compatibility out of the box.

The goal of our research was to develop a lightweight micro framework middleware for rapid IoT solution development that encompasses a minimum set of necessary components, both for embedded systems on a sensor network side, and for the cloud infrastructure. The solution is based on combination of various widely used web, networking, and embedded open source libraries and frameworks that are loosely coupled via thin common data exchange protocol layer. The proposed solution has been developed with modular approach in mind that would enable easy replacement of existing modules and development of future plug-ins. The system supports not only Wi-Fi wireless sensor nodes because it includes a bridge for transparent translation of any proprietary radio protocol into the common IP-based data exchange protocol. One of the important design goals was scalability in scenario when the size of the sensors network grows. Although the framework is suitable for a wide range of applications, the main practical design goal was to provide an easy implementation for remote building management; therefore we decided to name it IBMS - *Intelligent Building Management System Framework*.

III. IBMS System Architecture

The proposed IBMS system is logically divided into the following main system parts: *Sensors, Aggregators*, *Cores*, and *Generic Data Acquisition System* (GDAS). The Fig. 1 depicts the proposed system architecture and how the components are organized in the overall system.

Sensors are low-power embedded devices that collect sensory data from the physical environment, and they send readings to the other components of the system. Sensors are typically battery-powered devices that are connected to the local data *Aggregators* through wireless communication protocol in order to avoid the necessity for building the wired infrastructure. *Sensors* are equipped with low power microcontroller which is capable of implementing simple

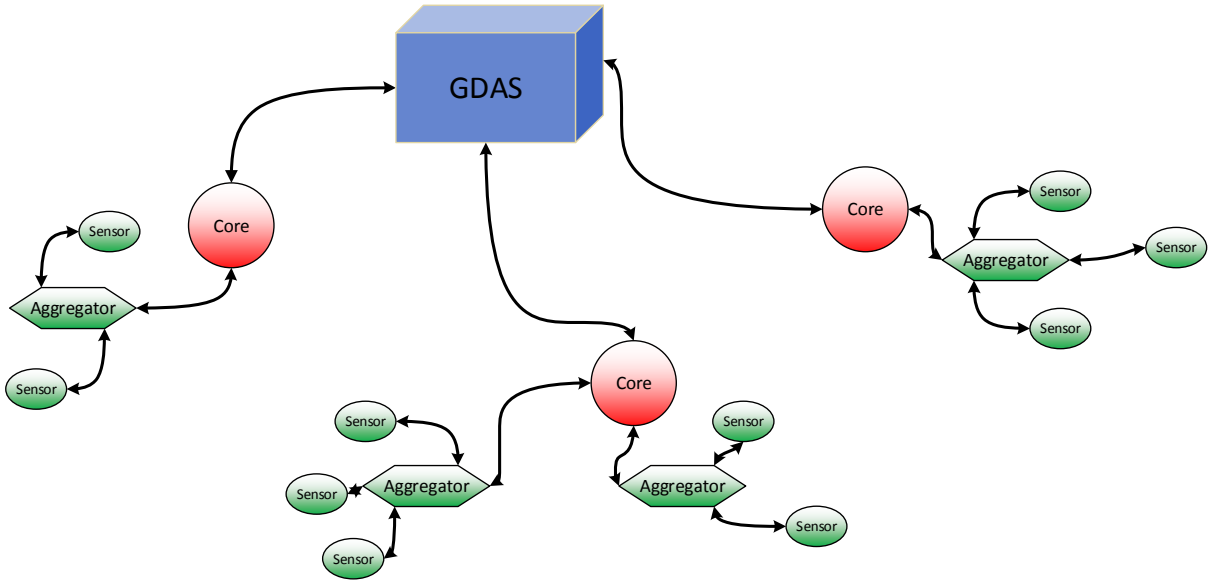


Fig. 1 The IBMS system architecture

protocol for data exchange with local *Aggregator*. The main design goal of *Sensor* is to provide ultra-low power operation, and possibly several years of autonomy from a single battery power source. Therefore, in some cases Wi-Fi might not be a viable solution, while 802.15.4 protocol may be also overkill in practice for very simple sensor nodes. System is designed to accept a *Sensor* equipped by any kind of wireless communication interface.

The role of the local *Aggregator* is to interact directly with wireless sensor nodes, collect raw data, monitor sensor status, update individual sensor configuration, etc. However, one of the most important roles of the *Aggregator* is serving as a *protocol bridge* between any type of wireless sensor radio communication protocol, and the rest of the system. Protocol translation in the *Aggregator* is transparent and all *Sensors* are seen in the exactly same fashion from the other higher level system components. Local *Aggregator* is typically based on microcontroller and equipped with the necessary communication interfaces (Wi-Fi, 802.15.4, proprietary RF, Ethernet etc.). Radio modules can be added in a modular fashion through standard microcontroller peripheral communication interfaces (USART, I2C, SPI, etc.). The *Aggregator* is typically connected to a continuous power supply and does not have to be designed for ultra low power operation. In the context of building management, an *Aggregator* will cover the wireless range of sensors that are assigned to it. In the large building, there would be typically a need for more than one *Aggregator* to cover the whole range of scattered sensors.

The role of the local *Core* is to monitor and control all local data *Aggregators* and to connect the local network consisting of several *Aggregators* to the remote cloud service for data collection and monitoring (*Generic Data Acquisition System*, GDAS). In a typical usage scenario, one *Core* would be assigned for each building, communicating with the *Aggregators* that are in range of wireless *Sensors*. In a special case of spatially limited sensor installations and small buildings, the *Core* can be designed to include also the described functionality of the *Aggregator* in a one single hardware module. The *Aggregator* can communicate with a local *Core* over any media supporting IP protocols (Ethernet, Wi-Fi etc.). *Core* is typically based on higher computational power embedded system that is capable of running general-purpose operating system (such as a Linux). Examples of such embedded systems are *Beagle Boneblack*, *RasperryPi* etc. The *Core* also needs to provide local data storage capabilities, for caching sensor data in case of limited or unavailable connectivity. Therefore, Linux-based embedded systems provide an excellent choice by enabling the use lightweight database managements system (DBMS).

The central part of the whole system is *Generic Data Acquisition System* (GDAS). The main purpose of GDAS is to provide means for scalable sensor data collection, processing, and storage. GDAS communicates with *Cores* as mediators for managing local sensor networks. Each sensor can be reached from GDAS via *Core*, who takes care of routing the messages between GDAS and *Sensor* through the right *Aggregator* to which sensor is connected. GDAS is implemented as a web service, and it is built upon widely used open source frameworks to provide minimum server side functionality needed for managing the whole system. An example of open source solution providing similar functionality is Zabbix[14].

Proposed system architecture is not limited only to applications for intelligent building management systems, and it can be used other similar usages scenarios (for example, environmental monitoring in forests, fields, smart agriculture applications etc.). The *Core* serves as a local central place where we can access information about each connected sensor, without a need to know how to directly communicate with it, and which end protocol to use. *Core* also takes a role of immediate local data storage and acts as a bidirectional message forwarding node. In order to provide a simple and consistent mean of communication and message exchange among all components in a system, a thin unified data exchange layer was proposed. We use the IP layer for low-level packet exchange between components to provide maximum independence of physical layer implementation. The communication between *Aggregators* and *Sensors* is the only place when other protocols for data exchange are used, but it is transparent to the rest of the system. Components have assigned IP addresses through which they can be reached, except for *Sensor* nodes, as they are not necessarily connected with *Aggregators* through the IP layer. *Sensor* nodes are assigned meaningful virtual addresses through which they can be reached, and *Aggregators* take care of routing the messages to access correct connected *Sensor*. Data is transferred by using TCP transport layer, and HTTP application layer is used to achieve protocol-independent execution with defined start, stop, transfer, and sync moments. HTTP layer is also a convenient choice because it enables direct communication with a web framework, as a part of GDAS. Messages sent between system components are in the *Java Script Object Notation* (JSON) format. Modularity, simplicity of parsing procedure, and widespread use in web technologies makes it

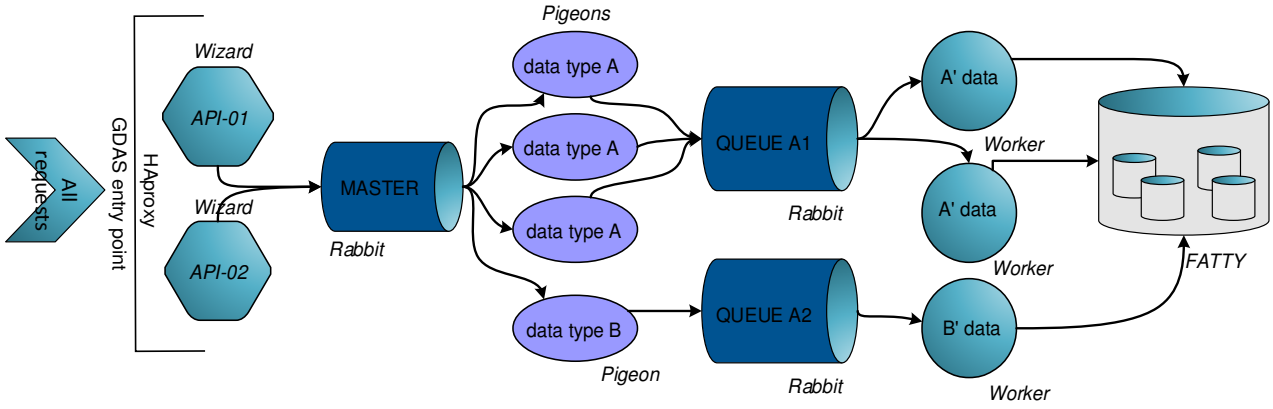


Fig. 2 Structure of the GDAS system

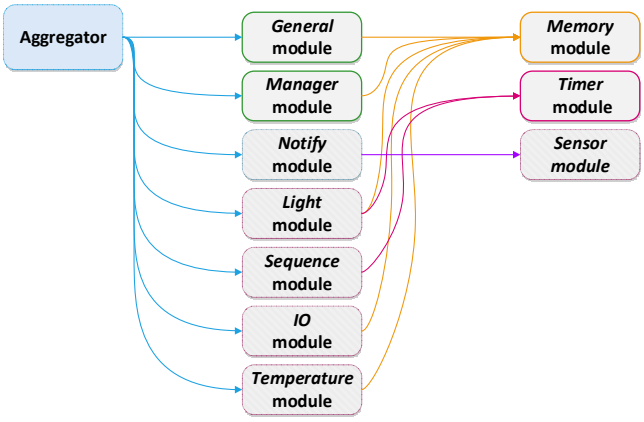


Fig. 3 Aggregator software architecture

an ideal choice, even from the embedded system standpoint. Lightweight, human readable format makes also system easier to debug and maintain.

A minimum required set of components consists of one *Core* and one *Aggregator*. For sake of simplicity for the rest of the system, *Sensors* assigned to some *Aggregator* are mapped into the *Aggregator Ports*. They are not seen as separate entities but as *Ports* of the *Aggregator,* with defined type and attributes. Local *Aggregator Ports* are mapped into sensor and actuator nodes.

The dataflow can be between *Core* and *Aggregator* is bidirectional, in a request-response manner. When the Aggregator sends data to the *Core*, there are two basic cases, synchronous and asynchronous. The synchronous case is when the *Core* requests a data from the *Aggregator* by issuing a HTTP GET request, and gets a response from the *Aggregator*. It is used when the state of the whole array of *Ports* is requested in one chunk. In the asynchronous case, the *Aggregator* sends HTTP POST packets to the *Core*, and it exchanges only a subset of *Ports*, including a case of information about a single *Port*. The *Core* sends data to the *Aggregator* in HTTP POST packets, only in case when some parameter on the *Aggregator* has to be changed. JSON formatted data is exchanged in the body of the message.

IV. Practical Implementation of Test Environment

Following the specification of the IBMS architecture from the previous section, an implementation of the proposed system and test environment for simulation of various use cases were designed. Three types of networks were included in a test environment setup to demonstrate the interoperability in merging IP protocol with an example of proprietary solution: IEEE 802.3(Ethernet), IEEE 802.11 (Wireless Ethernet) and NRF24L01+ Nordic Semiconductor proprietary physical RF layer for ultra-low power transceivers on wireless sensor nodes.

The *Core* module was implemented as an embedded system, which needs to satisfy the following requirements: embedded web server (with capability of serving dynamic content), Ethernet connectivity, database support for local sensor data caching, possibility of hardware extension via external plug-in modules, portability between different platforms, and simplicity of required software install. Requirements for small size and power consumption, above already mentioned requirements, lead to the conclusion that it is best to employ some Linux-based embedded system. Therefore, a *BeagleBoneBlack* platform was selected due to compliance with all previously stated requirements and low price.

The software for *Core* module was developed in JavaScript programming language, due to simplicity, and native support in most of modern Internet browsers. Additionally, JSON data format which is a base for thin common data layer is natively supported in JavaScript environment. As a development environment, the NodeJS was used due to the support over wide variety of platforms, and because it is a proven technology that it can deal with large number of concurrent requests. MongoDB was selected as a database backend because it supports JSON format natively, and it stores data into NoSQL database. Storing data into collections instead of tables like in classical SQL databases perfectly fits our proposed model for data collection and storage.

The functionality of the *Core* module itself can be divided into three main logical parts, that we call *Mainframe*, *Crawler* and *Notifier* module. The *Mainframe* module is in charge of serving the user request. User can be either a person that connects directly to the *Core* module via embedded web server running on the *Core*, or a GDAS subsystem running in the cloud. The *Mainframe* module is in charge for running web services and tracking the user sessions. It also has an ability to communicate with local NoSQL database in order to be able to present data to the user. The *Mainframe* module is implemented in *Express.js* environment, which has all necessary ingredients for agile development of web application. The *Crawler* module is in charge for collecting data from the *Aggregators*. It searches for changes in data for all *Aggregators* and copies data to the local database storage, aggregating them at the single point. *Crawler* is querying the *Aggregators* periodically, but in some cases (e.g. alarms) there is a need for instant event notification to user or GDAS subsystem. *Notifier* module is in charge of listening to asynchronous messages from the *Aggregators* and forwarding them to the users or GDAS subsystem. This is useful for processing high priority information, which needs to be forwarded further instantly.

GDAS subsystem is running at remote server or cloud infrastructure, and it uses the similar subset of technologies as a *Core*. Contrary to the requirements for a *Core*, GDAS is not intended to be implemented on embedded system. GDAS subsystem uses Python-based Flask web micro framework for handling incoming requests, RabbitMQ for storing data during processing, and MongoDB as a main

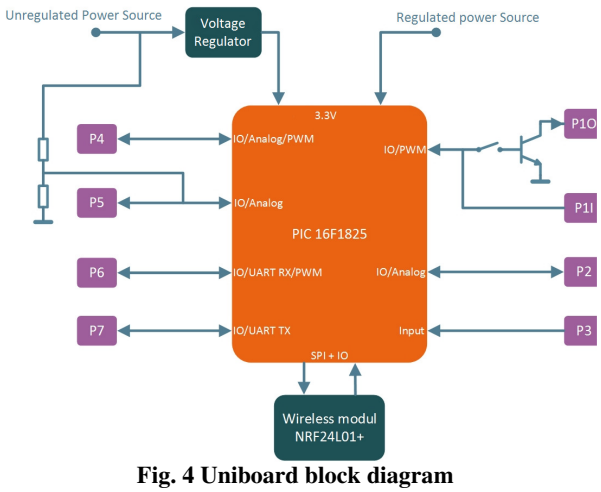


Fig. 4 Uniboard block diagram

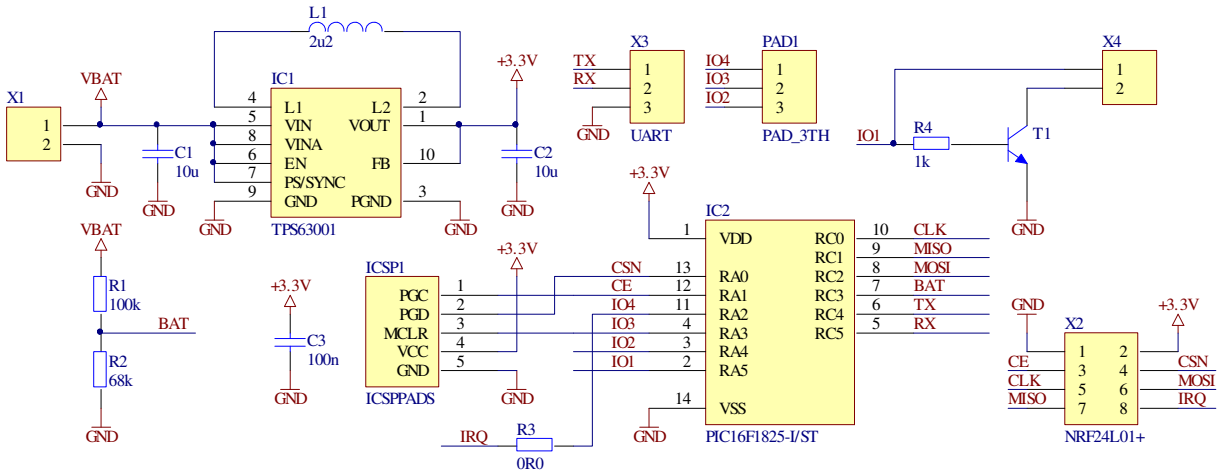


Fig. 5 Uniboard schematic

database storage facility. Flask micro framework was chosen because of excellent balance between low resource demand and high modularity and extensibility. RabbitMQ is a GDAS backbone, capable of storing messages up to the size of system random-access memory (RAM) and passing them from the publishers to consumers at large speed. GDAS has a simple structure that contains several separated modules, as depicted in the Fig. 2. *Wizard* is an entry point of the GDAS system, which is in charge of handling the incoming requests and distributing those requests to the *Workers* equally (load balancing). *Wizard* additionally validates HTTP requests format, and forward valid requests to the *Rabbit* module. *Rabbit* module is based on readily available RabbitMQ message queue implementation, which ensures that this critical part of the system sorts out incoming requests reliably by their keywords. Requests are transferred from the queue to the *Worker* module by means of *Pidgeon* microservice. The role of the *Worker* is to process and store incoming data. If the data passes necessary check, the *Worker* stores it in serialized format into scalable database called *Fatty*, based on MongoDB. All these GDAS segments are micro services, and they can be executed on the same or distributed machines, in a limitless number as independent instances. This ensures system scalability and modularity from the standpoint of resource management.

On the sensor side of the system, custom hardware solutions for both *Aggregators* and *Sensors* were developed. The *Aggregator* hardware must satisfy the following requirements: Ethernet connection with the embedded TCP/IP software stack, I/O connectivity with different buses (GPIO, USART, SPI, I2C etc.) for connection with sensors, actuators, and wireless modules, and microcontroller capable of serving all necessary workload with low latency in a real-time. For purposes of demo system implementation, the *Aggregator* hardware was realized by using Microchip PIC18F26K22 microcontroller and ENC28J60 Ethernet module. Software for the *Aggregator* was developed in a modular way, which enables easy inclusion of new functionality through *Modules*. Some examples of implemented modules are the temperature module (for managing temperature sensors), I/O module (for direct control of digital I/O pins), light module (controlling the lights, with fading capabilities), NRF24L01+ sensor module (communication with wireless sensor nodes with on-board NRF24L01+transceiver), timer module (timing services for other modules), sequence module (sequential actions control), memory module (permanent configurations management), manager module (synchronous communication with *Core*), notify module (asynchronous communication with *Core*), general module (basic information about the *Aggregator*, such as IP address, virtual ID etc.) etc. Fig. 3 shows all the modules in the *Aggregator* software architecture. In order to enable *Aggregator* to exchange messages in JSON format, we used a parser *Jsmn*, which is suitable for embedded systems with limited resources.

The wireless *Sensor* nodes were designed as ultra-low power hardware units, based on NRF24L01+ RF transceivers modules (operating at 2.4 GHz ISM band), and PIC16F1825 microcontrollers. Some of the important requirements for *Sensor* nodes were as follows: ultra low power operation, battery supply, possibility to apply external power supply from renewable sources (such as solar panel), capability of connecting different types of sensors, programmable output capable of driving higher levels of power (e.g. switch relays). To provide maximum flexibility of *Sensor* mode, we developed custom printed circuit board (*Uniboard*), which block diagram is shown in Fig 4. and implementation schematic in Fig. 5. The *Uniboard* is powered from Li-Ion battery (with 1000 mAh capacity and nominal voltage 3.7 V), it has embedded temperature and relative humidity sensor (DHT11), and LEDs for visual indication. The *Sensors* can be parameterized through the serial interface, what includes the local node address, frequency channel, *Aggregator* address, duty cycle (for controlling average power consumption) etc. *Sensor* nodes

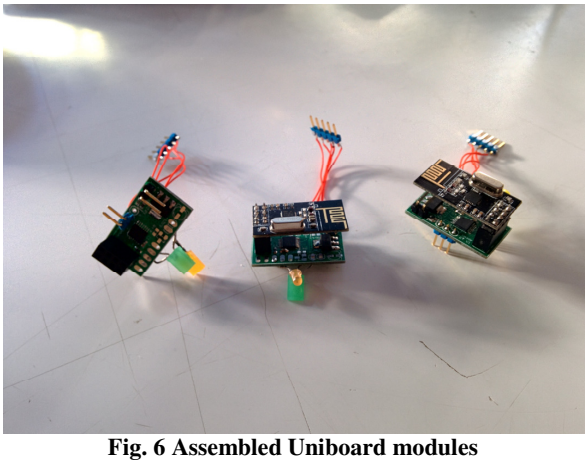


Fig. 6 Assembled Uniboard modules

based on *Uniboard* are shown in Fig. 6. Current consumption depends on application, and it can be reduced under 100 μA for typical periodical sensor polling applications.

V. Conclusion

In this paper we presented a novel framework for building and managing large interconnected sensor networks, with a primary focus on intelligent building management and monitoring. The system was designed as s lightweight middleware of loosely coupled components, distributed across very different set of computational platforms, ranging from a simple smart sensor low power devices, microcontroller-based local data aggregators, embedded Linux-based cores, and dedicated cloud-based service for remote management of the sensor network, and data collection and storage.

The proposed framework is based on widely used open source technologies that simplified and accelerated the system development and facilitated the interoperability among other open platforms. The proposed system architecture is modular, and it enables easy integration of new functionality or exchanging existing subsystems with the new ones. The efficient interoperability across the system components was achieved by use of the lightweight common message exchange protocol, based on HTTP application layer and JSON formatted packets, the use of the IP protocol to achieve the maximum independence of the physical layer, and bridging the proprietary wireless sensor network protocols at the point of the sensor connection to the rest of the system.

Experimental test environment was set up, and all system components' functionality was tested and verified in practical demo installation. The future research will encompass the further extension of the common data exchange protocol, functionality of the system components, adding support for wider set of embedded platforms, and testing the system behaviour under realistic loads expected for larger scale sensor networks.

Acknowledgment

The research leading to these results has received funding from the project "Information and communication technology for generic and energy-efficient communication solutions with application in e-/m-health (ICTGEN)" co-financed by the European Union from the European Regional Development Fund.

References

[1] C. Wang, M. Daneshmand, M. Dohler, X. Mao, R. Q. Hu, and H. Wang, "Guest Editorial - Special Issue on Internet of Things (IoT): Architecture, Protocols and Services," *IEEE Sens. J.*, vol. 13, no. 10, pp. 3505–3510, Oct. 2013.

[2] S. Tozlu, M. Senel, W. Mao, and A. Keshavarzian, "Wi-Fi enabled sensors for internet of things: A practical approach," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 134–143, Jun. 2012.

[3] N. Khalil, M. R. Abid, D. Benhaddou, and M. Gerndt, "Wireless sensors networks for Internet of Things," in *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, 2014, pp. 1–6.

[4] Q. Zhu, R. Wang, Q. Chen, Y. Liu, and W. Qin, "IOT Gateway: BridgingWireless Sensor Networks into Internet of Things," in *2010 IEEE/IFIP 8th International Conference on Embedded and Ubiquitous Computing (EUC)*, 2010, pp. 347–352.

[5] I. Bose and R. Pal, "Auto-ID: Managing Anything, Anywhere, Anytime in the Supply Chain," *Commun ACM*, vol. 48, no. 8, pp. 100–106, Aug. 2005.

[6] "ITU Internet Reports 2005: The Internet of Things." [Online]. Available: http://www.itu.int/osg/spu/publications/internetofthings/. [Accessed: 05-Mar-2015].

[7] D. Evans, "The Internet of Things: How the Next Evolution of the Internet Is Changing Everything," Apr. 2011.

[8] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey," *Comput Netw*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[9] Y. Liu and G. Zhou, "Key Technologies and Applications of Internet of Things," in *2012 Fifth International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 2012, pp. 197–200.

[10] "LSIR/gsn," *GitHub*. [Online]. Available: https://github.com/LSIR/gsn. [Accessed: 05-Mar-2015].

[11] "OpenIoT project | Open Source Solution for the Internet of Things into the Cloud." [Online]. Available: http://www.openiot.eu/. [Accessed: 05-Mar-2015].

[12] "TinyOS Home Page." [Online]. Available: http://www.tinyos.net/. [Accessed: 05-Mar-2015].

[13] S. Hong, D. Kim, M. Ha, S. Bae, S. J. Park, W. Jung, and J.-E. Kim, "SNAIL: an IP-based wireless sensor network approach to the internet of things," *IEEE Wirel. Commun.*, vol. 17, no. 6, pp. 34–42, Dec. 2010.

[14] "Homepage of Zabbix :: An Enterprise-Class Open Source Distributed Monitoring Solution." [Online]. Available: www.zabbix.com. [Accessed: 20-Mar-2015].

A Hybrid Local Method for Stereo Correspondence based on Three-Dimensional Recursive Search

Miroslav Rožić
Mikroprojekt d.o.o.
Zagreb, Croatia
Email: miroslav.rozic@mikroprojekt.hr

Matea Đonlić and Tomislav Pribanić
Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia
Email: matea.donlic@fer.hr; tomislav.pribanic@fer.hr

***Abstract*—Obtaining dense disparity maps of stereo image pairs is a common problem in computer vision. We present a new hybrid method for dense stereo correspondence based on the well-known winner-take-all optimization and the three-dimensional recursive search (3DRS) algorithm. Originally devised for motion estimation and frame rate conversion in high definition televisions, the 3DRS algorithm employs predictive block matching to generate a coarse correspondence map between two images in a sequence. The algorithm is widely used in ASIC and FPGA implementations, achieving real-time performance in processing of 60Hz high definition video signals. In our proposed hybrid method, the 3DRS algorithm is applied to a stereo pair in order to generate a coarse disparity map, which is further optimized using a local, winner-take-all driven coarse-to-fine approach. The selected approach greatly reduces the local disparity space to be searched, yielding significantly better run-time performance while retaining or even improving the accuracy of the results. The method is evaluated using the Middlebury stereo datasets and compared with both the reference winner-take-all algorithm and coarse-to-fine hierarchical optimizations. The obtained results favor the proposed method in terms of accuracy and performance. The proposed method is suitable for implementation in custom parallel computing hardware, ASIC or FPGA.**

I. Introduction

Dense stereo matching, or estimation of disparity in a stereo image pair, has been a heavily investigated topic and an important problem within the field of computer vision [1]. The methods which are employed can generally be classified as either local or global, depending on the strategy used for the estimation. While in the global methods the disparity computation involves the minimization of a global energy function for the entire disparity map, the local methods choose the disparity associated with the optimum cost value at each pixel - a "winner-take-all" (WTA) optimization. While it has been shown that the global approaches deliver the highest quality results [2], they are also typically very slow and impractical [3]. The local methods, while faster and much more compatible to a practical implementation, are still behind the global methods quality-wise, nevertheless their computational complexity is still considerable. Research efforts have been therefore directed to both improving the quality [4] [5] of WTA algorithms, and reducing their computational requirements [6] [7].

This work has been supported in part by Croatian Science Foundation's funding of the project IP-11-201-3717.

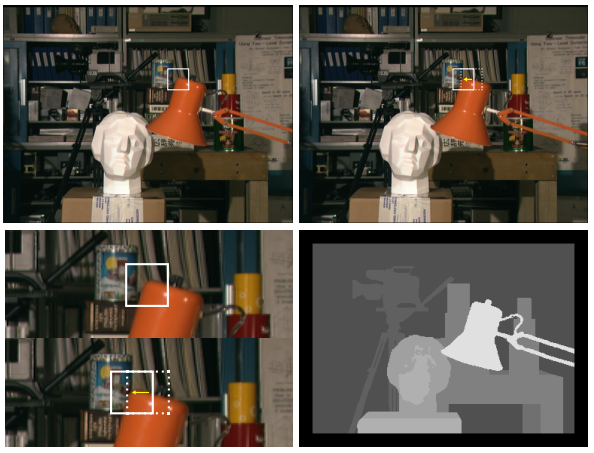


Fig. 1. Disparity estimation in stereo images using block matching.

II. Related Work and Motivation

A lot of research efforts have had a goal of improving the WTA optimization. Most of the approaches in improving the WTA optimization have focused on improving either the matching cost [8] or the cost aggregation [6], whether they are using traditional square windows, shiftable windows [9], variable or adaptive windows [10], or, more recently, non local aggregation through methods such as bilateral filtering [11]. Additionally, hierarchical and hybrid methods have been proposed with the goal of improving the performance and quality of local stereo methods.

The problem of computing correspondent features in two images has also been extensively studied in the field of motion estimation, where a motion map (also called the optical flow) of objects is generated for a sequence of images [12]. The goal is to establish a correspondence map of areas within a source image similar to areas within a reference image in a sequence. Two categories of motion estimators are present: the residue-minimizing, used in

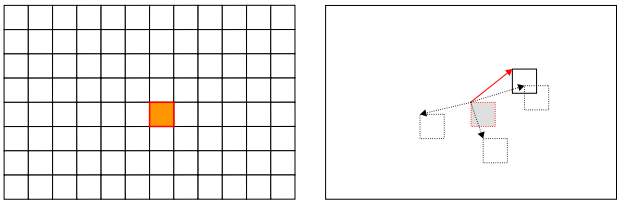


Fig. 2. An illustration of block matching. Left: Source image divided in blocks. Right: reference image with multiple matching candidates.

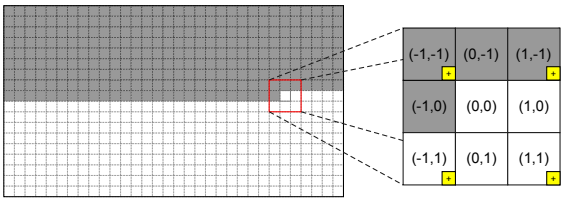


Fig. 3. The spatio-temporal neighborhood used in the 3DRS estimation. The vector for the block at (0,0) is currently being estimated. Vectors for the blocks at (-1,-1), (0,-1), (1,-1) and (-1,0) have already been estimated for this image. The remainder of the blocks used are from a previous estimation, making them temporal predictors. The predictors coming from the blocks marked with a '+' are additionally modified with update vectors.

video compression algorithms, and true motion estimators, used in video up-conversion and deinterlacing. One of the most successful true motion estimation methods is the Three-Dimensional Recursive Search (3DRS) [13], widely used due to its simplicity, low cost implementation, and the ability to provide good true motion estimates. The method has been successfully applied for real-time processing of 60 Hz high definition video and frame-rate conversion with frame rates up to 400 Hz, with specialized ASIC implementations.

It has been shown [14] that true motion estimation is necessary to produce good disparity estimation results, as well as that the 3DRS algorithm can be successfully applied in real-time stereo matching. The 3DRS, however, produces coarse results, while the goal of passive stereo matching is to produce dense disparity maps. A number of research efforts has successfully employed a coarse-to-fine hierarchical approach [15] [16]. In this paper, we will present a hybrid approach to dense disparity estimation combining the coarse disparities obtained from the 3DRS algorithm, with the subsequent refinement using the WTA optimization. By using 3DRS results to narrow the disparity space and applying the 3DRS principles on the WTA estimation to gain accuracy, the proposed method manages to vastly reduce the computational complexity and improve run-time, while also improving the quality of the final disparity map.

III. Algorithm Description

A. Three-dimensional Recursive Search (3DRS)

The 3DRS algorithm is a *block-matching algorithm*; the input source image is divided into blocks, for each of which the location of a best-matching block is searched in the reference image. A two-dimensional vector defining the distance between the original block position and the position of the matching block in the reference image is assigned, as shown in Figure 2. The map of vectors for all of the image blocks is the optical flow of the image and describes either the motion or disparity. For disparity estimations of epipolarly rectified stereo pairs, the vertical component can be ignored.

The blocks are matched through a *matching cost function*, such as the *sum of absolute differences* (SAD) or the *sum of squared differences* (SSD) or *normalized cross-correlation*(NCC). Finding the proper matching block involves finding the extreme value of the matching cost function. To find the extreme, an optimization method is usually applied, where the 3DRS algorithm reduces the number of required candidate blocks (or candidate vectors) by making the following two assumptions:

1) Objects are larger than blocks.
2) Objects have inertia.

The objects referred to are the features in the actual image, the movement (or disparity) of which is being estimated. Assumption 1) therefore implies that the vectors of the neighboring blocks provide a good predictor, or assumption, about the block being estimated. Assumption 2), on the other hand, implies that a temporal predecessor - a previously estimated motion vector, is also a good candidate for the new vector. The application of these assumptions to the estimation process significantly reduces the number of candidates which have to be evaluated. To introduce variation, *update vectors* which are either generated randomly, from a distribution, or a predefined set, are added to a subset of the predictors. The variation introduced by the update vectors ensures that the algorithm will progress towards a solution regardless of the initial conditions. The typical initial conditions are that the vector field contains only zero vectors. An example of the update vector set is:

$$U_v = \left\{ \begin{bmatrix} \pm 2^k \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \pm 2^k \end{bmatrix} \right\} \; k = 0, 1, 2, 3, ..., k_{max} \tag{1}$$

Sequentially for every block in the image, 3DRS selects the output motion (or disparity) vector from a set of prediction vectors chosen from a spatio-temporal neighborhood, with the addition of updates in order to evaluate alternate solutions, as seen in Figure 3. For an individual block, the matching cost function is evaluated for each predictor, and the predictor with the best matching cost is selected and assigned as the new vector for the current block, which then also becomes a part of the candidate set for the next location as the estimation progresses to the adjacent block.

The block evaluation order is defined by the *scanning direction*. The most common order is from top-left to bottom-right, with each line beginning on the left. This order defines which of the predictors are spatial, and which have also a temporal component, as seen in Figure 3. It

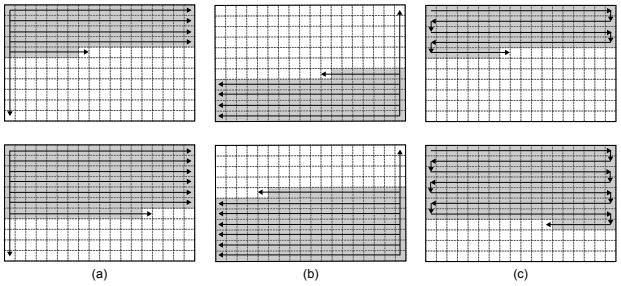


Fig. 4. Various scanning orders for 3DRS, shown in time progression from the upper to the lower image; (a): Top to bottom, left to right; (b): Bottom to top, Right to left; (c): Boustrophedonic/meandering. The darker areas illustrate the progression of current estimates throughout the image.

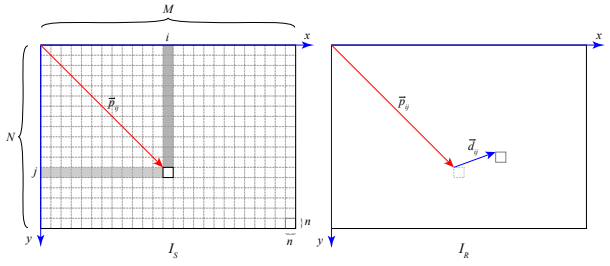


Fig. 5. 3DRS image space and coordinate system

also defines the direction in which the newly estimated, "good" values propagate within the image. Changing the scanning order from bottom-right to top-left on each new image, or alternating the scanning direction in a *boustrophedonic* or *meandering* way as shown in Figure 4, yields the propagation of good estimates in all four directions, helping the 3DRS to quickly converge on the final vector map. For the correct estimates to propagate even faster, multiple scans can be done for one image pair, alternating the direction and the meander of the scan on each pass. The *temporal* component of the 3DRS vectors refers to the results of the previous estimation, either for the current pair of images, or for a previous pair in a sequence of images. The initial state of the 3DRS vector field is that all components are set to zero, from where they converge to the proper value by aggregating update vectors during the estimation process.

The 3DRS algorithm works by analyzing the blocks in two images:the source image I_S and the reference image I_R, shown in Fig.5.

The source image I_S is divided into a rectangular grid of $M \times N$ blocks of $n \times n$ pixels as shown in the figure. For each block B_{ij}, i denotes the row index and j denotes the column index of the block. A motion vector $\vec{d}_{ij}$ is assigned to each block.

The pixel position vector $\vec{p}_{ij}$ for each block in I_S is defined as:

$$\vec{p}_{ij} = n \cdot \begin{bmatrix} i \\ j \end{bmatrix} \tag{2}$$

This vector points from the origin to the upper-left pixel of each block.

The 3DRS algorithm consists of the following steps:

1) For each block B_{ij} in I_S, form a set of predictors:

$$P = \begin{cases} \vec{d}_{ij} \\ \vec{d}_{i+k\,j} \\ \vec{d}_{i\,j+l} \\ \vec{d}_{i+k\,j+l} + \vec{u}_v \end{cases} \qquad k = -1, 1;\ l = -1, 1 \tag{3}$$

In the set above, $\vec{u}_v$ is the update vector, and a new one is selected either randomly of from a predefined set, for each of the predictors it is applied to.

2) Evaluate the matching cost for each of the predictors, and assign to $\vec{d}_{ij}$ the predictor with the minimum matching cost.

$$\vec{d}_{ij} = arg \min_{\vec{d}_P \in P} (MC(\vec{p}_{ij}, \vec{p}_{ij} + \vec{d}_P)) \tag{4}$$

where the matching cost $MC(\vec{s}, \vec{r})$ is the sum of absolute differences of pixels within the compared blocks

$$MC(\vec{s}, \vec{r}) = \sum_{y=0}^{n-1} \sum_{x=0}^{n-1} |I_S(\vec{s} + \begin{bmatrix} x \\ y \end{bmatrix}) - I_R(\vec{r} + \begin{bmatrix} x \\ y \end{bmatrix})| \tag{5}$$

3) Repeat the previous two steps for every block in the image based on the selected scanning direction.

B. Winner-Take-All optimization

In local approaches, correlation between intensity values inside a matching window $N(p)$ assigned to a reference pixel p is analyzed, based on the assumption that the pixels within the matching window have similar disparities. The generalized procedure of the WTA optimization [6] of a left disparity map in the case where *absolute difference* is used is as follows. For each of the disparity hypotheses, a per pixel cost $e(p, d)$ is calculated by using the image pair where the right image is shifted by the disparity d. An aggregated cost $E(p, d)$ is then computed via a summation of the per-pixel cost multiplied by a weighting function $w(p, q)$. The WTA optimization is finally performed for seeking the best disparity among all of the hypotheses:

$$e(p, d) = min(|I_l(x, y) - I_r(x - d, y)|, \sigma)$$

$$E(p, d) = \frac{\sum_{q \in N(p)} w(p, q) e(q, d)}{\sum_{q \in N(p)} w(p, q)} \tag{6}$$

$$d(p) = arg \min_{d \in [0, \ldots, D-1]} E(p, d)$$

where I_l and I_r are the left and the right images of the stereo pair. The per-pixel cost is additionaly truncated with a threshold σ to limit the influence of outliers on the dissimilarity measure.

Many research efforts have been directed at improving the WTA algorithm in terms of speed and accuracy. Diverse matching cost functions have been explored which bring accuracy improvements at varying computational

TABLE I
NUMBER OF EVALUATED HYPOTHESES PER CASE

r	Input Disparities D	Best Case	Worst Case
5	256	11 (-95,7%)	99 (-61,3%)

loads. However, the cost aggregation step remains the most important and the most time-consuming part of the calculation in the local approaches. Therefore, most of the efforts focus on defining the weighting function $w(p, q)$ which defines how the costs are aggregated within the window. For the classic square window, a noteworthy enhancement is the use of integral images [10], which greatly accelerates the algorithm's execution in comparison with the direct aggregation. A great part of the computational complexity, however, remains in the requirement to test all of the disparity hypotheses $d \in [0, ..., D-1]$ for each pixel p. The goal of the proposed hybrid method is to reduce the computational complexity by narrowing the disparity range to be tested per each pixel p.

C. The hybrid 3DRS-WTA method

In the proposed method, shown in Algorithm 1, a coarse analysis is performed first using the 3DRS algorithm, with the image divided into blocks. This yields a coarse map of estimated disparities d_{ij} assigned to each of the blocks. For each block, a disparity range to be tested using the WTA method is obtained by re-applying the 3DRS assumption that the correspondence vectors of a block's spatio-temporal neighborhood can be good estimates of the final result. Based on the user-defined parameter r, an initial disparity range is assigned to the estimated block and the blocks in its 8-neighborhood. Finally, a union of individual disparity ranges of blocks within the 8-neighborhood is performed to obtain a list of disparity hypotheses to be tested within the WTA optimization step.

The parameter r limits the range of disparity hypotheses evaluated per each block within the 8-neighborhood to $[d - r, ..., d + r]$, where d is the result of the 3DRS disparity estimation for each given block. In the optimum case where all of the 3DRS results in the 8-neighborhood are equal and the ranges perfectly match, the total number of hypotheses to test is $2 * r + 1$, the size of a single initial range. In the extremely unlikely worst case where the disparity ranges of within the 8-neighborhood do not intersect, the total number of disparity hypotheses to be evaluated is $9 * (2 * r + 1)$.

A comparison of the best and worst case for a typical range of input disparities is shown in Table I. While most local methods expect the user to specify a certain disparity search range to optimize the performance, the proposed hybrid method will address any range of disparities which occurs in the input stereo pair with constant performance.

Algorithm 1 Coarse-3DRS-to-fine-WTA optimization

Require: r: integer > 0; {disparity range per block}
Require: b: integer > 0; {3DRS block size}
Require: I_l: image; {Left image}
Require: I_r: image; {Right image}
Require: I_{dc}: array; {3DRS coarse output}
Require: I_d: image; {Disparity image}

```
3DRS(I_l, I_r, I_dc, b); {Compute rough 3DRS map}
for all block disparities dc_ij ∈ I_dc do
  rct := Rectangle(i*b, j*b, (i+1)*b, (j+1)*b);
  L = NewList;
  for v = −1 to 1 do
    for h = −1 to 1 do
      if i + h and j + v is a valid block index then
        AddToList(L, dc_{i+h,j+v})
      end if
    end for
  end for
  SortAscending(L)
  dmin := dmax := 0;
  for all dc in L do
    dmin_n := dc − r;
    dmax_n := dc + r;
    if dmin_n > dmax then
      WTA(I_l[rct], I_r, dmin, dmax, I_d);
      dmin := dmin_n;
    end if
    dmax := dmax_n;
  end for
  WTA(I_l[rct], I_r, dmin, dmax, I_d);
end for
```

IV. EVALUATION

The proposed method was implemented in software within a simulation environment built specifically for testing stereo algorithms. The motivation of designing our own simulation environment was to reduce the dependency of the algorithm on external libraries as much as possible, in order to facilitate subsequent hardware implementations of the tested methods. Individual 3DRS and WTA implementations were developed and tested first. All methods operate on 8-bit grayscale images.

The 3DRS implementation does not include advanced optimizations such as hierarchical processing or penalties [17], but supports multiple iterations to improve algorithm convergence on static images. SAD and ZSAD [8] cost functions can be utilized for block matching. The WTA implementation implements the basic algorithm as denoted in Eq.6 with $w(p, q) = 1$. Traditional square windows, without shifting or specific filtering, are used to aggregate the cost values, for which both SAD and SSD cost functions can be used. Cost aggregation is accelerated with the use of integral images similar to the technique in [10]. The proposed method implementation follows Alg.1 directly. A WTA matching algorithm utilizing an image pyramid and a coarse-to-fine disparity optimization was additionally implemented to compare the proposed method with a different coarse-to-fine optimization method. The environment does not implement a sub-pixel disparity refinement step after the optimization, as the goal of the

TABLE II
MEASURED ACCURACY OF TESTED METHODS

Method	Tsukuba	Venus	Teddy	Cones	Overall
3DRS only	11,79%	9,65%	35,15%	32,22%	22,20%
WTA only	10,09%	7,64%	31,28%	26,87%	18,97%
WTA pyram.	15,21%	21,95%	38,34%	34,70%	27,55%
Proposed	**9,59%**	**7,40%**	**28,79%**	**23,68%**	**17,37%**

TABLE III
MEASURED PERFORMANCE OF TESTED METHODS

Image	Tsukuba	Venus	Teddy	Cones
Resolution	384 x 288	434 x 383	450 x 375	450 x 375
3DRS only	0,024 s	0,037 s	0,042 s	0,035 s
WTA only	5,558 s	8,412 s	8,659 s	9,713 s
WTA pyram.	2,558 s	3,635 s	4,831 s	4,816 s
Proposed	**1,275 s**	**1,961 s**	**2,799 s**	**2,281 s**

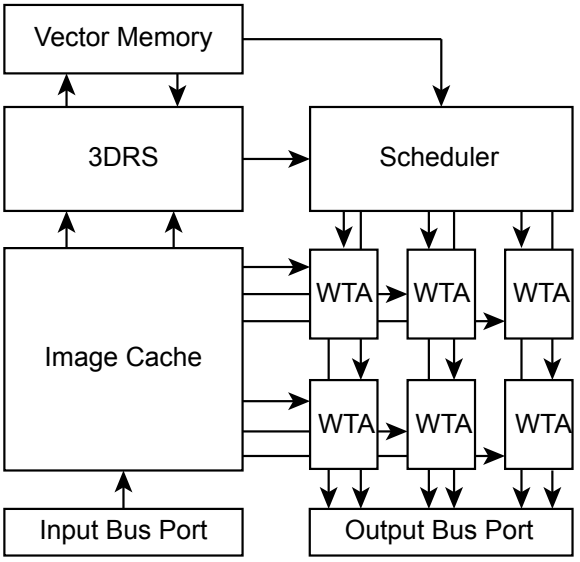


Fig. 7. Envisioned Hardware Architecture for the proposed method.

evaluation was to compare raw results of the optimization methods.

In our evaluation of the algorithm, we have measured two aspects of quality for the implemented methods. The first is the accuracy of the output results indicated by the percentage of incorrect pixels

$$B = \frac{1}{N} \sum_{(x,y)} (|d_C(x,y) - d_T(x,y)| > \delta_D) \qquad (7)$$

where d_C is the estimated disparity map, and d_T is the ground truth disparity map [1]. The threshold δ_D is 1. The second aspect is the algorithm run-time performance, indicated by the computation time. The method was tested on reference images from the Middlebury set [2]: "Tsukuba", "Venus", "Teddy" and "Cones". The resultant disparity maps are shown in Figure 6.

The measured accuracy results in Table II show that the 3DRS algorithm can provide coarse results with a level of accuracy comparable to the dense stereo optimizations such as the WTA, at a fraction of the computational complexity. The accuracy of the WTA method is typical for a fixed window implementation with the chosen window size of 15 ([1], Figure 11), although the results for the images "Teddy" and "Cones" exhibit an offset due to a large number of occluded pixels not excluded from evaluation. However, it is obvious that the proposed hybrid method either retains the accuracy of the WTA method or slightly improves it. The results of the WTA Pyramid method are similar to WTA while slightly lower in quality due to the lower resolution of reference images at the higher pyramid levels.

The computational time shown in Table III greatly favors the hybrid method in comparison to both the classic and the pyramid version of the WTA. The 3DRS execution time, although an order of magnitude faster, is provided for reference only, as it provides only coarse results.

V. FUTURE WORK AND HARDWARE IMPLEMENTATION

In our future research, we plan to explore the properties of the proposed method to increase accuracy of the results, as well as further improve the performance by evaluating different algorithm settings and cost functions in order to obtain the optimum configuration. The ultimate goal of the research is to develop a computing architecture leading to an efficient hardware implementation of the proposed method in an ASIC or an FPGA integrated circuit. As the WTA evaluations which yield the final dense disparity map are performed over discrete chunks of the image, the method is very suitable for a parallel computing implementation. The envisioned architecture is shown in Fig.7. Our idea is to combine an existing efficent 3DRS hardware implementation with a parallel computing array implementing multiple processing nodes, equipped with local memories required to store the local algorithm data. A shared cache memory should provide left and right image data to the processing units in order to exploit the locality imposed by the 3DRS scanning order. To enhance performance, a pipelined operation is envisioned, with the WTA operations dispatched as soon as the 3DRS results are available. We expect to achieve a real-time performance on high definition video at 60 frames per second.

VI. CONCLUSION

We have presented a new hybrid method for dense stereo matching which combines the 3DRS motion estimator and the classic WTA approach. The proposed method performs significantly better than the classic WTA implementation, while also being insensitive to the range of input disparities. The method is suitable for parallel hardware implementations. Our future work will focus on defining a hardware architecture for the proposed method.

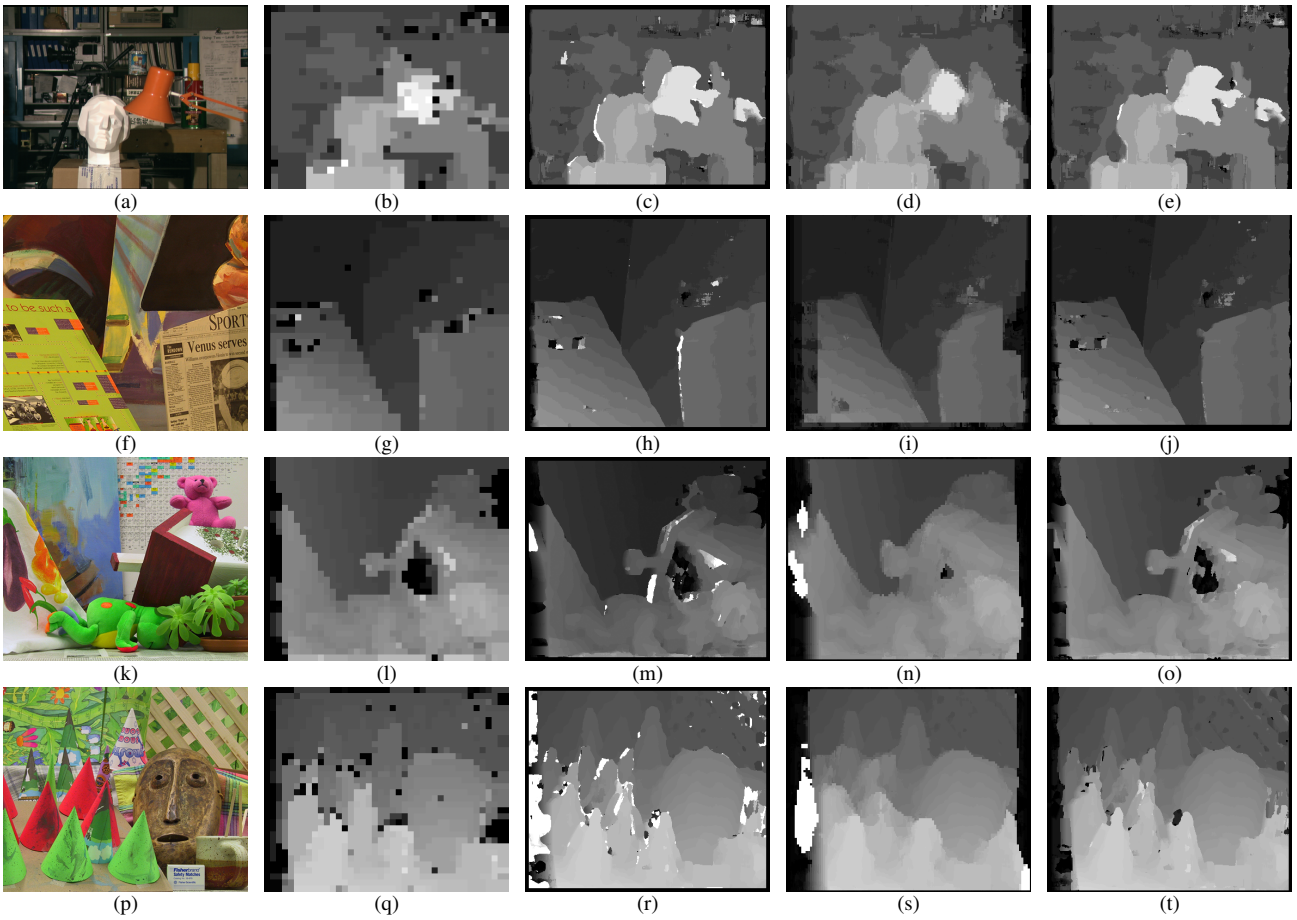


Fig. 6. Simulation results. From top to bottom: Tsukuba (6a-6e), Venus (6f-6j), Teddy (6k-6o), Cones (6p-6t). From Left to Right: Left Image, 3DRS Disparities, WTA Disparities, WTA Pyramid Disparities, Proposed Method Disparities.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, pp. 7–42, Apr. 2002.

[2] D. Scharstein and R. Szeliski, "Middlebury Stereo Vision Page." http://vision.middlebury.edu/stereo, 2014. [Online; accessed 31-Jan-2014].

[3] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Proceedings of the 10th Asian Conference on Computer Vision - Volume Part I*, ACCV'10, (Berlin, Heidelberg), pp. 25–38, Springer-Verlag, 2011.

[4] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, June 2007.

[5] C. Stentoumis, L. Grammatikopoulos, I. Kalisperakis, and G. Karras, "Implementing an Adaptive Approach for Dense Stereo-Matching," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 309–314, July 2012.

[6] D. Min, J. Lu, and M. Do, "A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy?," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1567–1574, Nov 2011.

[7] T. Pribanic, N. Obradovic, and J. Salvi, "Stereo computation combining structured light and passive stereo matching," *Optics Communications*, vol. 285, no. 6, pp. 1017–1022, 2012.

[8] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1582–1599, 2009.

[9] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–103–I–110 vol.1, 2001.

[10] O. Veksler, "Fast variable window for stereo correspondence using integral images," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, pp. I–556–I–561 vol.1, June 2003.

[11] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1402–1409, June 2012.

[12] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[13] G. de Haan, P. Biezen, H. Huijgen, and O. Ojo, "True-motion estimation with 3-d recursive search block matching," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 3, pp. 368–379, 388, Oct 1993.

[14] E. Hendriks and G. Marosi, "Recursive disparity estimation algorithm for real time stereoscopic video applications," in *Image Processing, 1996. Proceedings., International Conference on*, vol. 1, pp. 891–894 vol.2, Sep 1996.

[15] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, pp. 321–334, March 2004.

[16] J. Cai, "Fast stereo matching: Coarser to finer with selective updating," in *Image and Vision Computing New Zealand 2007*, (Hamilton, New Zealand), pp. 266–270, Image and Vision Computing New Zealand, 2007. For more information, please refer to the conference website (see hypertext link) or contact the author.

[17] A. Heinrich, C. Bartels, R. van der Vleuten, C. Cordes, and G. de Haan, "Optimization of hierarchical 3drs motion estimators for picture rate conversion," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, pp. 262–274, April 2011.

Two Approaches to Derive Approximate Formulae of NILT Method with Generalization

L. Brančík* and N. Smith*
* Brno University of Technology/Department of Radio Electronics, Brno, Czech Republic
brancik@feec.vutbr.cz

Abstract - The paper deals with relationship between two approaches to derive approximate formulae of one specific numerical inverse Laplace transform (NILT) method, which is based on the approximation of the exp(st) function in the definition Bromwich integral, and the method based on the direct numerical integration of this ILT integral. It is shown that respective approximate formulae can also be derived by integrating the Bromwich integral numerically provided the integration path and the step are properly chosen as time-dependent. The generalization of the NILT formulae is also suggested leading to possibility to predict a limiting absolute error. The experimental error analysis is performed in the Matlab program for properly chosen Laplace transforms, and modified usage of Euler transformation to accelerate convergence of infinite series is tested successfully.

I. INTRODUCTION

The Laplace transform is used in many scientific and engineering areas to solve relevant differential equations. In field of the electrical engineering and electronics it is usually the solution of transient processes in linear time-invariant systems. After a given problem is resolved in the Laplace domain, $F(s)$, the inevitable step is to find the original in the time domain, $f(t)$. This is the most difficult part in general as finding original in its analytical form is either impractical, hard or even impossible in some cases. We can e.g. mention transcendental and irrational Laplace transforms which occur in systems with distributed parameters [1]. The numerical inverse Laplace transform (NILT) can thus be an effective tool to overcome these difficulties, when determining the original quite fast and accurately enough for practical purposes.

There is many various NILT methods that distinguish from one another in its accuracy, computation fastness, difficulty at the correct choice of the method's parameters, and its versatility to cover wide range of transforms which can be processed [2] - [6]. This paper is focused on the NILT method in which the inverse kernel of the Laplace transform, that is the exponential function $\exp(st)$, is approximated by relations containing hyperbolic functions sinh(.) and cosh(.), or their combinations [5], [6]. This method does not require the determination of poles of the image $F(s)$ to be inverted, and its accuracy can be improved by increasing the number of terms in the sum and by using the method for acceleration of convergence of the infinite series in the formula for computation of the original $f(t)$. It will be shown that the same NILT formulae can be derived by the other way, namely by direct numerical integration of the ILT Bromwich integral under condition both the integration path and step are chosen properly as time-dependent [7], [8]. Based on this fact the modification of the NILT [6] is suggested while keeping predictability of the absolute error. Experimental error analysis is performed on chosen Laplace transforms in the Matlab language environment, and modified version of the Euler transformation is tested to improve the accuracy. Finally, a practical example of the application to electrical systems with distributed parameters is presented.

Research described in this paper was financed by Czech Science Foundation under grant 15-18288S, by IGA BUT in Brno under grant FEKT-S-14-2281, and by Czech Ministry of Education in frame of National Sustainability Program under grant LO1401. For research, infrastructure of the SIX Center was used.

II. RESUMPTION OF ORIGINAL NILT METHOD

In [6] there is developed the NILT method based on the approximation of the Laplace transform's inverse kernel, $\exp(st)$, in the ILT definition Bromwich integral

$$f(t) = \frac{1}{2\pi \mathrm{j}} \int_{c-j\infty}^{c+j\infty} F(s) e^{st} ds \,, \tag{1}$$

considering assumptions

a) $F(s)$ is regular for $\mathrm{Re}\{s\} > 0$,

b) for $|s| \to \infty$ is $F(s) \to 0$,

c) $F^*(s) = F(s^*)$, by the relations,

by the relations

$$E_s(st, a) = \frac{e^a}{2\sinh(a - st)} \,, \tag{2}$$

$$E_c(st, a) = \frac{e^a}{2\cosh(a - st)} \,. \tag{3}$$

After replacing reciprocal hyperbolic functions by their infinite series expansions (below in a variable z)

$$\frac{1}{\sinh z} = \frac{1}{z} + 2z \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2\pi^2 + z^2} \,, \tag{4}$$

$$\frac{1}{\cosh z} = 2\pi \sum_{n=0}^{\infty} \frac{(-1)^n \left(n + 1/2\right)}{\left(n + 1/2\right)^2 \pi^2 + z^2} \,, \tag{5}$$

and successive application of residual theorem, see [6] for details, it is possible to derive approximate formulae

$$f_s(t,a)=\frac{e^a}{t}\left(\frac{1}{2}F\left(\frac{a}{t}\right)+\sum_{n=1}^{\infty}(-1)^n\,\mathrm{Re}\left\{F\left(\frac{a}{t}+\mathrm{j}n\frac{\pi}{t}\right)\right\}\right), \quad (6)$$

$$f_c(t,a)=\frac{e^a}{t}\sum_{n=1}^{\infty}(-1)^n\,\mathrm{Im}\left\{F\left(\frac{a}{t}+\mathrm{j}\left(n-\frac{1}{2}\right)\frac{\pi}{t}\right)\right\}. \quad (7)$$

On the other hand, by expressing hyperbolic functions by their definition exponential functions, (2) and (3) can be arranged into the form of the sum of infinite geometric series, from which the absolute errors of the formulae (6) and (7) follow, respectively

$$\varepsilon_s(t,a)=\sum_{n=1}^{\infty}e^{-2na}\,f\left((2n+1)t\right), \quad (8)$$

$$\varepsilon_c(t,a)=\sum_{n=1}^{\infty}(-1)^n e^{-2na}\,f\left((2n+1)t\right). \quad (9)$$

The accuracy of the method can further be increased by considering the arithmetic mean of (6) and (7), which leads to an approximate formula

$$f_a(t,a)=\frac{e^a}{2t}\left(\frac{1}{2}F\left(\frac{a}{t}\right)+\sum_{n=1}^{\infty}(-1)^n\left(\mathrm{Re}\left\{F\left(\frac{a}{t}+\mathrm{j}n\frac{\pi}{t}\right)\right\}+\right.\right.$$
$$\left.\left.+\,\mathrm{Im}\left\{F\left(\frac{a}{t}+\mathrm{j}\left(n-\frac{1}{2}\right)\frac{\pi}{t}\right)\right\}\right)\right), \quad (10)$$

and gives an absolute error

$$\varepsilon_a(t,a)=\sum_{n=1}^{\infty}e^{-4na}\,f\left((4n+1)t\right). \quad (11)$$

Designating M as a maximal absolute value of an original $f(t)$, and using a formula for the sum of infinite geometric series, the limiting absolute errors of (6), (7) and (10) are

$$\varepsilon_{sM}(a)=\varepsilon_{cM}(a)=\frac{Me^{-2a}}{1-e^{-2a}}\approx Me^{-2a}, \quad (12)$$

$$\varepsilon_{aM}(a)=\frac{Me^{-4a}}{1-e^{-4a}}\approx Me^{-4a}. \quad (13)$$

It is evident that the absolute errors are adjustable by a proper choice of the parameter $a > 0$. Due to a finite number of terms which can be enumerated at the infinite series in practice, the real errors are higher. To suppress this effect an Euler transformation can be applied which is being used to accelerate alternating infinite series [5], [6]. A principle is that to the sum of N_S terms of the original series another N_E terms is added, but weighted by factors

$$2^{-N_E}V_{N_E-k+1}, \quad (14)$$

$k = 1, 2, ..., N_E$, and where V_r is given recursively as

$$V_{r+1}=V_r+\binom{N_E}{r}, \quad V_1=1. \quad (15)$$

III. Bromwich Integral Numerical Integration

Let's consider the ILT Bromwich integral (1), and the assumptions a) to c). Letting the contour be a vertical line, $s = c + \mathrm{j}\omega$, with c as its distance from the imaginary axis, such that $F(s)$ has no singularities on or to the right of this line, the original function can be evaluated by

$$f(t)=\frac{e^{ct}}{2\pi}\int_{-\infty}^{\infty}F\left(c+\mathrm{j}\omega\right)\left(\cos\omega t+\mathrm{j}\sin\omega t\right)d\omega. \quad (16)$$

Replacing the differential $d\omega$ for the difference Ω and the integration for the summation we get a formula

$$\hat{f}(t)=\frac{e^{ct}\Omega}{2\pi}\sum_{k=-\infty}^{\infty}F\left(c+\mathrm{j}k\Omega\right)\left(\cos k\Omega t+\mathrm{j}\sin k\Omega t\right). \quad (17)$$

In the case that a distance of the integration path from the imaginary axis, c, and the integration step, Ω, are chosen as time-dependent, namely inversely proportional to time in which the original is to be computed

$$c=\frac{a}{t}, \quad (18)$$

$$\Omega=\frac{\pi}{t}, \quad (19)$$

excluding beginning of the time axis, $t \neq 0$, (17) leads to

$$\tilde{f}(t)=\frac{e^a}{2t}\sum_{k=-\infty}^{\infty}F\left(\frac{a}{t}+\mathrm{j}k\frac{\pi}{t}\right)\left(\cos k\pi+\mathrm{j}\sin k\pi\right). \quad (20)$$

In the following two cases will be considered:

1) The summation index in (20) gathers integer values $k = \ldots -2, -1, 0, 1, 2, \ldots$, when this equation becomes

$$\tilde{f}_1(t)=\frac{e^a}{2t}\sum_{k=-\infty}^{\infty}(-1)^k F\left(\frac{a}{t}+\mathrm{j}k\frac{\pi}{t}\right). \quad (21)$$

Now, under assumption c), i.e. $F^*(s) = F(s^*)$, and validity $F(s)+F^*(s)=2\,\mathrm{Re}\{F(s)\}$, we get from (21)

$$\tilde{f}_1(t)=\frac{e^a}{t}\left(\frac{1}{2}F\left(\frac{a}{t}\right)+\sum_{k=1}^{\infty}(-1)^k\,\mathrm{Re}\left\{F\left(\frac{a}{t}+\mathrm{j}k\frac{\pi}{t}\right)\right\}\right), \quad (22)$$

which is, however, a formula corresponding to (6).

2) The summation index in (20) gathers fractional values $k = \ldots -3/2, -1/2, 1/2, 3/2, \ldots$, or substituting $k = l - ½$, it is possible to take integer index $l = \ldots -2, -1, 0, 1, 2, \ldots$. again, when this equation becomes

$$\tilde{f}_2(t)=\frac{e^a}{\mathrm{j}2t}\sum_{l=-\infty}^{\infty}(-1)^l F\left(\frac{a}{t}+\mathrm{j}\left(l-\frac{1}{2}\right)\frac{\pi}{t}\right). \quad (23)$$

Now, under assumption c), i.e. $F^*(s) = F(s^*)$, and validity $F(s)-F^*(s)=\mathrm{j}2\,\mathrm{Im}\{F(s)\}$, we get from (23)

$$\tilde{f}_2(t) = \frac{e^a}{t}\sum_{l=1}^{\infty}(-1)^l \operatorname{Im}\left\{F\left(\frac{a}{t} + \mathrm{j}\left(l-\frac{1}{2}\right)\frac{\pi}{t}\right)\right\}, \quad (24)$$

which is, however, a formula corresponding to (7).

From a geometric viewpoint, (22) and (24) correspond to a rectangular integration method, with the difference in a shift by half of the integration step (19).

Now, we will consider the integration step halfway in relation to (19), nameky

$$\Omega = \frac{\pi}{2t}, \quad (25)$$

while the Bromwich integration path is kept according to (18). Then (17) results in

$$\hat{f}(t) = \frac{e^a}{4t}\sum_{k=-\infty}^{\infty} F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{2t}\right)\left(\cos k\frac{\pi}{2} + \mathrm{j}\sin k\frac{\pi}{2}\right). \quad (26)$$

Let the summation index be chosen again as integer, i.e. k = … –2, –1, 0, 1, 2, …. Then we get from (26)

$$\hat{f}(t) = \frac{e^a}{4t}\sum_{k=-\infty}^{\infty} \mathrm{j}^k F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{2t}\right). \quad (27)$$

Considering again assumption c), $F^*(s) = F(s^*)$, identities $F(s) + F^*(s) = 2\operatorname{Re}\{F(s)\}$, $F(s) - F^*(s) = \mathrm{j}2\operatorname{Im}\{F(s)\}$, and collecting pairs of terms with neighbouring indexes, we get a result

$$\hat{f}(t) = \frac{e^a}{2t}\left(\frac{1}{2}F\left(\frac{a}{t}\right) + \sum_{k=1}^{\infty}(-1)^k\left(\operatorname{Re}\left\{F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{t}\right)\right\} + \right.\right. \\ \left.\left. + \operatorname{Im}\left\{F\left(\frac{a}{t} + \mathrm{j}\left(k-\frac{1}{2}\right)\frac{\pi}{t}\right)\right\}\right)\right), \quad (28)$$

which is, however, a formula corresponding to (10).

If shifting the integration step by its half a resulting formula does not directly correspond to the original method [6].

IV. Proposed NILT Generalization

Some examples of the numerical inversion of typical Laplace transforms through approximate formulae (6), (7), and (10), incl. the experimental verification of limiting absolute errors (12) and (13), are presented in [6]. As was shown in previous chapter, however, these formulae are implicitly present in (17), while properly choosing parameters c, Ω, and a set of summation indexes. Let's adopt generalization when the integration step is further decreasing in compliance with

$$\Omega = \frac{\pi}{mt}, \quad (29)$$

m integer, while keeping c as in (18). Then (17) results in

$$\hat{f}(t) = \frac{e^a}{2mt}\sum_{k=-\infty}^{\infty} F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{mt}\right)e^{\mathrm{j}k\frac{\pi}{m}}. \quad (30)$$

In the following we restrict the summation index to run integers k = … –2, –1, 0, 1, 2,…. The reason is that at the integration step (19) the limiting error is almost the same for fractional indexes, see (12), the difference is at a sign of absolute errors, cp. (8), (9). It is therefore expectable, when shifting the index k by ½, no significant change in errors occurs neither for the other Ω. Then (30) leads to

$$\hat{f}(t) = \frac{e^a}{2mt}\left(F\left(\frac{a}{t}\right) + \sum_{k=1}^{\infty}\left(F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{mt}\right)e^{\mathrm{j}k\frac{\pi}{m}} + \right.\right. \\ \left.\left. + F\left(\frac{a}{t} - \mathrm{j}k\frac{\pi}{mt}\right)e^{-\mathrm{j}k\frac{\pi}{m}}\right)\right), \quad (31)$$

when, after considering assumption c), $F^*(s) = F(s^*)$, and identity $F(s) + F^*(s) = 2\operatorname{Re}\{F(s)\}$, a result is

$$\hat{f}(t) = \frac{e^a}{2mt}\left(F\left(\frac{a}{t}\right) + 2\sum_{k=1}^{\infty}\operatorname{Re}\left\{F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{mt}\right)e^{\mathrm{j}k\frac{\pi}{m}}\right\}\right). \quad (32)$$

After considering $\operatorname{Re}\{F_1(z)\cdot F_2(z)\} = \operatorname{Re}\{F_1(z)\}\cdot\operatorname{Re}\{F_2(z)\} - \operatorname{Im}\{F_1(z)\}\cdot\operatorname{Im}\{F_2(z)\}$, (32) leads to

$$\hat{f}(t) = \frac{e^a}{mt}\left(\frac{1}{2}F\left(\frac{a}{t}\right) + \sum_{k=1}^{\infty}\left(\operatorname{Re}\left\{F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{mt}\right)\right\}\cos k\frac{\pi}{m} - \right.\right. \\ \left.\left. - \operatorname{Im}\left\{F\left(\frac{a}{t} + \mathrm{j}k\frac{\pi}{mt}\right)\right\}\sin k\frac{\pi}{m}\right)\right). \quad (33)$$

When substituting $m = 1$ we can see the sine function is zero for all the k, while the cosine function alternates its sign, a result is (22). In case of $m = 2$, either the sine or cosine function is always zero, then after collecting pairs of terms with neighbouring indexes and taking into account alternation of the sign, we get (28).

Taking into account relations between the integration steps (19), (25), and the limiting absolute errors (12), (13), respectively, we suggest to generalize these error formulae to cover also the above NILT (33) as

$$\varepsilon_M(a,m) \approx Me^{-2ma}. \quad (34)$$

As this formula does not have direct support in the theory of the original NILT [6], it will be in the first instance verified numerically on chosen Laplace transforms with known originals in the time domain.

V. Experimental Error Analysis

For the experimental error analysis with the intention to confirm (34) we choose the Laplace transforms whose originals reach limiting absolute values $M = 1$ to be able to easily conclude about errors, see Tables I to III, and Fig.s 1 to 6. In Table IV and Fig. 7, examples for time delayed originals with discontinuities are shown.

TABLE I. TESTING LAPLACE TRANSFORMS AND THEIR ORIGINALS.

i	$F_i(s)$	$f_i(t)$
1	$\frac{1}{s}$	unit step, $1(t)$
2	$\frac{s}{s^2+\omega^2}$	$\cos \omega t$
3	$\frac{e^{-\sqrt{s}}}{s}$	$\operatorname{erfc}\left(\frac{1}{2\sqrt{t}}\right)$

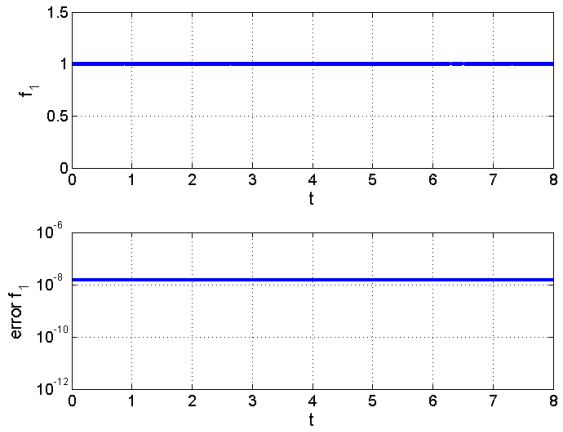


Figure 1. NILT results for $F_1(s)$ and $m = 3$

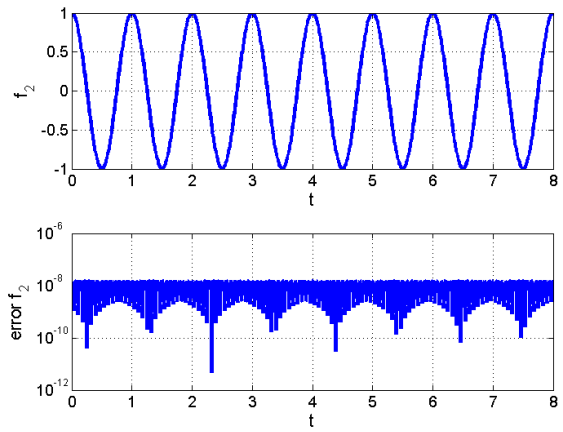


Figure 2. NILT results for $F_2(s)$ and $m = 3$

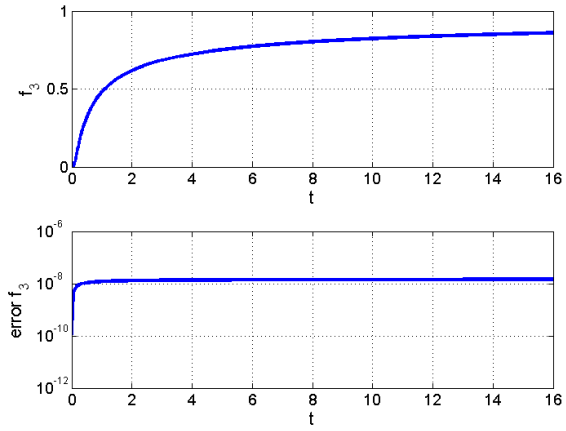


Figure 3. NILT results for $F_3(s)$ and $m = 3$

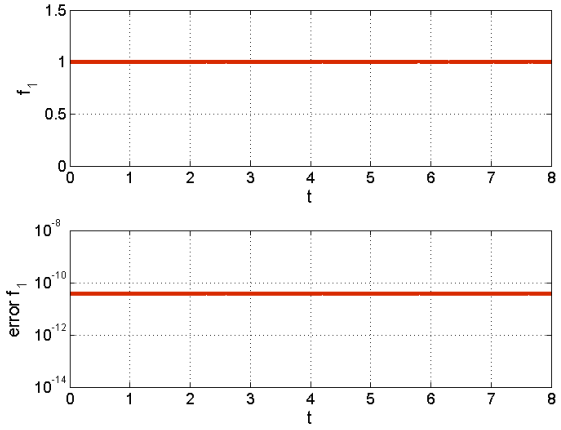


Figure 4. NILT results for $F_1(s)$ and $m = 4$

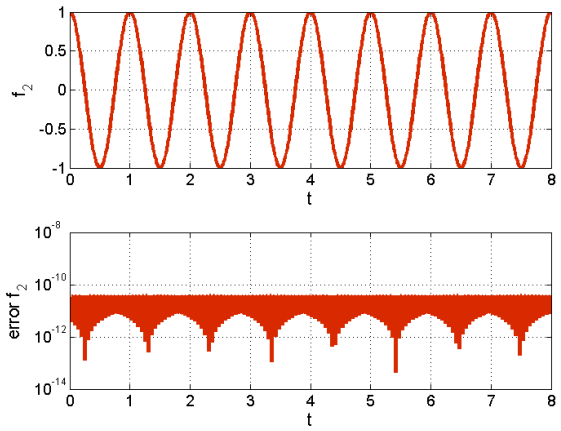


Figure 5. NILT results for $F_2(s)$ and $m = 4$

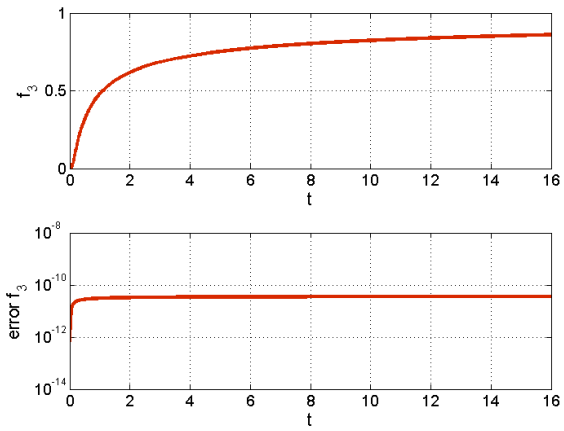


Figure 6. NILT results for $F_3(s)$ and $m = 4$

The other parameters of the NILT (33) are as follows: $a = 3$, $N_S = 100$ and $N_E = m \times 19$. The last item, a number of terms for the Euler transformation to accelerate the convergence of the infinite series, must be explained in more details. This transformation is efficient enough for alternating series only which is well fulfilled in case of $m = 1$ when its basic version (14), (15) can be used. In case of $m > 1$, however, the terms of series (33) do not alter with each change of the summation index k, therefore its modified version was applied when individual weights

are used repeatedly on neighbouring terms, namely m times. Based on a number of experiments, just such a choice seems to be well applicable. Even if the graphical results are shown only for $m = 3, 4$, to better quantify the errors, the computation was performed to get mean errors over 1000 time points, taking into account m from 1 to 4, see Table II.

TABLE II. MEAN VALUES OF ASOLUTE ERRORS ($a = 3$).

m	1	2	3	4
$\bar{\varepsilon}_{M1}$	$\mathbf{2.48\cdot10^{-3}}$	$\mathbf{6.14\cdot10^{-6}}$	$\mathbf{1.52\cdot10^{-8}}$	$\mathbf{3.78\cdot10^{-11}}$
$\bar{\varepsilon}_{M2}$	$1.58\cdot10^{-3}$	$3.91\cdot10^{-6}$	$9.69\cdot10^{-9}$	$2.40\cdot10^{-11}$
$\bar{\varepsilon}_{M3}$	$2.11\cdot10^{-3}$	$5.41\cdot10^{-6}$	$1.37\cdot10^{-8}$	$3.43\cdot10^{-11}$
theory	formulae (22), (24), see [6]		confirmation of (34)	
	$\mathbf{2.48\cdot10^{-3}}$	$\mathbf{6.14\cdot10^{-6}}$	$\mathbf{1.52\cdot10^{-8}}$	$\mathbf{3.78\cdot10^{-11}}$

We can see that for a unit step function these mean errors correspond almost exactly with the theory (highlighted by bold font styles), for the other functions the mean errors are even slightly less. If (34) is valid, the same or at least very close results should be obtained when keeping the product ma constant, see Tables II and III where is kept $ma = 3, 6, 9$, and 12.

TABLE III. MEAN VALUES OF ASOLUTE ERRORS ($m = 1$).

a	3	6	9	12
$\bar{\varepsilon}_{M1}$	$\mathbf{2.48\cdot10^{-3}}$	$\mathbf{6.14\cdot10^{-6}}$	$\mathbf{1.52\cdot10^{-8}}$	$\mathbf{1.88\cdot10^{-11}}$
$\bar{\varepsilon}_{M2}$	$1.58\cdot10^{-3}$	$3.91\cdot10^{-6}$	$9.69\cdot10^{-9}$	$3.02\cdot10^{-11}$
$\bar{\varepsilon}_{M3}$	$2.11\cdot10^{-3}$	$5.21\cdot10^{-6}$	$1.29\cdot10^{-8}$	$3.34\cdot10^{-11}$

From theoretical viewpoint is interesting that there are two possibilities how to improve the NILT accuracy, either by increasing the parameter a, which is in fact also a distance from the imaginary axis when applying a residue theorem in [6], and while keeping the integration step constant, or by decreasing the integration step while conversely fixing the parameter a on a constant value.

Hereafter, examples of the application of the NILT method to Laplace transforms leading to delayed original functions with discontinuities are presented, see Table IV.

TABLE IV. LAPLACE TRANSFORMS AND RELATED DELAYED ORIGINALS.

i	$F_i(s)$	$f_i(t)$
4	$\frac{e^{-s}}{s}$	delayed unit step $\underline{1}(t-1)$
5	$\frac{1}{s\left(1-e^{-s}\right)}$	staircase function

In this case, due to a Gibbs' effect, it is not possible to suppress effectively swings in a vicinity of discontinuities, as is obvious from Fig. 7. The acceleration of convergence of infinite series has lower effect here, and to improve the accuracy, the increasing number of terms in the sum to be evaluated has to be considered. As was already discussed in [6], slopes of delayed functions at their discontinuities depend just on a value of the delay. That is why, to keep the same precision when inverting such a kind of Laplace transforms a number of terms in the sum can be increasing according to the actual time.

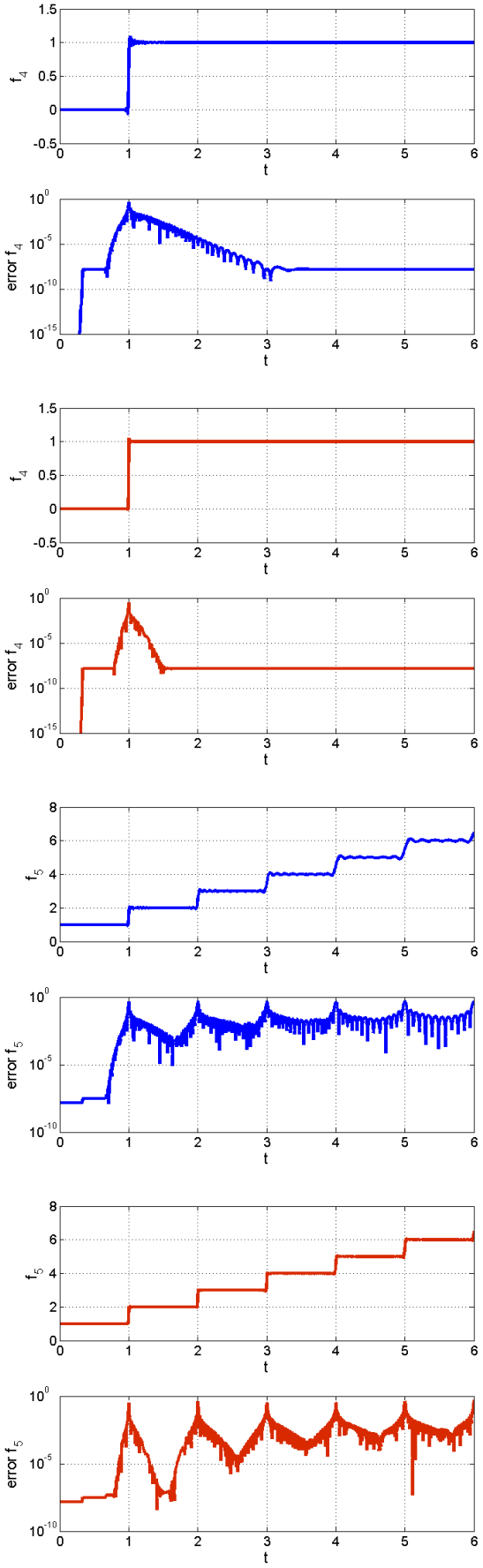


Figure 7. NILT results for $N_S = 50$, $N_E = 25$ (blue), $N_S = 200$, $N_E = 100$ (red), and $ma = 9$.

VI. Application in Electrical Engineering

Often applications of the NILT methods are to process Laplace transforms arising from the solution of systems with distributed parameters in which transcendental and irrational functions occur. As the example, let's consider a uniform lossy transmission line (TL) of the length l, being terminated in both ends by linear networks (or their Thévenin equivalents), see Fig. 8 [9].

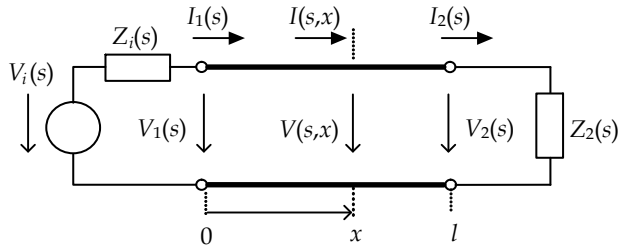


Figure 8. TL's Laplace-domain model with terminations

Our task is to determine a voltage waveform in the middle of the TL, i.e. $v(t,x)$, for $x = l/2$. The Laplace transform of the solution is [9]

$$V(s,x) = V_i(s)\frac{Z_c(s)}{Z_i(s)+Z_c(s)}\cdot\frac{e^{-\gamma(s)x}+\rho_2(s)e^{-\gamma(s)[2l-x]}}{1-\rho_1(s)\rho_2(s)e^{-2\gamma(s)l}}, \quad (35)$$

where $Z_c(s)$ and $\gamma(s)$ are the characteristic impedance and the propagation constant, respectively, equal

$$Z_c(s) = \sqrt{\frac{R_0+sL_0}{G_0+sC_0}}, \quad (36)$$

$$\gamma(s) = \sqrt{(R_0+sL_0)(G_0+sC_0)}, \quad (37)$$

with R_0, L_0, G_0, and C_0 as per-unit-length parameters, and $\rho_1(s)$, $\rho_2(s)$ as reflection coefficients in the TL's beginning and end, respectively, equal

$$\rho_1(s) = \frac{Z_i(s)-Z_c(s)}{Z_i(s)+Z_c(s)}, \quad (38)$$

$$\rho_2(s) = \frac{Z_2(s)-Z_c(s)}{Z_2(s)+Z_c(s)}. \quad (39)$$

Generally, in case of a lossy TL, an analytical form of the original cannot be found, thus just a proper NILT method should be utilized. Let's consider parameters as follows: $l = 1$ m, $R_0 = 1$ mΩ/m, $L_0 = 600$ nH/m, $G_0 = 2$ mS/m, $C_0 = 80$ pF/m, $Z_i = 10$ Ω, and $Z_2 = 1$ kΩ. The excitation voltage has a waveform $v_i(t) = \sin^2(\pi t/2\cdot10^{-9})$, if $0 \le t \le 2\cdot10^{-9}$, $v_i(t) = 0$, otherwise, with the Laplace transform

$$V_i(s) = \frac{2\pi^2\left(1-\exp(-2\cdot10^{-9}s)\right)}{s\left((2\cdot10^{-9}s)^2+4\pi^2\right)}. \quad (40)$$

The graphical result is presented in Fig. 9, when using the NILT parameters as in Fig. 7 (red).

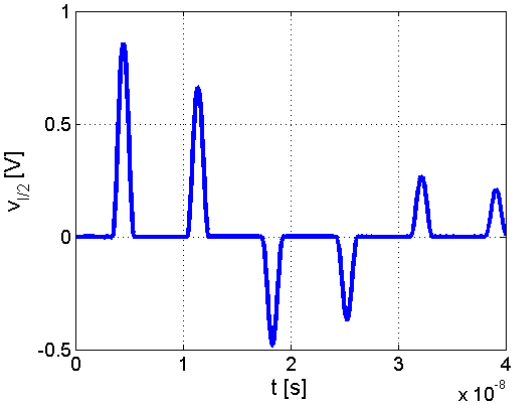


Figure 9. Voltage waveform in the middle of TL

VII. Conlusion

It is shown in the paper that the NILT formulae in [6] can also be derived by another way, through the direct numerical integration of the ILT Bromwich definition integral, namely if both a distance of the integration path from the imaginary axis and the integration step are suitable chosen as time-dependent. Further generalization was done towards the NILT formula which enables to increase the accuracy by decreasing the integration step, while allowing to predict a limiting absolute error. This approach can be regarded as an alternative way how to improve the accuracy, as in the NILT [6], this is already possible by increasing a, which is, on principle, a distance to the imaginary axis. In fact, both ways can be combined as is supposed by (34). In this case, however, the Euler transformation had to be applied with some modifications to ensure its efficiency. Thus other acceleration techniques will be investigated for this purpose, especially those also applicable on non-alternating series. Using the Matlab, CPU times were around tens of miliseconds for individual originals, and last example took about 400 ms.

References

[1] J. Pírko, J. Veit, Laplace Transform. Basics of Theory and Usage in Electrotechnics. Prague: SNTL, 1970 (in Czech).

[2] A. M. Cohen, Numerical Methods for Laplace Transform Inversion. New York, NY: Springer, 2007.

[3] V. Zakian, "Optimisation of numerical inversion of Laplace transforms," in Electronic Letters, 1970, no. 6, pp. 677-679.

[4] Vlach, J., Singhal, K., Computer Methods for Circuit Analysis and Design. New York, NY: Van Nostrand Reinhold, 1994.

[5] T. Hosono, "Numerical inversion of Laplace transforms and some applications to wave optics," in Radio Science, 1981, vol. 16, no. 6, pp. 1015-1019.

[6] J. Valsa, L. Brančík, "Approximate formulae for numerical inversion of Laplace transforms," in International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, 1998, no. 11, pp. 153-166.

[7] L. Brančík, "Comment to one method for numerical inversion of Laplace transforms," in Proceedings of the Seminary of Circuit Theory, Brno: Military Academy, 2002, pp. 38-41.

[8] L. Brančík, N. Smith, "Commentary on method of numerical inverse Laplace transformation - possible generalization and error analysis," in Electronic Horizon, 2014, vol. 70, no.4, pp. 6-10 (in Czech).

[9] L. Brančík, "Numerical Inverse Laplace Transforms for Electrical Engineering Simulation," in MATLAB for Engineers - Applications in Control, Electrical Engineering, IT and Robotics. Edited by Karel Perutka, Rijeka: InTech, 2011, pp. 51-74.

Laboratory exercise for class B/AB power amplifier

Ž. Stojanović*, S. Stojanović * and I. Činikas**
* Polytechnic of Zagreb, Zagreb, Croatia
** Vilniaus Gedimino technikos universitetas, Vilnius, Lithuania
zeljko.stojanovic@tvz.hr

Abstract - Laboratory exercise for amplifier class B/AB for sophomore is described. In the first part, the influence of quiescent current and input signal amplitude on output signal waveform distortion and amplifier efficiency is examined. Output signal waveform is observed in time domain qualitatively and in frequency domain quantitatively. In second part, amplifier's output resistance is measured. In each measurement students have to reason out measurement results, the difference between measurement results and analytical results and measurement uncertainty. Students learn concepts of quiescent point, harmonic distortion, power, output resistance and measurement uncertainty by encompassing different areas of electrical engineering – electronics, circuit and signal analysis, power analysis and measurements.

I. INTRODUCTION

Amplification of a signal is one of the main issues of analog electronics. Teaching amplifiers takes the most of the time in general analog electronics courses. It is not just that amplifiers play important, usually hidden role in many areas of human activities, but it is very interesting from educational point of view because it encompasses a variety of circuit analysis techniques related with different amplifier characteristics and levels of analysis. There are also different demands for amplifier characteristics. They depend on amplifiers purpose and they are often opposed one to each other. Essential characteristics for power amplifiers are output power, efficiency and linearity. Considering the quiescent point, power amplifiers are divided into classes like A, B, AB, C, D and others [1, 2, 3]. Amplifier classes are usually taught at sophomore level. On Polytechnics of Zagreb it is taught in courses Analog circuits and Electronic circuits. Freshmen are already introduced with concepts of quiescent point, power and efficiency, distortions and Thevenin theorem.

In this article laboratory exercise of class B/AB amplifier for sophomore [4] will be described and the main learning outcomes of the exercise will be pointed out.

Laboratory work generally helps students to become good engineers in many ways. It helps in proving and improving theory concepts, improves procedural knowledge, social skills and so on. The aim of the article is to remind teachers on varieties of circuit analysis techniques applicable for amplifier class B/AB analysis. Class B amplifiers are especially important from educational point of view, because they exhibit crossover distortion, and it is easy to approximately calculate powers and efficiency. Both classes are appropriate to learn adjustment of quiescent point. We think that power analysis in analog circuits is somewhat neglected. So, it is important to exercise power calculus, and these amplifiers are simple and concrete example for calculating powers. Importance of harmonic analysis is also one of our aims. This exercise is suitable as an introducing exercise for class D on higher level courses. One can compare class B, AB and class D efficiency and harmonic distortions. Further, impact of temperature on quiescent point and its consequences on power and harmonic analysis is also suitable way of investigation. At the end, adjusting, measuring and matching/bridging output impedance is important as an example of Thevenin theorem application or for power and efficiency consideration, small and large signal considerations and etc. One should be aware that there is a risk of overloading students with too many or too complex theory and measurement skills [5]. So, we will specify "basic measurements" which are obligatory to all students and "extensions of measurement" which is optional and suitable for earning higher grades as homework, project or as a thesis.

In exercise description all obligatory tasks during laboratory exercise will be presented and all parts that are considered as important will be commented. It is estimated 90 minutes to finish whole exercise. Some of optional tasks will be suggested as extensions of measurements. The extensions of measurement will not be analyzed in details. We think it is too broad for this article. We want to use those extensions to attract students and to increase their interest for electrical engineering.

II. DESCRIPTION OF THE EXERCISE

A. General considerations

Students have to test and compare three different amplifier configurations. One of them is class B amplifier, and the other two are low-bias and high-bias class AB amplifier [4].

There is always a compromise between benefits of assembling a circuit and loss of time for doing measurements and conclusions. We think that assembling a circuit is more important only for introductory courses and it is not so in this case. So, according to our and others [5] experience we decided to base the exercise on

conducting measurements and making conclusions, but we forced students to identify each component of physical realization of amplifier at the beginning of the laboratory exercise.

Before each measurement students have to assume the measurement results. After measurement they have to explain and comment measurement uncertainty and eventual differences between assumption and measurement results. If they are intrigued and interested they are encouraged to elaborate any part of exercise as a homework or research.

At the beginning of the exercise students are aware of its specificities and about differences between principal schematic and physical realization of the amplifier. The physical realization of amplifier is presented in Fig. 1 and its schematic is presented in Fig. 2.

B. Measurement setup

There are some differences between principle schematics and physical realization of the amplifier:

- Additional resistors R_D, R_{E1} and R_{E2} are used to fulfill thermal stability of quiescent point.
- Diodes D_1 and D_2 are also used to fulfill thermal stability of quiescent point. They are mounted together with transistors on the same heat sink [1, 5].
- Emitter resistors R_{E1} and R_{E2} are additionally useful for emitter current measurement by measurement of resistors voltage.
- Diode D_3 is used for protection of wrong polarity of supply voltage.

Change of amplifier class was realized by mechanical three-position switch. For the switch in position 1, 2 and 3 the amplifier works as a class B, low-bias class AB and high-bias class AB respectively. The single power supply was chosen because it was a smaller equipment requirement than dual power supply. The frequency of 500 Hz is chosen in order to put attention on frequency dependent and "partly accurate" measurement results obtained using ordinary multimeter. We think that each teacher should choose frequency according to his/her learning outcomes.

C. Quiescent point

Students are often unsecure in definition, and what is more important in meaning of quiescent point. The meaning of quiescent point becomes clearer in distortion considerations and in power considerations so main learning outcomes of this measurement are connected with distortion and power measurement.

Basic measurements: Depending on switch position there are three different amplifier configurations and three quiescent points. Students have to measure base-emitter voltage and emitter current for each configuration. Then, they have to determine which switch position is associated with particular amplifier configuration and to argue their conclusion. The teacher should immediately check their conclusion and comment it. The intention is to accentuate low-bias class AB amplifier and to keep it on mind and compare it with high-bias class AB and class B amplifier in later measurements. One another intention is to notice differences in base-emitter voltages between themselves and difference from the default value of 0,7 V.

Extensions of measurements: To embed potentiometer instead of resistor R_D in order to make measurements for any quiescent point. We think it is not suitable when group of students is doing exercise at the same time, because it is difficult to achieve equable quiescent point for each student in a group so it is not so easy to compare their measurement results.

D. Distortions of output signal

The amplifier is very suitable for study of various distortions of output signal. There are several sources of distortion [1, 2, 3]. Students are scenically introduced with crossover distortion, clipping distortion and distortion caused by nonlinearities of semiconductor components. It is important to underline to students that distortion of output signal is inevitable, because amplifier is consisted of transistors which *v-i* characteristics is nonlinear.

Basic measurements: There are two values of input

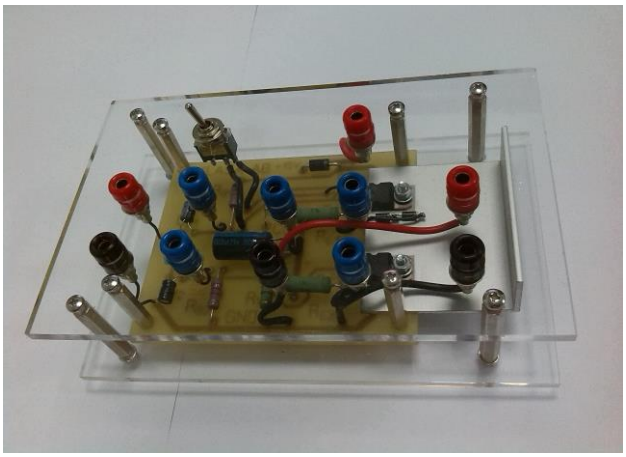


Figure 1. Physical realization of amplifier used in laboratory exercise.

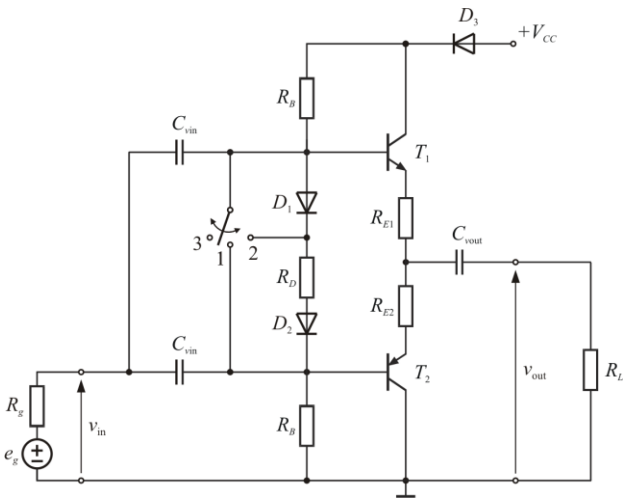


Figure 2. Schematic of amplifier used in laboratory exercise.

voltages. The smaller amplitude is 0,5 V in order to emphasize impact of threshold on signals having amplitude around base-emitter threshold voltage. The bigger amplitude was 5 V in order to reduce impact of crossover distortion on output voltage measurement. For each value of input voltage students have to measure output voltage for every switch position. Then, they have to reason differences between input and output voltage and between output voltages each other. The teacher should immediately check their conclusion and comment it. Typical measurement results are presented in Fig. 3. Students should associate these results with quiescent point determined in previous measurement.

The main learning outcomes are:

- To recognize various types of distortion and to associate them with their cause.
- To notice distortion dependence of quiescent point.
- To raise curiosity about signal distortion.

It is decided to use input voltage as a reference parameter instead of output voltage or output power because the goal was to direct the attention to crossover distortion.

Extensions of measurements: Although the distortion of output signal is very obvious it is useful in several ways to make harmonic analysis of output signal. Typical measurement results are presented in Fig. 4. The amount of higher harmonics is the biggest in "pure" class B amplifier and the smallest in high-bias class AB amplifier. It is obvious that decreasing of distortion is related with biasing. Students should associate these results with quiescent point determined in previous measurement. This measurement will also help students in understanding of Fourier analysis. It can be an introduction to Fast Fourier transform (FFT) and a way to properly choose type and width of FFT window. Students can also calculate total harmonic distortion (THD). Furthermore, it is very suitable to compare results obtained using FFT and using ordinary multimeter and True-RMS multimeter. Under certain conditions it is even better to use loudspeaker and hear the distortion. Obstacles in broader involving of FFT is that students learn FFT based signal analysis first time in later courses, but deficiency of measurement equipment also. So, this part of measurement is used only for earning higher grades through project or as graduation thesis.

E. Power and efficiency

High efficiency is one of essential amplifier characteristics especially nowadays, because energy consumption is becoming more and more important. Transistor power loss is easy to notice by touching the heat sink because the temperature of heat sink will be changed by varying the amplitude of input signal [2]. Students should be aware that efficiency in real use of

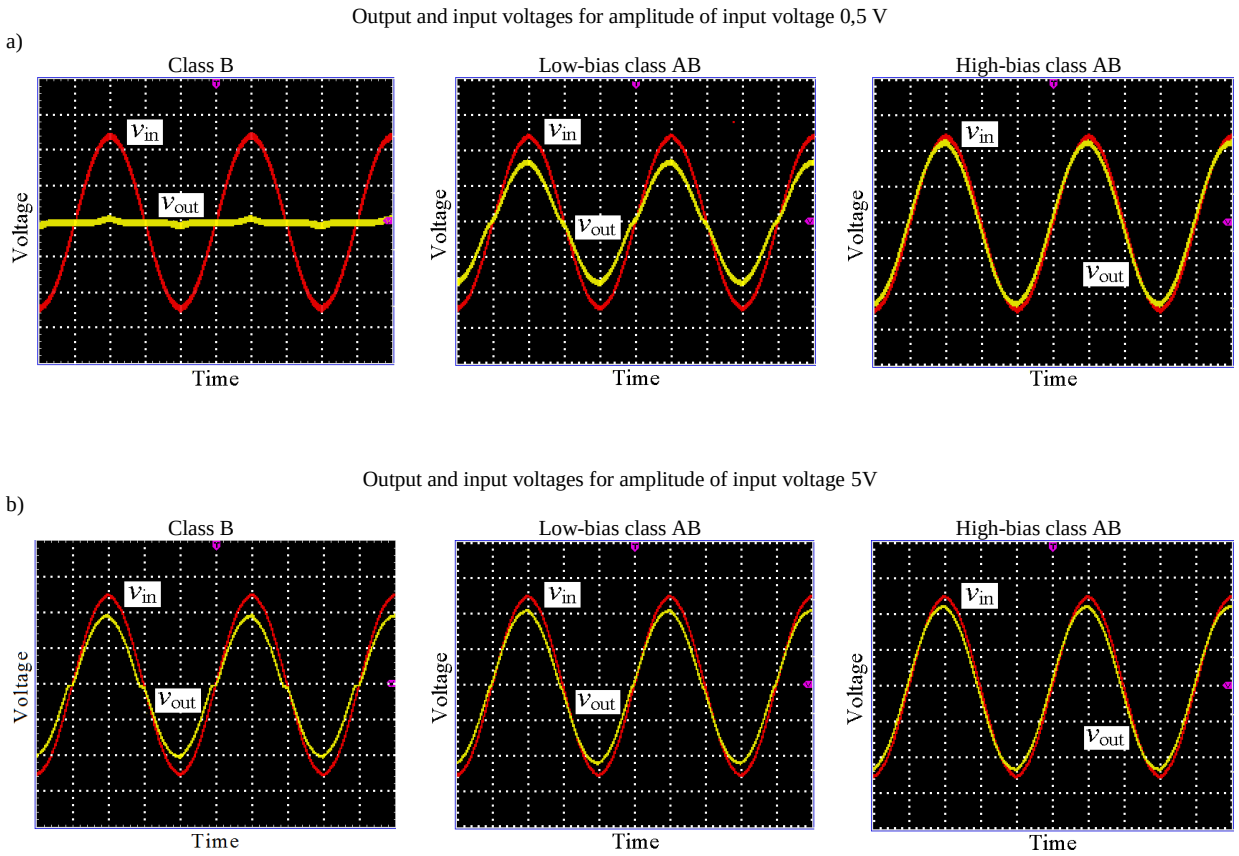
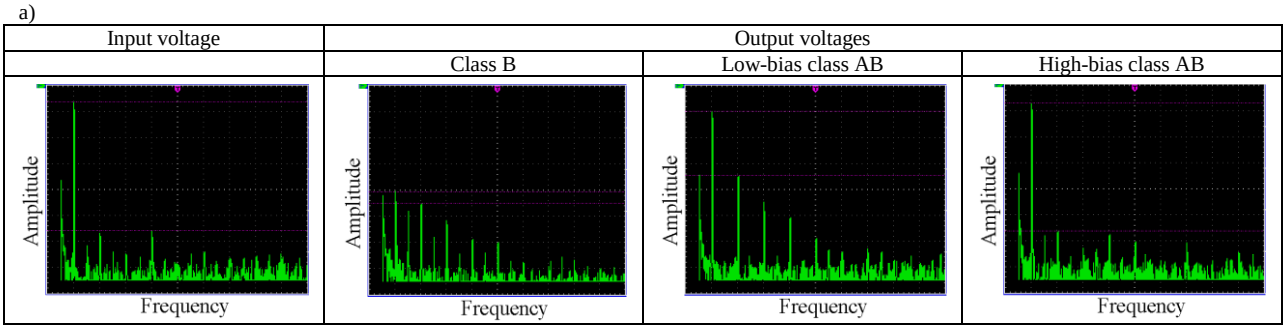


Figure 3. Input and output voltage waveforms:
a) Input voltage $\hat{V}_{in}$ = 0,5 V . Sensitivities 500 µs/div, 200 mV/div
b) Input voltage $\hat{V}_{in}$ = 5 V . Sensitivities 500 µs/div, 2 V/div

Output and input voltage spectrum for amplitude of input voltage 0,5 V

a)



Output and input voltage spectrum for amplitude of input voltage 5 V

b)

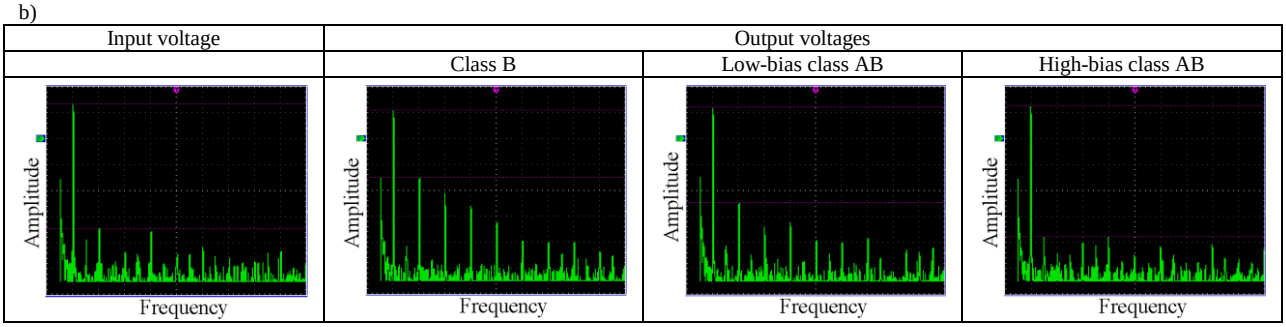


Figure 4 Input and output voltage spectrum:

a) Input voltage $\hat{V}_{\text{in}} = 0{,}5$ V . Sensitivities 1 kHz/div, 10 dB/div, Blackman window

b) Input voltage $\hat{V}_{\text{in}} = 5$ V . Sensitivities 1 kHz/div, 10 dB/div, Blackman window

amplifiers is much lower than maximum theoretical efficiency, not only because of inevitable power losses, but because input signal is most of the time much lower than the value which gives maximum output swing.

Basic measurements: There are three values of input voltages and two amplifier configurations – class B and high-bias class AB. The biggest value is the one give the maximum output swing in order to achieve the highest efficiency. The other two values are one half and one fourth of the maximum output swing. For each amplifier configuration and each input voltage students have to measure output voltage and calculate load power and supply voltage power. There are several ways to calculate/measure powers. It is possible to find load power using oscilloscope in time or frequency domain (FFT) or by using ordinary multimeter or True-RMS multimeter. Before making power calculations students have to comment the method they suggest to use. The teacher should immediately check their comments and recommend or approve method according to his/her learning outcomes and available instruments. Considering the time consumption, authors suggest use of waveforms obtained by oscilloscope or ordinary multimeter readings.

The main learning outcomes are:

- To exercise and learn converting from effective to average values of current/voltage in power calculus of supply voltage.
- To notice the difference between DC and AC components of power.
- To notice maximum efficiency and efficiency dependence on input voltage amplitude and to compare it with maximum theoretical efficiency.
- To notice the difference between maximum efficiency and efficiency feasible in real class AB amplifier excited with real signals which are usually signals of variable amplitude.

At the end of measurement, students have to compare and comment class B and class AB efficiency differences. The differences are too small for making reliable conclusion because they are affected by measurement uncertainty.

Extensions of measurements: Actual values of emitter resistors are $R_{E1} = R_{E2} = 10\ \Omega$, and load resistor is $R_L = 100\ \Omega$ in order to make lower costs and easier amplifier construction. Use of load $R_L = 4\ \Omega - 8\ \Omega$ and appropriate emitter resistors will make exercise more realistic. Under certain conditions it is even better to use loudspeaker. Another task is to check power balance of DC and AC power components separately and altogether.

F. Output resistance

The main learning outcomes are:

- To connect term of output resistance with concepts of Thevenin theorem and source resistance.
- To realize that these concepts are just a way how to descript real components or devices and to use that description (model) to do analysis.

- To notice the difference in small-signal and large-signal analysis or to do considerations about impedance matching/bridging.

These concepts are difficult to master even to the best students, but they do understanding of electrical engineering much easier.

Basic measurements: Output resistance is determined on the base of change of output voltage when the amplifier is loaded with two different loads. Measurement is conducted for default load $R_L = 100\ \Omega$ and another load $R'_L = 100\ \Omega \parallel 10\ \Omega$. Students have to comment the dependence of output voltage on load current and measurement uncertainty of the method.

Extensions of measurements: To use potentiometer as a load and to find the value which gives the half of the swing obtained for unloaded amplifier. Then, to compare results obtained with that method with results obtained by method used in exercise. Further, for the small/large signal considerations it is interesting to compare results obtained for different values of input signal. For general amplifier considerations it is convenient to examine how impedance matching/bridging will affect on whole structure of any amplifier.

III. ACTUAL STUDENT PERFORMANCE

There are many perceived obstacles for successful completion of the exercise. They are connected with general concepts and skills. We will mention and comment obstacles characterizing most of the students during basic measurements:

- Identification of components of physical realization of amplifier at the beginning of the laboratory exercise. This is not usual type of task in regular teaching in Polytechnics of Zagreb.
- Writing of explanations and conclusions. We think that writing skills are low for many students of technical area, not only in our Polytechnics but in whole country and we believe it is the same in whole world. So, if the teacher insists on proper writing students are not able to finish whole exercise.
- Correct notation of circuit variables and correct and clear notation of waveforms. We think that teachers should be more consciousness about that problem and they should increase their effort in correcting student's mistakes.
- Many of students do not make difference between effective and average value.
- Students believe that $P = UI$ works always regardless of the electrical component. It is hard to change first knowledge about electrical power acquired in their previous education or courses.
- Understanding of concept of output impedance. Understanding of Thevenin theorem is big obstacle for most of students. Dichotomous role of amplifier as load for input signal and as source for output signal also complicate understanding of amplifier parameters.

Successful parts of the exercise are connected with concepts specific to amplifiers and power amplifiers. We will mention and comment perceived one:

- Measuring of quiescent point. It is classic procedure in every laboratory exercise.
- Using of oscilloscope. It is classic procedure in every laboratory exercise, but still can be an obstacle for minority of class.
- Linking quiescent point with distortions.
- Recognition of various types of distortion and linking them with their cause.
- Linking efficiency with amplitude of signal.
- Noticing the difference between DC and AC power components.

Mentioned obstacles are more general concepts (except of identification of components) than successful parts of the exercise, so we believe they are more difficult to master. It is a task and duty for many courses and teachers to improve student's knowledge in these areas.

IV. CONCLUSION

Laboratory exercise of class B/AB amplifier was presented. There are four main topics – quiescent point, distortions of output signal, power and efficiency and output resistance.

Suggestion for basic measurements and extended measurements for each subject are given. The aim was to remind teachers on varieties of circuit analysis techniques applicable for amplifier class B/AB analysis and to create basis for (under)graduate thesis.

Comments of main obstacles in exercise and successful parts are given. We believe that obstacles are of very general nature, so it is a task for many teachers to improve student's knowledge in these areas.

Future work will include simulation of amplifier and implementation of suggested extended measurements in (under)graduate thesis and articles.

REFERENCES

[1] A. Šarčević, Elektroničke komponente i analogni sklopovi, COUO Ruđer Bošković, Zagreb, 1987, pp. 235–264

[2] Ž. Butković, Elektronika 2, Fakultet elektrotehnike i računarstva - ZEMRIS, Zagreb, 2010, pp. 65–85

[3] D. Self, Audio power amplifier design handbook, 3rd ed., Newnes, Oxford, 2002

[4] Ž. Stojanović, Elektronički sklopovi – Upute za laboratorijske vježbe, Tehničko veleučilište u Zagrebu – ELO, Zagreb, 2014, pp. 23–27

[5] O. H. Graven, D. A. H. Samuelsen, "Remote laboratory exercise for class B/AB power amplifier," IEEE EDUCON, Jordan, 2010, pp. 734–738

Integration of High-Level Synthesis to the Courses on Reconfigurable Digital Systems

I. Skliarova*, V. Sklyarov *, A. Sudnitson**, and M. Kruus**
* University of Aveiro/Department of Electronics, Telecommunications and Informatics, IEETA, Aveiro, Portugal
** Tallinn University of Technology/Department of Computer Engineering, Tallinn, Estonia
iouliia@ua.pt, skl@ua.pt, alsu@cc.ttu.ee, kruus@cc.ttu.ee

Abstract - Digital system design constitutes an essential part of engineering education. Many universities worldwide target the respective courses to field-programmable gate arrays and the relevant design flow. Recent programmable systems-on-chip integrating a processor-based computing system and a programmable logic rely on quite a different design philosophy constituting a new challenge for future engineers. These new trends have to be certainly reflected in the pedagogical activity so that the current students would, in a few years, satisfy the growing industry requirements. In this paper we analyze the new design flow and explore in detail one particular part of it: a possibility of directly generating hardware modules from software specifications through high-level synthesis. The reported teaching method is to be applied in two tightly collaborating Universities: University of Aveiro in Portugal and Tallinn University of Technology in Estonia.

I. INTRODUCTION

Any electrical engineering or computer science/engineering curriculum includes as least one course on digital design [1]. Many universities worldwide target the respective course(s) to Field-Programmable Gate Arrays (FPGA). There are many reasons that certainly justify such a choice. First of all, FPGA are easily reconfigurable by the user and can be employed to implement very different functionalities, from a simple logical gate to a complete microprocessor system. Secondly, different computer-aided design (CAD) tools are readily available that can be used by the students free of charge (frequently, in slightly limited versions, which are nevertheless well suited to educational purposes) [2, 3]. Very same CAD systems are employed in industry, so the students get adequately prepared for future work as engineers. And last but not least, there are many prototyping boards available on the market that include low-cost FPGA and simple peripheral devices, such as buttons, LEDs, switches, 7-segment displays, etc. [e.g. 4, 5]. Therefore the students can design circuits and interact with them through the existing simple peripherals. Moreover, frequently the boards have expansion connectors through which more complex peripheral devices can be connected such as temperature sensors, keypads, organic LED graphic displays, Wi-Fi interfaces and others [6], permitting interesting and varied projects to be proposed to the students. Many of such boards might be acquired for a special academic price and therefore can relatively easily be afforded by both educational institutions and students.

Recently, programmable system-on-chip (SoC) devices have appeared on the market that incorporate on a chip a processing system and a programmable logic. The processing system (PS) is based on general-purpose processing cores and is well suited to support operating systems and to execute software routines. The programmable logic (PL) is a traditional FPGA and is intended to be used for implementing high-speed co-processors/subsystems. Both the PS and the PL are linked through a variety of high-speed connections. This kind of chips brings a number of advantages over both traditional FPGAs and general-purpose processors because a single chip can potentially replace two devices resulting in cost saving. This is also true that the new kind of chips, to be appreciated at fullest, is more difficult to program comparing to a standalone FPGA or a processor. Actually, the traditional FPGA design flow is being largely revised and these new trends have to be certainly reflected in the pedagogical activity so that the current students would, in a few years, satisfy the growing industry requirements.

In this paper we analyze the new design flow and explore in detail one particular part of it: a possibility of directly generating hardware modules from software specifications through high-level synthesis. The rest of the paper is organized in five sections. Section II briefly introduces the targeted programmable SoC and analyzes the SoC design flow. Section III explores high-level synthesis design flow. Section IV presents an example of a typical student project that makes use of high-level synthesis. Section V discusses difficulties and challenges that arise with inclusion of the latest programmable SoCs to the educational process. Concluding remarks are given in section VI.

II. PROGRAMMABLE SOC DESIGN FLOW

One example of a programmable system-on-chip is Zynq-7000 [7] which combines on the same micro-chip an ARM dual-core Cortex™-A9 processor (PS) with a gate array (PL) from either Artix-7 or Kintex-7 FPGA families. Zynq product range comprises a number of devices, all with slightly different features and sizes. The PS is the same - dual core ARM Cortex-A9 with NEON and

This research was supported by EU through European Regional Development Funds, by the institutional research funding IUT 19-1 of the Estonian Ministry of Education and Research, ESF grant 9251, and by Portuguese National Funds through FCT, in the context of the project PEst-OE/EEI/UI0127/2014.

floating point extensions; only the maximum processor clock frequency ranges from 666 MHz to 1 GHz. The main difference lies in the type and density of programmable logic [7]. The smallest Zynq-7000 family member (Z7010) is based on FPGA of Artix-7 family and includes 17,600 look-up tables (LUTs), 35,200 flip-flops, 240 KB of embedded block memory and 80 digital signal processing (DSP) slices. The largest device (Z7100) is an FPGA of Kintex-7 family and has roughly 15 times more logical resources, 2,020 DSP slices, and high-speed communication interfaces.

The design process for Zynq-based devices is supported by Vivado design suite [8-10]. For Zynq, a wide variety of intellectual property (IP) cores is available and the whole design process is mainly targeted to the appropriate selection, integration and interconnection of the IPs. Such an approach permits the design abstraction level to be raised resulting in reduced design cost.

A simplified programmable SoC design flow is presented in Fig. 1. First of all functional requirements are decided upon, followed by the detailed specification and software/hardware partitioning. The realized partitioning is usually not final and can be subsequently adjusted if the designers discover that it is desirable to delegate certain software functions to hardware or vice-versa. Then development and testing of both hardware and software is executed. The designers have at their disposal libraries of pre-tested hardware and software IP modules to be reused. If an own hardware IP has to be developed, a variety of design methods might be used, such as register-transfer level (RTL) design with hardware description languages (HDL) and specification in high-level languages. Whatever is the design specification, it is further synthesized, tested and packaged to an IP to be integrated with other modules. Afterwards verification of both hardware and software is required which would frequently force the designers to go back to correct connections or functionality of certain modules. Finally, the whole system is integrated and tested.

It should be mentioned that the programmable SoC design flow differs significantly from the traditional FPGA design flow [11] and requires novel materials to be integrated in the respective advanced courses on reconfigurable systems.

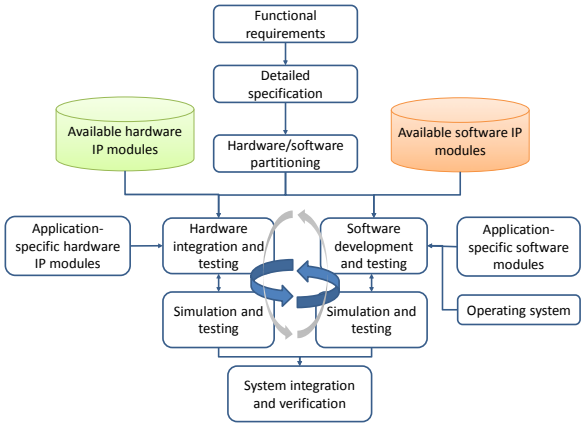


Figure 1. A typical programmable SoC design flow

III. HIGH-LEVEL SYNTHESIS

In this paper we pretend to explore one design possibility allowing hardware IP modules to be generated from high-level specifications. Commercial tools allowing digital circuits to be synthesized from high-level specification (HLS) languages, such as Handel-C and SystemC, appeared on the market in the mid-1990s - early-2000s. These languages are derived from the C language and introduce hardware-oriented extensions, libraries and compiler directives to specify features such as concurrency and timing. This fact allowed the designers to work at a higher level of abstraction. The recent high-level synthesis tools focus on using as many standard C/C++ constructs as possible to capture design intent. This makes the tools much more accessible to algorithm and system designers [12-14].

There are several reasons that force the designers to incorporate HLS languages in their practice [12, 15]:

- The growing design complexity pushes to raise the level of abstraction beyond RTL. HLS specifications allow code density to be reduced significantly compared to HDLs, resulting in increased design productivity.
- The latest HLS tools have made significant progress in providing wide coverage of language features and robust compilation technology that further lead to improved design productivity and reduction in FPGA resource usage compared to hand-coded RTL design.
- The design functionality may be specified for both embedded software and reconfigurable hardware within a programmable SoC. Therefore different software/hardware partitioning boundaries can be experimented and the respective area/performance trade-offs straightforwardly explored.
- HLS permits to leverage the experience of engineers comfortable with C/C++ languages for the purposes of digital design, greatly simplifying the process of converting a software function to a hardware module.
- The available system-level verification methods permit the RTL output to be simulated with the same test bench that was used to check the high-level code. Moreover, the verification and debug may be performed at earlier stages. This significantly reduces the verification effort because the simulation at higher levels runs much faster than at RTL level and the design errors are much easier to locate and fix.

As a result, a growing number of designs are produced using HLS tools in such application domains as video and image processing [16], wireless systems [17], aerospace applications [18], and others [19].

A. *High-level Synthesis Design Flow*

Fig. 2 illustrates a typical HLS design flow on the example of Vivado HLS tool [20, 21]. Basically, Vivado HLS transforms a C design specification into an RTL implementation that in turn can be synthesized [21].

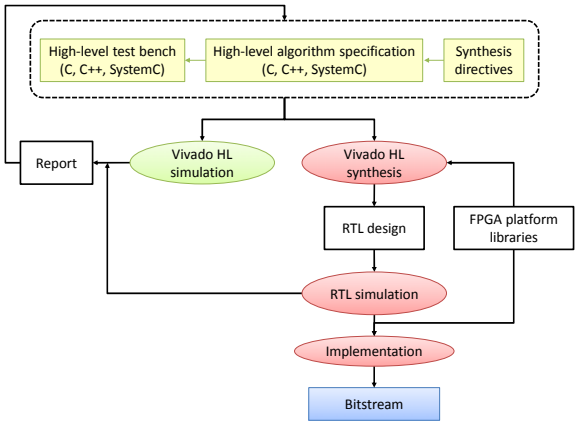


Figure 2. A typical HLS design flow on the example of Vivado

As input, specifications in synthesizable subsets of ANSI-C, C++ and SystemC are accepted [21]. Within further discussion we will limit to using C/C++ languages. The top-level of every C/C++program is the main() function and any function below the level of main() can be synthesized. The function to be synthesized is called the *top-level function* [21]. The main() function is used for verification and is therefore referred to as *test bench*. The test bench permits the behavior of the top-level function to be validated at both high and register-transfer levels.

HLS performs two types of synthesis on the design: algorithmic synthesis (which converts functional statements into RTL statements) and interface synthesis (which transforms function arguments and return values into RTL ports augmented with specific timing protocols) [21]. Synthesis directives permit the designer to improve the results of synthesis by driving the optimization engine towards the desired performance goals and RTL architecture [21]. For example, the designer might indicate which FPGA resources should be used, whether loops should be unrolled and pipelined and so on. The available profiling tool allows performance and resources to be analyzed for every level of design hierarchy and helps to apply the right synthesis directives.

The result of HLS is produced in one of the following languages: VHDL, Verilog, or SystemC [21]. The generated RTL might be verified using the original high-level test bench and is used by Vivado implementation tools to finally generate a bitstream to be uploaded to FPGA/programmable SoC or can be packaged as IP in a number of formats [21] and linked with other design modules developed using different specification methods.

IV. An Example of a Student Project

Fig. 3 illustrates a part of a software/hardware co-design project to be proposed to students which includes high-level synthesis of a single IP module. The designed system has to sort *N* *M*-bit data items with the aid of an iterative even-odd transition network [9, 22]. Sorting networks are very well suited for implementation in programmable logic because the data can be processed in parallel by a series of comparators/swappers as shown in Fig. 3 where an example of sorting *N*=8 integers is given. Every pair of connected black circles represents a comparator/swapper which compares items on its inputs and propagates the greater item to the top output and the smaller item to the bottom output. Two vertical lines of comparators/swappers (enclosed in a dashed rounded rectangle) form even and odd levels of the even-odd transition network. All the comparators/swappers that belong to the same even/odd level, function in parallel.

The data to sort are generated randomly in the PS and transferred to the PL through AXI-burst interface. Once the sorting accelerator implemented in the PL finishes, it forms an interrupt to the PS. Then, the PS either reads the sorted subsets from memory and displays the result or merges different subsets of sorted data enabling much bigger data sets to be sorted [9]. Note that a single even-odd level of the network can be reused iteratively (in different clock cycles) through a feedback register until the result of sorting is produced [22]. The maximum number of clock cycles for sorting *N* data items is equal to *N*/2. However, this can be less than *N*/2 [22, 23].

Let us describe in detail how the high-level design (in C++) of the sorting module can be accomplished in Vivado. The design entry is done with Vivado HLS tool [8] and includes two source files: ex_sort.h and ex_sort.cpp and a test bench file sort_test.cpp.

The sorter has to receive on its input the data to sort organized as a long vector of N×M bits. C/C++ built-in data types are supported by Vivado HLS, but these data types are on 8-bit boundaries (8, 16, 32, 64 bits). Since we need to process M-bit unsigned integers, we have to indicate this requirement explicitly so that only M-bit comparators will be synthesized. That is why the code will use ap_uint<M> data type instead of the unsigned int. In all other aspects standard C++ is used. C++ arbitrary precision types ap_uint<W> (for unsigned integers) and ap_int<W> (for signed integers) are defined in the file ap_int.h.

By default, the maximum width allowed for ap_uint<W> type is 1,024 bits (*i.e.* W=N×M cannot exceed 1,024). This default may be overridden by defining the macro AP_INT_MAX_W with a positive integer value less than or equal to 32,768 before inclusion of the ap_int.h header file [21]. Let us define the maximum allowed value as 16,384, which permits up to 512 32-bit data items to be received and sorted (16,384 / 32 = 512). The actual number of data to sort is defined with the constant N and the number of bits in each data item is specified with the constant M. The top-level function ex_sort will receive an N×M-bit vector of unsorted data and will return a N×M-bit vector of sorted data. As a result, the contents of the file ex_sort.h are the following (N is set to 16 in this example):

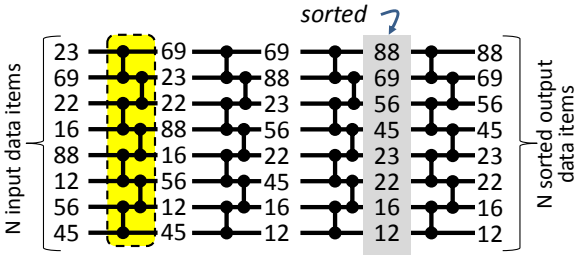


Figure 3. An example of even-odd transition sorting network for *N*=8

```
#define AP_UINT_MAX_W 16384 //must be defined before the next line
#include <ap_int.h>
const unsigned int M = 32; //number of bits in each data item
const unsigned int N = 16; //number of data items
ap_uint<N*M> ex_sort(ap_uint<N*M> input_data);
```

The parameterizable sorter functionality, based on the iterative even-odd transition network, can be specified as follows (file ex_sort.cpp):

```
#include "ex_sort.h"
ap_uint<N*M> ex_sort(ap_uint<N*M> input_data)
{  ap_uint<N*M> sorted_data;
   bool sorting_completed_odd, sorting_completed_even;
   ap_uint<M> work_array[N];

    //1. Fill in the work array
   ap_uint<M> mask = ~0;
   init_loop: for (unsigned i = 0; i < N; i++)
   {  work_array[i] = input_data&mask; //extract M Less Significant Bits (LSBs)
      input_data >>= M;                    // shift right M bits
   }

    //2. Sort the data
   sorting_completed_odd = sorting_completed_even = false;
   proc_loop:
   while (!sorting_completed_odd | !sorting_completed_even)
   {  ap_uint<M> swap;
      // processing even pairs of registers
      sorting_completed_even = true;
      sort_even: for (unsigned j = 0; j < (N / 2); j++)
            if (work_array[2 * j] < work_array[2 * j + 1])
            {        sorting_completed_even = false;
                     swap = work_array[2 * j];
                     work_array[2 * j] = work_array[2 * j + 1];
                     work_array[2 * j + 1] = swap;
            }
      // processing odd pairs of registers
      sorting_completed_odd = true;
      sort_odd: for (unsigned j = 0; j < (N / 2 - 1); j++)
            if (work_array[2 * j + 1] < work_array[2 * j + 2])
            {        sorting_completed_odd = false;
                     swap = work_array[2 * j + 1];
                     work_array[2 * j + 1] = work_array[2 * j + 2];
                     work_array[2 * j + 2] = swap;
            }
   }

    //3. Write the result
   write_res_loop: for (unsigned i = 0; i < N; i++)
   {        sorted_data <<= M;          // shift left M bits
            sorted_data |= work_array[i];    // write M LSBs
   }
   return sorted_data;
}
```

Data processing is organized in three steps:

1) First of all the work array work_array of N M-bit data items is filled in from the supplied data input input_data.

2) Then the data are iteratively sorted with the aid of two inner *for* loops: sort_even and sort_odd. The outer *while* loop proc_loop finishes processing when the data are sorted.

3) The result stored in work_array is written to the output N×M-bit vector.

Now the test bench has to be developed that would allow to verify the functionality of the sorter. The test bench has to be written respecting the following rules:

- The test bench has to call the function to be synthesized (ex_sort).
- The test bench has to save the output from the function under test and to compare the results with the known expected results.
- If the produced results do match the expected results, a message is issued confirming that the function has passed the test and the return value of the test bench main() function is set to 0. Otherwise, if the output is different from the expected results, a message indicates this, and the return value of main() is set to any value different from 0.
- The test bench must generate the same input stimuli every time it is executed to be used successfully for RTL verification (*i.e.* test values cannot be generated randomly).

Let us organize the verification procedure as shown in Fig. 4. The test bench will read data to sort from a text file input_data.dat, call the function ex_sort, store the result in the file sorted_data.dat and finally compare the results produced with the golden results kept in the file sorted_data.gold.dat and issue the verdict. Examples of files input_data.dat and sorted_data.gold.dat are given in Fig. 4. If the test fails, the students can debug the code [9].

When the behavior of the sorting function is verified we can proceed to synthesis. The synthesis report gives details on both the performance and resources of the RTL design. The results of the synthesis reveal that, besides of LUTs and flip-flops, one memory block is used. HLS tool may correctly estimate the design latency due to the involved resources but may not account net delays due to further RTL synthesis, place and route. The *latency* is the number of cycles required to produce the output. The *initiation interval* (II) is the throughput, *i.e.* the number of cycles between new input reads. The design has a latency of 51 cycles and II = 52. The latency is composed of the following values: 16 cycles to execute the loop init_loop, 8 cycles for the loop sort_even, 7 cycles for the loop sort_odd, 16 cycles for the loop write_res_loop and plus 4 cycles are required to enter each block. The initiation interval is 52 clock cycles, *i.e.* the next set of inputs is read after 52 cycles. This is one cycle after the result is written. This indicates that the design is not pipelined which is confirmed in the synthesis report.

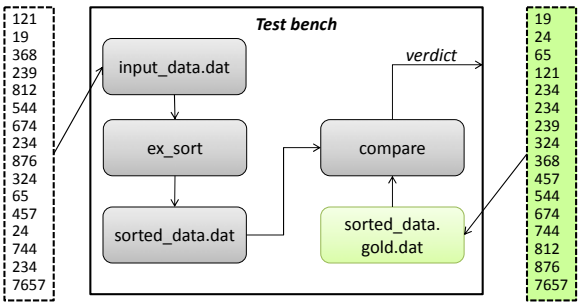


Figure 4. Verification procedure for the sorter project

Arrays in C/C++ are typically mapped to memories [21] and the array work_array was synthesized to a RAM. However, our idea is to process all even/odd data items in parallel and therefore all the array elements have to be accessed in parallel. This is obviously not possible when a RAM is used because the number of read/write ports is limited. Array optimization directives permit to map arrays efficiently to the available memory resources and to reshape or stream them [21]. The array work_array has to be implemented as registers and not RAM and this might be accomplished with the ARRAY_PARTITION directive.

ARRAY_PARTITION directive can be applied to the function ex_sort with *variable* set to work_array, *type* set to *complete* and *dim* set to 0. This directive permits arrays to be partitioned either in equally sized blocks containing consecutive elements of the original array (type *block*), in equally sized blocks containing interleaving elements (type *cyclic*), or in registers (type *complete*). The option *dim* specifies which array dimension to partition, if set to zero all dimensions are partitioned. If the design is synthesized, you can check that no memory is now used but the number of LUTs and registers has greatly increased. Moreover, latency was not improved. Therefore, additional optimization directives have to be applied to the design.

By default, all loops are kept *rolled* in HLS, *i.e.* all operations in the loop are implemented using the same hardware resources and each iteration is performed in a separate clock cycle. HLS provides the ability to unroll *for* loops, either partially or fully. When a loops is *fully unrolled* all the loop operations are performed in a single clock cycle by replicating the required for each iteration hardware resources. This leads to faster designs which however require more resources. We can fully unroll the loops init_loop and write_res_loop as well as the inner loops sort_even and sort_odd by applying the directive UNROLL to perform all comparison/swapping operations in parallel.

The final synthesis report indicates that the estimated number of LUTs and registers was reduced. The total circuit latency and the iteration latency and trip count of the loop proc_loop are not estimated correctly because the loop bounds are not constant, *i.e.* the number of times the loop is executed depends on the input data.

The results of *interface synthesis* are displayed at the end of the synthesis report. We have a sequential circuit therefore clock and reset signals have been added: ap_clk and ap_rst. Besides, signals ap_start, ap_done, ap_idle and ap_ready have also been included to support a block-level input/output (I/O) protocol ap_ctrl_hs. This is the default function-level handshaking I/O protocol. The input signal ap_start controls the block execution and must be asserted for the design to start. The output signal ap_done indicates that the function has finished and, if there is a return value, this value is valid and may be read. The output signal ap_idle, when asserted, indicates that the function is not operating (idle). The output signal ap_ready indicates when the design is ready to accept new inputs.

The remaining ports are implemented with ap_none protocol. This protocol indicates that no control signals will be included and therefore represents simple wires. By default, Vivado synthesizes input pass-by-value arguments and pointers as simple wire ports with no associated handshaking signal and output values/pointers with an associated output valid signal to indicate when the output is valid [21]. Besides of the mentioned protocols other types of interfaces can be synthesized such as two-way handshakes, RAM access ports and FIFO ports.

When the designer is satisfied with the results, RTL verification (cosimulation) can be initiated. By default, the cosimulation is performed using the generated SystemC RTL which uses the built-in C compiler. It is also possible to perform the verification in VHDL and Verilog using a third-party simulator. In cosimulation, first of all, the C test bench is executed to generate input stimuli for the RTL design. Second, an RTL test bench is created containing the input stimuli from the C test bench and the RTL simulation is performed with the selected simulation tool. And, finally, the output from the RTL simulation is fed back to the C test bench to check the results and a message is issued informing whether the design passed the test. Since the C test bench is executed twice (first to create the input stimuli and then to check the RTL simulation output), it is not possible to generate test values randomly. This is because, when the C test bench will be executed for the second time, it may create different (from the first run) test values, and a mismatch with RTL simulation output will be reported.

The final step in the HLS flow is to package the design as an IP block for use with other tools in the Vivado Design Suite [20]. As a result, an IP package will be created, which can be imported into the Vivado IP catalog and used in any Vivado design. When the final project is assembled in Vivado, which includes the communication modules using AXI-burst mode and the sorter, and implemented in the ZedBoard [24], we can see that the system occupies 4,950 LUTs (from 53,200 available) and 4,570 flip-flops (from 106,400 available).

High-level synthesis is only one of the topics covered within the course with either one or two dedicated labs. The remaining labs deal with the traditional FPGA-based design [11, 25, 26] and software/hardware co-design. In one of those labs the students implement the same sorting project from specification in HDL. After completing both projects (based on HDL and HLS) the students are asked to identify trade-offs of the used specification methods. For example, the sorter circuit synthesized from specification in VHDL (for N=16 and M=32) occupies 4,088 LUTs and 3,501 flip-flops. Such a teaching technique allows for students to try different design methods and broadens their experience.

V. DIFFICULTIES AND CHALLENGES

The new programmable SoC design methodology is being currently incorporated in courses on reconfigurable digital design in both University of Aveiro in Portugal and Tallinn University of Technology in Estonia [11, 25, 26]. Besides of the reported in the paper high-level synthesis, many other design and verification opportunities have to be covered within the courses, such as specifying design constraints, design of interfaces, analysis of synthesis results, simulation, power analysis, design integration, software development and real-time test. Many of these

topics are covered in [9, 10]. With the exception of these two books, there are no published materials in English on Zynq programmable SoCs which would be easily adaptable and suitable to the students. There are many technical reference manuals, but our experience shows that this kind of materials are not very easily adapted by the students because they need more guided references where the topics are explained in a step-by-step manner.

The second main difficulty that we currently envision is the relative complexity of the respective design and verification tools. Synthesis and test of even not-very complex systems takes too long time to be effectively used on not-the-latest generation computers connected in a network (which university labs are usually equipped with). Nevertheless we look optimistically into the future and are sure that the respective design tools will quickly advance and make the design and verification process of programmable SoC more straightforward and efficient. Because of this we are sure that the new trends have to be addressed right now by modernizing the respective course contents, design tools, and equipment.

VI. Conclusion

In this paper we reported the need to modernize courses on reconfigurable digital design due to the introduction of programmable systems-on-chip. The system-oriented design philosophy differs to the great extent from the traditional FPGA design flow and favors greatly the design reuse. One particular method of generating IPs from software specifications through high-level synthesis is analyzed in the paper in detail. Design reuse keeps aligned with the teaching methodology practiced within our universities during the last years which relies upon decomposition of the students' projects on *given to* the students and *developed by* the students components [11, 26]. This permits various components to be reused extensively leading to the increased project complexity without expanding timeslots allocated for the classes. Moreover, the students understand better the advantages and necessity of team working and are forced to document better their designs.

References

[1] The Joint Task Force on Computing Curricula, IEEE Computer Society, Association for Computing Machinery, available at http://www.acm.org/education/curricula-recommendations.

[2] Xilinx, Inc., ISE Design Suite, available at: http://www.xilinx.com/products/design-tools/ise-design-suite/.

[3] Altera, Corp., Design Software, Quartus II, available at: http://www.altera.com/products/software/sfw-index.jsp.

[4] Digilent, Inc., Nexys™4 Artix-7 FPGA Board, available at: http://www.digilentinc.com/Products/Detail.cfm?NavPath=2,400,1184&Prod=NEXYS4.

[5] Altera, Corp., DE2-115 Development and Education Board, available at: http://www.altera.com/education/univ/materials/boards/de2-115/unv-de2-115-board.html.

[6] Digilent, Inc., Sensors/Interfaces/Peripheral Modules (Pmods™), available at: http://www.digilentinc.com/Products/Catalog.cfm?NavPath=2,401&Cat=9.

[7] Xilinx, Inc. Zynq-7000 All Programmable SoC Technical Reference Manual, 2014. Available at: http://www.xilinx.com/support/documentation/user_guides/ug585-Zynq-7000-TRM.pdf.

[8] Xilinx Inc., Vivado Design Suite, available at: http://www.xilinx.com/products/design-tools/vivado/.

[9] V. Sklyarov, I. Skliarova, J. Silva, A. Rjabov, A. Sudnitson, C. Cardoso, Hardware/Software Co-design for Programmable Systems-on-Chip, TUT Press, 2014.

[10] L.H. Crockett, R.A. Elliot, M.A. Enderwitz, R.W. Stewart, The Zynq Book, Strathclyde Academic Media, 2014.

[11] V. Sklyarov, I. Skliarova, A. Sudnitson, "Methodology and International Collaboration in Teaching Reconfigurable Systems", Proceedings of IEEE Global Engineering Education Conference - IEEE EDUCON'2012, Marrakesh, Morocco, April 2012, pp. 1143-1152.

[12] J. Cong, B. Liu, S. Neuendorffer, J. Noguera, K. Vissers, Z. Zhang, "High-Level Synthesis for FPGAs: From Prototyping to Deployment", IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems, vol. 30, no. 4, 2011, pp. 473-791.

[13] P. Coussy, D.D. Gajski, M. Meredith, A. Takach, "An Introduction to High Level Synthesis", IEEE Design & Test of Computers, vol. 11, no. 4, 2009, pp. 8-17.

[14] P. Coussy, A. Moraweic (eds.). High-Level Synthesis: from Algorithm to Digital Circuit. Springer 2010.

[15] W. Meeus, K. Van Beeck, T. Goedemé, J. Meel, D. Stroobandt, "An overview of today's high-level synthesis tools", Design Automation for Embedded Systems, vol. 16, no. 3, 2012, pp. 31-51.

[16] C. Desmouliers, E. Oruklu, S. Aslan, J. Saniie, F.M. Vallina, "Image and video processing platform for field programmable gate arrays using a high-level synthesis", Computers & Digital Techniques, vol. 6 , no. 6, 2012, pp. 414-425.

[17] V. Bhatnagar, G.S. Ouedraogo, M. Gautier, A. Carer, O. Sentieys, "An FPGA Software Defined Radio Platform with a High-Level Synthesis Design Flow", Proceedings of the 77th IEEE Vehicular Technology Conference, Las Vegas, USA, September 2013.

[18] P.J. Pingree, L.J. Scharenbroich, T.A. Werne, C. Hartzell, "Implementing Legacy-C Algorithms in FPGA Co-Processors for Performance Accelerated Smart Payloads", Proceedings of the 2008 IEEE Aerospace Conference, Big Sky, USA, March 2008.

[19] P. Bishop, "A Look Into How a High-Level Synthesis Design Flow Benefits Verification Turnaround", Electronic Engineering Journal, October 31, 2013, available at: http://www.eejournal.com/archives/articles/20131031-cadence/.

[20] Xilinx, Inc., Vivado Design Suite Tutorial. High-Level Synthesis, 2013. available at: http://www.xilinx.com/support/documentation/sw_manuals/xilinx2013_4/ug871-vivado-high-level-synthesis-tutorial.pdf.

[21] Xilinx, Inc., Vivado Design Suite User Guide. High-Level Synthesis, 2013. Available at: http://www.xilinx.com/support/documentation/sw_manuals/xilinx2013_4/ug902-vivado-high-level-synthesis.pdf.

[22] V. Sklyarov, I. Skliarova, A. Barkalov, L. Titarenko, Synthesis and Optimization of FPGA-Based Systems, Springer, Switzerland, 2014.

[23] V. Sklyarov, I. Skliarova, "High-performance implementation of regular and easily scalable sorting networks on an FPGA", Microprocessors and Microsystems, vol. 38, no. 5, July 2014, pp. 470-484, doi: 10.1016/j.micpro.2014.03.003.

[24] Avnet, Inc. ZedBoard Hardware User's Guide, Version 2.2, 2014, available at: http://www.zedboard.org/sites/default/files/documentations/ZedBoard_HW_UG_v2_2.pdf.

[25] D. Mihhailov, M. Kruus, A. Sudnitson, "FPGA Platform Based Digital Design Education", ACM International Conference Proceeding Series, vol. 374: International Conference on Computer Systems and Technologies - CompSysTech'2008, Gabrovo, Bulgaria, June, 2008, pp. IV.1-IV.6, 2008.

[26] I. Skliarova, V. Sklyarov, A. Sudnitson, M. Kruus, "Teaching FPGA-based Systems", Proceedings of the IEEE Global Engineering Education Conference – EDUCON'2014, Istanbul, Turkey, April, 2014, pp. 460-469.

DC VIS

International Conference on

DISTRIBUTED COMPUTING, VISUALIZATION SYSTEMS AND BIOMEDICAL ENGINEERING

Steering Committee

Chairs: Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Roman Trobec, Jožef Stefan Institute, Ljubljana, Slovenia
Uroš Stanič, Biomedical Research and Innovative Society, Ljubljana, Slovenia

Members: Enis Afgan, Ruđer Bošković Institute, Zagreb, Croatia
Piotr Bala, Nicolaus Copernicus University, Toruń, Poland
Leo Budin, University of Zagreb, Croatia
Jelena Čubrić, Ruđer Bošković Institute, Zagreb, Croatia
Borut Geršak, University Medical Center, Ljubljana, Slovenia
Simeon Grazio, "Sisters of Mercy" University Hospital Centre, Zagreb, Croatia
Gordan Gulan, University of Rijeka, Croatia
Yike Guo, Imperial College London, UK
Ladislav Hluchý, Slovak Academy of Sciences, Bratislava, Slovakia
Željko Jeričević, University of Rijeka, Croatia
Peter Kacsuk, Hungarian Academy of Science, Budapest, Hungary
Aneta Karaivanova, Bulgarian Academy of Sciences, Sofia, Bulgaria
Charles Loomis, Laboratory of the Linear Accelerator, Orsay, France
Luděk Matyska, Masaryk University, Brno, Czech Republic
Željka Mihajlović, University of Zagreb, Croatia
Damjan Miklavčič, University of Ljubljana, Slovenia
Laszlo Szirmay-Kalos, Technical University of Budapest, Hungary
Tibor Vámos, Hungarian Academy of Sciences, Budapest, Hungary
Matjaž Veselko, University Medical Center, Ljubljana, Slovenia

PAPERS

DISTRIBUTED COMPUTING

BEST PAPER AWARD

Cloudflow – A Framework for MapReduce Pipeline Development in Biomedical Research

Lukas Forer[1,2], Enis Afgan[3,4], Hansi Weißensteiner[1,2], Davor Davidović[3], Günther Specht[2], Florian Kronenberg[1], Sebastian Schönherr[1,2],
[1]Division of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria
[2]Institute of Computer Science, Research Group Databases and Information Systems, Innsbruck, Austria
[3]Center for Informatics and Computing, Ruđer Bošković Institute, Zagreb, Croatia
[4]Department of Biology, Johns Hopkins University, Baltimore MD, USA
sebastian.schoenherr@i-med.ac.at

Abstract - The data-driven parallelization framework Hadoop MapReduce allows analysing large data sets in a scalable way. Since the development of MapReduce programs can be a time-intensive and challenging task, the application and usage of Hadoop in Biomedical Research is still limited. Here we present Cloudflow, a high-level framework to hide the implementation details of Hadoop and to provide a set of building blocks to create biomedical pipelines in a more intuitive way. We demonstrate the benefit of Cloudflow on three different genetic use cases. It will be shown how the framework can be combined with the Hadoop workflow system Cloudgene and the cloud orchestration platform CloudMan to provide Hadoop pipelines as a service to everyone.

The framework is open source and free available at https://github.com/genepi/cloudflow.

I. INTRODUCTION

Since the advent of high-throughput technologies in the field of molecular biology (i.e. Next Generation Sequencing (NGS)), even more data is produced and needs to be analysed. Thus, molecular biology has envolved into a big data science, where the bottleneck is no longer the production of raw data in the laboratory, but its analysis and interpretation [1]. To scale with the increasing data volume and the number of available resources, workflows need to be parallelized efficiently. MapReduce and Cloud Computing constitute an attractive alternative to deal with the large datasets [2]. However, writing a MapReduce job can be a challenging task that prevents domain experts from using such models in their daily work. Additionally, the reusability of the mapper and reducer functions is limited, resulting in a use-case specific implementation for every problem.

Existing high-level languages on top of Hadoop facilitate the development process and allow writing MapReduce jobs in form of queries. For example, Apache Pig (http://pig.apache.org/) provides a compiler to translate such queries into an execution plan, which is then automatically translated into a sequence of MapReduce jobs. Apache Pig provides interfaces for *filter*, *group* and *join* operations, extendible for application-specific logic using user defined functions (UDFs). However, Apache Pig has been developed to analyze datasets based on the relational model. The execution of complex calculations across several rows is limited. This implies the consequence that users have to write complex UDFs. Again, such functions need to be implemented by the end users and can be, similar to native MapReduce jobs, hard to implement, test and maintain. Several projects (e.g. SeqPig [3] or BioPig [4]) make use of UDFs and provide a collection of Pig scripts to researchers in Genetics. The goal is to provide ready-to-use workflows to end users, which can be then adapted to their use-case. Nevertheless, combining different scripts to a pipeline or reusing existing blocks does not build a key feature.

To overcome this issue, FlumeJava [5] proposed a new concept to compose pipelines based on immutable parallel collections where several operations can be used to process them in a parallel way. This concept was successfully implemented in Apache Crunch (https://crunch.apache.org/), which executes the pipelines as Hadoop MapReduce jobs. However, the utilization of such pipelines in Bioinformatics is still limited.

In this paper we present Cloudflow, a MapReduce pipeline framework, which is based on a similar concept as proposed by [5]. In contrast to existing approaches, Cloudflow was developed to simplify the pipeline creation in biomedical research, especially in the field of genetics. For that purpose Cloudflow supports a variety of NGS data formats and contains a rich collection of built-in operations for analyzing such kind of datasets (e.g. quality checks, mapping reads or variation calling). The main concept behind our approach is to break complex data analysis steps into three basic operations. All further use-case specific operations are built by implementing or extending one of the basic operations. Pipelines are then composed by creating a sequence of these operations. The framework itself translates the set of pipeline operations into one or more MapReduce jobs and decides which of the operations are executed in the *map* or in the *reduce* phase. Thus, Cloudflow hides the complexity and the implementation details of MapReduce jobs allowing scientists to build pipelines via an intuitive method. Moreover, Cloudflow can be utilised in combination with the workflow system Cloudgene [6]. To validate our approach, we developed three Genetics data-analysis pipelines with Cloudflow. The results demonstrate that our contribution (a) helps minimizing the development time, (b) increases the reusability of code, and (c) creates only a minimal overhead in terms of execution time compared to an identical MapReduce implementation.

II. MapReduce Background

MapReduce is a parallel programming model introduced by Google in 2004 with the aim to develop a simple and scalable method to process large datasets on several machines in parallel. The main idea behind this distributed programming model is that a long-time calculation is split into a map and a reduce phase which contains all the logic behind the calculation and is specified by the user. The underlying framework itself takes care of parallelization, task scheduling, load-balancing and fault-tolerance. Due to its simplicity, MapReduce is used to solve many scientific problems where large-scale computing is needed. Moreover, MapReduce is ideal for parallel batch processing of terabytes of input data. The data-flow of a MapReduce program consists of several steps, where only the map and reduce function are problem-specific and the other steps are loosely coupled with the problem and generalized. In the first step, the input data set is split into key/value pairs and a user defined map function is executed for each pair:

$$map(key, value) \rightarrow list((key_i, value_i)) \text{ , where } i = 0 \ldots n$$

The map function reads a pair, performs some problem-specific calculations and produces zero or n intermediate key/value pairs for each input pair. In the next step, the intermediate key/value pairs are grouped by similar keys and a merged list of all values for this key is created. Finally, the user-defined reduce function is applied to the intermediate key/list pairs:

$$reduce(key_i, list(value_i)) \rightarrow list(outvalue)$$

The values created by the reduce function are the final outputs of a MapReduce job.

A. Apache Hadoop

Apache Hadoop is an open source project including several sub-projects for distributed computing. It includes the most widely used open-source implementation of Google's MapReduce framework (simply called MapReduce), the distributed file system (HDFS), and several other sub-projects. In general, all previous discussed features of the original MapReduce approach are implemented in Apache Hadoop. Writing a MapReduce job with Hadoop can either be done directly in Java by implementing the relevant methods or by using Hadoop's streaming mode. This allows specifying a script or any executable as the mapper or reducer. Two different versions of MapReduce are available. MRv1 includes a namenode (centerpiece of HDFS), a secondary namenode (merging logs into a file system snapshot), a job-tracker (job assignment) and several task-trackers (workers). MRv2 or Hadoop YARN, splits the work of the job-tracker (resource management, job scheduling) into two different daemons, namely a resource manager and an application master. Additionally a node manager is introduced to manage the user processes per node. The application master describes a framework that works together with the node manager and the resource manager to monitor and to execute tasks. The idea behind this new model is that applications using frameworks different than MapReduce (e.g. Apache Spark) can be executed via YARN as well.

B. High-Level Languages based on MapReduce

To simplify the implementation of MapReduce jobs, Apache Pig has been introduced. It is based on HDFS and MapReduce and allows a fast implementation of data flows with already available data operations such as *join*, *filter* or *group by*. Data flows are specified in the Pig Latin language that describes a directed acyclic graph (DAG) and defines how data should be processed. A further advantage of Apache Pig is the ability to check the data flow on optimizations, e.g. if two grouping statements can be combined. The costs to write code in Apache Pig are lower than setting up a Java project and implementing the functions. But as stated earlier, Apache Pig includes only a limited number of operators and writing Apache MapReduce jobs directly in Java yields to advantages in speed compared to Apache Pig.

C. MapReduce Pipelining

The pipeline framework FlumeJava [5] is based on the concept of immutable parallel collections. This kind of data-structure can be used to process and analyze their items in parallel. The end user can either use one of the predefined functions or can combine them with its own. Each function is implemented as parallel for-each loop, which is than translated by the framework into a series of MapReduce jobs. During the translation, the framework itself decides if such a function should be executed locally (i.e. sequentially) or remotely (i.e. parallel). Apache Crunch is a freely available open source implementation of FlumeJava.

III. Cloudflow

The overall idea behind Cloudflow is to simplify the creation of analysis pipelines by encapsulating complex data analysis steps in simple operations. This approach helps hide the complexity and the implementation details of complex data-parallel pipelines. Moreover, the concept of using basic operations increases the reusability and enables the testing of the operation logic on a local workstation by using existing unit testing frameworks.

Since Cloudflow uses the Hadoop framework for pipeline execution, it offers parallel data processing, data reliability and fault tolerance out of the box. This fact is especially important in the field of Cloud Computing, where infrastructure often relies on commodity hardware and nodes can fail on a regular basis (e.g. due to miss-configuration or hardware failures). At the same time, the architecture of Cloudflow is independent from MapReduce; it provides parallelization constructs and abstraction interfaces that can be used to extend the system by implementing other parallel programming models in the future (e.g. translating the operations into Apache Spark jobs).

Instead of developing a new declarative language for the pipeline composition, we developd a clear Java API. The proposed framework implements different patterns to speed up the pipeline creation, to be extensible and to support test-driven pipeline development. The following

section gives an overview on the abstraction, explains the basic operations in detail, and shows how pipelines are created.

A. Data Types and Basic Operations

Cloudflow operates on records consisting of a key/value pair, whereby different record types are available (e.g. TextRecord, IntegerRecord, FastqRecord). A *loader* class is responsible to load the input data and to convert it into an appropriate record type. As mentioned earlier, Cloudflow supports three different basic operations. These are used to analyze and transform records; specifically, *transform*, *summarize*, and *group* operations exist.

The *transform*-operation is used to analyze one input record and to create between *0* and *n* output records. The user implements the computational logic for this operation by extending an abstract class. This class provides a simple function, which is executed by the Cloudflow framework for all input records in parallel:

```
Class MyTransformer extends Transformer {
    public void transform(Record) {
       doSomethingInParallel();
       emit(new Record());
    }
}
```

The *summarizer* operates on a list of records, whereby records with the similar key are grouped. Thus, the signature of the process method has the key and a list of records as an input:

```
Class MySummarizer extends Summarizer {
   public void summarize(Key, List<Record>) {
       doSomethingInParallel();
       emit(new Record());
    }
}
```

The *group*-operation is a special operation, which takes a list of records as an input and creates a *Record* group with the same key. Our framework inserts automatically a group-operation between a transform- and a summarize-operation. This ensures, that output records of the *transform* operation are compatible with the input records of the *summarize* operation. The *group*-operation is realized by using the shuffle phase of a MapReduce job.

Based on these three operations, the user defines pipelines by building sequences of operations. A pipeline has to start with a transform-operation, all further operations are optional and can be used in arbitrary order.

B. Extended Operations

Complex operations are built by combing one or more basic operations. We already implemented several standard operations that are helpful for the analysis of text or numerical data records:

- *Filter*: this operation is a special *transform*-operation, which emits the record to the subsequent operation iff a user-defined condition is fulfilled.
- *Split*: the *transform*-operation calculates for each input record a new split level (i.e. a new key). This key is used by the group-by operation to create chunks, which can then be analyzed by a user-defined summarize-operation.
- Aggregation (sum, mean)*:* This defines a *group-by*-operation followed by a *summarize*-operation to aggregate all values with the same key (e.g. calculates the mean of all values). One record with the aggregated value is then emitted to the subsequent operation.
- *Executor:* this *summarize*-operation writes all grouped values into a file on the local disk. It is then used as the input of an external UNIX command line program. Based on the lines of the output file, new records are created and emitted to the subsequent operation.

Since Cloudflow's operations are based on the Composite pattern, all these extended operations can also be used as a basis for new operations. In addition, this enables to split complex operations into several sub-operations, which improves testing and maintenance.

C. Pipeline Composition

The user builds pipelines by connecting several operations with compatible interfaces. For this purpose our framework implements the Builder pattern, which enables (a) building complex pipelines, (b) providing type safety and (c) the implementation of domains specific builders (see Section III.D). In addition, the Builder pattern ensures that only a valid sequence of operations can be created (i.e. after the group-by operation a summarize operation has to be added).

```
Class LineToWords extends Transformer {
   public void transform(TextRecord rec) {
    String[] words = rec.getValue().split()
    for (String word: words){
      emit(new IntegerRecord(word, 1));
    }
   }
}

pipeline.loadText(input)
        .transform(LineToWords.class)
        .sum()
        .save(output);
```

Listing 1. WordCount Example using Cloudflow

To help the user and accelerate the pipeline composition process, Cloudflow provides already a set of useful operations. This has the advantage that even a default WordCount example can be broken down into a few simple operations and is defined in a single line of code (see Listing 1). In a first step, the text file is loaded from HDFS (*loadText*). Then, for each record (i.e. line of input) the application-specific *LineToWords* operation is executed, which splits the line into words and creates a new record efor ach word. In the last step a predefined *sum* operation is executed. It extends the pipeline by a *group-by* operation and a *summarize*-operation in order to sum up all the values for a certain key.

For frequently used operations (e.g. sum, mean or count), we created special builder functions, which extend the pipeline and improve the code readability by keeping the code simple.

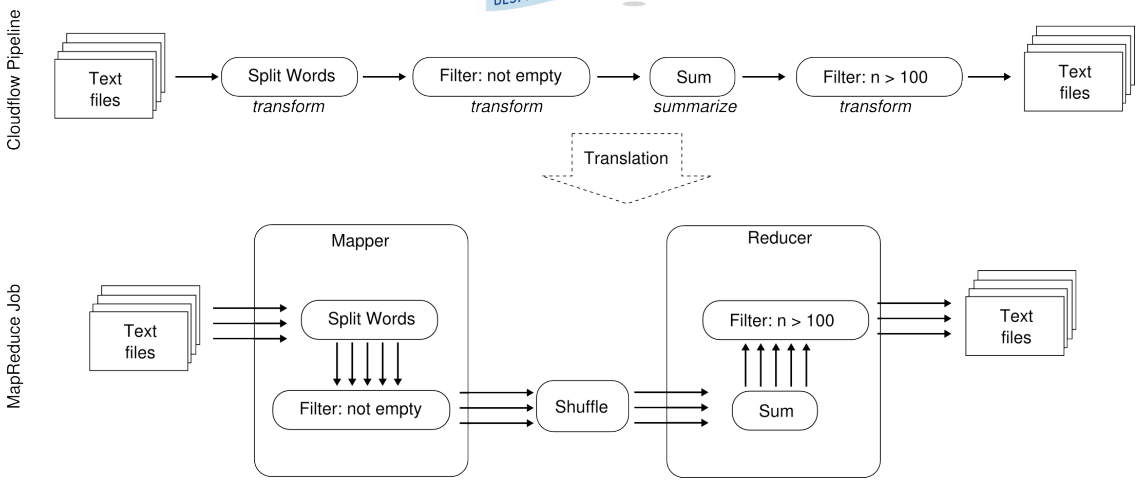


Figure 1. Cloudflow translates the operation sequence automatically into an executable MapReduce job

D. *Pipeline Execution*

Before the execution, Cloudflow checks the compatibility of input and output records of consecutive operations. This ensures that only valid and executable pipelines are submitted to a Hadoop cluster.

If the pipeline is executable and valid, then the operation sequence is translated into an execution plan, that decides if an operation is executed in the *map* or in the *reduce* phase. Based on this plan, Cloudflow creates one or more MapReduce jobs and configures them to execute the user-defined operations in the correct order. In this translation step, Cloudflow tries to minimize the number of MapReduce jobs by combining consecutive *transform*-operations and by executing all *transform*-operations after a *summarize*-operation in the same reducer instance (see Figure 1).

For additive *summarize*-operations (e.g. sum), Cloudflow takes advantage of Hadoop's combiner functionality. The idea of this improvement is to combine the key/value pairs that are generated by all map tasks on the same machine, into fewer pairs. Thus, the number of pairs that are transferred between mapper and reducer are minimized, which results in a positive effect on the network bandwidth since useless communication is avoided.

IV. CLOUDFLOW FOR BIOINFORMATICS

Cloudflow provides a variety of already implemented utilities, which facilitate the creation of pipelines in the field of Bioinformatics (especially for NGS data in Genetics). For that purpose, we implemented, based on HadoopBAM [7], several record types and loader classes in order to process FASTQ, BAM and VCF files. Moreover, we created several operations and filters for the analysis of biological datasets (see Table I for an overview of all currently implemented operations and filters).

For example, a typical quality control pipeline for VCF files can be implemented by simple combining of several built-in operations. First, we apply predefined filters to discard variations that are monomorphic, marked as duplicates, or are Insertions or Deletions (InDels). For all records passing the filters, Cloudflow applies a *summarize*-operation that calculates the call rate for each variation (see Listing 2).

```
Class CallRateCalc extends Transformer {
  public void transform(VcfRecord record) {
    VariantContext snp = record.getValue();
    float call = callRate(snp);
    emit(new FloatRecord(snp.getID(), call);
  }
}

pipeline.loadVCF(input)
        .filter(MonomorphicFilter.class)
        .filter(DuplicateFilter.class)
        .filter(InDelFilter.class)
        .transform(CallRateCalc.class)
        .save(output);
```

Listing 2. VCF Quality Control Pipeline using Cloudflow

V. DEPLOYING PIPELINES AS A SERVICE

Cloudgene [6] is a web-based platform to create and execute workflows consisting of Hadoop MapReduce, Apache Pig and command line-based programs. It can be seen as an additional layer between Hadoop MapReduce and the end user that hides the complexity of the MapReduce framework. Therefore, Cloudgene is the perfect candidate to provide Cloudflow pipelines as a service. Such pipeline can be integrated into the workflow platform by utilizing Cloudgene's plugin interface. No adaptation to the source code is needed, while only a simple plain text file including a header, input parameters, output parameters and the definition of the workflow itself need to be created. When launching Cloudgene, the manifest file is loaded and the client interface is automatically rendered using information from the file. As Cloudgene supports different technologies, it is possible to parallelize the calculations using Cloudflow and to visualize the results using R.

Cloudflow requires a compatible MapReduce cluster for executing pipelines. CloudMan [8] makes it possible to easily procure and configure a functional data analysis platform on a cloud infrastructure. The procured platform delivers a scalable cluster-in-the-cloud and a data analysis environment preconfigured with a number of applications. With its ability to be launched and managed via a web browser on a number of clouds, customized as necessary, and easily shared with collaborators, CloudMan makes it

TABLE I. CURRENTLY SUPPORTED DATA FORMATS AND OPERATIONS

Data Format		Pipeline Operation	Description
Fastq	Split	`split()`	Find pairs (for paired-end reads)
	Filter	`filter(LowQualityReads.class)`	Filters reads by quality
		`filter(SequenceLength.class)`	Filters reads by sequence length
	Other	`findPairedReads()`	Detects read pairs
		`align(referenceSequence)`	Aligns sequences against a reference (using jBWA for alignment)
BAM	Split	`split()`	Creates fixed size chunks (e.g. 64 MB)
		`split(5, BamChunk.MBASES)`	Creates logical chunks (e.g. 5MBases)
	Filter	`filter(UnmappedReads.class)`	Filters unmapped reads
		`filter(LowQualityReads.class)`	Filters reads by map.quality
	Other	`findVariations()`	Finds variations in aligned reads (using samtools)
VCF	Split	`split()`	Creates fixed size chunks (e.g. 64 MB)
		`split(5, VcfChunk.MBASES)`	Creates logical chunks (e.g. 5MBases)
	Filter	`filter(MonomorphicFilter.class)`	Filters monomorphic site
		`filter(DuplicateFilter.class)`	Filters duplicates
		`filter(InDelFilter.class)`	Filters inDels
		`filter(CallRateFilter.class)`	Filters by call rate
		`filter(MafFilter.class)`	Filters by MAF
	Other	`checkAlleleFreq(reference)`	Allele frequency check with external reference (e.g. 1000 genomes)

possible to readily utilize cloud resources in a research environment. The approach on how Cloudgene and CloudMan can be combined efficiently have already been demonstrated [9].

VI. EVALUATION

To evaluate our approach, we implemented three different Bioinformatics data-analysis pipelines using Cloudflow and integrated them into Cloudgene. The results of our experiments demonstrate that Cloudflow has only a minimal overhead in the execution time compared to an identical pure MapReduce implementation. However, the performance of Cloudflow is better than Apache Crunch's (see Figure 2).

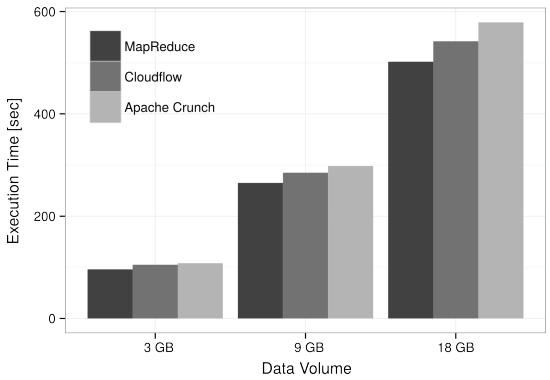


Figure 2. Execution Time of a Cloudflow pipeline, an Apache Crunch pipeline and a pure MapReduce implementation of WordCount

A. Preprocessing and Mapping

When working with NGS data, the quality of raw data needs to be checked before a successful subsequent downstream analysis (e.g. read mapping/alignment) can be achieved. The overall goal of alignment is to dock the vast amount of short reads, mostly in the FASTQ file format, to a reference genome. Factors such as read errors, insertions or deletions of bases must be considered that finally results in the most accurate genome position for each read.

The goal of the following pipeline is the alignment of paired-end data to a reference genome. Therefore, the FASTQ data is loaded using the FastqLoader, which creates records for each sequence. The records are then filtered by an average base quality of 30. Numerous other quality metrics (such as sequence length or C/G content) can be filtered as well. Since paired-end reads are used, read pairs are detected using a predefined *transform*-operation. The aligner step is implemented as a *summarize*-operation that calls a parallelized version of BWA-MEM [10]. This has been achieved by using JNIs (https://github.com/lindenb/jbwa). Similar to BWA-MEM 99,100 reads are aligned to the user-specified reference genome in one batch. Aligned reads are saved in HDFS in the BAM file format (see Listing 3.A).

B. Variation Calling

After data has been aligned and cleaned (e.g. removing duplicates, quality recalibration), the next step of NGS pipelines is the detection of reliable variants that can be used e.g. in association studies. A widely used pipeline for variant detection is GotCloud, developed at the Center of Statistical Genetics (University of Michigan) and utilized in the 1000 Genomes (1000G) project.

In this example pipeline, the aim is to find variations without a statistical model by implementing a simple counting approach of the four bases. This is only possible when using high coverage data. Therefore, in the first step the BAM file (created in the previous pipeline) is loaded and chunked in user-specified splits (e.g. 5 MBases). Variations are then detected for each chunk by counting

the occurrences of A, C, G, T on each position. Finally, the detected variations are stored in VCF files (see Listing 3.B).

C. Genome-Wide Association Studies

Many genome-wide association studies (GWAS) have identified associations between various phenotypes and common sequence polymorphisms, which might play a role for disease development. Technologies like microarrays have made it possible to measure millions of single nucleotide polymorphisms (SNPs) of one individual simultaneously and for low cost. Since the costs of microarrays is much lower than next-generation sequencing (NGS), it is today the cheapest method to genotype large-scale population studies. Such datasets are combined with collected phenotypes (e.g. diseases and measured data) in order to detect if one of these variations has a high impact on the value of a phenotype. Since the size of such datasets grows rapidly, a parallelization at the data-level is necessary to analyze the data in appropriate time.

```
//A. Preprocessing and Mapping
pipeline.loadFastq(input)
        .filter(LowQualityReads.class, 30)
        .findPairedReads()
        .align(refSeq)
        .save(output);

//B. Variation Calling
pipeline.loadBam(input)
        .split(5, ChunkSize.MBASES)
        .groupByKey()
        .findVariations(refSeq)
        .save(output);

//C. Genome-Wide Association Study
pipeline.loadText(input)
        .split(1000, ChunkSize.LINES)
        .execute(SnpTestExecutor.class).
        .filter(FilterHeader.class)
        .filter(FilterInvalidSnps.class)
        .save(output);
```

Listing 3. Complete NGS pipeline using Cloudflow

The parallelization of the association analysis was realized by splitting the list of markers into chunks. In detail, the mapper splits all input SNPs into chunks with a fixed number of SNPs (e.g. 1000). Then, the reducer executes the linear regression model for each chunk by using SNPTest. Finally, the reducer collects the results and merges them into a single file. The corresponding Cloudflow pipeline loads the text input file and automatically creates records for each line. On these records we apply the *split* operation, which creates chunks containing a fixed number of lines. For the execution of the SNPTest program, we can implement a special operation called *BinaryExecutor,* which enables us to write the chunks automatically to the POSIX file system. In the next step, we can use this file as the input file for SNPTest. After the execution the operation creates text records for each line of results (see Listing 3.C).

VII. Conclusion

Cloudflow's overall aim is to simplify the development of complex MapReduce pipelines by abstracting the map and the reduce function from end users. Therefore, operations need only be written once and can be re-used for future MapReduce usage. The major advantage of Cloudflow lies in the provision of validated operations, especially in the area of genetics, and its extensibility. Combining Cloudflow with CloudMan (cluster orchestration) and Cloudgene (Hadoop workflow system) allows users to use Hadoop without a deeper knowledge of the internal MapReduce concepts and could yield to a boost of Hadoop in genetics.

Acknowledgment

This work was, in part, supported by the "Scalable Big Data Bioinformatics Analysis in the Cloud" grant from the Croatian Ministry of Science, Education, and Sport and the Austrian Federal Ministry of Science and Research (BMWF) and by the FP7-PEOPLE programme grant 277144 (AIS-DC).

References

[1] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, pp. 255–260, 2013.

[2] S. Yazar, G. E. C. Gooden, D. A. Mackey, and A. W. Hewitt, "Benchmarking undedicated cloud computing providers for analysis of genomic datasets.," *PLoS One*, vol. 9, no. 9, p. e108490, Jan. 2014.

[3] A. Schumacher, L. Pireddu, M. Niemenmaa, A. Kallio, E. Korpelainen, G. Zanetti, and K. Heljanko, "SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop.," *Bioinformatics*, vol. 30, no. 1, pp. 119–20, Jan. 2014.

[4] H. Nordberg, K. Bhatia, K. Wang, and Z. Wang, "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data.," *Bioinformatics*, p. btt528–, Oct. 2013.

[5] C. Chambers, A. Raniwala, F. Perry, S. Adams, R. R. Henry, R. Bradshaw, and N. Weizenbaum, "FlumeJava," *ACM SIGPLAN Not.*, vol. 45, no. 6, p. 363, May 2010.

[6] S. Schönherr, L. Forer, H. Weißensteiner, F. Kronenberg, G. Specht, and A. Kloss-Brandstätter, "Cloudgene: A graphical execution platform for MapReduce programs on private and public clouds," *BMC Bioinformatics*, vol. 13, no. 1, p. 200, 2012.

[7] M. Niemenmaa, A. Kallio, A. Schumacher, P. Klemelä, E. Korpelainen, and K. Heljanko, "Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud.," *Bioinformatics*, p. bts054–, Feb. 2012.

[8] E. Afgan, B. Chapman, and J. Taylor, "CloudMan as a platform for tool, data, and analysis distribution.," *BMC Bioinformatics*, vol. 13, no. 1, p. 315, Jan. 2012.

[9] L. Forer, T. Lipic, S. Schonherr, H. Weisensteiner, D. Davidovic, F. Kronenberg, and E. Afgan, "Delivering bioinformatics MapReduce applications in the cloud," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, 2014, pp. 373–377.

[10] H. Li, "Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM," vol. 00, no. 00, pp. 1–3, 2013.

Science Gateway for Distributed Multiscale Course Management in e-Science and e-Learning — Use Case for Study and Investigation of Functionalized Nanomaterials

Yu. Gordienko[*,**], S. Stirenko[**], O. Gatsenko[*] and L. Bekenov[*]
[*] G.V.Kurdyumov Institute for Metal Physics, National Academy of Sciences, Kyiv, Ukraine
[**] National Technical University of Ukraine "KPI"/ High Performance Computing Center, Kyiv, Ukraine
e-mail: gord@imp.kiev.ua

Abstract - The current tendency of human learning and teaching is targeted to development and integration of digital technologies (like cloud solutions, mobile technology, learning analytics, big data, augmented reality, natural interaction technologies, etc.). Our Science Gateway (http://scigate.imp.kiev.ua) in collaboration with High Performance Computing Center (http://hpcc.kpi.ua) is aimed on the close cooperation among the main actors in learning and researching world (teachers, students, scientists, supporting personnel, volunteers, etc.) with industry and academia to propose the new frameworks and interoperability requirements for the building blocks of a digital ecosystem for learning (including informal learning) that develops and integrates the current and new tools and systems. It is the portal for management of distributed courses (workflows), tools, resources, and users, which is constructed on the basis of the Liferay framework and gUSE/WS-PGRADE technology. It is based on development of multi-level approach (as to methods/algorithms) for effective study and research through flexible selection and combination of unified modules ("gaming" with modules as with LEGO-bricks). It allows us to provide the flexible and adjustable framework with direct involvement in real-world and scientific use cases motivated by the educational aims of students and real scientific aims in labs.

I. INTRODUCTION

The current tendency of human learning and teaching is targeted to development and integration of digital technologies like cloud solutions, mobile technology, learning analytics, big data, augmented reality, natural interaction technologies, etc. But the available e-Science and e-Learning frameworks (for example, in Massive Open Online Courses (MOOCs) [1] and OpenCourseWare [2] initiatives, like TIMMS [3], SlideWiki [4], Wikiversity [5], edX [6], Coursera [7], et al.) are: (a) very heterogeneous (the available learning content is quite different in various institutions), (b) often far away from the real-life learning situations, (c) in general, static like stationary web-pages with online access from desktop PCs only, (d) not flexible to the very volatile demands of end users (students, pupils, ordinary people) and real-world, (e) not-adjustable by complexity, duration, and range (e.g. for taking some modules, but not the whole course), especially in the life-long learning or vocational training, (f) the feedback is not available about the quality, value, and necessity of some parts of the course or combinations of some modules.

The work presented here was partially funded by EU FP7 SCI-BUS (SCIentific gateway Based User Support) project, No. RI-283481, EU TEMPUS LeAGUe (A Network for Developing Lifelong Learning in Armenia, Georgia and Ukraine) project No. 543839-TEMPUS-1-2013-1-SE-TEMPUS-SMHES, and partially supported in the framework of the State Targeted Scientific and Technical Programs in 2013-2015.

The paper is aimed on the description of the approach to close cooperation among the main actors in learning world (teachers and students) with industry and academia. Here, the new frameworks and interoperability requirements are considered for the building blocks of a digital ecosystem for learning (including informal learning) that develops and integrates the current and new tools and systems.

II. BACKGROUND

A vast amount of e-Science and e-Learning frameworks briefly enlisted in "Introduction" is published online to make their educational content accessible to a wide range of audiences. Many profound researches were dedicated to their estimation in the view of awareness of these courses among users and increase of popularity of these frameworks [7-9]. However, from a subjective experience of many users, online courses are mainly dull, outdated, and/or non-attractable. As it was observed in one of the researches [8]: ~28% of the courses are indeed open, ~12% are available in a language other than English, ~16% are available in a format facilitating reuse and repurposeability, ~30% were updated in the last 3 years, and <50% comprise self-assessment questions, only ~65% have at least one example and one illustration, ~60% have been objectively determined as low attractive, and only 10% have been categorized as highly attractive. In the context of this investigation (on 100 various online courses), the immediate conclusion is that the quality of these e-Learning courses and frameworks should be improved significantly, especially in the sense of better attractiveness, feedback (more interactive examples) and applicability (more practical use cases).

A possible solution for improving the quality can be found by leveraging the collaboration and sharing experience and e-Learning courses/frameworks. Several platforms (like SlideWiki [4] or Wikiversity [5]) were mentioned in "Introduction", which can be used to support the collaborative creation and usage of e-Learning courses by a wide range of authors (teachers, developers, providers) and users (students, ordinary people) communities.

In addition to this, the big potential of the recent "science gateway" (SG) ideology can be used [10]. In general, SG means a user-friendly interface between users (for example, scientists or scientific communities), various software tools and hardware resources, usually distributed computing infrastructures (DCIs). The aim SGs is to decrease the time and effort spent in setting up the environment and DCI for a specific scientific application. By means of SGs researchers can concentrate their efforts on their scientific goals and thus spending less time and effort on assembling and managing the required DCI components. The most important aspects of SG are: a simplified intuitive graphical user interface (GUI) that is highly tailored to the needs of the given scientific community; a smooth access to national and international computing and storage resources; collaborative tools for sharing scientific data. Nowadays various SG frameworks exist (for example, ASKALON [11], KNIME [12], MOTEUR [13], gUSE/WS-PGRADE [14]), which use different enabling components and technologies: web application containers (Tomcat, Glassfish, etc.), portal or web application frameworks (Liferay, Spring, Drupal, etc.), database management systems (MySQL, etc.), and workflow management systems. Here, our SG for distributed multiscale course management in e-Science and e-Learning, which is based on gUSE/WS-PGRADE framework, is described. The detailed description of SG ideology on the basis of gUSE/WS-PGRADE (https://guse.sztaki.hu) can be found elsewhere [14]. In short, WS-PGRADE portal is a web based front end of the gUSE infrastructure. gUSE (grid and cloud User Support Environment) is a framework that provides the user-friendly access to DCI resources. SG contains the workflow-oriented graphical user interface (GUI) to create and manage workflows for tasks executed on various DCI resources. The DCI resources can include PCs, clusters, service grids, desktop grids, clouds [15-17].

III. SCIENCE GATEWAY FOR COURSE MANAGEMENT

The aim of the approach is to increase attractiveness of e-Learning courses by direct involvement in design of learning process, feedback by more interactive examples and applicability by more practical use cases. It could be performed by development of multi-level approach (methods/algorithms) for effective learning through flexible selection and combination of unified modules ("gaming" with modules as with LEGO-bricks) under SG. This can provide the flexible and adjustable framework with the elements of direct involvement in real-world and scientific use cases. It should be motivated by the educational aims of students and real scientific aims in labs [16].

A. *Objectives*

The objectives of the approach are to design, create and implement:

- smart learning environments providing students with adaptive and personalized learning and assessment through multi-modal/multi-sensory interaction technologies and advanced interfaces;
- interactive learning systems with more efficient and natural ways of delivering content to users especially for the real scientific tasks;
- technologies to match and operate multiple applications in real environments, while research is expected to be based on freely available and re-usable tools and resources.

B. *General Structure and Main Components*

Due to SG technology several sets of modules can be linked together by the end users (teachers or students) along the desirable track, which can form the course customized by the end users, for example, for molecular dynamics (MD) simulations in computational physics (Fig. 1). To support such linking some basic requirements should be formulated and satisfied for wrapping and connecting the modules. They should be considered as meta-information (Table 1), which are obligatory for successful combination of various modules (mainly for their compatibility). In addition to this, the similar vocabulary should be used among designers and creators of the modules, which is the simple task for the developers from the same institution, but could be headache for combination of modules from various institutions.

The main components of the approach are: portal(s), portlets(s), content, tools, users, and their repositories. "Portal" is assumed to be collection of components like tools, users, and courses (workflows). The courses could be installed as the separate portlets. It is the technological backbone constructed on the basis of the Liferay framework with customizable scientific gateways (gUSE/WS-PGRADE) based on workflows. "Content" is a set of available "legacy" passive modules and new

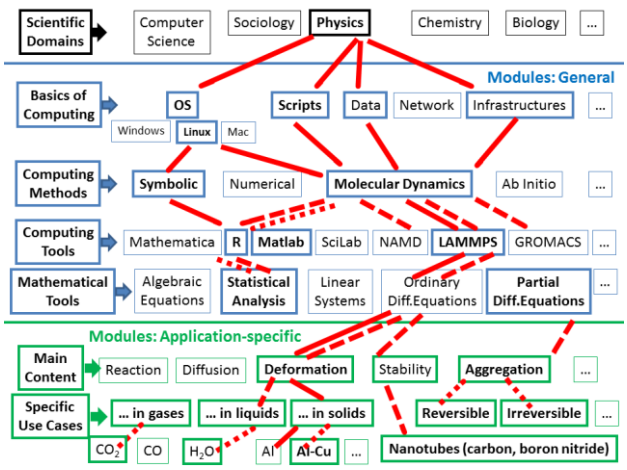


Figure 1. Example of the multiscale course hierarchy in materials science: beginner ("macro") tracks (solid lines), advanced ("mini") tracks (dash lines), expert ("micro" and "nano") tracks (dot lines)

active nano-, micro-, mini-, and macro- modules. The legacy modules should be "wrapped" by standardized meta-information about them to facilitate their combination and construction of workflows (customized tools). The legacy passive modules are assumed to be some sets of web-pages (with multimedia content), which can be enclosed in the portlets and attached to portals. The new active modules are based on the so-called "Virtual Labs" [16], i.e. software workflows (i.e. domain-specific applications connected to some research infrastructures) and enclosed in portlets, which can be attached to portals also. "Repositories" are considered as collections of content in the shape of multiscale modules (like macro-, mini-, micro-, and nano- courses) in the shape of workflows and portlets, which can be uploaded/downloaded and attached to portals.

C. *Meta-information*

The compatible meta-information for modules (Table 1) will be necessary for successful combination and compatibility of various modules. The information about "Previous module(s)" is considered as an input socket, i.e. a list of the available modules, which are pre-requisites (obligatory) to use (learn) this module. In similar way "Next module(s)" plays the role of an output socket, i.e. a list of the available modules, which can be used as continuation of this module. "Alternatives" is a list of the available modules, which can be selected as alternative versions of this module. "Categories" describes the place of the current module in the hierarchy of learning disciplines. "Complexity" is a subjective measure of the relative efforts that should be applied by students (it could be used to estimate grades, points, costs, and learning curve). "Scale" is a relative level of the module, i.e. macro-, mini-, micro-, or nano- level. "Duration" is a unit of the module length, which can be measured by time (hours, days, etc.) or lessons. "Workload" is a subjective estimation of time needed for studying the proposed content. "Exercises" are some questions and tasks with answers. "Keywords" are tags, which can be used for search of this module. "Languages" is a measure of universality of the content, i.e. multilingual content is desirable, but English should be obligatory for universal compatibility. Addition of any other languages will increase usability at a local geographical scale. "Rating" is a subjective estimation of the module attractiveness and usefulness, i.e. its quality. "Certificate" is a proof of any formal recognition marks. "Price" is a possible cost for this module (which is usually related with certificate costs, tuition fee, etc.)

TABLE I. EXAMPLE OF META-INFORMATION

Category	***Format***	***Example***
Title	String	MD Simulation of Metal Nanocrystals under Deformation
Previous module(s)	List of strings	MD Simulation of Metal Nanocrystals
Next module(s)	List of strings	MD Simulation of Defect Evolution in Al-Cu Alloys with Nanoinclusions
Alternatives	List of strings	MD Simulation of Non-metal Solids
Categories	List of strings	Physics:Computational Physics
Complexity	Number (1-5)	4
Scale	String from the set (macro, mini, micro, nano)	Mini
Duration	Number (in weeks)	2 (weeks)
Workload	Number (per week)	8-10 (hours/week)
Exercises	Number	5
Keywords	List of string	MD, metal, alloy, Al-Cu,
Languages	List of string	English, Ukrainian
Rating	1-5 stars	4
Certificate	String (Yes/No)	No
Price	Number (in $)	0

IV. USE CASES FOR STUDY AND INVESTIGATION OF FUNCTIONALIZED NANOMATERIALS

Here we demonstrate the capabilities of this approach applied for MD simulations and further data post-processing on the basis of LAMMPS package for MD simulation [18], and other packages like R for statistical analysis [19], Pizza.py Toolkit for manipulations with atomic coordinates files [20], AtomEye for visualization of atomic configurations [21], debyer for simulation of X-ray diffraction (XRD), and neutron diffraction (ND) analysis (https://code.google.com/p/debyer/), etc.

These use cases (Fig. 2) were targeted on MD simulation of nanocrystals under uniaxial tension (Fig. 3a) and functionalized nanomaterials like carbon nanotubes (Fig. 4a). Their user communities in the field of nanotechnologies include the user groups (teachers and students) from academy (G.V.Kurdyumov Institute for Metal Physics, National Academy of Sciences) and education (National Technical University of Ukraine "KPI"). MD simulation workflow (Fig. 2) was created and used for investigation of relaxation behavior of nanocrystals for different nanocrystals (Al and Al-Cu), physical conditions (tensile rate, size and orientation of the nanocrystals) and methodological parameters

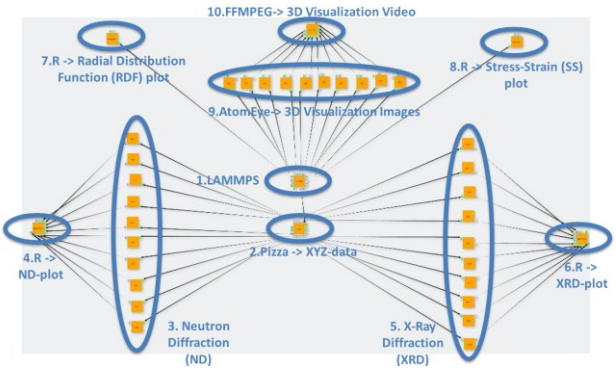


Figure 2. Example of workflow for MD simulations of functionalized nanomaterials. It is created by means of WS-PGRADE package: tools (yellow crates) inside the modules (thick ovals) with connecting links (thin lines)

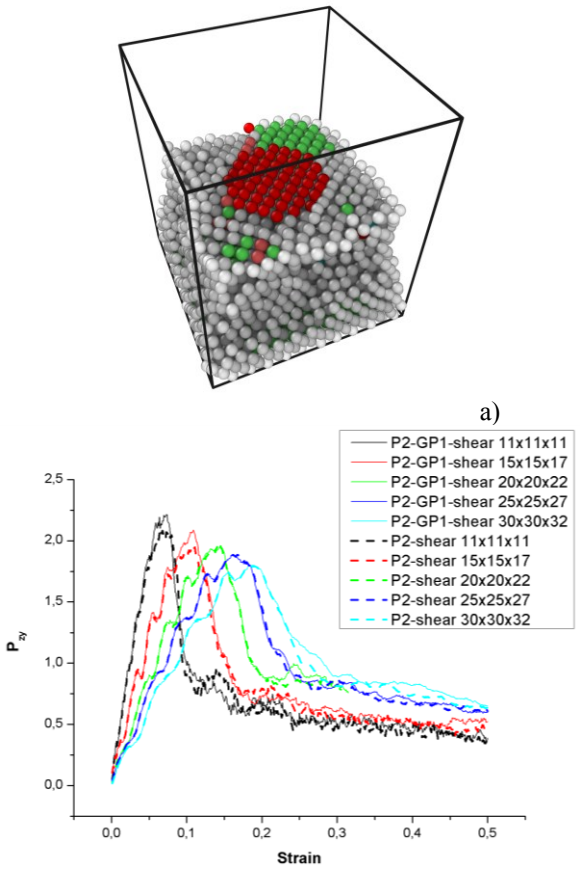


Figure 3. Examples of output data for MD simulations of functionalized nanomaterials: (a) 3D visualization by AtomEye: cross-section of nanocrystal with Cu-nanodisk (Guinier-Preston zone noted by red color), (b) post-processing by R environment: stress-strain plot

(potential type, boundary conditions, and others). The actually used workflow (Fig. 2) includes several subset workflows, which can be considered as separate tracks on the basis of separate modules (crates in SG terminology). The simulations included the following subset workflows: 1) LAMMPS + R-package — for MD simulations and plotting some post-processed values, like stress-strain dependencies; 2) LAMMPS + R-package + AtomEye + FFMPEG — for the same operations + visualization of intermediate states of atoms and final video visualization; 3) LAMMPS + R-package + AtomEye + FFMPEG + debyer — for the same operations + processing data for X-ray and neutron diffraction and their plotting, like XRD-plots. From the physical point of view it was shown that nanocrystal size and complex nanoscale defect substructure influence the strength properties of Al-Cu alloys (Fig. 3b). These results confirm and illustrate the known experimental data described in [22-24].

The potential for variability of this approach can be illustrated by possibility to change input/output scripts (for LAMMPS, AtomEye, and R crates-components) and parameters (for LAMMPS, AtomEye, FFMPEG, and debyer crates). By change of these scripts and parameters the workflow can be applied for the quite different use case, namely, for MD simulations of the nanosubstrates functionalized by carbon nanotubes (Fig. 4a).

As it schematically shown in Fig. 1 the general backbone structure (Fig. 2) can be used in the shape of some subsets for various scientific domains (Physics, Materials Science, Chemistry, Biology, etc.). These subsets can include various sets of modules (user can select the modules connected and emphasized by red color).

Among general modules, for example, “Basics of computing” set could include the crates with some executables dedicated to various computing aspects (basic operations in Linux, Windows, Shell Scripts, Data Operations, Network Basics, ...). “Computing Tools” set can comprise the correspondent binaries (Octave, R, SciLab, Matlab, Mathematica) and input/output scripts for them. “Mathematical Tools” can include static or dynamic content dedicated to some basic and more complex mathematical approaches (like Algebraic Equations, Linear Systems, Ordinary Differential Equations, Partial Differential Equations, etc.). The similar content can be in included “Computing Methods” set (like Symbolic Computation, Numerical Computation, Molecular Dynamics, Ab Initio).

Among application-specific modules, for example, “Main Content” can be related with the specific processes: physical (Diffusion, Aggregation, Reaction, Coalescence, etc.), mathematical (Random Walk, Fractals, etc.), biological, financial, geological, astronomic, and others.

Finally, “Specific Use Cases” can be concentrated on the more practical examples, which are illustrated in this section (Fig. 2 and 3).

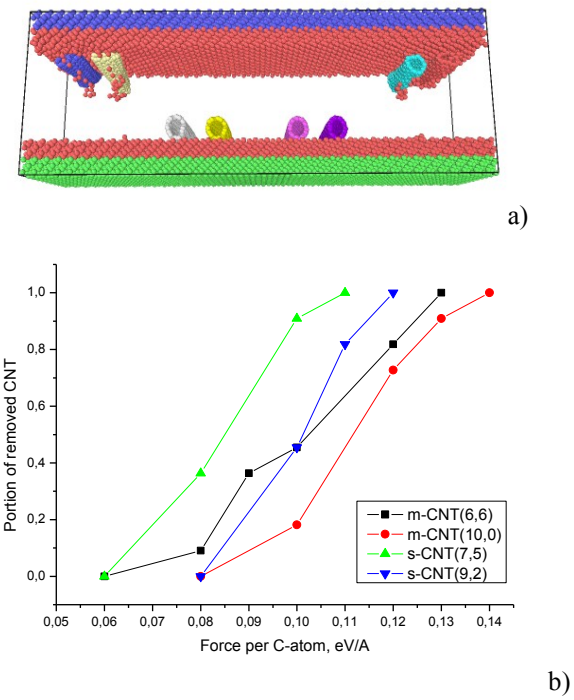


Figure 4. Examples of output data for MD simulations of the nanosubstrates functionalized by carbon nanotubes: (a) 3D visualization by AtomEye: nanotubes with various adhesion properties, (b) post-processing by R environment: the portion of removed nanotubes vs. the force applied to the nanotube

TABLE II. EXAMPLES OF SCALE LEVELS IN THE USE CASE OF MOLECULAR DYNAMICS SIMULATIONS OF NANOMATERIALS

Level	*Duration*	*Student Work*	*Topic*
Nano (skill)	10-30 min.	Input values	Basics of simulation (LAMMPS, NAMD, GROMACS, debyer), visualization (AtomEye, Ovito), simple plotting (R, ipython)
Micro (labwork)	1-8 hours	Input script	Scenarios of MD simulations (LAMMPS, NAMD, GROMACS), statistical processing (R), complex plotting (R, ipython)
Mini (module)	1-14 days	Set of scripts + parameters	Simple experiment (set of modules linked by scripts)
Macro (course)	1-6 months	Tree of scripts + parameters	Complex experiment, even up to diploma work

The additional way for variability of this approach can be illustrated by possibility to change the scale of the learning track. It is schematically depicted by connecting lines between modules in Fig. 1. The various scale levels are illustrated in Table 2 for the use case of molecular dynamics simulations of functionalized nanomaterials. All of these scale levels are actually constructed around various numbers of crates (SG components) and connecting links between them, which can be easily arranged and tuned in SG framework [16]. Our SG (http://scigate.imp.kiev.ua) in collaboration with High Performance Computing Center (http://hpcc.kpi.ua) was aimed on the close cooperation among the main actors in learning and researching world (teachers, students, scientists, supporting personnel, volunteers, etc.) with industry and academia to propose the new frameworks and interoperability requirements for the building blocks of a digital ecosystem for learning (including formal, non-formal, and informal learning) that develops and integrates the current and new tools and systems. This multi-level approach improves the effective study and research through flexible selection and combination of unified modules.

V. CONCLUSIONS

Usage of SG ideology for development and integration of real-world and user targeted digital technologies for learning will foster the academia to use them and stimulate the market for innovations in educational technologies. Industrial actors can create "supermarkets" of modules and courses, created by academia or in cooperation with academia, and provide "menus" to obtain the customized learning "meals" (learning courses) created by teachers or/and users. It could boost the formal learning (schools, universities, etc.), the life-long and vocational fields of learning, and, especially, in-formal learning.

The novelties and advantages of the proposed paradigm (in comparison to the numerous available online learning solutions) consist of:

- homogenization and smoother integration of the very heterogeneous available learning content from various institutions (by the standard requirements for the new modules and the wrappers of legacy modules),
- close approach to the real-life learning situations,
- dynamic content, i.e. some modules in the shape of Virtual Labs around the simulations or remote experimentations (if the latter will be found among the potential partners),
- flexible combinations of various modules on demands of end users,
- modules with various versions differed by complexity (beginner, advanced, professional, etc.), scale (duration) like "nano" (10-30 min), "micro" (1-8 hours), "mini" (1-14 days), "macro" (1-6 months), and range,
- "billing" system around modules to provide the feedback for teachers about the quality, value, and necessity of some parts of the course or combinations of some modules,
- open or commercial repositories of the "bricks": modules, Virtual Labs, remote experimentations,
- portal(s) for content, tools, and user management.

This flexible and multiscale approach can help to implement the paradigm of ubiquitous learning (which development is usually restricted in time and space), and it can be feasible and efficiently implemented in the modern education system.

ACKNOWLEDGMENT

The authors would like to thank Prof. P.Kacsuk, D.G.Iglesia and A.Pester for fruitful discussions of this approach and valuable critics.

REFERENCES

[1] H.Macleod, J.Haywood, A.Woodgate, M.Alkhatnai, "Emerging patterns in MOOCs: Learners, course designs and directions", TechTrends, 59(1) pp. 56-63, 2015.

[2] P.D.Long, "OpenCourseWare: Simple Idea, Profound Implications", Syllabus, 15(6) pp. 12, 2002.

[3] Tubingen Internet Multimedia Server (TIMMS) (http://timms.uni-tuebingen.de/archive/sose99.aspx) 1999.

[4] A.Khalili, S.Auer, D.Tarasowa, I.Ermilov, "SlideWiki: elicitation and sharing of corporate knowledge using presentations". In Knowledge Engineering and Knowledge Management, pp. 302-316, Springer Berlin Heidelberg, 2012.

[5] N. Friesen, J.Hopkins, "Wikiversity; or education meets the free culture movement: An ethnographic investigation", First Monday, 13(10), 2008.

[6] L.Breslow, D.E.Pritchard, J.DeBoer, G.S.Stump, A.D.Ho, D.T.Seaton, "Studying learning in the worldwide classroom: Research into edXs first MOOC", Research & Practice in Assessment, 8, pp. 13-25, 2013.

[7] S.Audsley, K.Fernando, B.Maxson, B.Robinson, K.Varney, "An Examination of Coursera as an Information Environment: Does Coursera Fulfill its Mission to Provide Open Education to All?", The Serials Librarian, 65(2), pp. 136-166, 2013.

[8] S.Vahdati, C.Lange, S.Auer, "OpenCourseWare Observatory - Does the Quality of OpenCourseWare Live up to its Promise?", arXiv preprint arXiv:1410.5694, 2014.

[9] J.Reich, "Rebooting MOOC Research", Science, 347(6217), pp. 34-35, 2015.

[10] P.Kacsuk, Science Gateways for Distributed Computing Infrastructures, Springer International Publishing, 2014.

[11] T.Fahringer, A.Jugravu, S.Pllana, R.Prodan, C.Seragiotto, H.L.Truong, "ASKALON: a tool set for cluster and Grid computing", Concurrency and Computation: Practice and Experience, 17(2-4), pp. 143-169, 2005.

[12] M.R.Berthold, N.Cebron, F.Dill, T.R.Gabriel, T.Kotter, T.Meinl, P.Ohl, C.Sieb, B.Wiswedel, "KNIME: The Konstanz Information Miner", Data Analysis, Machine Learning and Applications, pp. 319-326, 2008.

[13] T.Glatard, J.Montagnat, D.Lingrand, X.Pennec, "Flexible and efficient workflow deployment of data-intensive applications on grids with moteur", Int. J. High Perf. Comp. Appl., 22(3), pp. 347-360, 2008.

[14] P.Kacsuk, Z.Farkas, M.Kozlovszky, G.Hermann, A.Balasko, K.Karoczkai, I.Marton, "WS-PGRADE/gUSE generic DCI gateway framework for a large variety of user communities", J. Grid Comp., 10(4), pp. 601-630, 2012.

[15] Yu.Gordienko, L.Bekenov, O.Gatsenko, E.Zasimchuk, V.Tatarenko, "Complex Workflow Management and Integration of Distributed Computing Resources by Science Gateway Portal for Molecular Dynamics Simulations in Materials Science", Proc. Third Int. Conf. “High Performance Computing” HPC-UA 2013, (Ukraine, Kyiv, 2013), pp. 148-155, 2013.

[16] Yu.Gordienko, L.Bekenev, O.Baskova, O.Gatsenko, E.Zasimchuk, S.Stirenko, "IMP Science Gateway: from the Portal to the Hub of Virtual Experimental Labs in Materials Science", In Proc. 6th International Workshop on Science Gateways, IEEE Xplore Digital Library, DOI: 10.1109/IWSG.2014.17, pp. 61-66, 2014.

[17] Yu.Gordienko, O.Baskova, O.Gatsenko, "Enabling High-Performance Distributed Computing to e-Science by Integration of 4th Generation Language Environments with Desktop Grid Architecture and Convergence with Global Computing Grid", Proc. of Cracow Grid Workshop’10, Cracow, Poland, pp. 234-243, 2011.

[18] S.Plimpton, "Fast parallel algorithms for short-range molecular dynamics", Journal of Computational Physics, 117(1), pp. 1-19, 1995.

[19] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (http://www.R-project.org) 2008.

[20] S.Plimpton, M.Jones, Pizza.py Toolkit (http://pizza.sandia.gov/) 2005.

[21] J.Li, "AtomEye: an efficient atomistic configuration viewer", Modelling and Simulation in Materials Science and Engineering, 11(2), pp. 173-177, 2003.

[22] O.Gatsenko, L.Bekenev, E.Pavlov, Yu.Gordienko, "From Quantity to Quality: Massive Molecular Dynamics Simulation of Nanostructures Under Plastic Deformation in Desktop and Service Grid Distributed Computing Infrastructure", Computer Science, 14, pp. 27-44; doi:10.7494/csci.2013.14.1.27, 2013.

[23] Yu.Gordienko, "Molecular Dynamics Simulation of Defect Substructure Evolution and Mechanisms of Plastic Deformation in Aluminium Nanocrystals", Metallofiz. i Noveishie Tekhn., 33(9), pp. 1217-1247, 2011.

[24] A.S.Gatsenko, E.E.Zasimchuk, P.O.Teselko, S.G.Stirenko, Yu.G.Gordienko, "Computer Simulation of Mechanisms for Appearance of Localized Synergetic Defect Substructures under Plastic Deformation of Metal Nanocrystals", Metallofiz. i Noveishie Tekhn. 36(9), pp. 1001-1018, 2014.

Evaluation of POS Scheduling Algorithm for Distributed Computing Environment

Anastasia Shamakina and Leonid Sokolinsky
South Ural State University, Chelyabinsk, Russian Federation
shamakinaav@susu.ru, Leonid.Sokolinsky@susu.ru

Abstract - One of important classes of computational problems is the class of problem-oriented workflow applications executed in distributed computing environment. The problem-oriented workflow application can be represented by the oriented graph with the tasks as the nodes and the data flows as the edges. For the problem-oriented workflow application, we can predict the execution time of the task and the amount of data to be transferred between the tasks. Nowadays, the significant number of scheduling algorithms for the distributed computing environment are proposed. Some of them (like DSC algorithm) take into account the peculiarity of problem-oriented workflow applications. Others (like Min-min algorithm) take into account the many core structure of the node of computational network. However, no one of them take into account both these properties. We developed the new Problem-Oriented Scheduling (POS) algorithm for distributed cluster computing environment, which is able to plan the launch of a task on several processor cores, taking into account the limit of task scalability.

I. Introduction

Today a large number of scheduling algorithms for distributed computing environments are suggested [1, 2]. Only a small part of these algorithms takes into account the problem-oriented specificity of workflows in complex applications. This specificity is expressed in the form of setting an execution time of a task and an amount of transmitted data.

In this paper, we propose Problem-Oriented Scheduling (POS) algorithm for distributed cluster computing environment, which is able to plan the launch of a task on several processor cores, taking into account the limit of task scalability.

This paper presents the results of computational experiments to investigate the adequacy and efficiency of the POS algorithm for problem-oriented computing environments. The article contains the following sections. In section II, the basic idea of the POS algorithm is proposed. In section III, we present the evaluation of the POS algorithm. We describe a methodology of the experiments, methods of generating test job graphs and results of the comparison of algorithms POS, DSC and Min-Min. In section IV we summarize the results of our research.

The present work was supported by Russian Foundation for Basic Research (RFBR), Research Project No. 14-07-31159-mol-a "My first grant".

II. Problem-Oriented Scheduling Algorithm POS

We have developed a new scheduling algorithm for distributed computing environments, which enables to plan the launch of a task on several processor cores with a limitation of task scalability. A computing application, which represents a workflow, is presented as a weighted labeled directed acyclic graph. Each task (that is a graph node) is labeled by a pair of natural numbers: the executing time of the task on a single processor core and a maximum number of cores on which the task demonstrates a nearly-linear speedup. In future, we will consider tasks which have a nearly-linear speedup. The edge weights define the amount of transferred data. The proposed scheduling algorithm constructs a sequence of graph configurations in order to minimize a cost of the critical path in this graph. In the initial configuration, the graph is presented by a canonical tiered-and-parallel form (multilevel structure). Initial task clustering is defined. For each task a separate compute node is allocated. Respectively, each cluster includes only one task. All task are marked as non-fixed. Then a schedule of task execution is defined. All tasks belonging to the same tier have the same start time. The start time for all tasks of the first tier equals zero. Tasks of the second tier will able to execute when all tasks from the first tier were done and all necessary data for all tasks of the second tier were transmitted. Similarly, the start time of tasks from other tiers is calculated. The tasks of the first tier are marked as fixed on their compute node. The next step is a conversion to a new graph configuration. To reduce the total execution time of the graph, we can zero the communication cost of an edge between a fixed and a non-fixed tasks by joining these tasks to a particular cluster at the same compute node. To do it, we move the non-fixed task along its tier to the fixed task on another compute node. Also a new additional tier in the graph for the non-fixed task can be created. The task is marked as fixed. Thus, the clustering and the schedule are changed. The conversion to the new configuration is performed only in case when the cost of the critical path does not increase. The process of constructing a sequence of graph configurations is completed when all tasks are fixed. The detailed description of POS algorithm can be found in paper [3, 4].

III. Evaluation of the POS algorithm

A. Methodology of the experiments

We evaluate the POS algorithm using the following two classes of jobs:

1) multi-objective optimization (MOO);
2) random jobs (RJ).

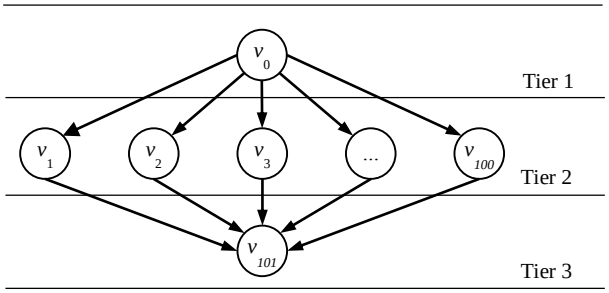


Figure 1. An example of a job graph of the MOO class

The MOO class is one of the most important classes of modern distributed computing system workload. Job graphs of the MOO class have the following structure: a tiered-and-parallel form consists of three tiers. The first and third tiers of the tiered-and-parallel form contain one vertex. The second tier contains *w* vertices. The number *w* is specified when the job graph is generated. The tasks in second and third tier are multiple copies of the same task in first tier. In practice, all tasks in different tiers are different. A vertex of the first tier is connected with all vertices of the second tier by edges. A vertex of the third tier is also connected with all the vertices of the second tier by edges. Each vertex is labeled by a pair of natural numbers (m_v, t_v). The first one is a maximum number of cores on which the task demonstrates a nearly-linear speedup. The second one is the time of task execution on a single processor core. For this experiment, numbers m_v and t_v are constants and they have the same value for all tasks. Practically numbers m_v and t_v may have different values. Similarly, each edge is labeled by a constant δ, which defines the amount of data transferred between the tasks. The constant δ has the same value for all edges. Fig. 1 shows an example of a job graph of the MOO class with a width of the tiered-and-parallel form *w* equals to 100.

The RJ class represents random jobs with different numbers of vertices and edges. A qualitative characteristic of a job graph in the RJ class is a ratio T/Δ, where T is the average executing time of all tasks on a single processor core, Δ is an average weight of all edges. According to this parameter in the RJ class Gerasoulis et al. [5] distinguish the following three important groups.

1) The group M1 is a set of balanced jobs graphs with the ratio $T/\Delta \in (0.8, 1.2)$. The average weights of computation and communication are close.
2) The group M2 is a set of coarse grain graphs with the ratio $T/\Delta \in (3, 10)$. In these job graphs, most of the time is spent on the computation.
3) The group M3 is a set of fine grain graphs with the ratio $T/\Delta \in (0.1, 0.3)$. In such jobs, significant time is expended on transmission data and the calculation requires a little time.

B. Results of the experiments

In the first series of experiments, we investigated a density of a schedule generated by the POS algorithm. Here, the density of a schedule is called the inverse of the number of compute nodes involved for a job.

TABLE I. PARAMETERS FOR CONSTRUCTING A JOB GRAPH OF MOO CLASS

Parameter	Semantics	Value
m_v	Maximum number of cores on which the task demonstrates a nearly-linear speedup	10
t_v	Executing time of the task on a single processor core	100
δ	Amount of data transferred between the tasks (conventional units)	50

TABLE II. PARAMETERS FOR CONSTRUCTING A JOB GRAPH OF RJ CLASS

Parameter	Semantics	Value
l	Height of a job graph	10
m_v	Maximum number of cores on which the task demonstrates a nearly-linear speedup	10
t_v	Executing time of the task on a single processor core	100
δ	Amount of data transferred between the tasks (conventional units)	50

TABLE III. PARAMETERS FOR CONSTRUCTING A JOB GRAPH FOR THE SECOND SERIES OF EXPERIMENTS

Parameter	Semantics	Value
m_v	Maximum number of cores on which the task demonstrates a nearly-linear speedup	20
t_v	Executing time of the task on a single processor core	40
δ_{M1}	Average weight of an edge for group $M1$ (conventional units)	40
δ_{M2}	Average weight of an edge for group $M2$ (conventional units)	10
δ_{M3}	Average weight of an edge for group $M3$ (conventional units)	400
d	Number of cores in a compute node	100

Firstly, the density of the schedule has been investigated for jobs from the MOO class. The experiments were carried out for three values of the parameter *w*, which determines a width of a job graph: 100, 200, 300. The same values of edge weights and vertex labeling were used for all edges and vertices of a graph. They are presented in Table I.

The experimental results are shown in Fig. 2 (A) as the dependence of the number of involved compute nodes on the number of processor cores on a single compute node. The plots show that if the number of cores in a compute node is increased, then the density of the schedule for the MOO class will increase and tend to the maximum value 1 in all considered cases. Herewith, the schedule density increases significantly with an increase in the width of the job graph.

Secondly, the density of the schedule has been investigated for jobs from the RJ class. The experiments were carried out for three values of the parameter *w*, which determines a width of a job graph: 30, 50, 70. The same values of edge weights and vertex labeling were used for all edges and vertices of a graph. They are presented in Table II. The experimental results are shown in Fig. 2 (B). The plots show that if the number of cores in a compute node is increased, then the density of the schedule will also

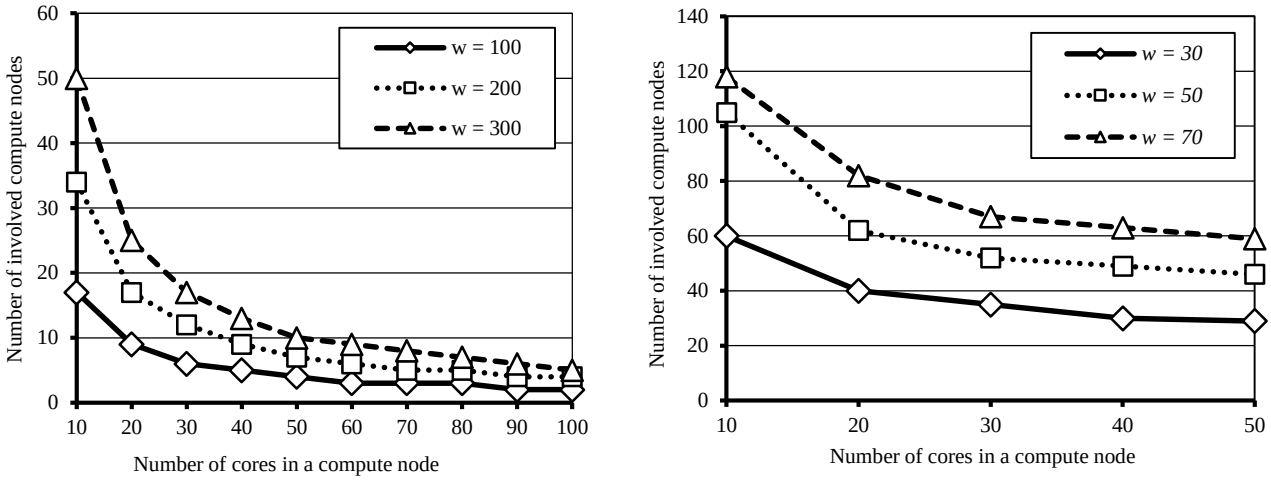


Figure 2. Dependence of the number of involved compute nodes on the number of processor cores on a single compute node: (A) for the MOO class; (B) for the RJ class

increase and tend to the maximum value 1 in all considered cases. Herewith, the schedule density increases significantly with an increase in the width of the job graph.

In the second series of experiments, we investigated the ratio of the execution time of a job graph by the POS algorithm in comparison with the other two scheduling algorithms: DSC [6] and Min-Min [7]. For this purpose three groups M1, M2 and M3 of job graphs have been prepared with the parameters given in Table III. The experimental results are shown in Tables IV-VI. The POS algorithm significantly exceeds algorithms DSC and Min-min by total time of a job in all cases. For example, an average acceleration of the POS algorithm in comparison with the Min-min algorithm exceeds 11, and with the DSC is more than 5 for balanced jobs graphs. The POS algorithm minimizes the number of involved compute nodes better, than the DSC algorithm.

IV. CONLUSION

The POS algorithm for distributed cluster computing environments allows to plan the launch of one task on multiple cores with the limitations on its scalability. This scheduling algorithm is designed to create a system of intelligent planning and scheduling a workflow in order to increase the efficiency of the supercomputer complexes with cluster architecture. This algorithm was implemented in Java as a service based on UNICORE platform. Computational experiments demonstrated that the POS algorithm significantly outperforms DSC and Min-min by the total time of job execution.

TABLE IV. COMPARISON OF ALGORITHMS POS, DSC И MIN-MIN FOR THE GROUP M1

No.	l	w	$\|V\|$	$\|E\|$	Number of involved compute nodes			The ratio of the execution time of a job graph	
					POS	DSC	MinMin	DSC/POS	MinMin/POS
1	5	10-20	51	126	7	21	4	3.78	8.86
2	10	10-20	117	296	4	36	4	3.67	9.53
3	10	20-30	211	505	16	68	6	3.77	8.36
4	20	5-10	137	252	2	46	2	10.19	17.74
Average value								5.35	11.13

TABLE V. COMPARISON OF ALGORITHMS POS, DSC И MIN-MIN FOR THE GROUP M2

No.	l	w	$\|V\|$	$\|E\|$	Number of involved compute nodes			The ratio of the execution time of a job graph	
					POS	DSC	MinMin	DSC/POS	MinMin/POS
1	5	10-20	49	113	3	25	4	6.67	5.67
2	10	10-20	130	390	12	42	4	4.79	2.51
3	10	20-30	206	464	17	73	6	4.05	3.47
4	20	5-10	141	290	13	41	2	4.79	1.47
Average value								5.08	3.28

TABLE VI. COMPARISON OF ALGORITHMS POS, DSC И MIN-MIN FOR THE GROUP M3

No.	l	w	$\|V\|$	$\|E\|$	Number of involved compute nodes			The ratio of the execution time of a job graph	
					POS	DSC	MinMin	DSC/POS	MinMin/POS
1	5	10-20	59	190	4	21	4	4.50	13.15
2	10	10-20	123	378	3	34	4	4.89	42.46
3	10	20-30	201	453	9	61	6	2.06	10.78
4	20	5-10	133	287	2	29	2	3.13	7.88
Average value								3.64	18.57

REFERENCES

[1] F. Dong, S.G. Akl, "Scheduling algorithms for grid computing: State of the art and open problems". Technical Report No. 2006-504, Queen's University, Canada, 2006. P. 55.

[2] J. Yu, R. Buyya, K. Ramamohanarao, "Workfow Scheduling Algorithms for Grid Computing", In "Metaheuristics for scheduling in distributed computing environments". Springer, 2008. Vol. 146, pp. 173-214.

[3] A. Shamakina, L. Sokolinsky, "Resource Scheduling Algorithm in Distributed Problem-Oriented Environments", Proceedings of the UNICORE Summit 2014. IAS Series, 2014. Vol. 26. P. 49-60.

[4] A. Shamakina, "Brokering Service for Supporting Problem-Oriented Grid Environments", Proceedings of the UNICORE Summit 2012, Forschungszentrum Julich, 2012. P. 67-75.

[5] A. Gerasoulis, T. Yang, "A Comparison of clustering heuristics for scheduling directed acyclic graphs on multiprocessors", Journal of Parallel and Distributed Computing, 1992. Vol. 16, No. 4. P. 276-291.

[6] T. Yang, A. Gerasoulis, "DSC: Scheduling Parallel Tasks on an Unbounded Number of Processors", IEEE Transactions on Parallel and Distributed Systems, 1994. Vol. 5, No. 9. P. 951-967.

[7] M. Maheswaran, S. Ali, H.J. Siegal, D. Hensgen, R.F. Freund, "Dynamic Matching and Scheduling of a Class of Independent Tasks onto Heterogeneous Computing Systems", Journal of Parallel and Distributed Computing, 1999. Vol. 59, No. 2. P. 107-131.

Performance evaluation on a Grid platform

Asif Ali Banka[1]
Deptt. Of Computer Science & Engineering
Islamic University of Science & Technology
Awantipora, J&KIndia
asifbanka@icloud.com

Muneer Ahmad Dedmari[2],
Mufti Tufail Masood[3]
Deptt. Of Computer Science & Engineering
Islamic University of Science & Technology
Awantipora, J&K India
ermuneer.ahmad@hotmail.co.uk, tufailcs0929@gmail.com

***Abstract*— The technological advancements have resulted in substantial increase of commodity computing, mainly as the outcome of faster hardware and more sophisticated softwares. In spite the presence of super computers in present age, all the problems in the fields of science, engineering and business cannot be efficiently and effectively dealt with. This is mainly because of complexity factor and cost margin. For a complex program presented by any of the fields above, the data provided requires a number of heterogeneous resources that are scattered across the globe hence making the problem very cumbersome to handle. To address this; concept of grid computing evolved which combines and connects all the required heterogeneous resources to form a single entity which can resolve the problem at hand. With the implementation of grid at practical level lot of problems arise which need to be addressed before the system is put to use. One of the main aspects that must be kept under check while executing a problem set on grid is the security, which mainly includes the privacy and the integrity of the data that becomes vulnerable due to its distributed nature. This paper mainly focuses on the implementation of grid and gives the idea of the exponential difference between the performances of a stand-alone system in comparison to grid. In its broader view we intend to discuss analysis of a specific problem set on both the platforms, and provide the analysed data to support the high end performing nature of the grid in contrast to a stand-alone system.**

Index Terms*— Grid computing ,grid components, grid setup, performance evaluation.(keywords)*

I. INTRODUCTION

Grid Computing can be considered as an infrastructure service that makes computation available on demand like water, gas or electricity by joining resources spread over the globe at different locations. Grid is virtualization of resources which is a hardware software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to computational capabilities, as described in Ian Foster and Carl Kesselman [1]. In grid computing sharing of computational resources (CPU cycles, disk space etc.) rather than data is be attained. Its key values are in the underlying distributed computing infrastructure technologies that evolve in support of cross-organizational application and resource sharing (virtualization) across platforms and organizations. This kind of virtualization is achievable through use of open standards, which ensure that applications can transparently take advantage of whatever appropriate resources made available to them. Grid computing could be defined as any of a variety of levels of virtualization along a continuum. Exactly where along that continuum one might say that a particular solution is an implementation of grid computing versus relatively simple implementation using virtual resources is a matter of opinion [2].

In most organizations there are lots of underutilized resources, such as computing and data resources. With the help of grid these unused resources can be properly utilized. Often, resources may have enormous unused disk capacity. Grid computing (more specifically, a data grid) can be used to aggregate the unused storage into a larger virtual data resource.

Grid also offers resource balancing by scheduling jobs on machine with low utilization.

On the basis of use, grid computing can be divided into different types:

- **Computational grids:** These type of grid are meant to provide secure access to computational resources, sufficient enough to perform processing of computational problems which otherwise would have required high computing power machines.
- **Collaboration grid:** With the advances in network hardware resources and Internet services, demand for better collaboration has increased. Such desired collaboration is best possible with these kinds of grids.
- **Utility Grid:** In this type of grid not only CPU cycles are shared, also other software's and special peripherals like sensors are also shared.
- **Network grid:** Even if we have computational machines with enough computational power as a part of grid but with poor network communication one can't utilize those machines optimally. Network grid provides high performance communication using data caching between nodes there by speed-up communication with each cache nodes acting as router.
- **Data grid:** There are two things, data and computation over that data. Data grid provides the support for data storage other data related services like data discovery, handling, publication, etc [3].

Globus: A software infrastructure that enables applications to handle distributed, heterogeneous computing resources as a single virtual machine. The Globus project is a U.S. multi-institutional research effort that seeks to enable the construction of computational grids. A computational grid, in this context, is a hardware and software infrastructure that provides dependable, consistent, and pervasive access to high-end computational capabilities, despite the geographical distribution of both resources and users. Globus is constructed as a layered architecture in which high-level global services are built upon essential low-level core local services. The Globus toolkit is modular, and an application can exploit Globus features, such as resource management or information infrastructure, without using the Globus communication libraries [4].

SETI@HOME: The seminal Internet distributed computing project, SETI@home, originated at the University of California at Berkeley. SETI stands for the "Search for Extra-terrestrial Intelligence," and the project's focus is to search for radio signal fluctuations that may indicate a sign of intelligent life from space. SETI@home is the largest, most successful Internet distributed computing project to date. Launched in May 1999 to search through signals collected by the Arecibo Radio Telescope in Puerto Rico (the world's largest radio telescope) [5].

II. IMPLEMENTATION

For realizing grid following steps should be taken into consideration:

- Integration and communication of various individual software and hardware components to act like a single system.
- Implementation of middle-ware so that the resource sharing and access becomes secure.
- Development of grid tools so that grid applications and resources can be easily monitored and managed.
- Developing applications that can be used to take the advantages of resources available in grid [6].

The components required to establish a grid are:

- **Grid Fabric:** It includes geographically distributed and accessible from anywhere on the Internet resources. It comprises computers (workstations or PC's running on different operating systems), databases, networks, resource management systems such as LSF, CONDOR or PBS and sensors or special scientific devices.
- **Grid Middle-ware:** It acts as intermediator between components of software with the applications, allowing them to communicate. It is made up of services sets such as remote process management, co-allocation of resources, information exchange, and security. Interoperability between two or more networked computers is enabled by grid middleware with the help of common protocols.
- **Grid Tools:** These tools helps in realizing grid, developing grid enabled applications and brokers that acts as user agents which can manage or schedule computational resources across grid.
- **Grid Applications and Portals:** Grid portals provide us user interface for accessing grid services. It can be used for submission of jobs and getting result back in grid environment [7].

While creating the grid environment, we walked through the following processes/tasks

1) Globus Toolkit Configuration: The Globus Toolkit is an open source software toolkit used for building computational grids and grid based applications. It allows sharing of computing power, databases, and other tools securely online across corporate, institutional and geographic boundaries without sacrificing local autonomy. It is being developed by the Globus Alliance. It is widely used as a basis for grid implementations all over the world, including Grid Ireland. The Globus toolkit has a modular design which contains at its core an implementation of the Open Grid Services Infrastructure (OGSI) specification. Globus provides resource management, data management and information services. Built upon this, there is a security infrastructure which provides security at a message level as well as authentication and authorization [8].
2) Network Setup: Configuring DNS: Domain Name System (DNS) converts the name of website (www.gridboss.mquad.com) to an IP address (192.168.0.6). This is important, because the IP address of a Web site's server, not the Web site's name, is used in routing traffic over the Internet. Everyone in the world has a first name and a last, or family, name. The same thing is true in the DNS world: A family of Web sites can be loosely described a domain. For example, the domain linuxhomenetworking.com has a number of children, such as www.linuxhomenetworking.com and mail.linuxhomenetworking.com for the Web and mail servers, respectively. The DNS queries from remote machines can be responded by BIND. It is an acronym for the Berkeley Internet Name Domain project, which is a group that maintains the DNS-related software suite that runs under Linux [9].
3) Security Implementation: For implementing security in grid environment one must consider the following security aspects.
    - Authentication: Is the process by which a resource or user's identity is verified. It is extremely important to correctly establish the correct identity of an entity.
    - Authorization: Once the identity of an entity is correctly established authorization is used to find what actions the entity is entitled to perform.

- Encryption: Any data passing over public networks is open to be intercepted. A sniffer has the ability to listen in on this data and extract sensitive information as well as usernames and passwords. To avoid this encryption is used to encrypt all transmissions. To a party sniffing the data it would appear as garbage.
- Data Integrity: This involves ensuring that the integrity of the data transferred across the Grid is maintained. This is extremely important in guaranteeing accurate results from jobs.
- Data Confidentiality: A lot of grid applications involve processing data. In many cases where this data is of a sensitive nature it is imperative to ensure that only those authorized to view this data be allowed.
- Single Sign On and Delegation: It is often the case that an entity A will act on behalf of entity B and will need to contact another entity C on B's behalf. Delegation of entities' credentials enables this to be possible [10, 11].

The attempt to fulfill these security requirements systems based on the Public Key Infrastructure (PKI) were used. The Globus Toolkit offers a Grid Security Infrastructure (GSI) that provides a PKI based security system that can be used to address most of the security problems encountered when dealing with grids [12].

For the Globus toolkit, X509 certificates were used for authenticating users, hosts and resources. This package uses OpenSSL commands to create public and private keys for the Certifying Authority (CA) [11].

III. RESULT

The implemented grid system is analyzed to see how it performs in contrast with stand-alone system. The performance of both the systems can be analyzed by the time taken for a job (operation) to get completed. Difference between the times taken by two systems to complete a job can be evaluated by:

$$t_d = t_s - t_g \quad (1)$$

where t_s is the time taken by stand-alone system to complete the job, t_g is the time taken by grid system to complete the job, and $t_{d\ is}$ the difference of time taken by the grid system and the stand-alone system to get job completed.

The jobs (operations) this paper is focused on are:

i) Copy operation: In this operation different set of data of varying size has been copied from one location to another on the same system both by the grid and stand-alone system. As shown in Fig1. the time taken by a stand-alone system for carrying out the copy operation is more than that taken by the grid system for the same job because in grid, computational power of different systems are utilized for carrying out the operation.

TABLE I. REPRESENTS THE TIME TAKEN BY GRID AND STAND ALONE SYSTEM TO COPY THE DATA OF DIFFERENT SIZES.

Time Taken By A Stand Alone System(In Sec)(Ts)	Time Taken In Grid(In Sec)(Tg)	Size(In GB)
0.86	0.79	0.1457
1.452	1.252	0.1958
13.236	7.134	0.3730
14.121	10.975	0.5689
27.049	25.59	0.746
31.023	28.263	0.7649
41.999	45.649	1
50.979	47.32	1.4
92.596	86.081	2
135.42	129.26	3
180.13	174.46	4
220.56	201.16	4.8
355.12	345.79	8
1661.68	1499.36	15.5
3265.68	2956.66	32

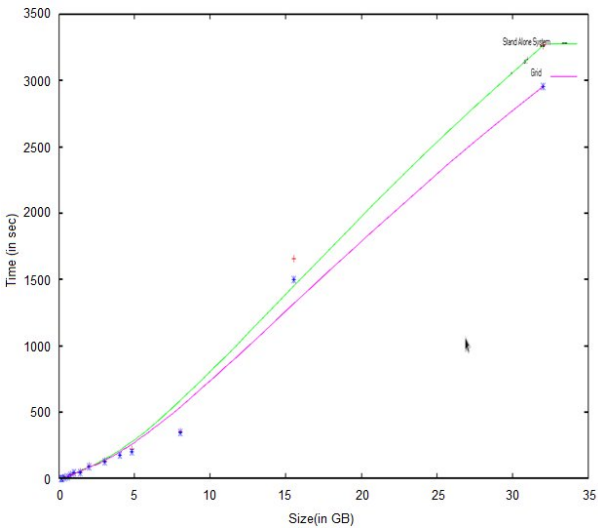


Fig1.Copy operation submitted to stand alone system and grid

From graph shown on Fig 1, it is evident that the copy operation processed on grid takes less time when compared with a standalone system.

ii) Tar operation: It is concerned with the archiving of different files into one compressed entity. On grid these

different files are compressed using resources of multiple components so as to enhance the overall performance for carrying out the mentioned operation.

TABLE II. REPRESENTS THE TIME TAKEN BY GRID AND STAND-ALONE SYSTEM FOR DOING TAR OPERATION ON DIFFERENT SIZES OF DATA.

Time Taken By A Stand-Alone System(In Sec)(Ts)	Time Taken In Grid(In Sec)(Tg)	Size(In GB)
0.032	0.028	0.0079
1.448	0.475	0.1457
1.754	1.542	0.1958
7.542	7.139	0.3730
12.045	11.477	0.5689
27.233	24.212	0.7649
38.145	35.968	1
57.19	50.15	1.4
100.011	95.79	2
147.56	140.67	3
192.16	182.76	4
236.06	217.56	4.8
370.72	352.16	8
1168.21	989.32	15.5
2502	1573	32

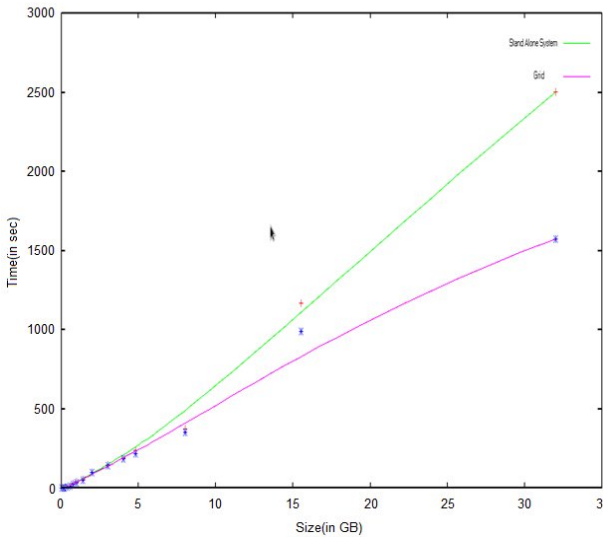


Fig 2 .Tar operation submitted to stand-alone system and grid.

Fig 2 clearly signifies that performance of operation on the stand-alone system is less than the same operation performed on the grid.

iii) Untar Operation: It deals with expansion of a compressed file back into its original form. The decompression is done over different machines by the distribution of the compressed data on the grid and then collected to form the original. The comparison of performing the operation on both the platforms reviewed that the Untar operation on the stand-alone system took more time than the same operation done on the same file on the grid.

TABLE III. REPRESENTS THE TIME TAKEN BY GRID AND STAND-ALONE SYSTEM FOR DOING UNTAR OPERATION ON DIFFERENT SIZES OF DATA.

Time Taken By A Stand-Alone System(In Sec)(Ts)	Time Taken In Grid(In Sec)(Tg)	Size(In GB)
0.046	0.027	0.0079
1.855	0.510	0.1457
2.301	1.627	0.1958
8.403	7.564	0.3730
29.56	27.52	0.5689
38.145	34.331	0.7649
47.230	42.778	1
65.45	58.65	1.4
107.517	99.70	2
158.77	150.76	3
199.79	187.66	4
249.24	238.56	4.8
395.96	360.23	8
1405.11	1107.56	15.5
2939	1893	32

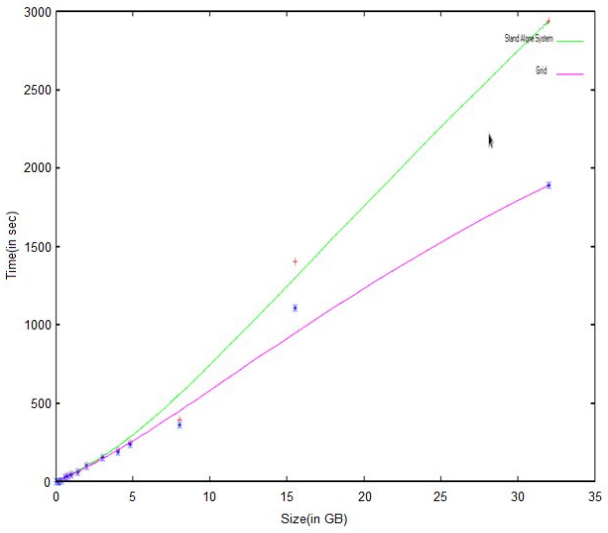


Fig 3 . Untar operation submitted to stand-alone system and grid

In consideration to Fig 3, it is consistent that grid performs faster than a standalone system.

When compared, the results of operations done on both the system, it is concluded that grid does no doubt perform a specific task much faster than the same operation performed on the stand alone system. These operations being COPY, TAR and UNTAR.

This can be verified by calculating the average time taken by the stand alone system and the grid system in TAR, UNTAR and COPY operations individually (The average time taken by grid (t_g) being 6091.905 sec, 4862.11 sec, 5648.377 sec for copy, tar and untar operations respectively and the time taken by stand alone system (t_s) being 5559.744 sec, 3682.251 sec, 4210.477 sec for copy, tar and untar operations respectively). Substituting the values of each in equation I:

COPY operation:

$t_d = 6091.905 – 5559.744 = 532.161$sec

TAR operation:

$t_d = 4862.11 – 3682.251 = 1179.859$sec

UNTAR operation:

$t_d = 5648.377 – 4210.477 = 1437.9$ sec

The positive values of 't_d' for each operation above show that grid actually performs better than the stand alone system.

Furthermore, performance of grid system is dependent on certain factors and this can be shown by:

$$t_p \propto \frac{S}{\sum_{i=1}^{n} Pi} \quad (2)$$

were t_p is time taken by grid system to complete an operation, S is size of the archiving operation and $P_1, P_2, P_3 \ldots\ldots P_n$ are the available processing power of n nodes participating in grid system.

It can also been seen that the variance in the time taken for the TAR and UNTAR operations in comparison with the stand-alone system is much larger than the variance observed in case of the COPY operation.

IV. CONCLUSION

With each passing nano second the complexity of problems provided to the stand alone computers is increasing exponentially so for now the preexisting super computers focused on a specific computational method may handle the problem but in the near future the one stand-alone super computer may not be sufficient to solve the high complexity problems. As we have presented in this paper, in support by the results that the factor by which the performance of computation on a problem set by utilizing the various resources around the globe is actually very high as compared to the same computation being performed on the stand alone computer. The fact that a technology can connect all the specific resources around a globe required for a problem and then single handedly solve the same by a high performance factor and in an affordable manner as that of a super computer of present or future gives us the glimpse of the importance of grid computing.

ACKNOWLEDGMENT

The work of course would not have been completed without support from family and guidance provided by the supervisor. It is worth mention that Mr. Naiman Altaf and Mr. Mohammad Nayeem Shah has provided support and vital role in completion of this research paper.

REFERENCES

[1] Ian Foster, Carl Kesselman, "The GRID: Blueprint for a new Computing Infrastructure," Morgan Kauffman Publishers, 1999.

[2] Bart Jacob, Michael Brown, Kentaro Fukui and Nihar Trivedi, "Introduction to grid computing", listed in IBM RedBooks, 2005,pp.03–04.

[3] Ovaise Khan, "Types of Grid", http://thegridweblog.blogspot.in/2005/10/types-of-grids.html.

[4] Luis Ferreira, ViktorsBerstis, Jonathan Armstrong, Mike Kendzierski, Andreas Neukoetter, Masanobu Takagi, Richard Bing-Wo, Adeeb Amir, Ryo Murakawa, Olegario Hernandez, James Magowan, Norbert Bieberstein, "Introduction to Grid Computing with Globus", IBM RedBook.

[5] SETI@home: Search for Extraterrestrial Intelligence at home, "http://setiathome.ssl.berkeley.edu/"

[6] Mark Baker, Rajkumar Buyya and Domenico Laforenza Grid, "International Efforts in Global Computing"

[7] JC Desplat, Judy Hardy, Mario Antonio letti, Jarek Nabrzyski, Maciej Stroinski, Norbert Meyer, "Grid Service Requirements", Enacts, 2002

[8] Ian Foster and Carl Kesselman, "Globus: A Metacomputing Infrastructure Toolkit", International Journal of Supercomputer Applications, 11(2): 115-128, 1997

[9] Paul Albitz and Cricket Liu, "DNS and BIND", 4TH edition, 2001, pp. 04-05.

[10] Ian Foster, Carl Kesselman, G. Tsudik, S. Tuecke, "A Security Architecture for Computational Grids", 5th ACM Conference on Computer and Communications Security Conference, San Francisco, CA, USA.

[11] S. Tuecke, V. Welch, D. Engert, L. Pearlman, M. Thompson, "Internet X.509 Public Key Infrastructure (PKI) Proxy Certificate Profile", http://www.hjp.at/doc/rfc/rfc3820.html

[12] Shushan Zhao, A. Aggarwal, R.D. Kent, "PKI-Based Authentication Mechanisms in Grid Systems", IEEE Conference on Network, Architecture and Storage, 2007.

Synergy of Volunteer Measurements and Volunteer Computing for Effective Data Collecting, Processing, Simulating and Analyzing on a Worldwide Scale

N. Gordienko*, O. Lodygensky**, G. Fedak*** and Yu. Gordienko****
* Phys.-Math. Lyceum 142, Kyiv, Ukraine
** LAL, University Paris South, Orsay, France
*** University of Lyon, Lyon, France
**** G.V.Kurdyumov Institute for Metal Physics, National Academy of Sciences, Kyiv, Ukraine
e-mail: assasin.nik@gmail.com

Abstract - The paper concerns the hype idea of "Citizen Science" and the related paradigm shift: to go from the passive "volunteer computing" to other volunteer actions like "volunteer measurements" under guidance of scientists. They can be carried out by ordinary people with standard computing gadgets (smartphone, tablet, etc.) and the various standard sensors in them. Here the special attention is paid to the system of volunteer scientific measurements to study air showers caused by cosmic rays. The technical implementation is based on integration of data about registered night flashes (by radiometric software) in shielded camera chip, synchronized time and GPS-data in ordinary gadgets: to identify night "air showers" of elementary particles; to analyze the frequency and to map the distribution of "air showers" in the densely populated cities. The project currently includes the students of the National Technical University of Ukraine "KPI", which are compactly located in Kyiv city and contribute their volunteer measurements. The technology would be very effective for other applications also, especially if it will be automated (e.g., on the basis of XtremWeb or/and BOINC technologies for distributed computing) and used in some small area with many volunteers, e.g. in local communities (Corporative/Community Crowd Computing).

I. INTRODUCTION

The proposed topic concerns the new hype idea of "Citizen Science" and volunteer involvement in science [1]. The main principle here is the paradigm shift: to go from the passive "volunteer computing" (widely used now in many fields of sciences [2-5]) to other volunteer actions under guidance of scientists as "volunteer measurements", "volunteer data processing and visualization", "volunteer data mining", etc. They can be carried out by ordinary people with modern standard computing units (wearable PC, smartphone, tablet, smart watch, etc.) with various operational systems (Android, iOS, Windows, Tizen, etc.) and the standard sensors in them or easily accessible external measuring units.

The work presented here was partially supported by EU FP7 SCI-BUS (SCIentific gateway Based User Support) project, No. RI-283481.

The main components of the idea or conditions for effective implementation of volunteer measurements are as follows: (a) they should be used by ordinary people – not scientists only, (b) they should be performed by easily accessible measuring units – like modern smartphones with numerous sensors, and (c) in unobtrusive way.

The most important expectations (and potential criteria of success) are possibilities to:

- leverage the available "crowdsource" resources: machine (personal CPUs + sensors) and human (brains, manual operations of sensors) ones,
- get huge number of volunteers (up to millions),
- obtain the new scientific "quality" from these huge "quantities",
- involve ordinary citizens in scientific process,
- report to society about the current scientific activities and priorities.

II. PREVIOUS WORK AND STATE OF THE ART

The feasibility of the proposed idea was already tested in several awarded volunteer projects with the several types of activities in different fields of science, which is illustrated by some use cases below in this section.

A. Monitoring Environmental Conditions

In 2012 "My Green City" project was initiated by the group of students for monitoring environmental conditions and their effects on everyday life [6]. It was motivated by the famous problem that vehicle exhausts are ones of the most dangerous sources of air pollution in cities, but global attempts to assess how emissions impact local city ecology have proven to be ineffective.

The aims of "My Green City" project were as follows:

- to estimate personal dosage of accumulated dangerous emissions by volunteers,

Figure 1. The typical volunteer gear: CO-gas-USB logger (top left), which can be attached to rucksack, and smartphone with GPS-navigator in the pocket (bottom left) of the volunteer (right)

- to determine real-time distribution of pollutants by vehicle exhausts in cities.

The technical implementation was based on integration of pollution data by vehicle emissions sensor (CO-gas-USB logger) and location data by GPS-navigator in ordinary smartphones (Fig. 1) for: estimation of personal dosage of accumulated dangerous emissions (so called "personal air health monitor"), and continuous air pollution data processing and construction of interactive air pollution map (Fig. 2).

To check feasibility of this proposal the student pilot project "My Green City" was implemented with construction of the air pollution map for the city of Kyiv (Ukraine) in 2012. The social incentive behind this pilot project can briefly be expressed by the slogan: "from volunteer scientific calculations to volunteer scientific measurements". It became one of the 15 global finalists in Google Science Fair 2012 [7].

Scientists can develop tools (personal air health monitor and online virtual air pollution map) by integration of location-based data from sensors of numerous volunteers. It can allow volunteers and communities to make real time estimations about the actual impact of vehicle emissions on a city environment.

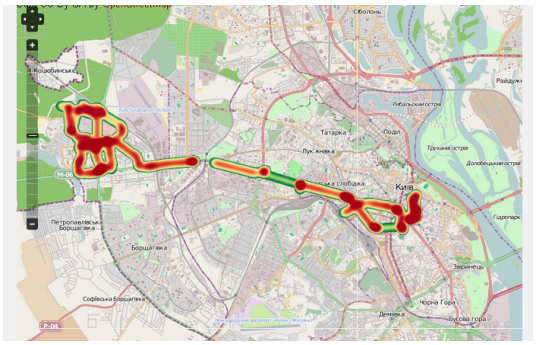


Figure 3. The interactive air pollution map for the city of Kyiv along the regular tracks of children acting as volunteers (http://dg.imp.kiev.ua/slinca/my-green-city/CO)

For municipal authorities, these tools could improve overall city traffic planning and city management accountability. For people, these tools could stimulate better personal planning and implement healthier behavior due to: (A) the greater public awareness of air pollution distribution by virtual air pollution map; and (B) the targeted estimation of personal dosage of accumulated dangerous emissions by personal air health monitor.

B. Monitoring Personal Health

The other volunteer measurement project «How to Make Old Age More Comfortable, Less Boring and Active for Your Nearest and Dearest» (2013) was dedicated to the following problem [8]. The human population become older and will continue to do so in the coming decades; elderly people suffer from various diseases. The project activities were based on the fact that elderly people suffer from various diseases, but their progress can be monitored by decrease of mobility. The aim of the project was to explore the use of sensors in the usual smartphone to monitor the health of older people by analysis of their locomotor activity. The similar method was used in this project, namely, embedded personal sensors (accelerometers) and smart phones were used for real time monitoring the motion pattern (mean, standard deviation, skewness, curtosis, etc.) for different types of motion (sleep, walk, run, etc.).

The technical implementation (Fig. 3) consisted in integration of personal measurement tools and attachment points on the basis of the following available gadgets:

- smartphone with accelerometer (G-sensor) on the arms (during writing, web-surfing, etc.) and legs during cycling (walking, running, etc.);
- "smart watch" (the arm wrist device ez430 by Texas Instruments) with accelerometer and wireless data broadcasting to a PC.

Multi-parametric moment analysis was applied for data analysis. And the previous results shown that

Figure 2. The typical location of measurement gear on volunteers: HTC smartphone with G-sensor on the arm during writing, web-surfing, etc. (top left) and on the leg during biking (top right), and "smart watch" (ez430, Texas Instruments) (bottom)

activities can be roughly classified, i.e. divided into groups (colored ellipses) with the similar values of the acceleration distribution parameters (Fig. 4, see legend in the color electronic version):

- passive (web surfing, read, sleep) (brown ellipse),
- moderate (writing, sitting) (green ellipse),
- active (sports, housework, walking) (blue ellipse).

To check feasibility of this idea the student project "How to Make Old Age More Comfortable, Less Boring and Active for Your Nearest and Dearest" (2013) was implemented in the local community. Again, the social incentive behind this pilot project can be expressed by the slightly different slogan: "from general to personal volunteer scientific measurements". It became one of the 90 Regional Finalists in Google Science Fair 2013 [9].

The tests have shown that some essential features of health and activity patterns can be distinguished and classified. By analysis of sensor (accelerometer) data scientists can develop tools (so called "personal activity monitor") that allow volunteers and communities to make real time estimations of the actual locomotor activity decrease and subsequent health decay. For health care authorities, these tools could improve overall elderly care and stimulate personal online elderly care services. For communities, these tools could stimulate closer attention to locomotor activity and its automated monitoring due to: (A) the greater public awareness of dangerous locomotor activity decrease and probable subsequent health decay; (B) the targeted estimation of personal dangerous decrease of locomotor activity by personal activity monitor; (C) increase of efficiency of elderly care in some communities by means of automatic detection of dangerous decrease of locomotor activities among neighbors, (D) decrease of burden on communities, which is related with a high costs of elderly care.

III. CURRENT PROJECT - AIRSHOWER@HOME

A. *Scientific Problem*

Cosmic rays originate mainly outside the Solar System. They may produce showers of secondary particles ("air showers") that penetrate and impact the Earth's atmosphere and sometimes even reach the surface. "Air showers" of secondary elementary particles affect electronics, health, environment. Their study is expensive (>$50 mln.) and limited in scale. The most energetic cosmic rays (>5×10^{19} eV) are most important, but very rare were observed, for example, 1 high-energy event is registered per month by Pierre Auger observatory. Pierre Auger Observatory — is an international cosmic ray observatory designed to detect ultra-high-energy cosmic rays [10,11]. It detects cosmic rays in the range of 10^{17}-10^{21} eV, costs >$50 mln., and has a detection area >12 km^2 and huge installation area >3000 km^2 (like Luxemburg). Cosmic rays affect electronics, because they have sufficient energy to alter the states of circuit components in electronic integrated circuits, causing transient errors like corrupted data in electronic memory devices, and incorrect performance of CPUs, often referred to as "soft errors" (not to be confused with software errors caused by programming mistakes/bugs!). They affect health, especially the health threat from cosmic rays is the danger posed by galactic cosmic rays and solar energetic particles to astronauts on interplanetary missions. They are one of the most important barriers standing in the way of plans for interplanetary travel by crewed spacecraft [12,13]. Cosmic rays affect environment directly or via solar-induced modulations in climate change, because solar variations modulate the cosmic ray flux on Earth, they would consequently affect the rate of cloud formation and hence the climate [14].

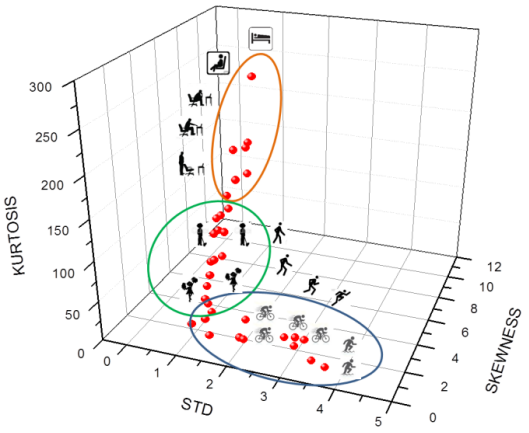


Figure 4. The moments (standard deviation, skewness, kurtosis) plot: activities (denoted by pictograms) can be classified by groups (colored ellipses in the color electronic version) with the similar values of the acceleration distribution parameters: passive (brown ellipse), moderate (green ellipse), active (blue ellipse)

B. *AirShower@Home Project Description*

As far as "air showers" of elementary particles created by cosmic rays affect electronics, they can be monitored by some available devices (like CMOS-camera chips).

The aims of AirShower@Home project were:

- to estimate possibility to identify "air showers" of secondary elementary particles,
- to determine frequency and distribution of "air showers" in the densely populated cities – by massive involvement of volunteers with creation of virtual online map of "air showers".

The volunteer measurements could be performed on the basis of the following tools:

- camera chips as embedded sensors in GPS-enabled smartphones;
- volunteer communities in densely populated locations for real time monitoring the "air showers" and analysis of their effects (on environment, electronics, health).

The technical implementation is based on integration of data about registered flashes (by radiometric software) in shielded camera chip, synchronized time and GPS-data in ordinary smartphones/tablets/other gadgets. The current workflow is shown in Fig. 5, where the size of the available academic solution, Pierre Auger observatory, is

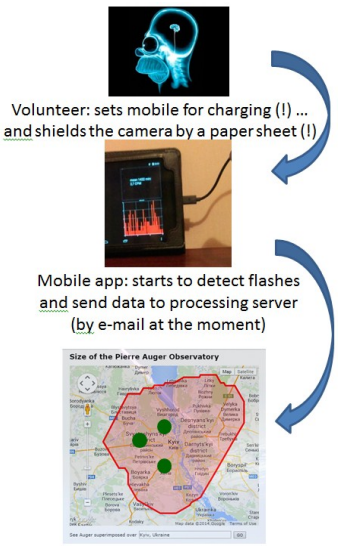


Figure 5. The diagram of the current workflow, where the sizes of Pierre Auger observatory are shown for comparison on the map of measurements in the city of Kyiv (bottom)

shown for comparison on the map of measurements in Kyiv. It includes the following components:

- hardware (various gadgets were tested: Samsung SIII (with 8 megapixels camera), HTC Amaze smartphone (8 Mpx), Nexus 7 tablet (1.2 Mpx), Nexus 7 tablet (5 Mpx), ASUS TF201 tablet (8 Mpx),
- "brainware", because some human actions are needed (volunteer should shield the camera, and perform manual post-processing, analyzing and mapping "air showers"),
- software (various mobile apps): radiation counters (used for tests only), apps for time synchronization, and e-mail clients for submissions of logs,
- volunteer community: >80 pupils and students were involved already (during the lab works in "Distributed Computing Course" at the National Technical University of Ukraine "KPI" in Kyiv).

IV. Results of the Current Project

One of the examples of the registered radiation results in counts per minute ("cpm") vs. time (within ~5 days) is shown in Fig. 6. The four smooth peaks corresponds to slow increases of background radiation (caused by the temperature increase during days), and numerous sharp peaks potentially correspond to cases of air showers caused by cosmic rays. The different colors indicate different values of filters used in the measurements. These results lead to conclusion that seasonal and other temperature changes could crucially influence the measured values of radiation and should be taken into account during such crowd measuring/computing projects.

The other example is related to tests performed on a board of the commercial air flight. The idea was to check feasibility of volunteer measurements of air showers on

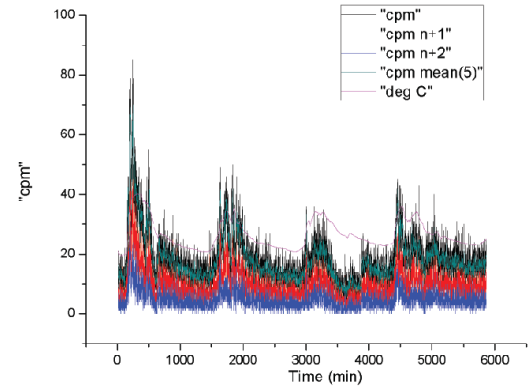


Figure 7. Registered radiation (counts per minute — cpm) vs. time

various heights using the standard commercial flights (with the usual safety precautions as to usage of electronics during commercial flights). Registration of air showers was performed simultaneously by two Nexus tablets (with different sensitivity) on a board of the Airbus 320 plane during flight Kyiv-Budapest on 10/03/2014. The radiation (in cpm) as a function of time was measured (Fig. 7), where smooth growth corresponds to increase of the registered radiation counts with altitude (9 km), and the same sharp peaks correspond to the events of simultaneous registration of the air showers caused by cosmic rays. As to possible pitfalls in such experiments, the problems related with gadget overheating can be also observed in the end of flight.

The initial stage of this crowd measuring/computing idea was performed among numerous volunteers from Lyceum 142 and the National Technical University of Ukraine "KPI" (Fig. 8). Mainly they were compactly located in the city of Kyiv (Fig. 8, inset), and this local concentration of volunteers is especially useful for the current scientific application. The measured data were send by mobile application to server, where they were collected, post-processed (in R environment for statistical analysis), mapped (by means of googleVis and Shiny packages for R environment), and published on the AirShower@Home project web-page at ShinyApps.io server (http://yocto.shinyapps.io/Height_test).

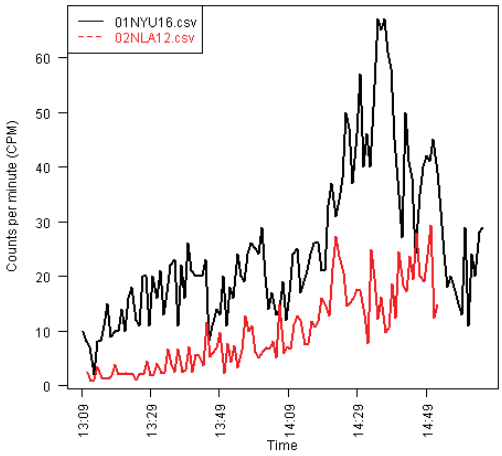


Figure 6. Registration of air showers during a commercial flight

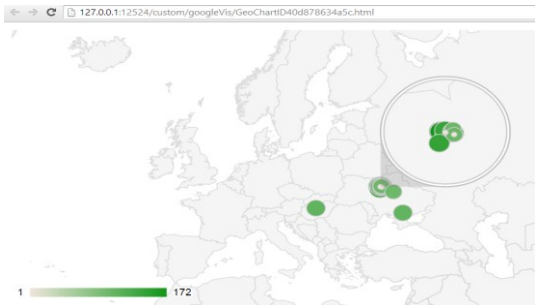


Figure 8. The map of heights at volunteers' locations among other maps with post-processed data of AirShower@Home project (http://yocto.shinyapps.io/Height_test)

Actually, these interactive maps contain information about volunteer locations, date and time of volunteer measurements, height of their locations (which are important for comparison), and the registered radiation counts.

In Table 1 the short comparative analysis is given for several solutions, namely, fdor academic solution (Pierre Auger observatory), do-it-yourself (DIY) solution with external sensor (customized solution on the basis of Raspberry PI single-board computer), and the current AirShower@Home solution with embedded sensor (widely available solution like smartphone). From this analysis it is clear that the AirShower@Home project provides an additional (not alternative!) way to study cosmic ray (by analysis of air showers) and their influence on civilization without huge installations, numerous high-qualified scientists, high basic and maintenance costs. Moreover, its crowd measuring/computing infrastructure can be effectively used for many other applications, for example, in two previous projects to upgrade and scale them up to worldwide scale.

TABLE I. COMPARISON WITH OTHER SOLUTIONS

	Academic solution (Pierre Auger observatory	*DIY solution with external sensor (Raspberry PI + sensor)*	*This "Citizen Science" solution (smartphone with sensors)*
Size	3000 km^2	1 cm^2	4 mm^2
Scalability	<30%	<10 times	> 10^6 (unlimited)
Maximal Area	~3000 km^2	~1 cm^2	>10^6 km^2
Staff (per installation)	500 scientists	1-3 scientists	1 ordinary human ("Citizen Scientist")
Cost	> $50 000 000	~$500	<$200 (smartphone)

V. CONCLUSIONS

Some lessons were learned during these three different volunteer measuring/computing projects. First of all, the positive aspects consist in feasibility:

- to leverage the available "crowdsource" resources (hardware and "brain-ware") even at this level – it is feasible, if sensors are embedded already (integration with external sensors is hardly can be attractive for potential volunteers),
- to get huge number of volunteers (millions) – it is feasible for project without sensitive data, but it could be hardly feasible for projects related with personal location and health data due to privacy issues,
- to obtain the new scientific "quality" from these huge "quantity" of data – it is feasible, and porting to the standard platforms for volunteer computing should help, because the limited automation is used at the moment,
- to involve ordinary citizens in scientific process – it is feasible, if task, results, and volunteer management will be fully automated, for example, by porting to BOINC [15] and XtremWeb environments [16],
- to report to society about the current scientific activities and priorities – it is feasible, again after porting to BOINC [15] and XtremWeb environments [16].

Taking into account these considerations and comparative analysis (Table 1) the immediate ways for improvement are seem to be based on porting this crowd measuring/computing infrastructure to the standard platforms for volunteer computing. Some of them (like BOINC and XtremWeb) are proved to be effective for distributed computing applications. For example, XtremWeb was successfully used for AIRES application used in relation to results obtained at Pierre Auger observatory [17]. Additional potential for improvement is related with optimization of the current workflow, for example, on the basis of the Science Gateway ideology. Now it is very promising way for integration and management of tasks, users, and resources [18,19], for example, on the basis of gUSE-WS-PGRADE package [20].

Such porting will allow better integration with embedded sensors and automatic cooperation with them (without manual operations), post-processing, analyzing and mapping results, and effective volunteer community management (job submission, data storage and analysis, forum, etc.) at server, especially, in open local and corporative communities (like schools, colleges, universities). As to potential contribution to science, many scientific phenomena can be investigated by ordinary people and in any place, not just in some secluded labs, which is essential for "Citizen Science" paradigm. This approach could be very useful, if it will be used in some areas with many volunteers, e.g. in local communities (which could be considered as Corporative/Community Crowd Computing)

ACKNOWLEDGMENT

The authors would like to thank numerous volunteers from Lyceum 142 and the National Technical University of Ukraine "KPI" who took part in the volunteer measurements during the initial stage of the project.

REFERENCES

[1] A.Irwin, Citizen science: a study of people, expertise, and sustainable development, Psychology Press, 1995.

[2] D.P.Anderson, G.Fedak, "The computational and storage potential of volunteer computing", In. Proc. Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID 06), Vol. 1, pp. 73-80, May, 2006.

[3] O.Baskova, O.Gatsenko, Yu.G.Gordienko, "Enabling High-Performance Distributed Computing to e-Science by Integration of 4th Generation Language Environments with Desktop Grid Architecture and Convergence with Global Computing Grid", Proc. Cracow Grid Workshop'10, Cracow, Poland, pp. 234-243 (http://www.researchgate.net/publication/261361640) 2010.

[4] O.Gatsenko, O.Baskova, O.Lodygensky, G. Fedak, Y.Gordienko, "Statistical properties of deformed single-crystal surface under real-time video monitoring and processing in the desktop grid distributed computing environment", Key Engineering Materials, 465, pp. 306-309, 2011.

[5] O.Gatsenko, L.Bekenev, E.Pavlov, Yu.Gordienko, "From Quantity to Quality: Massive Molecular Dynamics Simulation Of Nanostructures under Plastic Deformation in Desktop and Service Grid Distributed Computing Infrastructure", Computer Science, Vol. 14, 1, pp. 27-44, 2013.

[6] M.Klimenko, O.Kozlov, "My Green City" project, Google Science Fair 2012 (http://goo.gl/o2qmL1).

[7] "Google Announces 15 Finalists for 2012 Google Science Fair", BuzzIT: Buzz in Technology (http://goo.gl/3PcbW7) June 6, 2012.

[8] N.Gordienko, "How to Make Old Age More Comfortable, Less Boring and Active for Your Nearest and Dearest", Google Science Fair 2013 (http://goo.gl/2Tl200).

[9] J.Weinberg, "Google Science Fair 2013: Banana breast implants and squid submarines - the best ideas from teen science competition", Mirror (http://goo.gl/WzJdFt) 26 September 2013.

[10] P.Auger, P.Ehrenfest, R.Maze, J.Daudin, R.A.Fréon, Extensive Cosmic-Ray Showers, Rev. Mod. Phys. 11 (3–4), pp. 288–291, 1939.

[11] J.W.Cronin, T.K.Gaisser, and S.P.Swordy, Cosmic Rays at the Energy Frontier, Scientific American, pp. 44 (1998).

[12] Biomedical Results From Apollo - Radiation Protection and Instrumentation (http://lsda.jsc.nasa.gov/books/apollo/S2ch3.htm).

[13] Evaluation of the Cosmic Ray Exposure of Aircraft Crew (http://cordis.europa.eu/documents/documentlibrary/75331981EN6.pdf).

[14] Yu.V.Balabin, A.V.Germanenko, B.B.Gvozdevsky, and E.V.Vashenyuk, "Variations of gamma radiation spectra during precipitations", Journal of Physics: Conference Series, Vol.409, 1, 012243, pp. 1-4 (2013).

[15] D.P.Anderson, "BOINC: A system for public-resource computing and storage", In Proc. Fifth IEEE/ACM International Workshop on Grid Computing, pp. 4-10 (2004).

[16] F.Cappello, S.Djilali, G.Fedak, T.Herault, F.Magniette, V.Néri, O.Lodygensky, "Computing on large-scale distributed systems: XtremWeb architecture, programming models, security, tests and convergence with grid", Future Generation Computer Systems, 21(3), 417-437, 2005.

[17] O.Lodygensky, G.Fedak, V.Neri, A.Cordier, F.Cappello, "Auger & XtremWeb: Monte Carlo computation on a global computing platform", In Proc. Computing in High Energy and Nuclear Physics (CHEP2003) Vol. 40 (2003).

[18] Yu.Gordienko, L.Bekenov, O.Gatsenko, E.Zasimchuk, V.Tatarenko, "Complex Workflow Management and Integration of Distributed Computing Resources by Science Gateway Portal for Molecular Dynamics Simulations in Materials Science", In Proc. Third International Conference "High Performance Computing" HPC-UA 2013 (Ukraine, Kyiv, October 7-11, 2013), pp. 148-155.

[19] Yu.Gordienko, L.Bekenev, O.Baskova, O.Gatsenko, E.Zasimchuk, S.Stirenko, "IMP Science Gateway: from the Portal to the Hub of Virtual Experimental Labs in Materials Science", In Proc. 6th International Workshop on Science Gateways, IEEE Xplore Digital Library, DOI: 10.1109/IWSG.2014.17, pp. 61-66, 2014.

[20] P.Kacsuk, Z.Farkas, M.Kozlovszky, G.Hermann, A.Balasko, K.Karoczkai, I.Marton, "WS-PGRADE/gUSE generic DCI Gateway Framework for a Large Variety of User Communities", Journal of Grid Computing, 10(4), 601-630, 2012.

Successful Implementation of L3B: Low Level Load Balancer

Kiril Cvetkov, Sasko Ristov, Marjan Gusev
University Ss Cyril and Methodius, FCSE,
Rugjer Boshkovic 16, Skopje, Macedonia
Email: kiril_cvetkov@yahoo.com, {sashko.ristov, marjan.gushev}@finki.ukim.mk

***Abstract*—Cloud computing paradigm offers instantiating and deactivating the virtual machine instances on demand according to the clients requirements. When some customer's application or a service needs more resources than a physical server can supply, the cloud provider offers a certain load balancing technique to distribute the load among several servers that will host the analyzed application or service. Additionally, the cloud provider should offer a resource broker to make the application scalable and elastic. In this paper we present a new solution for a low level load balancer, working on a network level. Our load balancer maps the memory addresses of the balancer and the target physical servers (or virtual machines on the cloud) and thus balances the load. The experiments showed that it adds small latency of several milliseconds and thus it slightly reduces the performance when the distributed system is underutilized. However, there is a region of client requests where the system achieves a superlinear speedup (speedup greater than the number of scaled resources). Our case study with doubled resources achieves up to 6.5 speedup (maximum expected linear speedup is 2).**

***Index Terms*—Distributed Computing; HPC; Performance; Services.**

I. Introduction

Information and communication technology offers a huge support for the business today by speeding up the internal business processes. However, a company must be prepared for the business processes that interact with its customers and handle the peaks in the load [1]. Adding more computing resources and balancing the load can solve this issue and various load balancers with different approaches for balancing techniques exist. The efficient and effective load balancer should have satisfy several functional requirements: balancing all incoming clients' packets to the available endpoint servers, having a configurable port where the requests should be sent their, having a pool of active endpoint servers with IP addresses that will handle the incoming packets, and the introduced load balancer between the clients and endpoint servers should improve the overall response time of the clients' requests.

Horizontal scaling is impossible without load balancing. Even more, todays most effective paradigm, the cloud computing, can not be imagined without load balancing. It is one of the four layers in the cloud application architecture [**?**], [**?**]. Rimal et al. [2] describe the load balancing techniques that are used by various cloud service providers and solutions, including Amazon, Google, SalesForce, Azure, Eucalyptus, OpenNebula, etc. Most of these techniques are implementations of a conventional Round Robin schema, weighted selection mechanisms, HAProxy, Sticky session, SSL Least Connect, Source address, cluster server load equalization, high performance protocols over EC2, hardware load balancing, cloud controllers, etc.

Load balancing is a widely used technique long time ago before the introduction the cloud computing paradigm and the virtualization technique. The performance of load balancing techniques in cloud is analyzed by several authors [3], [4].

Radojevic and Zagar [5] grouped them into three groups: Session switching at the application layer, Packet-switching mode at the network layer and Processor load balancing mode. On the other hand, Heinzl and Metz [6] classified the load balancers as hardware and software, and commercial and open source.

There are two main techniques in load balancing: centralized on a specific server or distributed on several servers. The former is controlled by a single central node with all the other endpoint servers communicate with it. Bhadani and For example, Chaudhary [7] proposed the Central Load Balancing Policy for Virtual Machines (CLBVM) solution, while the CLBDM solution [8] considers server load and application performance on top of the Round Robin Algorithm. However, the centralized load balancing approaches in cloud computing have desing limmitation and communication overhead, thus failing to offer full scalability features [9].

Simjanoska et al. [10] proposed an architecture for a Low Level Load Balancer (L3B). Their architecture can distribute server load among several machines that are integrated over the communication link [11]. The proposed architecture is realized on a network level. It dynamically balances the network packets of the clients among all end point servers and after the computation, it forwards the responses to the clients that send the requests. All these transformations are conducted on network layer, that is, by changing the information in the incoming and outgoing packet's header.

Ristov et al. [12] developed a naive implementation of the proposed L3B architecture, which unfortunately did not yield the expected results. In this paper, we present the redeveloped implementation according the same L3B architecture. The results of the conducted experiments not only that confirmed the successfulness of the implementation, but they show a superlinear speedup, that is, we achieved speedup of up to 6.5 by using two endpoint servers compared to the case when

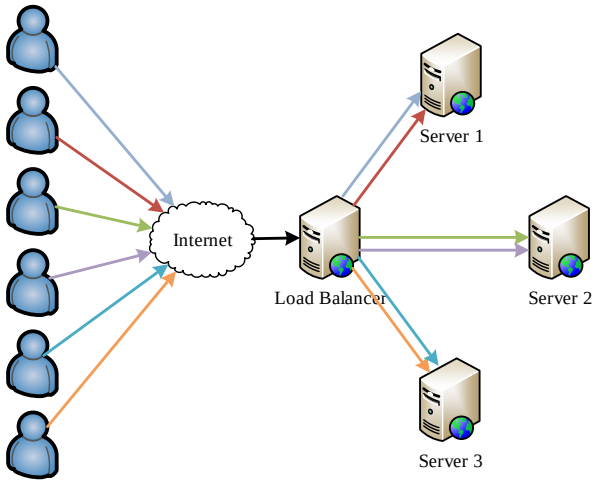


Fig. 1. L3B System architecture

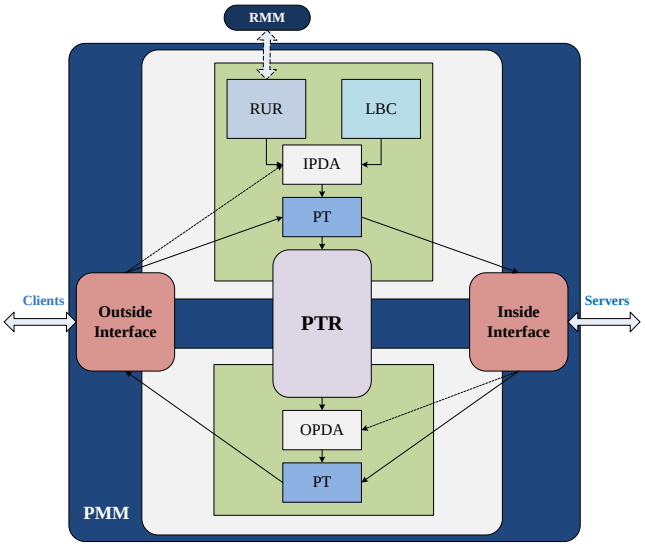


Fig. 2. Design of the PMM module

using only one endpoint server.

The rest of the paper is organized as follows. Section II briefly describes the background of L3B. The successful implementation of the L3B balancer is presented in Section III and experimentally confirmed in Section IV. Section V discusses the successful implementation of the architecture. Finally, Section VI concludes the work and presents future work.

II. L3B BACKGROUND

This section briefly describes the architectural concept and the naive implementation, which did not yield towards scalable balancer.

A. The architectural concept

The L3B's concept is presented in Fig. 1 [12]. The L3B balancer is hosted on a machine in front of a pool of endpoint servers and its main goal is to balance the clients' requests among the available servers. The latency due to balancer's additional layer should be compensated with faster response time of the endpoint servers.

The L3B balancer is designed with two main modules: *Resource Management Module (RMM)* and *Packet Management Module (PMM)* [10]. The main objective of the former is to provide new available endpoint resources according to the current clients' load and endpoint utilisation. PMM balances the incoming packets according to the received information of RMM about available and utilized resources. This paper focuses on the implementation of the PMM module. Its design is presented in Fig. 2.

PMM manages the client's requests, i.e., it forwards an incoming packet to a particular endpoint by using some balancing technique. After the target endpoint server responds back, PMM forwards again the response to the client that has initiated the request.

When a client request is received at the Outside Interface (OI), it sends information to the Input Packet Decision Agent (IPDA). which decides where to forward the request (which endpoint server), by using relevant information by the Resource Utilisation Repository (RUR) and Load Balancing Configuration (LBC) about current utilisation of the endpoint servers and load balancing configuration, correspondingly. The realized decision is sent to the Packet Translation (PT) agent, which proceed with the IP packet header translation using the NAT/PAT (Network Address Translation / Port Address Translation). These translations are stored in the Packet Translation Repository (PTR) as they can be used to forward the response back to the client. Finally, the transformed packet is forwarded to the Inside Interface (II) in order to be forwarded to the target endpoint server. The response packets are transformed similarly by the PT, by using the decision of the Output Packet Decision Agent (OPDA), according to the information stored in PTR. More details can be found in [10].

The L3B was developed as a Java implementation according to the proposed architecture, since modern Java virtual machines' performance is similar to C or C++ [13]. However, the results were not promising [12]. That is, introducing the L3B balancer reduced the performance even worse than the case without balancer. This motivated us to analyze various proposals how to make the L3B more efficient. The next section presents the improved L3B architecture that supports all functional requirements and provides better performance than its predecessor.

III. THE NEW IMPLEMENTATION

Since the naive implementation of L3B did not yield the expected results, we proceed to improve the L3B implementation. This section presents the development of the new improved implementation.

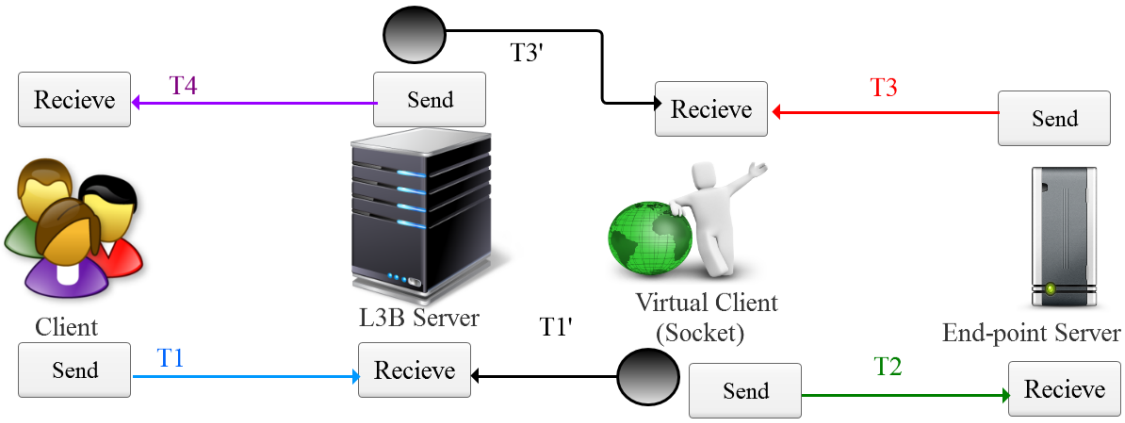


Fig. 3. Improved L3B implementation

A. The new implementation's concept

Fig. 3 presents the new improved implementation, based on creation of virtual client socket. When a packet arrives at the L3B server, L3B creates a "Virtual Client", which is a creation of a socket between new client on a certain port, and a port and IP address with the end-point server, presented as T_1 in the figure. Now, the L3B takes the memory (packet) from the received data in the state and prepares for transmission on the server (T_1'). The next step is to forward the packet to the end-point server, shown as T_2 in the figure. In the next step, denoted by T_3 in the figure, the server processes the request and returns the reply to the virtual client. Then the L3B prepares to send to the client everything that it has previously received and takes its memory (T_3 VirtualClient.Send = Balancer.Recieve). Further on, the final activity, denoted by T_4 is totally the same as previous case.

B. The implementation details

Robust, reliable and correct package transmitter from source to destination is used for complete conduction of our concept. Java Network Sockets features are extended in order to realize the load balancing concept.

The implementation is done in Java JDK 7.0, by using network sockets. The algorithm provides package forwarding using Sockets for TCP communication. It will be the same algorithm if a UDP communication is used, but in this case, the Datagram transfer should be used, instead of Sockets.

The details of the implementation are presented in Fig. 4. The L3B architecture is implemented with several classes:

- *Client* class, which sends the client's request to Balancer;
- *Balancer* class, which receives request from client;
- *ServerInput* class, which sends client's request to multiple servers;
- *ServerOutput* class, which receives the responses from the servers; and
- *PackageForwarder* class, which creates instances of ServerInput class and ServerOutput class as well as responds the outputed data to the client.

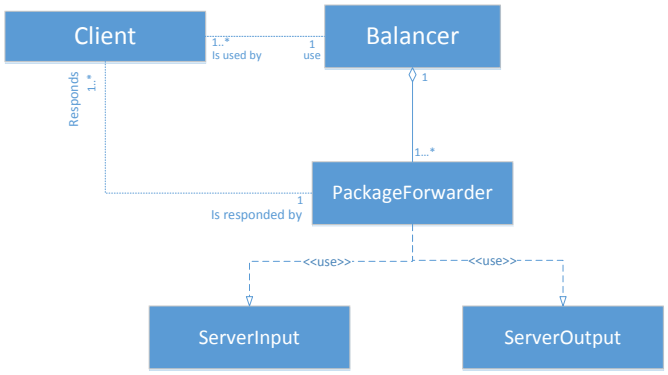


Fig. 4. The details of the implementation

IV. Experiments and Results

This section presents the results of the experiments to confirm the successfulness of the new implementation.

A. Testing methodology

Four workstations with the same hardware resources and platform are used as a testing environment. Each workstation has Intel(R) Xeon(R) CPU X5647 @ 2.93GHz with 8GB RAM memory, installed with Ubuntu 12.04. The client is on the same LAN as the L3B workstation and two end-point workstations to reduce the network latency.

A client requests a packet with constant web page content, which size is $56KB$. The different number of concurrent requests are then initiated to create various loads. We choose this page size in order to be smaller than the IP packet limit of 64KB. Usually the web requests and responses are smaller than $64KB$.

The concurrent requests are simulated with SOAPUI. Each test case consists of sending N concurrent requests per second, such that N varies in each test case starting from $N = 5$ until $N = 100$ requests, by increasing N with step 5. Each test case lasts 60 seconds.

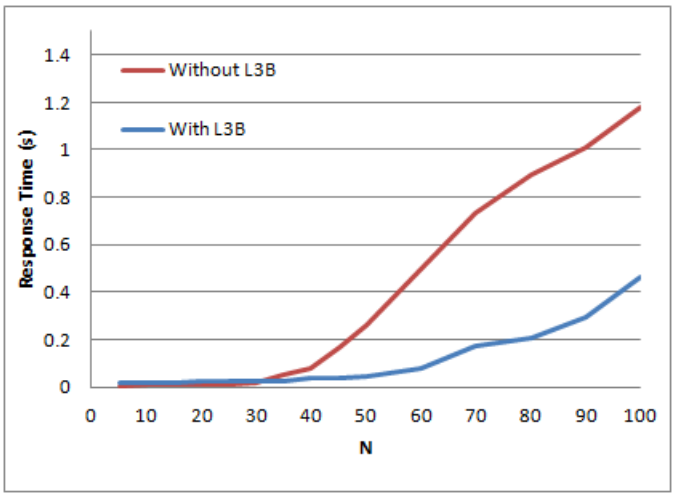


Fig. 5. Results of the experiments for the L3B implementation

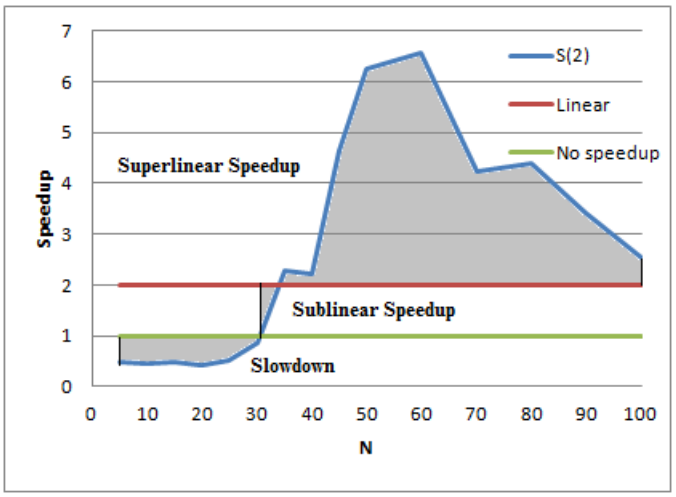


Fig. 6. The achieved speedup of the improved L3B when using two end-point servers

B. Performance analysis of the L3B implementation

Fig. 5 shows the results of the experiments that evaluated the performance of the new improved L3B implementation over the system which does not use L3B balancer. The results show that using the L3B is always better for each number of concurrent requests $N > 30$.

Let us explain why the scenario without the L3B balancer (4 CPU cores, 8GB RAM) is better comparing with the scenario with L3B (8 CPU cores and 16GB RAM) for $N \leq 30$, which uses double resources. Introducing L3B adds a constant delay per packet because a packet goes through the additional server (the L3B server) and virtual client (socket). For small server load N this delay is in the range of the server response time. But, when the total load is increased, the single server in the first scenario (without L3B) will be over-utilised, which will decrease its performance, while the second scenario (with L3B) will still work in the normal mode with better performance. The next section analyzes the scalability and achieved speedup by using the L3B balancer to balance the load among end-points in more detail.

V. Discussion about the L3B scalability

In this section we discuss the speedup of the improved L3B implementation. We conduct two experiments based on a configuration using two end-point servers.

A. Theoretical analysis

According to the Gustafson's law [14], adding more computing and processing resources (CPUs) to some parallel or distributed system will improve its performance limited with linear speedup (the speedup that is equal to the number of scaled resources). The *Speedup* $S(p)$ is defined in (1) as a ratio of the response time of the nominal system T_1 (without L3B with 1 end-point server) and the response time of the scaled system T_p (p end-point servers and the L3B in front of them).

$$S(p) = \frac{T_1}{T_p} \tag{1}$$

Our experiment is conducted using L3B with $p = 2$ end-point servers. That is, N client requests are sent to a single end-point server (as described in Section IV-A), while N client requests are sent to two end-point servers in the second scenario (with L3B), thus balancing the requests by $N/2$ per end-point server. According to the

B. Case study with scaling factor 2

Fig. 6 presents the speedup that the improved version of L3B achieves over the scenario without L3B. Three regions are observed:

- *Slowdown* region, where the speedup $S(p) = S(2) < 1$;
- *Sublinear* speedup region, where the speedup $S(p) = S(2) < p = 2$;
- *Superlinear* speedup region, where the speedup $S(p) = S(2) > p = 2$;

The slowdown region appears in the range $N < 30$ and the speedup in this region has a constant value. The servers in both scenarios are under-utilised, and the L3B reduces the system performance because it introduces a delay. When a load is increased, the first scenario is over-utilised, while the L3B balances the load and thus keeping both servers still in the normal working mode (going into Sublinear and Superlinear speedup regions). However, when the load reaches $N = 60$, that is, 30 requests per server, both servers become over-utilised and thus the speedup starts to decrease. Increasing the load $N > 100$ started to generate errors in the first scenario.

C. A superlinear speedup phenomenon

A superlinear speedup region is achieved in distributed environment. The superlinearity is a well known phenomenon in the cache intensive algorithms [15] because of using more CPU cache memory in parallel implementations compared

to the sequential, where the communication among CPUs is low. The superlinear speedup is achieved in the cloud virtual environment, despite the additional virtualisation layer [16], [17].

However, the reason why the superlinear speedup region appears in this distributed environment is totally different. The nature of web servers is such that increasing the incoming requests dive the server into CPU over-utilisation, while balancing the load among more servers keep the servers in the normal mode, and thus generating the superlinear speedup, despite the delay that the L3B introduces.

VI. Conclusion and Future Work

This paper continues the research in designing and implementing the new approach for a load balancer that works on a network layer of OSI model. Since the naive implementation of the L3B balancer did not yield the expected results, we have improved the implementation and redeveloped the improved L3B, which proved to be scalable with improved performance. Even more, a superlinear speedup region is achieved with the superlinear speedup up to 6.5 when scaling the resources by 2.

This paper presented the case study with introducing L3B in front of only two end-point web servers, loading them with a single static web page. Additional experiments and research will be conducted with larger number of end-point servers and different end-point services and algorithms, i.e., computation intensive, memory demanding or even cache intensive web services.

This implementation of L3B promises a lot and thus it will be upgraded even more. Additional plan is to upgrade the intelligent balancer to balance the load keeping the end-point web servers in the superlinear speedup region where the maximum speedup and the best performance are achieved.

Also, this version of L3B works well only for requests (messages) that do not exceed the size of IP packet. We will upgrade the L3B balancer to support an arbitrary message size, and measure its scalability. In this direction, the RMM module should be implemented to become an elastic load balancer, beside its proved scalability.

Acknowledgment

This work was partially financed by the CloudNet project (Cloud Networking) at Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University, Skopje, Macedonia.

References

[1] A. Murua, I. Gonzalez, and E. Gomez-Martinez, "Cloud-based assistive technology services," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, Sept 2011, pp. 985–989.

[2] B. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, Aug 2009, pp. 44–51. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5331755&isnumber=5331299

[3] K. Nuaimi, N. Mohamed, M. Nuaimi, and J. Al-Jaroodi, "A survey of load balancing in cloud computing: Challenges and algorithms," in *Network Cloud Computing and Applications (NCCA), 2012 Second Symposium on*, Dec 2012, pp. 137–142. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6472470&isnumber=6472451

[4] N. J. Kansal and I. Chana, "Cloud load balancing techniques: A step towards green computing," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 238–246, 2012.

[5] B. Radojevic and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 416–420. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5967092&isnumber=5967009

[6] S. Heinzl and C. Metz, "Toward a cloud-ready dynamic load balancer based on the apache web server," in *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2013 IEEE 22nd International Workshop on*, June 2013, pp. 342–345. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6570639&isnumber=6570561

[7] A. Bhadani and S. Chaudhary, "Performance evaluation of web servers using central load balancing policy over virtual machines on cloud," in *Proceedings of the Third Annual ACM Bangalore Conference*, ser. COMPUTE '10. ACM, 2010, pp. 16:1–16:4. [Online]. Available: http://doi.acm.org/10.1145/1754288.1754304

[8] B. Radojevic and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 416–420. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5967092&isnumber=5967009

[9] D. Ardagna, S. Casolari, and B. Panicucci, "Flexible distributed capacity allocation and load redirect algorithms for cloud systems," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, July 2011, pp. 163–170. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6008706&isnumber=6008659

[10] M. Simjanoska, S. Ristov, G. Velkoski, and M. Gusev, "L3b: Low level load balancer in the cloud," in *EUROCON, 2013 IEEE*, Zagreb, Croatia, 2013, pp. 250–257.

[11] L. Schubert, M. Assel, and S. Wesner, "Resource fabrics: The next level of grids and clouds," in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, Oct 2010, pp. 677–684.

[12] S. Ristov, M. Gusev, K. Cvetkov, and G. Velkoski, "Implementation of a network based cloud load balancer," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*, Sept 2014, pp. 775–780.

[13] D. A. Patterson and J. L. Hennessy, *Computer Organization and Design, Fourth Edition: The Hardware/Software Interface*. Morgan Kaufmann, 2009.

[14] J. L. Gustafson, "Reevaluating amdahl's law," *Communication of ACM*, vol. 31, no. 5, pp. 532–533, May 1988.

[15] M. Gusev and S. Ristov, "A superlinear speedup region for matrix multiplication," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 11, pp. 1847–1868, 2014. [Online]. Available: http://dx.doi.org/10.1002/cpe.3102

[16] ——, "Superlinear speedup in windows azure cloud," in *Cloud Networking (CLOUDNET), 2012 IEEE 1st International Conference on*, Paris, France, 2012, pp. 173–175.

[17] S. Ristov and M. Gusev, "Performance vs cost for Windows and linux platforms in Windows Azure cloud," in *2013 IEEE 2nd International Conference on Cloud Networking (CloudNet) (IEEE CloudNet'13)*, San Francisco, USA, Nov. 2013, pp. 214 – 218.

Scalable and Parallel Machine Learning Algorithms for Statistical Data Mining – Practice & Experience

M. Riedel*+, M. Goetz*+, M. Richerzhagen*, P. Glock*,
C. Bodenstein*, A.S. Memon*+, M.S. Memon*+
* Forschungszentrum Juelich, Juelich Supercomputing Centre, Juelich, Germany
+ University of Iceland/School of Natural Sciences and Engineering, Reykjavik, Iceland
m.riedel@fz-juelich.de

Abstract - Many scientific datasets (e.g. earth sciences, medical sciences, etc.) increase with respect to their volume or in terms of their dimensions due to the ever increasing quality of measurement devices. This contribution will specifically focus on how these datasets can take advantage of new 'big data' technologies and frameworks that often are based on parallelization methods. Lessons learned with medical and earth science data applications that require parallel clustering and classification techniques such as support vector machines (SVMs) and density-based spatial clustering of applications with noise (DBSCAN) are a substantial part of the contribution. In addition, selected experiences of related 'big data' approaches and concrete mining techniques (e.g. dimensionality reduction, feature selection, and extraction methods) will be addressed too. In order to overcome identified challenges, we outline an architecture framework design that we implement with open available tools in order to enable scalable and parallel machine learning applications in distributed systems.

I. INTRODUCTION

Statistical data mining methods can be roughly categorized in classification, clustering, or regression augmented with various techniques for data exploration, feature selection, or feature extraction. In many cases these methods are implemented with machine learning algorithms that have been mostly serially implemented in the past. Known tools that offer such algorithms are R, Matlab, Octave, Weka, or the Python-based scikit-learn.

As reported by Wood et al. in [18], a wide variety of scientific datasets we work with (e.g. earth sciences, medical sciences, etc.) increase with respect to their volume or in terms of their dimensions due to the ever increasing quality of measurement devices. One concrete example that motivates our work in this contribution is earth science data measurements being part of the PANGAEA collection [1]. This collection exceeds a total number of data sets of 349.871 with roughly 7.9 billions in terms of data items making manual quality control complicated today. Another example that motivates our work are hyperspectral remote sensing [24] images that are acquired with hundreds of channels over contigous wavelength bands by satellites (e.g. AVIRIS Indian Pine dataset [2]) . In this context, the application goal is to interpret the observed data and classify meaningful features or classes of land cover types.

In both cases, we observe limitations when analyzing the datasets with serial tools such as too long running times, or memory problems when loading samples of the dataset (even after feature extraction and dimensionality reduction techniques have been performed). More recently, we thus have been forced to analyze such aforementioned concrete examples of *'big data'* with parallel and scalable machine learning algorithms. Although we achieved a speed-up while at the same time maintaining accuracy of solutions, we observed a couple of challenges when working at the intersection of statistical data mining, data management/transfer, and High Performance Computing (HPC) infrastructures.

This contribution discussed these challenges by offerings lessons learned derived from two concrete data-driven science applications. We offer this with practice and experience obtained in the field of clustering time series data and classification of remote sensing images. In both cases, we observed a tremendous effort on working with a relatively high set of parameters, heterogenous job script, and complex HPC environments, that all in turn justify a demand for a data analytics framework that is less complex. This is important since the scientific domain scientists we typically work with as HPC experts are typically not experts in HPC. This paper therefore also outlines our approach of implementing such a framework that automates many processes and is implemented with open-source tools. By using open standard-based core building blocks in our outlined architectural design, we further demonstrate the advantages of using open standards in the concrete problem domain when using parallel and scalable machine learning algorithms.

This paper is structured as follows. After the introduction to the problem domain, Section II provides a a motivation in the light of recent activities around big data analytics and selected case studies. Section III offers the reader a survey of related work in the field of parallel and scalable machine learning frameworks, while Section IV describes lessons learned derived from a couple of scientific applications. Section V then outlines our architecture design and provides basic building blocks that we used for implementation. The paper ends with some concluding remarks of several findings and outlines future work elements.

II. MOTIVATION

Our motivation is driven by the needs of two concrete data-intensive science applications we introduced in earlier work [25] that require different parallel and scalable machine learning algorithms. Both raise a couple of joint requirements that we investigate in this paper and that motivate a more broader approach we refer to as a an *'open standards-based smart data analytics framework'* for parallel and scalable machine learning tasks.

The first data-intensive science application we refer to as PANGAEA time series analysis requires a parallel and scalable clustering algorithm. There are a wide variety of clustering algorithms available such as K-Means Clustering, K-Medoids Clustering, Sequential Agglomerative Hierarchic Nonoverlapping (SAHN), Clustering Using Representatives (CURE). We picked the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [11], because the algorithm itself is able to handle outliers inherently that fits nicely with our goal of automatically detecting outliers in large quantities of marine measurement data as a subset of the PANGAEA data collection.

The second data-intensive science application we refer to as remote sensing image analysis requires a parallel and scalable classification algorithm. There are also a wide variety of classification algorithms available such as linear learning models (i.e. Perceptron), decision trees with ensemble methods, artificial neural networks, or naïve bayes classifier. We picked the Support Vector Machines (SVM) algorithm [22] with non-linear 'kernels', because they are considered to be one of the best out-of-the-box classifiers and used in a wide range of applications that classify image-based datasets. The datasets [15] are shared within the B2Share tool of the European Data Infrastructure EUDAT [3] after using domain-specific features enhancement and various extraction/selection methods (e.g. principle component analysis).

Orthogonal to our both motivational data-intensive science applications our work is motivated by the fact that parallel resources are often geographically dispersed and offer only seldom the full suite of all required machine learning tools and techniques. Our interest includes, for example, parallel HPC machines that are part of the Partnership for Advanced Computing of Europe (PRACE) infrastructure [4] or the US Extreme Science and Engineering Discovery Environment (XSEDE) [5].

Hence, domain-specific scientists are forced to obtain manual accounts at a wide variety of sites, tune each of the systems to satisfy the demands of the corresponding machine learning package, and consider manual techniques to transfer their datasets. In particular the latter one is getting a major problem with large quantities of data while the other challenges above seriously hinder a high productivity of domain scientists that are not experts in parallel HPC environments. Although we obtained a good results in those scientific applications when using parallel and scalable algorithms, we observed a wide variety of issues. Since we also start to engage in other *'big data'* analysis in the field of biomedical sciences, we are require a framework that overcomes the issues today.

III. RELATED WORK

The wide variety of single related work per topic can be significantly reduced by taking into account the intersection of two different study fields: parallel and scalable machine learning algorithms and distributed computing infrastructures. Also, in the light of open and reproducable data-intensive science, we focus in our survey on open source frameworks and leave out selected commercial frameworks that might cover some functionality (e.g. SAS business analytics software, IBM Watson Analytics, etc.).

Selected approaches can be found in the field of bioinformatics such as described by Holl et al. in [6] that influenced earlier work described by Memon et al. in [7] using an open standards-based approach with parameter sweeps. Some of these applications are close to machine learning algorithms in terms of requiring a high number of parameters to get insights into biological datasets. We pick up the approach of using open standards such as the reported Basic Execution Service (BES) standard specification [8] and the Job Submission Description Language (JSDL) standard specification [9], including its normative parameter sweep extension [10].

Other pieces of related work can be found in the context of worfklow engines that partly are able to take advantage of distributed computing infrastructures (e.g. Taverna [26], Kepler [27], etc.). For example, in Da Silva et al. describes in [12] the use of workflow tools in conjunction of DBSCAN. We have been not able to retrieve the parallel DBSCAN implementation and in contrast to our approach that is considered for production usage in a parallel center, the paper rather describes a research project than a production implementation.

Related work in the field of SVMs and workflows are described again by Holl et al. in [13] but using Taverna for a bioinformatics workflow in conjunction with the UNICORE middleware [14]. The work describes a serial SVM that is in parallel executed using the parallelization capabilities of the Taverna workflow engine, while we found no evidence for an inherent parallel SVM implementation. Although Taverna is interesting in terms of workflows, its execution model is not standard-based nor tuned for parameter sweeps and the access to HPC systems is based on UNICORE in that paper.

Another area of related work in the field of *'open standards-based data analytics framework'* for parallel and scalabel machine learning are solutions based on the map-reduce paradigm [16]. While the underlying framework implementations like Apache Hadoop Version 1.0 or 2.0 are scalable, we did not found an inherent parallel implementation of the DBSCAN or SVM implementation in on-top frameworks like Apache Mahout, or the more recent frameworks known as Apache Spark with its machine learning library MLlib. Finally, we surveyed several individual parallel implementations, like a parallel implementation of DBSCAN as described by Patwari et al. in [20]. Although it is open source, our analysis [17] revealed that it is not scalable for our given problem while the GPU-based LibSVM [21] is based on a proprietary framework from NVidea and not straightforward to implement yet.

IV. LESSONS LEARNED AND REQUIREMENTS

In this section we reveal selected *'practice and experience'* when working with scalable and parallel machine learning using two distinct methods: clustering and classification both using each a concrete algorithm.

'Clustering' is a data mining method that divides the data into various subgroups whereby each member within a found group is similar to all others (i.e. based on a defined similiarity measure, e.g. euclidean distance). It belongs to *'unsupervised learning'* methods since no response/output data is given that could supervise the data analysis. Our first part of the analysis of practice and experience is based on working with a concrete algorithm known as DBSCAN in scientific applications.

Figure 1. Parameter challenges when using the parallel DBSCAN algorithm implementation HPDBSCAN on HPC machine JUDGE.

DBSCAN is a density-based clustering algorithm and its principal idea is to find cluster cores in a data set and subsequently expand these recursively having two distinct parameters illustrated in Figure 1. Cluster cores are defined as a region that contains at least a number *minPoints (i.e. parameter -m)* of neighboring points within a given search radius *epsilon (i.e. parameter epsilon)*.

We implemented a parallel DBSCAN version known as highly parallelizable DBSCAN (HPDBSCAN) [17] that overcomes the inherent sequential processing step of the recursive expansion through three major techniques. It offers a 'spatial indexing structure', a 'quadratic split heuristic', and 'differential merging scheme' to combine sub-results into a common result. For more information about HPDBSCAN please refer to [17]. As this code is a hybrid MPI/OpenMP code, it requires also another parameter (i.e. parameter –t) in order to specify the number of threads used for OpenMP elements.

We work with marine scientists in order to perform clustering of time series earth observation measurement data using this tool and given already the couple of these parameters makes the use of this tool in a job script not straightforward. This is particularly a challenge since marine scientists are not experts using HPC machines.

Furthermore, the code has dependencies to many new libraries such as the Hiearchical Data Format (HDF) 5 (i.e. >= 1.8.13), C++ libraries (i.e. >=4.9 for compilation), and MPI (i.e. >= 3.0) that also need to be specified in the job script (cf. Figure 1) and that need to be supported by the HPC machine in question that is hard to find out for not HPC domain scientists.

'Classification' is a data mining method that creates a model that aims to separate data into different groups that are known as classes based on predefined examples. It belongs to *'supervised learning'* methods since predefined response/output is given that could supervise the data analysis. Our second part of the analysis of practice and experience is based on working with a concrete algorithm known as SVM in scientific applications.

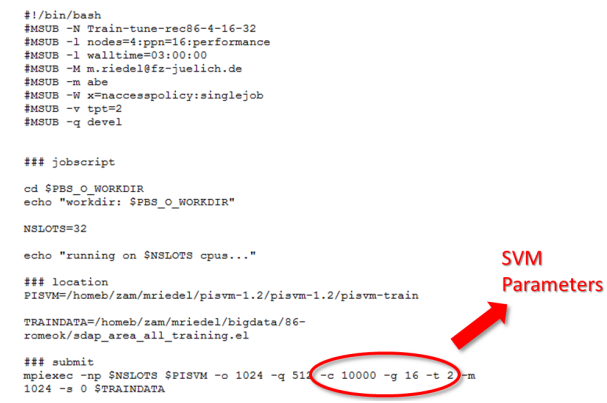


Figure 2. Parameter challenges when using the parallel SVM algorithm implementation piSVM 1.2.1 on HPC machine JUDGE.

SVMs are one of the commonly used out-of-the-box algorithms for classification and their fundamental idea lies on separating training samples which belong to different classes by tracing maximum margin hyperplanes in the space where the samples are mapped. A kernel-based SVM method maps input data instances into a high-dimensional feature space with a non-linear mapping function and then performs linear classification in the high-dimensional features space. The parameter set of the SVM and its used kernel such as the *Radial Basis Function (RBF) kernel (i.e. parameter –t 2)* require basically two parameters whereby the choice of the kernel itself is thus the third parameter shown in Figure 2. The first parameter enables violations of the maximum margin and can be considered as a *cost threshold of the violation (i.e. parameter –c)*. The second parameter is the Υ *parameter of the RBF kernel itself (i.e. parameter –g)*. Both latter parameters are typically obtained by performing cross-validation in order to perform model selection and are thus not known in advance and highly dependent on the dataset in use.

We tuned an existing parallel SVM version known as piSVM 1.2 (and recent 1.3) [19] that is in turn is based on the de-facto standard serial implementation libsvm and takes advantage of MPI. For more information about piSVM please refer to [19]. Tunings include the broader use of MPI collectives instead of individual MPI calls in loops and a better memory setup in order to scale much better in terms of *'big data'* (e.g. large number of samples, or large number of dimensions, etc.). The memory setup requires a couple of parameters (e.g. –o and –q) while some caching layout requires another (e.g. –m). The type of SVMs can be configured by parameter (e.g. –s).

We work with scientists that perform classification of hyperspectral remote sensing images using this tool and given all these parameters the use of the piSVM tool in a job script is not straightforward. This is particularly the

case since remote sensing scientists are not experts in HPC environments. Also, in many cases the scientists need to vary the above mentioned parameters in order to obtain good accuracies or have to apply feature selection/extraction mechanisms (e.g. principle component analysis, etc.) before the actual job run.

Based on these lesson learned we formulate the following four major requirements for a data analytics framework supporting machine learning tasks: *(a) abstraction from error-prone job scripts for specific HPC architectures; (b) standardized handling of job descriptions; (c) support for parameter studies; and (*d) *automation of data transfer into/out of a HPC system.*

V. Architectural Design & Implementation

In this section we outline the architectural design and describe the core building blocks of our implementation that is able to satisfy the requirements based on lessons learned raised in Section VI. Given the page restriction we concentrate on the overall design and only provide a brief description while referencing specific standard specifications and tools used for implementation.

As shown in Figure 3, HPC Resource A offers an MPI environment and deploys the parallel data analytics tool piSVM as part of its BES/JSDL implementation. Another center offers a HPC Resource B with an MPI environment and a preconfigured HPDBSCAN deployment as part of its BES/JSDL implementation. We can consider to use UNICORE as it offers a BES/JSDL implementation and this middleware is also deployed as part of the PRACE as well as XSEDE infrastructure. But any other BES/JSDL implementation (e.g. GENESISII [23]) can be part of the infrastructure deployment still being inline with the architecture.

By using BES/JSDL we satisfy requirement *(a) abstraction from error-prone job scripts for specific HPC architectures.* This is possible because implementations of these standard usually support different job scheduling systems underneath and thus job scripts will be automatically generated and forwarded to the corresponding scheduler of HPC resource A or B. The benefit is that low-level details such as specific dependencies of C++ libraries (e.g. HPDBSCAN with versions >=4.9) or HDF5 (e.g. HPDBSCAN with 1.8.13) can be easily abstractly used. This makes it more feasible to be used by domain-scientists since they do not need to manually edit job scripts. To work properly the JSDL implementation should support the JSDL HPC extension [X] illustrated in Figure 3 as part of the virtual workspaces using BES/JSDL functionality.

Another advantage of using these set of standards is to satisfy requirement *(b) standardized handling of job descriptions.* This enables that domain scientists can more easily choose a broader set of HPC resource of choice with the required data analytics tool for the given problem at hand. Domain scientists are able to use the same client for different machine learning packages and tools that even go beyond piSVM and HPDBSCAN both illustrated in Figure 3. One could also consider map-reduce-based implementations for example if needed (not illustrated).

We have to deeper explore another core building block illustrated as JSDL in Figure 3 in order to understand how our architectural design overcomes the requirement *(c) support for parameter studies.* The JSDL standard offers an extension that is very suitable to address the many complex parameter issues raised in Section IV. It is thus well suited to craft JSDL job descriptions that more easily are able to automatically sweep through different set of parameters. This is helpful for our given problem of HPDBSCAN and its parameters *epsilon* and *minPoints*, but also with respect of sweeping through different options of the piSVM parameters such as *C* and the *RBF kernel parameter* Υ.

Finding the right combination of these parameters is one of the most important elements of getting good results in the clustering application when analysing time series for outliers. It is also important for achieving good accuracy results in the classification application when analysing hyperspectral images. In both cases, the parameters are usually not known by hand and automation of this parameter search (also named gridsearch in machine learning or data mining) is very important. It this context it should be noted that cross-validation techniques benefit the most from running in parallel since it is usually a very time consuming process (e.g. Matlab runs over night for one 10 fold cross-validation run).

Finally, we would like to point out how we satisfy requirement *(*d) *automation of data transfer into/out of a HPC system.* In contrast to the automation meant by parameter sweeps, this requirement raise the demand for having a more automated process of *'staging'* [24] data into, or out of, an HPC environment. The JSDL specification supports this particular requirement very well thus avoiding the wide range of time consuming manual tools and approaches (e.g. using wget to obtain B2SHARE datasets). Of course, in this light, we need to mention the security challenge present in the architecture that we described with a focus on the needs of the scalable and parallel machine learning tools. Infrastructures such as EUDAT, PRACE, and others however have aligned to solve the security issue across infrastructures that is out of reach of this contribution to solve beyond Grid methods.

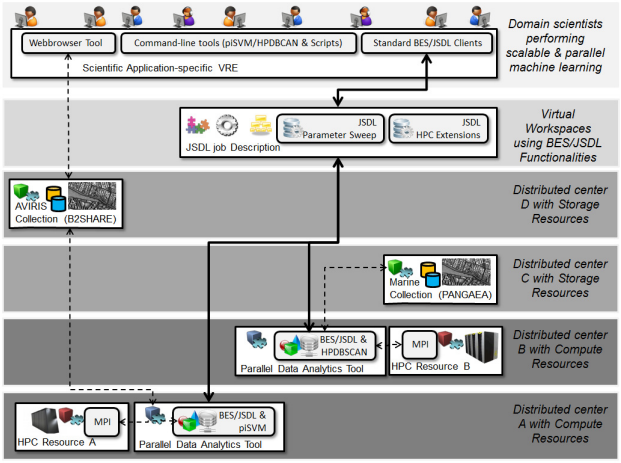


Figure 3. Architectural design with core building blocks of the tools we use to implement scalable and parallel machine learning tasks.

VI. CONCLUSIONS

The earth sciences have been always a major source of large quantities of data, because of the wide variety of measurement devices deployed to understand the world around us. We have performed parallel clustering and classification of concrete *'big data'* and obtained a wide variety of lessons learned in two distinct data-intensive earthy science applications. The practice and experience raised a couple of requirements that we have satisfied by our outlined *'big data analytics framework design'*, which implementation is based on open source tools.

We can conclude that the use of JSDL/BES, including its extensions for parameter sweeps and HPC, is a crucial design factor and enables to overcome many limitations of manually using different job scripts with parallel machine learning tools on heterogenous geographically distributed HPC machines. Most notably, because of the large number of parameters, but also with respect to data transfer/stagings and the overall automation provided by the corresponding JSDL/BES middleware.

In our case, we picked UNICORE as it is deployed on the PRACE and XSEDE infrastructure, but more generally, the architectural design building blocks can be implemented by other JSDL/BES implementations (e.g. GENESISII). We also focussed on only two particular machine learning tools (i.e. HPDBSCAN and piSVM), but we claim that also other parallel and scalable machine learning algorithm implementations for HPC systems can be equally well supported by the architecture.

Additionally, our design based on open standards bears a high potential when implementing new emerging regulations about open science, where scientific datasets and the means to get results (i.e. job configuration parameters, code, etc.) need to be publicly available in order to enable reproducable science. The JSDL specification and its normative extensions are perfectly suited to document this in a standard way being not influenced from underlying changes in HPC environments (i.e. different job scheduling scripts, file systems, etc.). Some of the publishers in science and engineering already have started to require such artefacts in order to publish a scientific article thus increasing trusted science output for decades to come.

By using the B2SHARE service of EUDAT we are inline with these requirements. All our datasets and processing results in the remote sensing classification application can be publicly viewed there and have been referenced in this paper. Hence, the findings can be reproduced when combining them with middleware tools like UNICORE and piSVM respectively. Also the HPDBSCAN tool is open source while the dataset is available in PANGAEA.

Future work includes an extension to the framework design in order to include work on GPUs since they become an important element in parallel and scalable machine learning. Another direction of future work is the support of in-situ analytics combining work from computational steering to augment a HPC simulation run with analytics during its runtime in order to offer on-the-fly statistical data mining (e.g. validation from data).

ACKNOWLEDGMENT

We thank Robert Huber and Andree Behnken (MARUM Bremen, Germany) for his close cooperation in the context of the PANGAEA automated outlier detection case study. We would like to express our thanks to Gabriele Cavallaro (University of Iceland) for his close cooperation in the context of the spectral image analysis.

REFERENCES

[1] Diepenbroek, M., Fuetterer D., Grobe H., Miller H., Reinke M., and Sieger R.,'PANGAEA information system for glaciological data management', Annals of Glaciology 27, pp. 655-660

[2] P.R. Foundation. Hyperspectral remote sensing scene. Online, avalable: https://engineering.purdue.edu/~biehl~MultiSpec/hyperspectral.html.

[3] Lecarpentier, D., Wittenburg, P., Elbers, W., Michelini, A., Kanso, R. Coveney, P., Baxter, R., 'EUDAT: A New Cross-Disciplinary Data Infrastructure for Science', International Journal of Digital Curation, 8(1), pp. 279 – 287

[4] Lippert, T., Eickermann, T., Erwin, D., 'PRACE: Europe's Supercomputing Research Infrastructure', in proceedings of PARCO 2011, pp. 7-18

[5] Bachmann, F., Foster, I., Grimshaw, A., Lifka, D., Riedel, M., Tuecke, S., 'XSEDE Architecture – Level 1 and 2 Decomposition', 2012

[6] Holl, S., Zimmermann, O., Palmblad, M. Mohammed, Y., and Hofmann-Apitius, M., 'A new optimization phase for scientific workflow management systems', Future Generation Computer Systems, 36, pp. 352-362, 2014

[7] Memon, S., Holl, S., Schuller, B., Riedel, M., Grimshaw, A., 'Enhancing the Performance of Scientific Workflow Execution in e-Science Environments by harnessing the Standards based parameter sweep model', in proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery, ACM Digital Library, 2013

[8] Foster, I., Grimshaw, A., Lane, P., Lee, W., Morgan, M., Newhouse, S., Pickles, S., Pulsipher, D., Smith, C., Theimer, M., 'OGSA Basic Execution Service – Version 1.0', OGF Recommendation GFD.108, 2007

[9] Anjomshoaa, A., Brisard, F., Drescher, M., Fellows, D., Ly, A., McGough, S., Pulsipher, D., Savva, A., 'Job Submission Description Language (JSDL) Specification – Version 1.0', OGF Recommendation GFD.136, 2008

[10] Drescher, M., Anjomshoaa, A., Williams, G., Meredith, D.,'JSDL Parameter Sweep Job Extension', OGF Proposed Recommendation GFD.149, 2009

[11] Ester, M., Kriegel, H.P., Sander, J., and Xu, X., 'A density-based algorithm for discovering clusters in large spatial databases with noise', in proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226 – 231

[12] Da Silva, R.F., Juve, G., Deelman, E., Glatard, T., Desprez, F., Thain, D., Livny, M., 'Toward fine-grained online task characteristics estimation in scientific workflows', in proceedings for the 8th workshop on workflows in Support of Large-scale Science, ACM Digital Library, pp. 58-67, 2013

[13] Holl, S., Zimmermann, O., Hofmann-Apitius, M., 'A UNICORE Plugin for HPC-enabled Scientific Workflows in Taverna 2.2', in Proceedings of 2011 IEEE World Congress on Services, 2011, pp. 220-223

[14] Streit, A., Bala, P. Beck-Ratzka, A., Benedyczak, K., Bergmann, S., Breu, R., Ziegler, W., et al., 'UNICORE 6 – Recent and Future Advancements', in Annals of Telecommunications, Vol. 65(11-12), pp. 757-762

[15] G. Cavallaro, B2Share, 'Indian pines: raw and processed', B2SHARE, Online: http://hdl.handle.net/11304/7e8eec8e-ad61-11e4-ac7e-860aa0063d1f

[16] Dean, J., Ghemawat, S., 'MapReduce: Simplified Data Processing on Large Clusters', in Communications of the ACM, Vol 51(1), pp. 107-113, 2008

[17] M. Goetz, C. Bodenstein, 'Highly Parallizable DBSCAN', Online: http://www.fz-juelich.de/ias/jsc/EN/Research/DistributedComputing/DataAnalytics/Clustering/Clustering_node.html

[18] Wood J, et al.*,'Riding the wave - How Europe can gain from the rising tide of scientific data'*, European Union, Report of the High Level Expert Group of Scientific Data, 2010.

[19] piSVM Webpage, Online: http://pisvm.sourceforge.net/index.html

[20] Patwary, Md Mostofa Ali, et al. , *'A new scalable parallel dbscanalgorithm using the disjoint-set data structure'*, High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for. IEEE, 2012.

[21] Athanasopoulos, A., Dimou, A., Mezaris, V., Kompatsiaris, I., 'GPU acceleration for support vector machines', in 12th International Workshop on Image Analysis for Multimedia Interactive Services, 2011

[22] Cortes, C., Vapnik, V., 'Support-vector networks', in Machine Learning, Vol. 20, No. 3, pp. 273-297, 1995

[23] Morgan, M., Grimshaw, A., 'GENESIS-II Standards based Grid Computing', in proceedings of IEEE International Symposium on Cluster Computing and The Grid, 2007, pp. 611-618

[24] Elachi, C., Van Zyll, J., 'Introduction to the physics and techniques of remote sensing', Hoboken, N.J., Wiley-Interscience, 2006

[25] M. Riedel, A.S. Memon, M.S. Memon, *'High Productivity Data Processing Analytics Methods with Applications'*, In proceedings of the 37th International Convention on Information and Communication Technology, Electronics, and Microelectronics, MIPRO, Opatija, Croatia, 2014

[26] Oinn, T., Greenwood, M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M.R., Senger, M., Stevens, R., Wipat, A., and Wroe, C., 'Taverna: Lessons in creating a workflow environments for life sciences: Research articles', Concurrency Computation: Practice and Experience, Vol. 18, No. 10, 2006

[27] Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., Mock, S., 'Kepler: An extensible system for design and execution of scientific workflows', in 16th International Conference on Scientific and Statistical Database Management, 2004, pp. 423-424

Decomposition of Natural Join Based on Domain-Interval Fragmented Column Indices

Elena Ivanova and Leonid Sokolinsky
South Ural State University, Chelyabinsk, Russia
ivanovaev@susu.ac.ru, Leonid.Sokolinsky@susu.ru

Abstract - The paper describes decomposition of natural join relational operator based on the column indices and domain-interval fragmentation. This decomposition admits parallel executing the resource-intensive relational operators without data transfers. All column index fragments are stored in main memory in compressed form to conserve space. During the parallel execution of relational operators, compressed index fragments are loaded on different processor cores. These cores unpack fragments, perform relational operator and compress fragments of partial result, which is a set of keys. Partial results are merged in the resulting set of keys. DBMS use the resulting set of keys for building the resulting table. Described approach allows efficient parallel query processing for very large databases on modern computing cluster systems with many-core accelerators. A prototype of the DBMS coprocessor system was implemented using this technique. The results of computational experiments are presented. These results confirm the efficiency of proposed approach.

I. INTRODUCTION

Nowadays, human scientific and practical activities create the new challenges that demand big data processing. According to IDC study [1], the amount of digital data is doubling in size every two years, and by 2020 the digital universe – the amount of digital data created and replicated – will reach 44 zettabytes, or 44 trillion gigabytes. In 2013, only 22% of the information in the digital universe would be a candidate for analysis (useful if it were tagged) and less than 5% of that was actually analyzed. By 2020, the useful percentage could grow to more than 35%, mostly because of the growth of data from embedded systems.

One of the popular ways to process efficiently big data is using the parallel database system, which are able to process data in parallel on the high performance system with distributed memory [2] – [5]. The traditional approach for database storing is row-oriented representation. However, column-oriented database systems have been shown to perform more than an order of magnitude better than row-oriented database systems ("row-stores") on analytical workloads such as those found in data warehouses, decision support, and business intelligence applications. The elevator pitch behind this performance difference is straightforward: column-stores are more I/O efficient for read-only queries since they only have to read from disk (or from memory) those attributes accessed by a query [6]. Column-oriented databases are particularly well suited for compression because data of the same type is stored in consecutive sections. This makes it possible to use compression algorithms specifically tailored to patterns that are typical for the data type [7].

In recent years, more and more many-core processors are superseding sequential ones. Increasing parallelism, rather than increasing clock rate, has become the primary engine of processor performance growth, and this trend is likely to continue. Particularly, today's GPUs (Graphic Processing Units) and Intel's MIC (Many Integrated Cores), greatly outperforming traditional CPUs in arithmetic throughput and memory bandwidth, can use hundreds of parallel processor cores to execute tens of thousands of threads [8]. Recent trends in new hardware and architectures have gained considerable attention in the database community. Processing units such as GPU or MIC provide advanced capabilities for massively parallel computation. Database processing can take advantage of such units not only by exploiting this parallelism, e.g., in query operators (either as task or data parallelism), but also by offloading computation from the Central Processing Unit (CPU) to these coprocessors, saving CPU time for other tasks [9]. The many integrated cores of the Xeon Phi make this hardware accelerator a natural computing platform for an in-memory database engine or server. The database tables reside in the memory space of the MIC thus supporting fast in-memory database applications [10].

Main memory as the primary storage location is becoming increasingly attractive as a result of the decreasing cost/size ratio [7]. Main Memory Database (MMDB) eliminates disk access by storing and manipulating entire database in main memory. For performance-significant systems MMDB offer very low response time and very high throughput [11]. According to Gartner's 2013 Hype Cycle for Emerging Technologies report, in-memory database management system have 2 to 5 years until widespread adoption [12].

According to this, the problem of developing new efficient methods of parallel database processing in main

memory on modern compute clusters with many-core accelerators using column-oriented representation and data compression is important. To meet this goal, we offer a special type of index structures called distributed column indices. Distributed column indices allow to perform a decomposition of relational operators, which admits the efficient parallel execution of them on computing cluster system, equipped with many-core accelerators. In this paper, we consider the decomposition of the natural join operator. In this paper, we use the notation from [13]. The symbol $^\circ$ we use to denote the operation of concatenation of the tuples.

II. COLUMN INDEX

Let $R(A,B_1,\ldots,B_u)$ be the relation R with *virtual key* (virtual record identifier) A and the following attributes: $B_1,\ldots,B_u$. Tuples of R have length of $u+1$ and form of $(a,b_1,\ldots,b_u)$, where $a\in\mathbb{Z}_{\geq 0}$ and $\forall j\in\{1,\ldots,u\}\left(b_j\in\mathcal{D}_{B_j}\right)$. Here, $\mathcal{D}_{B_j}$ is the domain of attribute B_j. Let $r.B_j$ denote a value of attribute B_j. Let $r.A$ denote a value of the virtual key of tuple r: $r=(r.A,r.B_1,\ldots,r.B_u)$. The *virtual key* of relation R has the property: $\forall r',r''\in R\left(r'\neq r''\Leftrightarrow r'.A\neq r''.A\right)$. Define *tuple address* as a virtual key value of the tuple. To get the tuple by its address, we will use $\&_R$ *dereferencing function*: $\forall r\in R\left(\&_R(r.A)=r\right)$.

Let $R\left(A^*,B,\ldots\right), T(R)=n$ be given. Let a linear order be defined on set $\mathcal{D}_B$. The *column index* $I_{R.B}$ for attribute B of relation R is an ordered relation, which satisfies the following requirements:

$$T(I_{R.B})=n \text{ и } \pi_A\left(I_{R.B}\right)=\pi_A\left(R\right); \quad (1)$$

$$\forall x_1,x_2\in I_{R.B}\left(x_1\leq x_2\Leftrightarrow x_1.B\leq x_2.B\right); \quad (2)$$

$$\forall r\in R\left(\forall x\in I_{R.B}\left(r.A=x.A\Rightarrow r.B=x.B\right)\right). \quad (3)$$

Condition (1) means that the sets of virtual keys of column index and indexed relation are equal. Condition (2) means that index elements are sorted in ascending order of values of attribute B. Condition (3) means that attribute A of an index element contains the address of tuple of R, which has the same value of B attribute as the corresponding element of column index has.

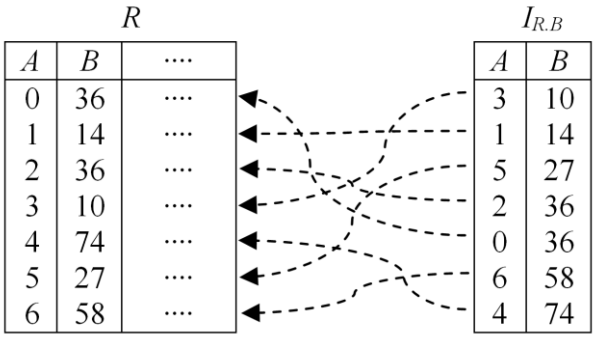


Figure 1. Column index

From the intensional point of view, the column index $I_{R.B}$ is a table with two columns A и B (see fig. 1). The number of rows in the column index is equal to the number of rows in the indexed table. Column B of index $I_{R.B}$ contains all the values of column B in table R (including duplicates). These values are sorted in ascending order inside column index.

III. DOMAIN-INTERVAL FRAGMENTATION

Let a total ordering relation be defined on domain $\mathcal{D}_B$. Divide $\mathcal{D}_B$ into $k>0$ nonintersecting intervals:

$$\left.\begin{array}{l} V_0=[v_0;v_1],V_1=(v_1;v_2],\ldots,V_{k-1}=(v_{k-1};v_k]; \\ v_0<v_1<\ldots<v_k; \\ \mathcal{D}_B=\bigcup_{i=0}^{k-1}V_i. \end{array}\right\}. \quad (4)$$

Define *interval fragmented function* on domain $\mathcal{D}_B$ as $\varphi_{\mathcal{D}_B}:\mathcal{D}_B\to\{0,\ldots,k-1\}$. This function satisfies the following requirement:

$$\forall i\in\{0,\ldots,k-1\}\left(\forall b\in\mathcal{D}_B\left(\varphi_{\mathcal{D}_B}(b)=i\Leftrightarrow b\in V_i\right)\right). \quad (5)$$

Let column index $I_{R.B}$ be given for relation $R\left(A^*,B,\ldots\right)$ with attribute B on domain $\mathcal{D}_B$. Let interval fragmented function $\varphi_{\mathcal{D}_B}$ be defined on domain $\mathcal{D}_B$. The function

$$\varphi_{I_{R.B}}:I_{R.B}\to\{0,\ldots,k-1\} \quad (6)$$

is *domain-interval fragmented function* for index $I_{R.B}$, if it satisfies the following requirement:

$$\forall x\in I_{R.B}\left(\varphi_{I_{R.B}}(x)=\varphi_{\mathcal{D}_B}(x.B)\right). \quad (7)$$

Define the ith fragment $\left(i=0,\ldots,k-1\right)$ of the index $I_{R.B}$ as

$$I^i_{R.B}=\left\{x\mid x\in I_{R.B};\ \varphi_{I_{R.B}}(x)=i\right\}. \quad (8)$$

It means that the ith fragment contains tuples, which have values of attribute B from the ith domain interval.
This fragmentation is called the *domain-interval fragmentation*. The number of fragments is the *degree of fragmentation*.

The domain-interval fragmentation has the following fundamental properties, which follow directly from its definition:

$$I_{R.B}=\bigcup_{i=0}^{k-1}I^i_{R.B}; \quad (9)$$

$$\forall i,j\in\{0,\ldots,k-1\}\left(i\neq j\Rightarrow I^i_{R.B}\cap I^j_{R.B}=\varnothing\right). \quad (10)$$

IV. Decomposition of the Natural Join Operator

Let two relations be given:

$$R\left(A^*, B_1, \ldots, B_u, C_1, \ldots, C_v\right)$$

and

$$S\left(A^*, B_1, \ldots, B_u, D_1, \ldots, D_w\right).$$

Let two sets of column indices be given for attributes $B_1, \ldots, B_u$:

$$I_{R.B_1}, \ldots, I_{R.B_u};$$

$$I_{S.B_1}, \ldots, I_{S.B_u}.$$

Let domain-interval fragmentation of degree k be defined for these indices:

$$I_{R.B_j} = \bigcup_{i=0}^{k-1} I^i_{R.B_j}; \quad (11)$$

$$I_{S.B_j} = \bigcup_{i=0}^{k-1} I^i_{S.B_j}. \quad (12)$$

Let

$$P^i_j = \pi_{I^i_{R.B_j}.A \to A_R,\, I^i_{S.B_j}.A \to A_S}\left(I^i_{R.B_j} \underset{I^i_{R.B_j}.B_j = I^i_{S.B_j}.B_j}{\bowtie} I^i_{S.B_j}\right) \quad (13)$$

for all $i = 0, \ldots, k-1$ and $j = 1, \ldots, u$. Define

$$P_j = \bigcup_{i=0}^{k-1} P^i_j. \quad (14)$$

Let

$$P = \bigcap_{j=1}^{u} P_j. \quad (15)$$

Define

$$Q = \{r.B_1, \ldots, r.B_u, r.C_1, \ldots, r.C_v, s.D_1, \ldots, s.D_w$$
$$r \in R \wedge s \in S \wedge (r.A, s.A) \in P\}.$$

Then

$$\pi_{*\setminus A}(R) \bowtie \pi_{*\setminus A}(S) = Q.$$

Note that calculation of P^i_j by (13) can be done in parallel on *k* different processors without data exchange. It ensures a near-linear speedup.

V. Performance Evaluation

The described approach was implemented as a prototype of DBMS coprocessor system. The source code of the program is openly available in the public GitHub repository [15]. Column indices and domain-interval fragmentation were evaluated using this prototype. We performed natural join operator over fragmented column indices $I_{R.B}$ and $I_{S.B}$. Attribute B of column index $I_{R.B}$ is a virtual key. Attribute B of column index $I_{S.B}$ is a foreign key. Join operator was implemented by merge join algorithm.

Number of column index tuples were following: $T(I_{R.B})$ = 600 000 and $T(I_{S.B})$ = 60 000 000. Column indices $I_{R.B}$ and $I_{S.B}$ were fragmented using domain-interval fragmentation. Every fragment of column indices was compressed by Zlib data compression library [14]. In described experiments, attribute B of column index $I_{S.B}$ had uniform and nonuniform (rule: 80/20, 65/20 and 45/20) data distribution.

The experiments were done on Intel Xeon Phi accelerator with 61 cores. Join operator was performed in 1, 2 and 4 threads per core (see fig. 2). The results of this experiment show that we have maximum speedup if we use 1 thread per core. The second experiment (see fig. 3) shows that load balancing can be effectively managed by increasing the number of fragments.

The experiments have shown that approach on base column indices allow to perform resource-intensive join operator for $T(I_{R.B})$ = 600 000 and $T(I_{S.B})$ = 60 000 000 during 1 second on one Intel Xeon Phi accelerator. The described approach eliminates data transfer, so we can expect a near-linear speedup on computing cluster systems with million nodes equipped with many-core accelerators.

VI. Related work

Binary table model was introduced in the paper [16]. On the basis of this model, several column-oriented DBMS were designed. As it was demonstrated by work [17] and [18], column-oriented systems offer an order-of-

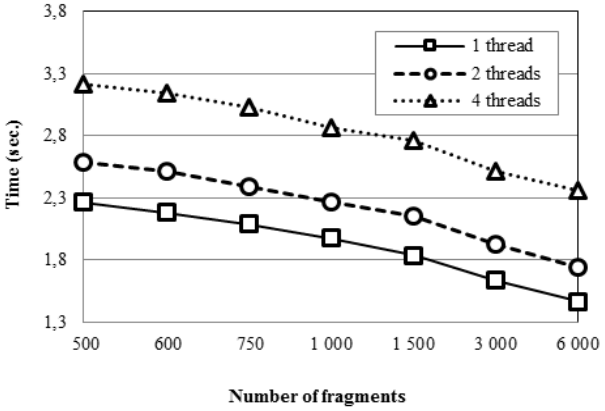


Figure 2. Dependence between time of calculation and number of fragments

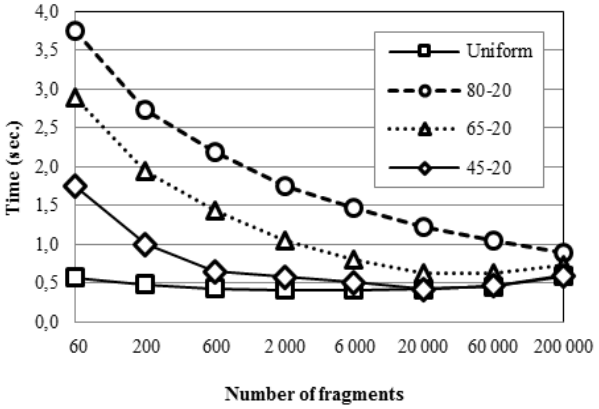


Figure 3. The influence of number of fragments on load balancing of Xeon Phi cores

magnitude performance improvement over traditional row-oriented systems for analytical processing workloads, such as those found in data warehouses or decision support systems. One of the main disadvantages of column-oriented DBMS is lacking the optimization technique, which is intrinsic to relational (row-oriented) DBMS. The work [19] investigated column-oriented simulation in a relational DBMS via the following techniques: vertical partitioning, index-only plans and materialized views. The investigation showed that such techniques do not improve the performance of row stores for analytical processing workloads. To overcome the problems faced with work [19], the work [20] introduced two new operators: Index Merge and Index Merge Join. The algorithms presented in this paper were designed specifically to take advantage of parallel processing whenever possible. Another approach was proposed in work [21]. This paper introduced a new index type, column store indexes, where data is stored column-wise in compressed form. Column store indexes are intended for data-warehousing workloads where queries typically process large numbers of rows but only a few columns. To further speed up such queries, the paper [21] also introduced a new query processing mode, batch processing, where operators process a batch of rows (in columnar format) at a time instead of a row at a time.

VII. CONCLUSIONS

In this article, we presented a decomposition of the natural join operator based on the column indices and the domain-interval fragmentation. Our approach was evaluated using the prototype DBMS coprocessor system. Experiments showed its efficiency for a resource-intensive natural join operator. Proposed approach can be used on computing cluster systems with many-core accelerators. Described technique is suitable for data warehouse workloads as well as for OLTP workloads.

As a direction of a future research, we are going to use described approach for the decomposition of another relational operators and compare speedup with existing DBMS.

ACKNOWLEDGMENT

The study was supported by the Ministry of education and science of Russia under Federal targeted program "Research and development in priority fields of scientific and technological complex of Russia in 2014-2020" (Governmental contract No. 14.574.21.0035).

REFERENCES

[1] V. Turner, J.F. Gantz, D. Reinsel, and S. Minton, "The digital universe of opportunities: rich data and the creasing value of the internet of things," IDC, White Paper, 2014. http://idcdocserv.com/1678 [January 20, 2015].

[2] L.B. Sokolinsky, "Survey of architectures of parallel database systems," Programming and Computer Software, vol. 30, No. 6, pp. 337-346, 2004.

[3] A.V. Lepikhov and L.B. Sokolinsky, "Query processing in a DBMS for cluster systems," Programming and Computer Software, vol. 36, No. 4, pp. 205-215, 2010.

[4] C.S. Pan, and M.L. Zymbler, "Taming elephants, or how to embed parallelism into PostgreSQL," in Processing of the 24th International Conference, DEXA 2013, Prague, Czech Republic, part I, vol. 8055, pp. 153-164, August, 2013.

[5] K.Y. Besedin, and P.S. Kostenetskiy, "Simulating of query processing on multiprocessor database systems with modern coprocessors," in Proceedings of the 37th International Convention, MIPRO 2014, Opatija, Croatia, pp. 1835-1837, May, 2014.

[6] D.J. Abadi, S.R. Madden, and N. Hachem, "Column-stores vs. row-stores: how different are they really?," in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, pp. 967-980, June, 2008.

[7] H. Plattner, and A. Zeier, In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, 2011.

[8] J. Fang, A.L. Varbanescu, and H. Sips, "Sesame: a user-transparent optimizing framework for many-core processors," in Proceedings of the 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid2013, Delft, Netherlands, pp. 70-73, May, 2013.

[9] S. Breß, F. Beier, H. Rauhe, K.-U. Sattler, E. Schallehn, and G. Saake, "Efficient Co-Processor Utilization in Database Query Processing," Information Systems, vol. 38, No. 8, pp. 1084–1096, 2013.

[10] M. Scherger, "Design of an In-Memory Database Engine Using Intel Xeon Phi Coprocessors," in Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA'14, Las Vegas, USA, pp. 21-27.

[11] P.A. Deshmukh, "Review on Main Memory Database," International Journal of Computer & Communication Technology, vol. 2, No. 7, pp. 54–58, 2011.

[12] H. LeHong, and J. Fenn, "Hype Cycle for Emerging Technologies," Gartner Inc, Research Report, 2013. http://www.gartner.com/doc/2571624 [December 16, 2014].

[13] H. Garcia-Molina, J.D. Ullman, and J. Widom, Database Systems: The Complete Book (2nd Edition). Prentice Hall, 2008. 1224 p.

[14] G. Roelofs, J. Gailly, and M. Adler. Zlib. http://www.zlib.net/ [April 4, 2015].

[15] A prototype of the DBMS coprocessor system using colomn indices based on domain-Interval fragmentation. https://github.com/elena-ivanova/colomnindices [April 4, 2015].

[16] G.P. Copeland, and S.N. Khoshafian, "A decomposition storage model," in Proc. SIGMOD, 1985, pp. 268-279.

[17] P.A. Boncz, M. Zukowski, and N. Nes, "MonetDB/X100: Hyper-Pipelining Query Execution," in Proc. CIDR, 2005, pp. 225-237.

[18] M. Stonebraker, D.J. Abadi, A. Batkin, X. Chen, M. Cherniack, M. Ferreira, E. Lau, A. Lin, S. Madden, E. O'Neil, P. O'Neil, A. Rasin, N. Tran, and S. Zdonik, "C-Store: A Column-Oriented DBMS," in Proc. VLDB, 2005, pp. 553-564.

[19] D.J. Abadi, S.R. Madden, and N. Hachem, "Column-Stores vs. Row-Stores: How Different Are They Really?," in Proc. SIGMOD, 2008, pp. 967-980.

[20] A. El-Helw, K.A. Ross, B. Bhattacharjee, C.A. Lang and G.A. Mihaila, "Column-oriented query processing for row stores," in Proc. DOLAP, 2011, pp. 67-74.

[21] P. Larson, C. Clinciu, E.N. Hanson, A. Oks, S.L. Price, S. Rangarajan, A. Surna and Q. Zhou, "SQL server column store indexes," in Proc. SIGMOD Conference, 2011, pp. 1177-1184.

Theoretical and statistical evaluation for approximate solution of large, over-determined, dense linear systems

Željko Jeričević* and Ivica Kožar**
* Department of Computer Engineering/Engineering Faculty, Rijeka, Croatia
** Department of Computer Modeling/Civil Engineering Faculty, Rijeka, Croatia
zeljko.jericevic@riteh.hr

Abstract - The solution of linear least squares system requires the solution of over-determined system of equations. For a large dense systems that requires prohibitive number of operations. We developed a novel numerical approach for finding an approximate solution of this problem if the system matrix is of a dense type. The method is based on Fourier or Hartley transform although any unitary, orthogonal transform which concentrates power in a small number of coefficients can be used. This is the strategy borrowed from digital signal processing where pruning off redundant information from spectra or filtering of selected information in frequency domain is the usual practice. For the least squares problem the procedure is to transform the linear system along the column to the frequency domain, generating a transformed system. The least significant portions in the transformed system are deleted as the whole rows, yielding a smaller, pruned system. The pruned system is solved in transform domain, yielding the approximate solution. The quality of approximate solution is compared against full system solution and differences are found to be on the level of numerical noise. Theoretical evaluation of the method relates the quality of approximation to the perturbation of eigenvalues of the system matrix.. Numerical experiments illustrating feasibility of the method and quality of the approximation at different noise levels, together with operations count are presented.

I. INTRODUCTION

The least squares method is one of the basic methods in data processing. In the case of linear models it comes down to solving over-determined systems of linear equations. For overdetermined system, usually none of the equations are satisfied exactly, but the sum of squared differences between observations and theoretical model is minimized.

Sparse systems are being solved by very efficient methods and problems with dimensions in millions are not uncommon. The dense problems are more problematic and rarely solved for dimension over 50000.

The paper extends the previous work [1] by extensive modeling and error analysis. Some equations developed previously [1] are briefly repeated here as final results without the details. The basic idea of constructing the approximate solutions for large, over-determined, dense systems using the Fourier space representation remains the same.

II. THEORY

The Fourier transform is formally done by multiplication with Fourier matrix but in actual computations FFT is used whenever possible.

$$b = Ac \tag{1}$$

Applying the Fourier transform on over-determined system of linear equations (1) is done along the columns. Premultiplying the (1) with Fourier matrix F, the transform along the column vector y and along the columns of matrix A is performed (2). It is a good idea that before the transformation, the elements of vector b and corresponding rows in matrix A are sorted in such a way that the transform has close to the most compact representation. That would concentrate the energy in the Fourier transform in as little number of frequencies as possible.

$$Fb = FAc \tag{2}$$

For the sake of simplicity Hartley transform [2] is used in the case of real systems to avoid computation with complex numbers. Noting that $Fb=d$ and $FA=D$ our original system (1) is now transformed with elements of vector d and columns in matrix D in frequency domain (3). This transformation does not affect the vector c which is still unknown solution for the systems.(1) & (3).

$$d = Dc \tag{3}$$

The over-determined system can be formally solved solved using normal equations and matrix inverse. Although that is not recommended for numerical reasons, it was formally used in [1] to develop feasibility of frequency selection and modeling of the linear algebra system [3-5].

III. METHOD

After the transformation the system (3) is of the same size as original system but is pruned down, possibly to a square matrix system. The selection of significant frequencies is accomplished by computing the magnitudes for frequencies in vector *d* and sorting them in decreasing order. Using that information, the vector *d* and matrix *D* can be shortened into significant parts to be retained and unsignificant parts to be discarded.

This would decrease the size (number of rows) of the system and represents the filtering out of non significant frequencies in order to build a smaller model system with less equations then the original system. In this respect our approach is different then Beylkin's [6] whose idea was to increase the sparsity of the matrix by using the wavelet transform. Using normal equations, residual matrix *R* and Taylor order expansion [1] it was shown that approximate inverse for system converges faster to exact normal equations inverse the smaller eigenvalues (λ) for (*I*-*R*) matrix are (*R* is residual matrix).

$$G = \left[D_0^T D_0 + D_1^T D_1 \right]$$

$$g_0 = D_0^T d_0 \quad G_0 = D_0^T D_0 \;\Rightarrow\; G_0 = G_0^T$$

$$G_0^{-1} \approx G^{-1}$$

$$R \equiv I - G_0^{-1} G \Rightarrow G_0^{-1} G = I - R$$

$$G = G_0 \left(I - R \right)$$

$$G^{-1} = \left(I - R \right)^{-1} G_0^{-1}$$

$$\delta c = \left(I - R \right)^{-1} G_0^{-1} \delta g$$

$$\left(I - R \right)^{-1} = \left(Q Q^{-1} - Q \Lambda Q^{-1} \right)^{-1}$$

$$= \left[Q \left(I - \Lambda \right) Q^{-1} \right]^{-1} = Q \left(I - \Lambda \right)^{-1} Q^{-1}$$

$$\lim_{\lambda \to 0} \left[Q \; diag \left(\frac{1}{1 - \lambda_i} \right)_{i=1,\ldots,n} Q^{-1} \right] \to I \qquad (4)$$

$$\left(I - R \right)^{-1} = I + R + R^2 + R^3 + \ldots$$

$$G_n^{-1} \equiv \left(I + R + \ldots + R^n \right) G_0^{-1}$$

$$G_\infty^{-1} \to G^{-1}$$

Where G_0 is approximation of $D^T D$ based on the most significant frequencies (computed using only the sub-matrix with index zero). Solving the model system will yield the approximate solution of original system (1). The approximate solution ($c+\delta c$) is different from true solution *c* for difference vector [7]. It is important to show how difference vector behaves depending on size of matrix G_0. The (4) shows that *R* should be a contraction mapping: its eigenvalues should all be non-negative, real and smaller then one. Smaller the eigenvalues are, closer $(I-R)^{-1}$ is to identity matrix as can be seen from the limit (4) and consequently closer G_0^{-1} is to G^{-1}.

The same system of linear equations as in [1] was again used here but with different levels of Gaussian noise. The least squares system [8] represents fitting of fourth degree polynomial. The numerical experiments are summarized in Tables I-V. Table I contains the results for data without noise, Tables II, III, IV and V contain the results for data with 1%, 5%, 10% and 20% noise respectively. Original system was of size 512 by 5.. The polynomial equation is:

$$P(x) = 24 - 14x - 13x^2 + 2x^3 + x^4 = 0$$

In the domain, 512 equidistant points in interval [-4, 3] were generated, and corresponding co-domain points were also generated and used as a vector of observations.

Normally distributed random noise mean equal zero and with variance of (1, 5, 10 and 20 percent) was added to vector y. For each noise level, the correct solution and series of approximative solutions using different number of frequencies (approximation levels) are presented.

TABLE I

#Rows	LSTSQ calculated polynomial coefficients 0% noise				
	C_0	C_1	C_2	C_3	C_4
512	23.9999	-13.9999	-13.0000	1.9999	1.0000
256	23.9999	-13.9999	-13.0000	1.9999	0.9999
128	24.0000	-14.0000	-13.0000	2.0000	1.0000
64	23.9999	-14.0000	-13.0000	2.0000	1.0000
32	23.9999	-13.9999	-13.0000	1.9999	1.0000
16	23.9999	-13.9999	-13.0000	1.9999	1.0000
8	23.9999	-13.9999	-13.0000	1.9999	1.0000

TABLE II

#Rows	LSTSQ calculated polynomial coefficients 1% noise				
	C_0	C_1	C_2	C_3	C_4
512	24.0078	-13.9948	-12.9967	1.9976	0.9991
256	24.0086	-13.9941	-12.9973	1.9975	0.9991
128	24.0089	-13.9936	-12.9974	1.9973	0.9991
64	24.0085	-13.9944	-12.9972	1.9975	0.9991
32	24.0072	-13.9962	-12.9963	1.9979	0.9991
16	24.0050	-13.9997	-12.9949	1.9987	0.9991
8	24.0053	-13.9952	-12.9945	1.9977	0.9987

TABLE III

#Rows	LSTSQ calculated polynomial coefficients 5% noise				
	C_0	C_1	C_2	C_3	C_4
512	23.8659	-13.9142	-12.8930	1.9787	0.9897
256	23. 8733	-13.9072	-12.8984	1.9772	0.9900
128	23. 8725	-13.9085	-12.8980	1.9774	0.9900
64	23. 8769	-13.9048	-12.9012	1.9766	0.9902
32	23.8682	-13.9162	-12.8953	1.9792	0.9900
16	23.8658	-13.9226	-12.8942	1.9806	0.9901
8	23.8658	-13.9117	-12.8924	1.9781	0.9896

TABLE IV

#Rows	LSTSQ calculated polynomial coefficients 10% noise				
	C_0	C_1	C_2	C_3	C_4
512	24.4527	-14.1579	-13.1657	2.0152	1.0080
256	24. 4699	-14.1363	-13.1776	2.0104	1.0084
128	24. 4663	-14.1358	-13.1743	2.0103	1.0081
64	24. 4633	-14.1439	-13.1729	2.0121	1.0082
32	24.4445	-14.1711	-13.1606	2.0182	1.0080
16	24.4489	-14.1766	-13.1654	2.0193	1.0085
8	24.4910	-14.1827	-13.2041	2.0204	1.0121

TABLE V

#Rows	LSTSQ calculated polynomial coefficients 20% noise				
	C_0	C_1	C_2	C_3	C_4
512	24.2089	-14.0300	-12.9389	2.0009	0.9925
256	24. 2056	-14.0615	-12.9411	2.0078	0.9936
128	24. 2486	-13.9938	-12.9684	1.9928	0.9940
64	24. 2471	-13..9951	-12.9673	19931	0.9939
32	24.2352	-14.0086	-12.9588	1.9962	0.9936
16	24.2353	-14.0229	-12.9612	1.9993	0.9941
8	24.3457	-13.8100	-13.0259	1.9513	01.9936

The data in Tables I-V show remarkable robustness with respect to reduction in system size. There is no sudden degradation in solution quality for series of approximations and it seems possible to approach the square system size as shown in Figure 1.

IV. CONLUSION

The approximate method for a fast solution of linear least squares problems has been proposed. The quality of the approximation can be tested and controlled. We developed the equation which shows how the eigenvalues of residual matrix determine quality of solution and extent of corrections to be applied in order to approach the true solution. The conclusions are also supported by extensive numerical simulations using different noise levels.

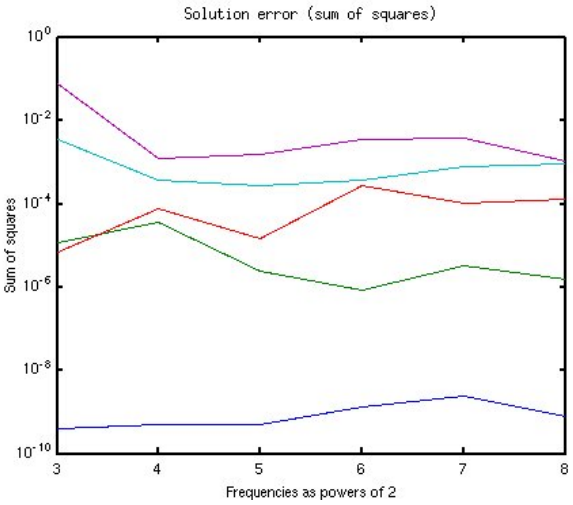


Figure 1. Plot showing the sum of squared errors in solution vector as function of the number of frequencies used to compute the approximate solution. The curves from bottom up are: no error, 1% , 5%, 10%, 20% error. Solution computed using all frequencies is used as correct solution

REFERENCES

[1] Jeričević Ž., Kožar I, "Faster Solution of Large, Over-Determined, Dense Linear Systems", DC VIS MIPRO (2013) 228-231.

[2] Bracewell, R.N., "The Hartley Transform", Oxford University Press, New York, 1986 p 160.

[3] Jeričević, Ž., "Approximate Solution of Linar Systems" Croatica Chemica Acta, Vol 78 (2005) 601-615

[4] Kožar I, Jeričević Ž and Pecak, T 2009 "Approximate solution of eigenvalue problem for modal analys Conf. of Croatian Society of Mechanics (ICCSM) : Book of Abstracts ed I Smojver and J Sorić (Dubrovnik Croatia) p 133

[5] Jeričević Ž., Kožar I., and Pecak, T. "Modeling of Linear Algebra System with the Fourier Transform", MIPRO (2010) GVS, 231-235

[6] Beylkin, G., Coifman, R., and Rokhlin, V, "Fast Wavelet Transforms and Numerical Algorithms", Com. Pure App Math. Vol. 44 (1991) 141–183

[7] Press, W.H. et all, "Numerical Recipes", 3rd Ed., 2007 Cambridge University Press, Cambridge, pp 1235

[8] Lawson, C.L. and Hanson, R.J. "Solving Least Squares Problems", 1974, Prentice-Hall, Englewood Cliffs, pp 340

DARIAH Meta Hosting: Sharing Software in a Distributed Infrastructure

Jedrzej Rybicki and Benedikt von St. Vieth
Forschungszentrum Jülich GmbH, Jülich, Germany
{j.rybicki, b.von.st.vieth}@fz-juelich.de

Abstract - Research infrastructures have become an everyday tool for doing science. They constitute a cost-effective, quick, and increasingly easy-to-use collaboration tool. So far the focus was on sharing resources (especially data) and on offering a (rigid) set of services for processing and accessing the resources. It becomes clear, however, that there is a demand from the users to share not only the data they have gathered or created but also the software they implemented. Such a sharing has the potential to speed-up the scientific discovery. But only if the software runs i.e., if it can be applied by other researchers to new problems. Unfortunately, it happens often that the software implemented in a project is only understandable and deployable by the authors. In this paper we will address the problem of sharing services (e.g. new data analysis tools) between users and also touch on the problem of sharing services between infrastructures, for instance to facilitate cross-disciplinary exchange. Our goal was to increase the sustainability of the developed research software by enabling an easy extension of research infrastructures beyond the rigid set of services towards flexible Software-as-a-Service (SaaS) solutions. We share the initial experiences gained during the implementation of such a meta hosting service for DARIAH-DE research infrastructure.

I. INTRODUCTION

Digital methods have become commonplace in modern science. Their popularity is amplified by the recent establishment of the so-called data-driven science, which is expected to provide scientific insights by applying software solutions on digital data. The obvious prerequisite for such approaches is the availability and accessibility of the data. A problem which is acknowledge and addressed (if yet-not-completely-solved). The less obvious problem is how to share the scientific software used for analyzing the data. Such software is often a custom-made solution, developed in the course of a research project as an after-effect of answering concrete scientific questions. Other researchers might, nevertheless, benefit from using it to solve similar problems or for the validation of the results by applying the software to a different set of data [1]. Sharing software has the potential to accelerate scientific discovery. But even more important is the fact that software which is not sharable is, in the long run, not sustainable. Scientific discoveries done with such a software are, in worst case, not reproducable and thus useless. We believe that sharing data and software is a corner stone for doing science today. In this paper we will brush aside the social challenges of sharing software and data, and concentrate on technical solutions to facilitate such sharing.

It should be stressed that the open software repositories like *GitHub* or *sourceforge* do not solve the stated problem completely. The software is only usable when it runs. Installing, configuring, and satisfying the software dependencies often proves to be harder than expected. Especially when it is conducted on a different platform than it was developed on. Furthermore the knowledge on how to perform these steps is often only present in the heads of the software authors or within the project it was developed for (e.g. closed wiki). Hence the first requirement for the effective software sharing is the availability of the abstract, platform-independent, yet executable deployment descriptions. Such descriptions could be, for instance, system images, executable installation scripts, configuration management tools, or container-based solutions. We will discuss the applicability of these approaches in Section II.

The proliferation of the distributed research infrastructures like [2] and [3] adds a new dimension to the problem of sharing scientific software. Such infrastructures offer a quick and cost-effective way of sharing resources. They excel at enabling access to data and often provide some compute resources. The later is offered either in a Infrastructure-as-a-Service or Software-as-a-Service manner (for the definition of the terms see [4]). None of this approaches is optimal from the user perspective. IaaS solutions put quite a large burden on the users: they have to install, configure, and maintain not only the software they require but also the whole system (e.g. do system upgrades). Furthermore, the created instances cannot be easily shared: this remains in power of the user who created them and is usually technically challenging. The higher-level abstractions in form of Software-as-a-Service are much more attractive for researchers: The effort on their part is reduced and they can directly apply the software in their scientific endeavors. But the effort still exists: The installation work has to be done by the computer centers in the infrastructure. From software sharing perspective the problems are twofold. Firstly due to the rapid development of digital methods, researchers constantly require new versions or new types of software in the SaaS portfolio. Secondly, the distributed research infrastructures span across multiple resource providers, thus high level of heterogeneity has to be coped with, making the installation harder. To this end, the demanded software deployment descriptions must be platform-independent and portable. To account for the constantly changing requirements, the descriptions should be executable. Finally the system administrators should be able to audit them to verify if they are correct, trustworthy, and secure. In this paper we will show how

the deployment descriptions can be used to build a dynamic Software-as-a-Service solution, which allows for an easy extension of the offered software portfolio

The high development's pace of cooperative science together with the increasing importance of interdisciplinarity, bears one more use case for sharing research software: exchanging software between e-Infrastructures. This use case reinforces the previously mentioned requirement of making software deployment descriptions executable, auditable, portable, and sharable. The potential users of research software from different disciplines will have to invest some time to understand how to use the software. The time should not be increased by making the user to understand how to install, and deploy the software.

This paper is structured as follows. Related work is discussed in Section II. We provide a short introduction in the technology which we will use as a basis for the software sharing in Section III. Section IV presents how we used executable software deployment descriptions to implement a extensible Software-as-a-Service solution for DARIAH-DE. We conclude our paper with a summary in Section V.

II. Related Work

Let us now reflect on how software reusing is typically achieved nowadays. In the first step the scientific software has to be made discoverable. This step involves some social challenges, for instance willingness to share and use software, or efficient communication between the parties, involving software citation. In this paper, however, we focus on the technical challenges and claim that open-source repositories are sufficient for discovery. As stated above, the software is only useful if it can be applied to a new problem, and this is only possible when the software runs. Hence, the discovered software has to be installed. The process is conducted either locally on the researcher's machine or (in a more modern fashion) on a remote machine running somewhere in the cloud e-Infrastructure. The installation usually involves the configuration of the operating system (OS) and the provision of the run-time environment, additional libraries, and actual software. The knowledge on how to achieve this might be, to some extend, conveyed in the service documentation. The experience shows that the documentation is seldom up-to-date and does not cover all the corner cases encountered in the real-life deployments e.g., library dependencies or support for differing operating systems. When it comes to a migration of the service, the process must be started from scratch again, often by a different person. Even such a down-to-earth, recurring situation like the OS upgrade for the underlying machine might render a service unusable and result in the need for a re-installation. Since the process described above is quite laborious, let us now discuss ways in which it can be either simplified or done in such a way that other researchers can use the software without repeating the installation process.

A. Virtual machines

One way of sharing "already-installed" software is to prepare images from which virtual machines can be created. The most obvious case for such applications are clouds offering an IaaS interface [5][6], although local deployments of images are conceivable as well. VM image preparation involves repeating the above steps in a virtual machine while accounting for special properties of the given cloud. The later is often subsumed under the term: contextualization. When an image is created and uploaded to a cloud image repository it can be instantiated by other users to obtain running version of the application. VM images suffer, however, under some limitations. First of all, images are black boxes: one has to trust the creator and has very limited capabilities of auditing the actual content of the image. Thus it is hard to safe-guard that the images do not include some malicious software or are not misconfigured so that running instances can be easily hacked or destroyed by malicious third parties. Virtual images include a complete software stack comprised of the required software, dependencies, and operating system. This results in a high overhead in terms of both image size (what makes the sharing of images harder) and performance. The inclusion of the operating system results in the need for periodic updates of the image: for instance to apply new security patches. The further problem with images is, that they are infrastructure-specific: they strongly depend on the hypervisor used and contextualization solution provided in the particular cloud. Clouds include their own image repositories but moving images between the infrastructures i.e. between different clouds, remains an unsolved problem. A problem with obvious ramification for software sharing.

B. Configuration management tools

One way to create trustworthy and reproducible service deployments that are, to some extent, infrastructure, hypervisor, and OS agnostic, is to use configuration management tools like Puppet [7] or Chef [8]. With these tools it is possible to describe the service deployment in a declarative way. One defines which packages should be installed, which configuration changes should take place or which commands should be executed instead of applying all these steps manually. The tools can be used to provide a generic description of the service deployment that interested partners can apply to virtual machines or servers as the configuration management tools do not offer any isolation nor virtualization themselves. Here the same limitations with respect to overhead of running VMs apply. On the upside, the descriptions are human-readable and can be reviewed before the actual deployment, this contributes to their trustworthiness. Configuration management tools enable reproducible service deployment, but they fail to some extend at abstracting the underlaying platform, for this the developer has to spent additional effort.

One example for this is the deployment of a webserver like Apache HTTP server [9], a software often used to provide webservices. Listing 1 shows an excerpt from a deployment description that is used to deploy the software by installing (section *package* on Listing 1), configuring (*file*), and starting (*service*) it. Puppet brings an abstractions for package managers and it also provides wrappers for different *init* systems, but it has problems with the different package, path, and service naming. To keep it platform-independent and portable, one has to

prepare deployment descriptions for all the supported platforms. The first lines of the Listing 1 show how to coped with the fact that the same software is available under different names in RedHat and Debian repositories.

```
if $::osfamily == 'RedHat'{
   $user = 'apache'
   $name = 'httpd'
 } elsif $::osfamily == 'Debian' {
   $user = 'www-data'
   $name = 'apache2'
 } else {
   fail("Unsupported os: ${::osfamily}")
 }
 file { "/etc/$name/$name.conf":
   owner => $user
   ensure => 'present'
 }
 package { $name:
   ensure => 'latest'
 }
 service { $name:
   ensure   => 'running'
 }
```

Listing 1. Puppet example for deploying Apache httpd,

We establish that tools like Puppet solve a slightly different problem. They are perfectly suited for managing the configuration of software running in a distributed infrastructure. For this purpose, they allow for configuration changes of the running machines (such changes are then automatically loaded on all relevant machines). The scientific software sharing, on the other hand, will be rather used to create quickly disposable instances of given services. This instances will be used e.g. to verify the researcher's hypothesis or validate previous results and they will be discarded afterwards. We argue that there will be no need to manage the configuration changes in one instance, for the modified setup new instance will be spun off. Researcher will not be willing to spent much effort into the configuration management when it would be possible to simply, quickly, and cheaply create new instances.

III. CONTAINER-BASED VIRTUALIZATION

Container-based virtualizations like Linux VServer [10] or Docker [11] became very popular for building, shipping, and running applications across many machines. They constitute the perfect basis for an efficient sharing of scientific software. In this section we will first provide the reader with some basic information on how Docker works, and argue why it is better than the previously described system images and configuration management tools. The section will be concluded with a proposal on how the software sharing life cycle could be implemented in a distributed infrastructure.

A. Introduction to Docker

Docker is used as the basis for our software-sharing solution. This section will equip the reader with enough information about this technology to understand the rest of the paper. Readers interested in more details are referred to the extensive Docker documentation [11].

Docker is a lightweight, open-source, virtualization solution. In opposite to the full-stack virtualization solutions, Docker images do not include a guest operating system kernel. It lowers the overhead in terms of both performance and size of the images allowing near-native performance. Docker is based on mature Linux kernel technologies (*cgroups* and *namespaces*, among the others) to isolate independent application containers. Docker images use AuFS which is a layered file system, enabling sharing libraries between images on one hand, and inspecting the changes done in the images, on the other. The later feature permits rudimentary provenance tracking of images.

Let us now briefly discuss the Docker terminology and usage. Applications running with Docker are called containers, they are created from images. Images should be understood as hierarchical templates. It is possible to start with one image, add some software, and save the result as a new image. It is also easily possible to share images by using image repositories either private or public ones, like the *Docker Hub*. There are two main ways of shipping applications with Docker. In the first one (let us call it interactive), the developer starts a basis Docker image (e.g.: official Debian Linux image) and installs application, its dependencies, and everything else by issuing ordinary Linux commands. As soon as the application is installed, a snapshot of the running container can be created. Such a snapshot is, in fact, an image from which new containers can be created. Another way of creating images follows a declarative approach. The developer can describe the installation steps in a special *Dockerfile*. An image is created from such a description by issuing the *docker build* command. A generalization of the declarative approach are automatic builds. To setup an automatic build one has to publishing the *Dockerfile* in a public code repository (like *GitHub*). Afterwards a web hook has to be created, so that each time the *Dockerfile* is modified, a new image build is automatically triggered. Regardless of the way in which an image was created, it is possible to view all the changes done in the image during the installation of the software. The automatic builds go an extra mile to improve the trustworthiness of images: each user can review not only the content of the image but also the way they were created, since the description is publicly available.

Docker has a very important abstraction of data volumes which allows to separate software and data. The data are not included in the image, a problem often encountered in VM images. Data volumes, on the other hand, enable an easy injection of the data into the running containers. Hence the software in the container can easily be used for analyzing various sets of data. The separation of the software and the data diminishes one of the common barrier for sharing software. Researcher are afraid of leaking out the data in the software they made available.

Docker is available on almost all Linux-based platforms, there is a support for MacOS and recently announced support on Windows platforms. Since the images encapsulates the application and all the dependencies, starting new containers on a new system is like like starting statically linked, self-contained programs: no modification in the host system are required.

B. DARIAH-DE software sharing life cycle

DARIAH, the Digital Research Infrastructure for the Arts and Humanities, aims to enhance and support digitally-enabled research and teaching across the arts and humanities. It comprises a number of national initiatives, there is also a Germany-based effort, DARIAH-DE [2], where the described work was done. Efficient software sharing is essential for digitally-enabled research of its users. Also sharing software between infrastructures is becoming relevant in this context [12]. Software developed in one part of DARIAH might be passed over to other national efforts.

The technology for creating executable and deployable descriptions of software must be integrated into the scientific workflows and play well with existing e-Infrastructures. On Fig. 1 we present how such a integration could be conducted. We differentiate between three roles: developers, infrastructure operators, and researchers. We have also two central components: a code repository, where the deployment descriptions are stored, and Docker Hub repository where the images are stored and pulled from.

The process of making scientific software available kicks off with the developer wanting to advertise and make his software available. He can propose the deployment description by forking the official *GitHub* repository of DARIAH-DE Docker files, adding a new description (or updating an existing one), and creating a pull request[1]. The operators of the repository are notified about the the pull request. They can review the proposed solution, adjust it when required, or even reject it if it does not work. The *GitHub* repository is connected to the Docker Hub via web hooks. In short, this mechanism triggers an HTTP request to the Docker Hub each time the *GitHub* repository is modified. Upon such a request, a new build of the provided Dockerfile is started. The infrastructure operators can view the details of the build and act accordingly in case of a failure. Successfully built images are published into the official DARIAH-DE image repository in the Docker Hub. The researchers can explore the repository, select the software they are interested in, pull the images and start it on their machines. Pulling and starting requires just one command (*docker run*). The Docker Hub allows for storing additional information about the software. The developer can describe the typical applications or provide a link to a more extensive documentation.

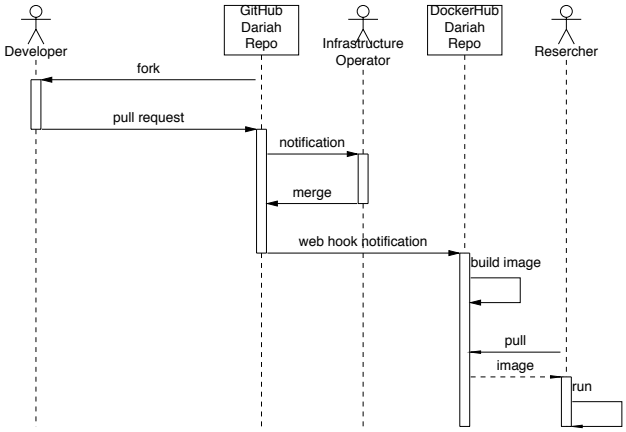


Figure 1. Proposed software sharing life cycle

The workflow described above can be repeated multiple times e.g., each time a new version of the software is made available. Moreover, the roles can be distributed or handed-over. Both repositories allow for registration of organizations. Such organizations can have multiple admins. Hence it is not required to have just one infrastructure operator but rather a group of people can be delegated for this task. Similarly, all the *GitHub* users can fork, and modify the Dockerfiles. The trustworthiness of the images is guaranteed by the reviews conducted by the operator who has to accept the pull requests. Of course each researcher can view the publicly available descriptions. The images (and Dockerfiles) are publicly available and they can be used by the researchers from outside of DARIAH-DE.

Fig. 1 depicts the simplest workflow, which ends with a local deployment of the software. It is sufficient to explain how the workflow works but clearly the local deployment is not always the optimal solution. In the next section we will show how a dynamic SaaS solution can be build, based on the proposed workflow.

IV. IMPLEMENTING META HOSTING

We have shown that Docker provides a means for a fast deployment of software. In the optimal case, the user has to issue just one command to obtain a running instance on her system. The optimal case requires some changes on the machine, at least the installation of Docker is required. Even if easily possible, the local deployments are not always the best option. One reason for remote deployments is efficiency: remote instances can run close to the resources they require. Many data-driven projects require long-lasting software runs. This is hard to perform on the local researcher's machine. Fortunately, e-Infrastructures like DARIAH-DE provide resources which can be used by the researchers for this purpose.

[1] Readers not familiar with the *GitHub* commands are referred to the documentation [13]

In this section we will show how the presence of deployable software description (as those provided by Docker) can facilitate a dynamic Software-as-a-Service solution. This service will offer all the benefits of the SaaS approach (i.e. very low effort is required to use scientific software shared by other researchers) whilst allowing for quick extensions of the software portfolio. We present our prototype solution which runs on low-level compute resources already available in the DARIAH-DE infrastructure. Our goal was a system which provides the researchers with an instance of the software she requires in just one request through a web page.

A. Actors

Fig. 2 depicts the architecture of the implemented dynamic Software-as-a-Service. It includes lots of components loosely coupled with each other by a messaging bus. Messaging is used for exchanging information between components in an asynchronous way. In production we use a *amqp*-based product (RabbitMQ [14]). Let us briefly discuss the roles and responsibilities of the single elements of our architecture.

The *Facade* is the entity facing the end-users. It is an entry point from which the commands (like create a new instance of a given software product) are issued. The Facade is stateless and it is possible to run multiple Facades. They can offer interfaces of different types, e.g.: a html-based for browser access or a json-based programmatic API.

The Facade has access to the messaging bus and a read-only access to the *Store*. Store is a system-wide cache with information about:

1. types of service registered in the system,
2. instances running in the system.

The content of the cache is updated by the *Store Updaters*. These entities subscribe for information about creation of new instances and registration of new types. Upon reception of such messages, Store Updaters write the information in the Store. It should be stressed that the content of the Store itself can be lost as it will be periodically republished. Store is solely used to speed-up the response times.

The *Workers* embody the main functionality of the system. Those are the entities which can manage instances of given service types (e.g. instances of eXist or neo4j databases). Each Worker upon its creation (and later periodically) informs other components about its capabilities (i.e. supported service type) and its current state (i.e. managed instances). Workers receive commands from users (via Facade) through the messaging system.

B. Communication with workers

The critical part of our system is the communication between the components. It is dynamic, asynchronous, and subject-based. We required the possibility to change the system capabilities during the run-time. In particular, to add new software to the portfolio of the offered Software-as-a-Service. This is done by adding (or removing) Workers. To become active in the system,

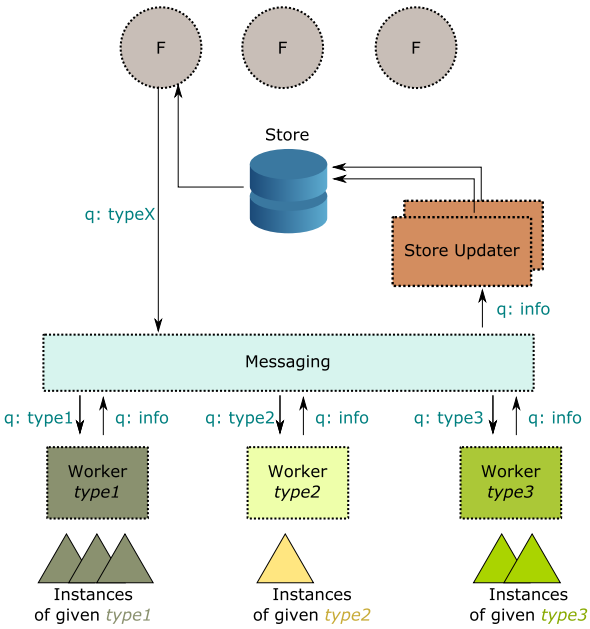


Figure 2. Architecture of the dynamic Software-as-a-Service

Worker needs access to the messaging bus. The incoming Worker publishes its "specialty" (i.e. instance type that it can provide) and subscribes for the messages in this instance-type-queue (like create new instance, delete instance). As can be seen on Fig. 2, for each instance type supported by the system, there is a separate queue. When a user requires a new instance of a given type it sends a message to the respective queue (via Facade). The message is routed to the proper Worker, it creates an instance and sends a response (via messaging) with details required for connecting to the just-created instance. Such details typically include IP address, protocol and port, username and password, etc. Workers republish their state periodically (i.e. instance types that they support and state of managed instances). Each worker has a dedicated queue for the incoming commands and all workers share a common response queue (the *info* queue on Fig. 2).

C. Implementation and operation

We have created a prototype of the dynamic Software-as-a-Service called DARIAH Meta Hosting. The implementation was done in Python and uses a set of frameworks and well-known products: Flask for the frontend, MongoDB as storage, and RabbitMQ for messaging. We offer two interfaces: one is a human-friendly web site, the other is a programmatic json-based API. We envision the later one to be used to script automatic deployments of software e.g. in data processing workflows.

One feature of our design was the division of the system into small, autonomous parts connected by messaging. The advantage of such a design was the ability to change the parts independently, for instance to scale up. We use Docker for deploying single parts and the Docker-based orchestration tool *fig* [15] to create test deployments including all the parts required. Fig is also able to scale out the system by spinning off additional

instances. In this self-test, the solution we use for the software sharing was successfully applied for the software we have implemented. Our prototype is deployed on OpenStack Infrastructure-as-a-Service cloud which we offer in DARIAH-DE. So far the complete system is placed in one data center but it should be possible to extend it beyond this one entity. This would only require to make the messaging accessible from the outside and deploy additional workers in different data centers.

For our prototype we have implemented a generic Docker worker. The deployment of Docker-based instances is usually conducted in the same fashion, regardless of what software is confined in the image. Hence the Worker is initialized with a name of the Docker image containing the software it will be responsible for. We create an instance of Worker for each software package offered in the dynamic Software-as-a-Service. To scale out the system it is also possible to have multiple workers offering the same type of the software. The requests will be then divided among the workers in a Round-robin fashion, allowing for fair load distribution.

At this stage it should be stressed, that the proposed architecture is extensible. The Workers abstract the technology used for creating the instances. It would be perfectly possible to write a worker which would use Puppet-based deployment descriptions or any other technology that might come around in the future.

V. CONLUSION

The motivation for this paper was the problem of scientific software sharing. A solution for that problem is crucial for the further development of reproducible data-driven science. We have argued that the problem is manyfold. It involves both technical and social challenges. In this paper we have focused on the former ones and firstly reviewed the technologies which could be used for the software sharing. Subsequently we have presented a workflow for sharing software in the context of e-Infrastructures like DARIAH-DE. In the second part of our paper we have shown how the presented software sharing solution can be used to leverage a dynamic Software-as-a-Service system. It enables a quick and easy deployment of the software on IaaS resources available in the e-Infrastructure. The proposed architecture is extensible and we have also shared some experiences gained during the implementation and operation of a working prototype.

ACKNOWLEDGMENT

The work has been supported by DARIAH-DE, which is partially funded by the German Federal Ministry of Education and Research (BMBF), fund number 01UG1110A-M.

REFERENCES

[1] C. Goble, “Better software, better research,” IEEE Internet Computing, vol. 18, no. 5, pp. 4–8, Sept 2014.

[2] T. Blanke, M. Bryant, M. Hedges, A. Aschenbrenner, and M. Priddy, “Preparing DARIAH,” in IEEE 7th International Conference on E-Science, Dec 2011, pp. 158–165.

[3] D. Lecarpentier, P. Wittenburg, W. Elbers, A. Michelini, R. Kanso, P. Coveney, and R. Baxter, “EUDAT: A new cross-disciplinary data infrastructure for science,” International Journal of Digital Curation, vol. 8, no. 1, pp. 279–287, 2013.

[4] P. Mell and T. Grance, “The NIST definition of cloud computing,” National Institute of Standards and Technology, Tech. Rep. 800-145, 2011.

[5] “Amazon Web Services,” http://aws.amazon.com/.

[6] “OpenStack,” http://openstack.org/.

[7] “Puppet,” http://puppetlabs.com/.

[8] “Chef,” https://www.chef.io/.

[9] “Apache HTTP server project,” http://httpd.apache.org/.

[10] “Linux-Vserver,” http://linux-vserver.org/.

[11] “Docker,” https://www.docker.com/.

[12] M. Hedges, H. Neuroth, K. M. Smith, T. Blanke, L. Romary, M. Küster, and M. Illingworth, “TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship,” Journal of the Text Encoding Initiative, Jun 2013.

[13] “GitHub,” https://github.com/.

[14] “RabbitMQ,” http://www.rabbitmq.com/.

[15] “Fig: Fast, isolated development environments using Docker,” http://www.fig.sh/.

Building and Provisioning Bioinformatics Environments on Public and Private Clouds

Enis Afgan[1,2,4], Konstantinos Krampis[3], Nuwan Goonasekera[4], Karolj Skala[1] and James Taylor[2]

[1]Center for Informatics and Computing (CIR)
Ruđer Bošković Institute (RBI)
Zagreb, Croatia

[2]Department of Biology
Johns Hopkins University
Baltimore, MD, USA

[3]Dept. of Biological Sciences, Hunter College
City University of New York,
New York, NY, USA

[4]Victorian Life Sciences Computation Center
University of Melbourne
Melbourne, Australia

Abstract - Unlike newly developed web applications that can be designed from the ground up to utilize cloud APIs and run natively within cloud infrastructure, most complex bioinformatics pipelines that are in advanced states of development can only be encapsulated within VMs along with all their software and data dependencies. To take advantage of the scalability offered by the cloud, additional frameworks are required to create virtualized compute clusters and emulate the most common infrastructure found on institutional resources where most of the existing bioinformatics pipelines are generally run. In this paper we describe one such framework, its compatibility with multiple Clouds and present an automated process for deploying the entire system so it can be made easily available on any Cloud.

I. INTRODUCTION

The adoption of cloud computing technologies has provided an opportunity to effectively democratize the field of computational biology, meaning that individual researchers and labs can now have access to the resources that were previously only available to the large sequencing and bioinformatics centers [1]. However, this presents challenges of its own, such as the need to configure and utilize bioinformatics software for next-generation sequencing, where significant expertise is required on UNIX based operating systems, programming languages, software compilation from source code, file systems, and large-scale data management [2]. For smaller laboratories this can become a significant obstacle, because in addition to coming up with the funds for building an informatics infrastructure with the capacity to handle large computations, they also need to hire trained bioinformaticians competent to install, configure and run sequence analysis software tools and manage the data. This can present a higher expense than that of acquiring the hardware.

The cloud computing model has showcased its ability to transform how access to compute resources is realized and has delivered on the notion of Infrastructure-as-Code [3]. However, this model presents the user with low-level resources that still need to be procured, configured, and managed. In the context of bioinformatics, this means installing required tools and reference data and setting up a virtualized cluster environment. As data volumes increase, this also requires utilizing compute and storage resources in a scalable manner. To address this, we have been developing CloudMan [4] as a versatile solution for enabling and managing compute clusters in cloud environments via a simple GUI web interface or a programmable API. As a full-fledged deployment, CloudMan integrates with the Galaxy framework [5], [6] and provides a comprehensive workbench for biomedical data analysis on the Cloud [7]; it enables one to seamlessly acquire cloud resources, assemble those into a compute cluster, upload data, perform data analysis by chaining output of one tool as the input of another, visualize the data, and finally share it. This is the end-user view. Internally, CloudMan manages cloud resources, operating system services, and application-level services.

As part of this article, we describe how we have extended CloudMan into an interoperable solution for multiple Clouds and built a framework for allowing anyone to deploy their own version. We describe the process for building the required cloud resources, how to launch an instance of the deployed components and how to configure CloudMan to manage a launched instance and correctly orchestrate any required runtime components. Once deployed, users will have a functional bioinformatics workbench and a dynamically scalable cluster for data analysis on a given cloud.

II. BACKGROUND

In recent years, cloud based bioinformatics data analysis systems such as Galaxy [7], CloVR [8], Cloud BioLinux [9] and bioKepler [10] were released, allowing smaller laboratories and institutions to perform large-scale data analysis with genomic datasets. All these platforms are available on the Amazon Elastic Compute (EC2) cloud, which provides data centers across the world, allowing any researcher worldwide to connect to the cloud end-point closest to their geographic boundaries.

Researchers or institutions with access to computing resources at their home institution have the option to implement their own private Cloud platform and run Virtual Machine (VM) servers, using the open-source Eucalyptus [11] or OpenStack cloud middleware. OpenStack is the official Cloud platform bundled with the Ubuntu Linux operating system [1] and can be readily installed on clusters running other Linux distributions[2],

[1] http://www.zdnet.com/article/ubuntu-syncs-up-with-
[2] http://docs.openstack.org/

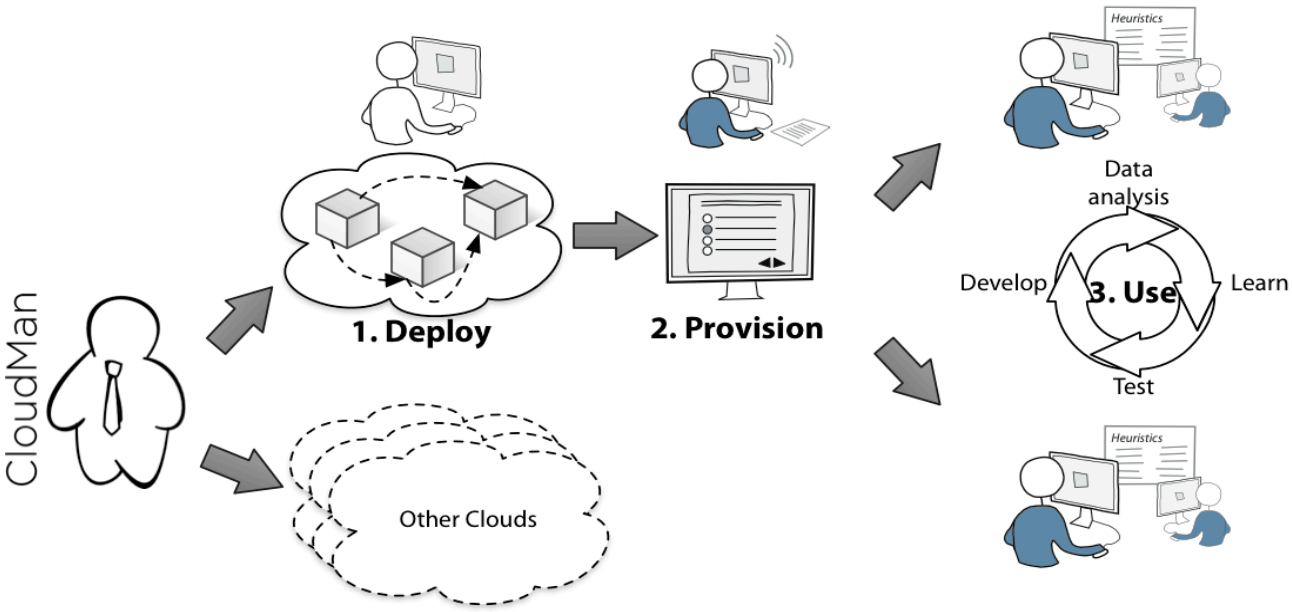


Figure 1. The effort described in this paper highlights the features of CloudMan as a platform for building and provisioning bioinformatics workbenches across multiple Clouds. With the multi-cloud compatibility of CloudMan (Section 3.A), **(1)** a deployer would use the provided automation methods to build and link the required cloud resources, tailored to the needs of the given research community (Section 3.B). Then, **(2)** a group leader, system administrator or even a domain researcher uses a launcher app to provision any number of instances of the workbench and possibly further customize those with custom data or tools (Section 3.C). Finally, **(3)** domain researchers use the workbench in their daily routine.

and Eucalyptus is similar. Both of these Cloud platforms offer Application Programming Interfaces (APIs) compatible with Amazon's EC2 service, and applications developed either on Eucalyptus or OpenStack work on the Amazon Cloud and vice-versa. Furthermore, users have the option to copy and execute VM server snapshots across installations of these Clouds as has been demonstrated with Cloud BioLinux.

For users who do not own local compute resources, nor have funds available to lease computing time from a public Cloud provider, there is the possibility of using government-funded Cloud infrastructures, such as the US-based Open Science Data Cloud[3], the Australian NeCTAR Research Cloud[4], or Helix-Nebula in the EU[5]. In addition, a number of academic computing centers in both the US and other countries have similar clusters with OpenStack installed[6], to which scientists can request access.

III. Deploying and Provisioning Virtual Environments

Despite the general availability of cloud computing resources around the world, due to the complexity of setting up a functional bioinformatics platform for data analysis, we have (1) expanded the capabilities of CloudMan to be interoperable with multiple cloud providers; (2) devised an automated method for deploying a complete system on compatible clouds; and (3) developed a launcher application allowing end-users to provision instances of the platform when and where needed. The overall concept is depicted in Figure 1 and each of these contributions is described in the following sections.

A. A Multi-cloud Cluster Manager

Architecturally, CloudMan is a standalone application written in Python that orchestrates cloud components (virtualized compute resources and software programs) to deliver persistent and functional user-level services (see [12] for the implementation details). To operate, CloudMan uses a number of cloud-level features and assembles those into a functional platform for performing data analysis [4]. Internally, CloudMan abstracts away the interaction with multiple cloud providers via a common cloud interface (`cloud_interface`). All the internal methods utilize the given interface providing uniformity throughout the code while supporting multiple cloud technologies. Compatibility with multiple clouds is achieved via a custom implementation of the cloud interface for each of the supported clouds (AWS, OpenStack and Eucalyptus). Adding support for future clouds is a matter of implementing the methods defined in the `cloud_interface`, with the option to inherit from the default AWS implementation and, again, identifying and implementing only the differences.

Because a running instance of the CloudMan platform represents a composition of multiple cloud services, in order to deploy the platform, it is necessary for the cloud middleware to provide the required features, as follows:

- *Customizable machine image*: this is the VM itself containing the operating system (Ubuntu) and required system level packages for applications such as Nginx, Slurm, RabbitMQ.
- *Instance user data*: this enables runtime contextualization [13], which is what makes it

[3] https://www.opensciencedatacloud.org/
[4] http://nectar.org.au/research-cloud
[5] http://www.helix-nebula.eu/
[6] https://www.tacc.utexas.edu/systems/rodeo

possible to have different and independent instantiations of the same platform. Sample data included as part of the instance data includes the API keys that CloudMan should use to connect to the cloud provider.

- *Security groups*: a virtual firewall offered by the cloud provider that permits selective traffic to one's compute instances based on user-defined access rules. CloudMan uses HTTP, SSH, FTP and several custom ports to enable access to the running services.
- *Persistent POSIX data storage*: this is where the user-uploaded and analyzed data is stored, as well as installed tools, databases, and tool configurations; it is leveraged to provide longevity beyond the uptime of an instance. Ideally, this storage is provided in the format of block storage devices of arbitrary size (i.e., data volumes) but other storage devices can be used instead, including mountable external file systems, such as NFS or Gluster, or the ephemeral storage on individual instances.
- *Block storage snapshots*: given that a significant amount of bioinformatics data is read-only reference genome data, the CloudMan platform maximizes data reuse by utilizing a single source of this type of data for multiple deployed instances of the platform. Technically, this is delivered via point-in-time block storage snapshots, which are converted into data volumes at runtime. To realize the reuse aspect of this data, the snapshots or archives need to be sharable. Alternatively, it is possible to use "tarballs" of file systems, which are archived in an object store and extracted at runtime onto an instance's ephemeral storage.
- *Object store*: persistent configuration data describing a given instance of the platform is stored in a globally accessible object store and it is used to recompose the CloudMan platform after it had been shut down. Information about running platform services (i.e., applications), utilized resources (e.g., volume, snapshot, machine image), and the CloudMan runtime code are stored here.
- *Per object ACL controls in the object store*: in order to enable instance sharing [14] with individual users, it should be possible to set Access Control List (ACL) permissions on each individual object inside the object store. Note that this is a requirement for the instance sharing functionality only.
- *Resource metadata*: popularly known as tags, assigning metadata to each of the utilized cloud resources helps identify all the components making up the complete platform. It also provides a fallback in case the platform needs to be recreated by hand following an unexpected corruption.

Given the feature-set of today's clouds, the above list is considered quite minimal and we strive towards minimizing it further. In the meantime, the required resources are available from a variety of cloud providers/middleware as bare-bones resources that behave consistently across those, making it possible to create the cross-cloud platform being described in this paper. This approach is in contrast to using custom (or high-level) services offered by a specific cloud provider and thus becoming locked-in to the given cloud or cloud features.

B. Automating Component Deployment

The previous section describes the multi-cloud support provided by CloudMan as a reusable framework and outlines the set of requirements a target Cloud must posses. In this section, we present a method for deploying the complete CloudMan platform. By deploying, we mean building the necessary components and making them available on the target cloud. These components will be joined at runtime (see next section) to deliver a functional, user-facing platform.

The process of deploying the required components has been automated using the Ansible automation framework. Ansible is a configuration management and provisioning tool; it uses files in YAML format to allow users to define tasks (e.g., `apt-get update`) and it then executes those tasks on target resources in an idempotent fashion. It uses SSH to communicate with target resources, hence imposing minimal requirements on setup. Conceptually, Ansible allows one to define reusable roles, which are composed of multiple tasks. It further allows multiple roles to be assembled into more comprehensive playbooks that can mix and match functionality offered by individual roles.

For the purposes of deploying a bioinformatics workbench backed by the CloudMan platform, we have developed a set of Ansible roles and assembled those into a playbook[7]. As stated above and described in detail in [12], the overall platform requires multiple components: the base machine image, an instance of the tools file system, and an instance of a file system containing indexed reference genomes. The playbook offers the capacity to build these components in an automated fashion. To start, it is necessary to build the machine image. This step leverages Packer[8], which is an additional layer of automation on top of Ansible. It handles provisioning of a VM on the target cloud, invoking the set of Ansible roles for building the machine image, creating the machine image, and cleaning up. Because Packer can run in parallel on multiple targets, it is possible to build the machine image with a single command (e.g., `packer build cloudman.json`) on multiple clouds at once. The built machine image will contain the required system packages and libraries, FTP server, Nginx proxy server, Slurm cluster manager and required configurations (e.g., users, environment settings).

Next, it is necessary to build the core file system, called `galaxyFS`. This file system holds all the tools, databases, the Galaxy application, and configuration

[7] https://github.com/galaxyproject/galaxy-cloudman-playbook
[8] http://packer.io

TABLE I. A SAMPLE FROM *SNAPS.YAML* CONFIGURATION FILE. THE FILE STRUCTURE ALLOWS MULTIPLE CLOUDS, MULTIPLE REGIONS, AND MULTIPLE DEPLOYMENTS WITHIN A REGION TO BE SPECIFIED THAT CAN BE READ BY THE LAUNCHER APPS AND CLOUDMAN TO PROPERLY ASSEMBLE AN INSTANCE OF THE PLATFORM.

```
version: 1
clouds:
 - name: amazon
   regions:
   - deployments:
     - name: GalaxyCloud
       filesystems:
       - name: galaxy
         roles: galaxyTools,galaxyData
         snap_id: snap-adad90fc
       - name: galaxyIndices
         roles: galaxyIndices
         snap_id: snap-5b030634
       default_mi: ami-118bfc78
       bucket: cloudman
     name: us-east-1
- name: nectar
<include details about other clouds/regions>
```

options that will be available via the workbench. Building *galaxyFS* is realized by launching an instance of the machine image built in the previous step, adding a file system of desired size (e.g., 10GB) via CloudMan's Admin page (note that CloudMan will be available because we are now using the image built in the previous step), and running the available Ansible roles. The roles will configure and initialize a PostgreSQL database as well as download and configure the Galaxy application with production settings. Installing the desired tools is realized via the ToolShed [15]. This step has been automated via a custom Python script that leverages the BioBlend API [16], [17] and installs more than a hundred tools at once. The list of installed tools is provided in a configuration file and can hence be easily modified. Once the process has been completed, it is necessary to create a snapshot of the resulting file system, which can be achieved via CloudMan's Admin page.

Similar to building the *galaxyFS*, it is necessary to build *indicesFS* - a file system containing desired reference genomes. The process is similar in that a new file system of adequate size needs to be added. Adding the reference genomes is performed via Galaxy Data Managers [18], which integrate with Galaxy and allow one to simply choose which genomes to include. Once the process has been completed, it is also necessary to create a snapshot of the resulting file system. Finally, it may be desirable to make the created snapshots public (or share them with select users) so others can launch instances of the deployed components (see next section). Also note that instead of creating volume snapshots, it is possible to create downloadable archives or shared file systems but, for brevity, we focus on volume snapshots alone.

C. Provisioning

Once the required components have been built, they need to be assembled into a provisioned instance of the platform. Instances can be provisioned by projects or labs to offer an 'always-on' ready-to-use workbench or by individual users on an as-needed basis (see Section IV). To make the process accessible, we have further automated the provisioning process via launcher web applications.

The build process described in the previous section produces a machine image and two file system snapshots. The provisioning process launches a VM instance based on the image, which starts CloudMan, and CloudMan then creates volumes based on the previously created snapshots to make functional file systems (along with setting up a virtual cluster and starting several service processes). Before an instance can be provisioned, it is necessary to compose the instance user data that contains access details for the given Cloud (e.g., region endpoints, ports) and information about the cluster (e.g., cluster name, access credentials). Once an instance of the image is launched, CloudMan will automatically start and continue the workbench contextualization process. To obtain the snapshot ID's, CloudMan refers to the *default_bucket* that is passed in with user data. This is an accessible object store container where a file (by default *snaps.yaml*) is stored and contains the details about the snapshots and corresponding file systems (see Table I). CloudMan will read the contents of the given file and proceed with setting up the system.

The above process can be somewhat detail-oriented and technical in nature. To simplify it, we have introduced a notion of a launcher web application that will handle the provisioning process automatically. The launcher app offers users a web form asking for cloud access credentials, a name for the instance, and a choice of VM type. The app then composes and formats the required information, instantiates an instance of the base machine image and informs the user when the launch process has completed. It also allows a previously created cluster to be readily recreated. There are two launcher applications available: one within the Galaxy application itself (CloudLaunch) and a standalone one (BioCloudCentral). CloudLaunch is automatically available on any Galaxy instance (at */cloudlaunch* URI) but requires a local instance of Galaxy (a public one is available at *https://usegalaxy.org/cloudlaunch*). Alternatively,

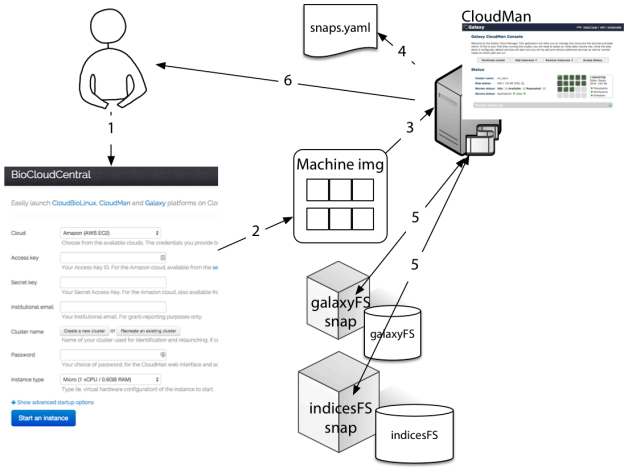


Figure 2. The process of deploying an instance of the platform, also showing all the required components.

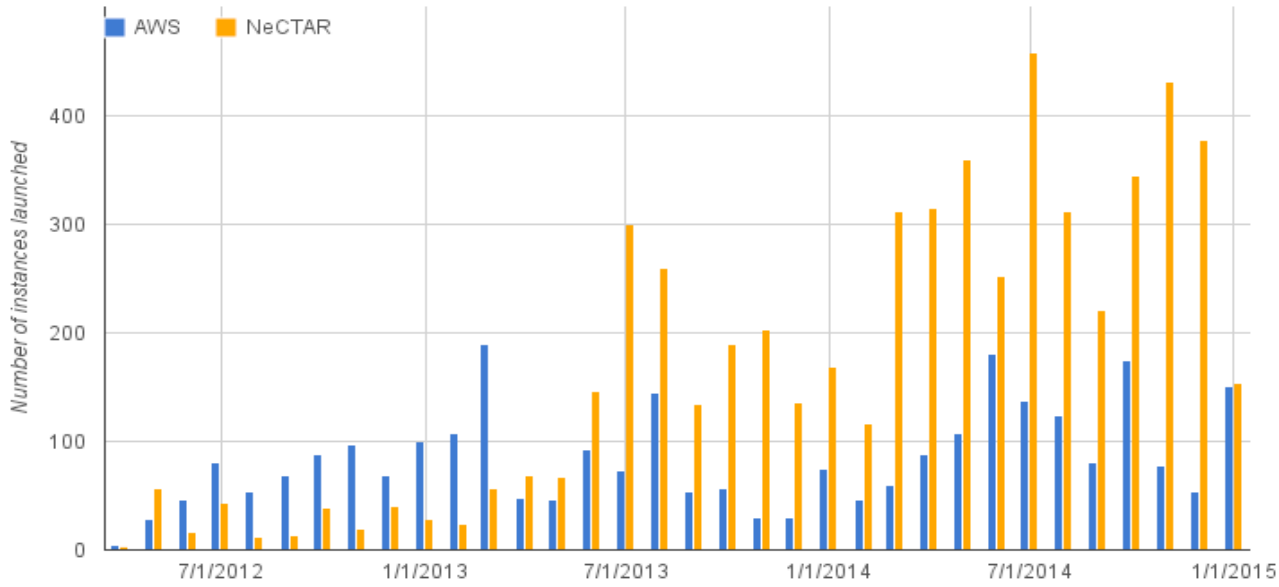


Figure 3. The process of launches per month of the CloudMan platform on two clouds, AWS and NeCTAR.

BioCloudCentral can be installed on a dedicated host and used independently of Galaxy (a public instance of this application is available at *https://biocloudcentral.org/*).

Figure 2 summarizes the described effort: after the required set of components has been built and registered in `snaps.yaml`, the launcher app is used to provision an instance of the platform. After the contextualization steps have completed, the user can start using the platform.

IV. DEMONSTRATION

The effort described in this paper allows anyone to build their own deployment of the CloudMan platform and the associated bioinformatics workbench. It further allows users to launch their own instances of the platform on the clouds on which it has been deployed. Currently, in addition to the long-standing deployment on AWS, the platform has been deployed on the Australian research cloud, NeCTAR; these represent two flagship deployments of the platform that are supported by the authors. Other deployments also exist but have been done by independent projects at their home institutions. Figure 3 captures the number of times the platform has been provisioned on the AWS and NeCTAR clouds over the past three years. The figure indicates that the NeCTAR cloud, despite the fact that it is available only to Australians, has been seeing approximately two to three times the number of instantiations as the world-wide deployment on AWS. We attribute this to the fact that the NeCTAR cloud is free of cost and that the Australian community has largely moved 'into' the cloud, especially for training and teaching purposes.

In addition to the count of provisioned platforms, the authors have been maintaining an always-on instance of the CloudMan platform on the NeCTAR cloud as part of the Genomics Virtual Laboratory (GVL) project[9]. This instance of the platform offers an instance of the Galaxy application that has been tailored for bioinformaticians training. During the course of the GVL project, a number of tutorials has been developed that demonstrate bioinformatics data analysis activities, such as variant calling or RNA-Seq analyses[10]. The instance, called *galaxy-tut* (available at *http://galaxy-tut.genome.edu.au/*), is where the tutorial data and tools are hosted, representing a fertile playground for training sessions. Besides being used during live training sessions, the instance is always available for anyone to follow the tutorials at his or her own pace. To accommodate these scenarios, the compute infrastructure behind the *galaxy-tut* user interface needs to scale with demand: as a lab full of students initiates a mapping job, the resource requirements rapidly grow. At the end of the compute-intensive portion of a tutorial, or during non-teaching periods, the resource requirements are much lower. The cloud model fits this scenario perfectly and the CloudMan platform is entrusted with keeping up with the cycle of requesting, configuring, managing, and releasing the necessary resources.

The *galaxy-tut* instance was launched as a default CloudMan instance on the NeCTAR research cloud and then manually customized to include all the data and Galaxy histories required for the tutorials. The master instance of the CloudMan cluster is running on an extra-large instance type (8 virtualized CPU cores, 16GB RAM). During training sessions, running of the jobs on the master instance is disabled via CloudMan's Admin UI and additional cluster nodes are explicitly launched. In addition, CloudMan's auto-scaling is left *on* to accommodate possible spikes in usage. The instance has nearly 500 registered users, stores 1TB of user data, and runs approximately 2000 jobs per month.

We have also deployed and run CloudMan and Galaxy on an experimental Eucalyptus cloud installation at the J. Craig Venter Institute[11]. The installation comprised of four

[9] https://genome.edu.au/

[10] https://genome.edu.au/wiki/Learn

[11] http://www.jcvi.org/

identical physical compute servers (16 core, 32GB memory, 1TB disk each). One of these servers was used as the cloud's head node to run Cloud Controller (CLC) service of Eucalyptus, and the remaining nodes running the passive Node Controllers (NCs) for worker VMs. The CloudMan framework has enabled us to bootstrap on-demand virtualized compute clusters with an easily accessible Galaxy interface on our private cloud. By simply starting with a pre-configured VM dropped on the CLC node, and configuring the Galaxy cluster endpoint to be handled by CloudMan, we were able to instantiate multiple transient VMs for parallel job execution across the Eucalyptus cloud nodes with CloudMan performing the initialization, management and termination of the VMs.

V. Conclusions

Dealing with the infrastructure and system level operations, such as acquiring cloud instances or managing operating system level services, requires a significant investment in time and expertise. Yet, for modern day bioinformatics it is an essential requirement. In our previous work, we have developed a system that provides a ready-to-use bioinformatics workbench with dozens of tools and gigabytes of reference genome data on the AWS cloud. This paper describes the effort of making the overall system underlying the workbench compatible with multiple Clouds and an automated process for deploying it. This makes it possible for anyone to deploy their own instance of the platform on a private cloud or under their own account on a public account. Each deployed instance can have a custom set of tools and reference genome data available, which are tailored to the specific needs of a group or an institution. Once deployed, others can launch the given instance of the workbench via provided launcher web applications.

Acknowledgments

The efforts of the Galaxy Team (Dan Blankenberg, Dave Bouvier, Martin Cech, Dave Clements, Carl Eberhard, Jeremy Goecks, Nitesh Turaga, Sam Guerler, Jennifer Jackson and Anton Nekrutenko) were instrumental in making this work happen. This project was supported through grant number HG005542 from the National Human Genome Research Institute, National Institutes of Health as well as grants HG005133, HG004909 and HG006620 and NSF grant DBI 0543285. Additional funding is provided by Huck Institutes for the Life Sciences at Penn State and, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions. Additional funding has been provided by the EU FP7-PEOPLE programme, under grant 277144. Further funding has been provided by the Australian NeCTAR project, under the Genomics Virtual Laboratory.

References

[1] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis.," *Nature Reviews Genetics*, vol. 11, no. 9, pp. 647–57, Sep. 2010.

[2] M. C. Schatz and B. Langmead, "The DNA data deluge," *IEEE Spectrum*, vol. 50, pp. 28–33, 2013.

[3] M. C. Schatz, B. Langmead, and S. L. Salzberg, "Cloud computing and the DNA data race.," *Nature Biotechnology*, vol. 28, pp. 691–693, 2010.

[4] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, "Galaxy CloudMan: delivering cloud compute clusters," *BMC bioinformatics*, vol. 11 Suppl 1, p. S4, 2010.

[5] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, p. R86, Jan. 2010.

[6] E. Afgan, J. Goecks, D. Baker, N. Coraor, A. Nekrutenko, and J. Taylor, "Galaxy - a Gateway to Tools in e-Science," in *Guide to e-Science*, X. Yang, L. Wang, and W. Jie, Eds. Springer, 2011, pp. 145–177.

[7] E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova, A. Nekrutenko, and J. Taylor, "Harnessing cloud computing with Galaxy Cloud," *Nature Biotechnology*, vol. 29, no. 11, pp. 972–974, Nov. 2011.

[8] S. V Angiuoli, M. Matalka, A. Gussman, K. Galens, M. Vangala, D. R. Riley, C. Arze, J. R. White, O. White, and W. F. Fricke, "CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing," *BMC Bioinformatics*, vol. 12. p. 356, 2011.

[9] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, and K. Nelson, "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC Bioinformatics*, vol. 13. p. 42, 2012.

[10] I. Altintas, "Distributed Workflow-Driven Analysis of Large-Scale Biological Data Using bioKepler," in *2nd International Workshop on Petascal Data Analytics: Challenges and Opportunities (PDAC)*, 2011, pp. 41–42.

[11] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The Eucalyptus Open-Source Cloud-Computing System," in *9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2009, pp. 124–131.

[12] E. Afgan, D. Baker, A. Nekrutenko, and J. Taylor, "A reference model for deploying applications in virtualized environments," *Concurrency Computation Practice and Experience*, vol. 24, pp. 1349–1361, 2012.

[13] K. Keahey and T. Freeman, "Contextualization: Providing one-click virtual clusters," in *4th IEEE International Conference on eScience*, 2008, pp. 301–308.

[14] E. Afgan, B. Chapman, and J. Taylor, "CloudMan as a platform for tool, data, and analysis distribution," *BMC Bioinformatics*, vol. 13, p. 315, 2012.

[15] D. Blankenberg, G. Von Kuster, E. Bouvier, D. Baker, E. Afgan, N. Stoler, J. Taylor, and A. Nekrutenko, "Dissemination of scientific software with Galaxy ToolShed," *Genome biology*, vol. 15, no. 2, p. 403, Jan. 2014.

[16] C. Sloggett, N. Goonasekera, and E. Afgan, "BioBlend: automating pipeline analyses within Galaxy and CloudMan," *Bioinformatics*, vol. 29, no. 13, pp. 1685–6, Jul. 2013.

[17] S. Leo, L. Pireddu, G. Cuccuru, L. Lianas, N. Soranzo, E. Afgan, and G. Zanetti, "BioBlend.objects: metacomputing with Galaxy," *Bioinformatics*, vol. 30, no. 19, pp. 2816–7, Oct. 2014.

[18] D. Blankenberg, J. E. Johnson, J. Taylor, and A. Nekrutenko, "Wrangling Galaxy's reference data," *Bioinformatics*, vol. 30, no. 13, pp. 1917–9, Jul. 2014.

Gathering useful programming data; analysis and insights from real-time collaborative editing

M. Rantala *, J. Soini * and T. Kilamo **
* Tampere University of Technology/Department of Software Engineering, Pori, Finland
** Tampere University of Technology /Department of Pervasive Computing, Tampere, Finland
mikko.r.rantala@tut.fi

Abstract – Traditionally, collaborative coding has been practiced in open source communities where cooperation has mostly taken place on a coordination level. Nowadays, web technology is sufficiently advanced to enable collaborative coding in real-time as group work, which eases communication in software development. In this paper this phenomenon has been studied from a knowledge transfer and learning perspective. With the aid of two different example cases (code camps), we have examined the possibilities and challenges in learning during real-time group work. Additionally, we have evaluated the effect of the structure of log data created during software development. The research frame for this study is the utilization of log data visualization in evaluating group work and further improvement of the visualization in order to support software development.

I. INTRODUCTION

Software development today is a collaborative team effort where software engineers work together to produce new software to meet changing customer needs in what is often an evolving business domain. Global software engineering [1] and open source software [2] illustrate how software is created by a team of people with versatile expertise and backgrounds working in a globally distributed environment. On a similar note, the wide adoption of agile approaches such as Extreme Programming [3] illustrates the need for individuals working in collaboration.

The rise of the Internet and social media has made a further significant change in the ways people work together and collaborate. Through the Web 2.0 technologies people can share data and services to work together from a distance. Software development is also benefiting from this trend – collaboration tools like wikis are commonplace in software projects [4, 5] and even the use of social media is on the rise [6, 7]. Collaborative development tools on the lines of document editing services such as Google Docs [8] or Overleaf [9] have also been proposed in recent years. They include the online development environment Cloud9 [10] and the real-time collaborative environment to advance close collaboration called Collabode [11]. Saros [12] allows distributed pair programming through an Eclipse plugin. In this paper, use of a collaborative cloud based IDE (Integrated development environment) called MideaaS and its predecessor CoRED [13], both developed at TUT, is described. MideaaS is integrated with a hosting solution for fast deployment of the developed software as a service. MideaaS offers a simple yet powerful web-based tool for implementing and publishing web applications written in Java using Google Web Toolkit [14] and an open source web application framework for rich Internet applications called Vaadin [15]. The development of new software only requires a browser and concurrent editing is supported. Hence, several developers can work on the software simultaneously and see all of the edits in real time.

This shift towards the web is also reshaping education, as learning is inherently an activity where both the learner's own effort and the social context play a key role. The theory of Communities of Practice (CoP) [16, 17] illustrates learning as an outcome of social participation. A rising trend in modern education, massive open online courses (MOOC), is one example of leveraging the social and collaborative possibilities of the web. In a standard classroom environment, however teaching activities focus mostly on knowledge transfer from teacher to students; collaborative learning activities are left for problem-solving activities and even then can be optional. Computer- supported collaborative learning is a pedagogical approach that aims at learning through computer-mediated social interaction either in the classroom on online [18].

In order to study learning and knowledge transfer in collaborating teams, two different code camps were organized by the Department of Pervasive Computing at the Tampere University of Technology. These code camps utilized MIDeaaS for teaching students web development through team collaboration and coding. In this paper we present a study of the log data produced during the two camps. Large volumes of data of great variety can be recorded; however, the problem lies in recognizing, displaying, and utilizing the correct data. Different approaches to recording and visualizing data have been taken by others [19, 20, 21]. Here we focus on the quality and suitability of the data collected for analysis, the use of coding patterns to recognize learning through collaboration, and the detection of programming patterns in the log data.

The structure of this paper is as follows. Section 2 elaborates on the focus of research and the structure of the log data. Section 3 examines the visualizations made from the log data and explores work by others on visualizations and new possibilities. Section 4 contains discussion on

some issues on the topic of measuring collaboration and learning. Section 5 summarizes the paper.

II. EXISTING DATA

Documented here is the study made of real-time collaborative coding log data recorded during two different code camps organized by the Tampere University of Technology. During the first code camp [22] data were gathered from the use of the browser based CoRED software [23] which enables multiple users to collaborate simultaneously on a web software project. The distinctive features of CoRED are the ability to allow two or more users to edit the same source code file and display those changes in real-time; in addition, edits and other user actions such as creating a new file are logged and can be visualized to enable recognition of patterns in user behavior and ways of working. User actions are explained below in the section on studied data. The second code camp [24] featured a cloud-based IDE called MIDEaaS, which evolved from CoRED. Both of these development environments have been developed at the Tampere University of Technology. For the remainder of the paper, the IDE will be called MideaaS.

A. Research Focus

The code camps were run on the lines of an empirical experiment [25, 26]. During the two code camps, software engineering data on the development work were collected on the teams' work.

The purpose of the study was to perform analysis of the recorded log data. In this case we were interested in the following three areas:

- Validate the quality and suitability of the data.
- Expand one of the original research questions (Does collaboration help transfer knowledge between learners?) with user coding patterns.
- Examine visualizations derived from said data, visualizations made by other research groups on the same topic and explore the detection of programming patterns. Coding patterns were chosen because they provide the benefit from an educational viewpoint by possibly uncovering bad programming habits.

B. Studied data

The data gathered in the two code camp events focuses on user actions on files and source code. The data logged during the first code camp are more varied compared to the second, which concentrates almost solely on user edits. The changes in recorded data between code camps are discussed in section 4. Table 1 (below) contains a list of recorded events, event data, and which code camp logged the event. Each log entry is preceded by a 13-digit timestamp providing millisecond precision.

TABLE I. LOGGED EVENTS

Event	Recorded data	Code camp
New file	File name	1
Delete file	File name	1
Open file	File name, collab id, user id, user name	1
Close file	File name, collab id, user id, user name	1
Add marker	File name, collab id, marker id, marker type, start position, end position	1
Remove marker	File name, collab id, marker id	1
Edit	File name, collab id (code camp 1), user id (code camp 2), user name (code camp 2), insert/delete, position of edit, number of characters inserted/deleted, number of characters in file after edit	1&2
Chat message	User id, message	1&2
Created		2
Loaded from disk	File path for project	2

Of the first six events that are exclusive to the first code camp, four contain user information in the form of two id numbers. The collab id is assigned to a text editor window and one user can therefore have multiple collab ids should they have multiple text editors open. A user id is assigned upon login, so if the user logs in again they are assigned a new user id. Both of these types of ids are problematic since their persistence is not guaranteed.

The first four events are not elaborated on, as the information they contain should be self-explanatory. The add and remove marker events record visual pointers, either signifying a code error from the automatic error checker or the cursor or selection of another user. The marker events add a considerable amount of lines to the log files since typing in code manually will produce errors from the Java error checker. The other type of marker, the cursor location of another user, seems redundant considering that edit events log the activity of other users and contain more detailed information. The edit event utilized during the first code camp contains only the collab id as identifying information, which makes distinguishing between users difficult. This deficiency was fixed for the second code camp with the addition of the user name and a persistent id number for each user. The last two events are exclusive to the second code camp and replace the more verbose file events of the first code camp. The created event signifies the start of a project and is the first event in a log file.

Overall, the data recorded into log files is suitable for examining user actions both alone and in collaboration with others. The changes made in recording between code camps reduced the volume and verbosity of data although some of these changes, such as the removal of a new/delete file event from code camp 2 logs, might prove

detrimental. Section 4 contains further discussion on changes in recording log data.

C. Observations on Log Data

The quality of the log files was analyzed with Excel as the original research publications [22, 24] made no mention of errors. Analysis of the files found 4 different types of errors, which are listed in table 2 and discussed below.

TABLE II. TYPES OF ERRORS IN LOG FILES

Event/Recorded data	Description of error	Code camp
Entire row, edit events only	Logged data contains duplicate rows.	1 & 2
Timestamp	Within one millisecond a user is logged as having performed multiple edits (inserts and/or deletions).	1 & 2
Number of characters in file	The number of characters in file-property is static for multiple edit events. This occurs independently and also in connection with the above timestamp error.	1 & 2
User id and name	Some edit events do not contain a user id and name.	2

The existence of duplicate lines was known before analysis of the log files began but, since no hard numbers existed, a count was performed. Code camp 1 material contained 18 log files of which 5 had no duplicate lines. The remaining 13 files contained the vast majority (99.8%) of lines logged during code camp 1. The average for duplicate lines per total lines is 0.02 with the minimum value being 0.01 and the maximum 0.07. Code camp 2 produced 7 log files, all of which contained duplicates. The amount of files is smaller because each project's logged edits were combined into a single file per project. Unlike code camp 1, where there were no significant outliers in duplicate lines per total lines, code camp 2 contains two logs that deviate significantly from the rest. Logs 3-7 have an average of 0.01 duplicate lines per total lines although logs 1 and 2 contain 0.18 and 0.07 duplicate lines per total lines.

An issue with timestamps and edits exists wherein a single user is logged as having performed multiple edits within a millisecond.

The number of characters in file property reflects the total size of the file. It is updated after each edit but an error exists where the count is not properly updated. This causes visualizations that depict changes in file size to appear flat with sudden jumps in size. This could cause a false impression of programming skills, which is explored further in section 3 "Different types of visualizations".

The last issue found in code camp 2 logs is the lack of user id and name for some edit events. The average for missing user ids and names per total lines is 0.03 with a minimum of 0.01 and a maximum of 0.08. No strong correlation existed between project size and this error.

Overall, we conclude that while log recording has problems, the errors present in the existing data are not plentiful enough to threaten the results of the previous studies [22, 24]. We recommend the log recording process undergo a thorough review in order to dispel any uncertainty regarding results.

III. IDENTIFYING AND VISUALIZING USEFUL DATA

The aim of this paper is to investigate the quality of the data collected and their suitability for further analysis of the development process. The collected data are visualized in order to point out development patterns and possible social learning through shared knowledge. Next, we discuss the collected data, the data visualization and the possible anti-patterns the visualizations can help to shed light on.

A. Edits Done by Users

In terms of responding to one of the original research questions "Does collaboration help transfer knowledge between learners?" [24], user actions must first be examined in two dimensions: time and position, in order to detect possible collaboration. The edit event recorded in the log files provides detailed information on user actions with a millisecond precision timestamp, character position, and number of characters changed. The first problem to be solved is examining the proximity of logged edits to detect collaboration between users, i.e., how much is the maximum time that can elapse for users to be considered to be collaborating? The original research group chose 30 seconds as the limit of collaborative edits 24]. The second problem is measuring a change in programming skills. These two problems were approached by the original research group from multiple directions: "statistical analysis on the teams' work, log data and interview answers". The problem of measuring a change in programming skills from log data is discussed in more detail later in section 4.

B. Introduction to MIDeaaS Visualizations

This brief chapter contains an explanation of the visualizations produced from MIDeaaS log data. Fig. 1 (below) is made from code camp 1 log data but its form is the same as those used to visualize code camp 2 log data.

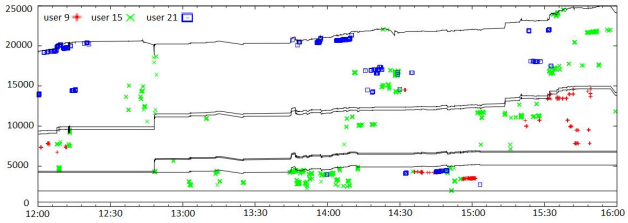


Figure 1 A four-hour long section of a log from one of the projects [22].

The above figure contains a four-hour segment of programming work done on a single project. The axes show the time and number of characters in the project. The gray lines separate different files and show changes in file size as the users perform edits. User edits are marked with a red plus sign, green x or blue square, as shown in the legend in the upper left-hand corner.

C. Scope of Data Visualizations

The visualizations produced from MIDeaaS log data focus on user actions on source code and interaction with other users. For the sake of readability, visualizations are limited to 2-hour segments as shown in fig. 2 and 4. This focus on user-level actions means that observing project-level patterns would require combining several 2-hour segments in order to form an overall picture of project-level progression. Additionally, neither IDE contains version control or issue tracking, thus limiting the options for detecting project level problems.

The currently available data focus on visible actions (edits) performed by users, however, it does not account for other kinds of user co-operation, such as advising done over communication software (Skype etc.) or live discussions. To determine whether this kind of 'invisible' co-operation has occurred, one possible solution could be to record which area of source code is being examined in the IDE. Should two or more users be examining and/or editing the same area of source code, knowledge transfer via problem solving or advising between users becomes possible. Fig. 2 [24] (below) has been edited to demonstrate this concept.

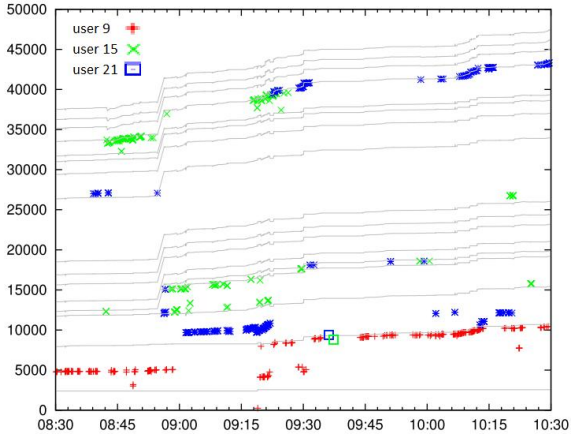


Figure 2 A typical two-hour-long editing session from one of the teams [24]. Edited to illustrate tracking of each user's focus.

Here the green and blue squares edited into the picture around the 09:35 mark illustrate the editor views of users 15 and 21 respectively, as they engage in problem solving with user 9. This method of observation is not without its weaknesses, however. Something as simple as a multi-monitor setup where two or more monitors have source code files open (or even one monitor with multiple source code files open) could provide another layer of complexity in determining exactly what file the user is actively following. Another problem with this approach to detecting 'invisible' collaborative activity is the fact that it is difficult to determine whether a user was actually co-operating with others or merely satisfying idle curiosity, unless they perform at least some edits in the observed area of source code. Nonetheless, this approach offers an additional method of detecting collaboration.

D. Different Types of Visualizations

Code size growth data have been examined by others as an indicator of programming skills by comparing the growth data curves between programmers of differing skill levels solving the same programming problems [19]. In the paper by Wang et al., the two best and worst performing programmers' code growth curves were compared, as shown in Fig. 3 (below).

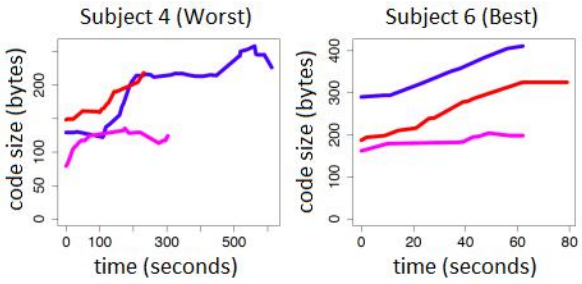


Figure 3 The traces of the worst and best performers [19]. Blue graph is problem A, red graph problem B, and orange graph is problem C.

The picture above contains code size growth curves for two different programmers whose programming skills were evaluated by making them individually solve the same problems. The colors in the figure correspond to different problems. The axes show code size in bytes and time in seconds, although the scales of the plots are different. However the main issue here is the behavior of the code growth rather than the scales of the plots.

The worst performing programmer's curve behaved in an "erratic" way compared to the best programmer, whose code growth curves were "very smooth and generally monotonically increasing". Wang et al. suggest that "Drastic drops imply code deletion indicating that corrections or refactoring occurred. Plateaus suggest "think time" possibly because users were experiencing challenges preventing them from making steady progress".

A direct comparison of code size growth curves between code camp participants cannot be performed since the participants worked on different projects and engaged in collaborative work. However, the approach by Wang et al. could be utilized together with the 'invisible' activity detailed above to observe knowledge transfer and its effectiveness. As an example, let us assume that a user in a collaborative project has been working on a file alone but at some point determines that they are unable to solve a problem on their own. Another team member is asked to provide assistance, which they supply either in the form of advice ('invisible' activity) or coding (edits). The outcome of this interaction could be observed from the code size growth curve of the user who requested help. If the curve behaves in a "smooth and generally monotonically increasing" way after interaction, we can say that collaboration has transferred knowledge successfully. Other possible areas of interest around this kind of scenario would be to examine how successful (does the advisee's code size growth curve indicate hesitation or confidence in their work?) different users are in this kind of situation in different roles (adviser or advisee). To summarize the benefits of the approach described above would help detect collaboration that would be missed by examining only edits and help determine the successfulness of collaboration.

As an addition to the suggestion of Wang et al. about drastic drops in code size, we propose that a drastic code insertion is either refactoring (if the insertion follows a deletion) or copy and paste programming (if the insertion occurs without a corresponding deletion). Examining edit events by a single user to determine the occurrence of either refactoring or copy and paste programming would not be prudent in this case due to the collaborative nature of the code camp projects. For example, two users working together but examined separately could lead to false conclusions regarding refactoring. Fig. 4 (below) contains collaborative work around a drastic change in code size [24].

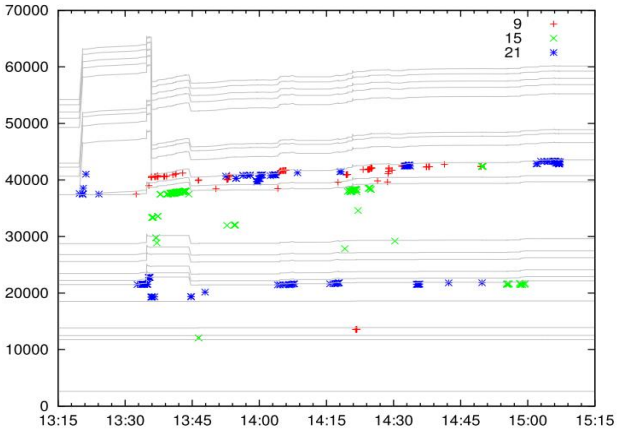


Figure 4 A typical two-hour-long editing session from one of the teams [24].

Here user 21 has inserted a large amount of code around 13:19. However, further work on the file is continued by user 9, whereas user 21 has moved on to work on another file. A task handover has been performed between users 9 and 21, and it is reasonable to assume that some knowledge must have been exchanged via advice regarding the code added by user 21 and possible further work needed on it.

E. Detecting a Potential Anti-pattern

Andrew Koenig first introduced the idea of an anti-pattern [27] in software engineering in 1995. He defines it as follows: "Anti-pattern is just like pattern, except that instead of a solution it gives something that looks superficially like a solution, but isn't one." Whereas a (design) pattern provides a solution to a design problem, an anti-pattern is a solution that appears to be beneficial but contains disadvantages that may negate any potential benefits. Since all existing visualizations produced from MIDeaaS log data focus on measuring and displaying user edits and code size growth, observable anti-patterns are coding, code, or work habit related. The anti-patterns presented below are rudimentary and not collaboration-specific but nonetheless offer an opportunity to detect problematic coding behavior.

Visualization of MIDeaaS log data can be used to detect at least one problematic coding-related anti-pattern: copy and paste programming [28], which is the reuse of code, possibly without fully understanding its purpose. The use of this technique is in some cases justified with increased productivity via code reuse. However, should one of the copied parts of code contain an error that is not immediately evident through inspection and/or testing, tracking down and fixing all instances of the copied code can become costly. Copying and pasting can be detected in the visualization graphs as vertical increases in code size that clearly differ in the scale of increase when compared to manually typed code. Fig. 4 above contains an example of a possible copy and paste occurrence, although the earlier edits should also be examined to determine whether refactoring was the case there.

A code-related anti-pattern that is observable from visualizations is the god object [29]. In programming, a god object breaks good object-oriented design by containing most of the information or methods of the entire program in a single class. In the MIDeaaS visualizations, a god object can be found by comparing file sizes; no exact definition for the size of a god object exists but should one file clearly exceed the size of the others in one project and the class in question be tightly coupled with the rest of the project's classes, the class should be inspected for the existence of this anti-pattern.

The visualizations produced by MIDeaaS allow not only a view of collaboration within a team but a window to individual coding habits and behavior. This presents a chance to help users by detecting possible bad coding habits.

IV. Discussion and Future Work

What cannot be derived from data or visualizations? The effects of real-time co-operation on learning and knowledge transfer in general are difficult to measure. In order to determine an increase in programming skills in an individual, one would first have to examine behavior over a long time in order to establish a baseline. Even determining a metric or a set of metrics to evaluate someone's skill is not a trivial task. Assuming that a suite of metrics could be chosen to represent a person's coding performance and a change in behavior (for better or worse) could be detected, the next step would be to determine exactly what caused this change. Pinpointing the exact event or group of events could be impossible; a human being engaging in creative work such as coding is not a glass box state machine with clearly labeled inputs and outputs. Changes in behavior could be a result of a host of factors all taken together. An additional problem is the fact that neither visualizations nor log data reveal the specific type of programming problem that is being solved. Collaboration might transfer knowledge that helps in solving a specific problem but another type of problem could again hinder a programmer. It is for the above reasons that we suggest that a combination of user edit information and 'invisible' activity should be used to study the effectiveness of collaboration. In collaboration situations, attention should be paid to the following facts: the time spent collaborating, its effectiveness (was work continued by the original programmer?) and any recurring pattern in collaborators (do some people mesh together better than others?). We believe that this could expand on one of the original research questions by the MIDeaaS team: Does collaboration help transfer knowledge between learners?

Changes in recorded data between code camps. The original research group chose to discard six logged events: new, delete, open and close file, in addition to add and remove marker. These events were possibly deemed not informative enough to provide insight into knowledge transfer among users. It is also possible that the file-related events were merely added to help in visualizing log files. Of the six discarded events, new and delete file could help provide information on a project level about library additions or refactoring. This could help clearly differentiate between adding a library file and pasting a large amount of code.

Collaboration thresholds. A possibly interesting avenue of further study would be to examine user actions after launching the application and especially situations where the application terminated with an error. A per user threshold for collaboration could be explored by examining possible patterns in a user seeking collaboration to help solve a problem with an application experiencing errors. This approach should be combined with that described in the first chapter of this section to maximize the amount of data available. MIDeaaS does not currently support deploying the application if another user is in the middle of editing code. This problem would have to be solved first as currently deploying and running the project is on the whole a group effort.

V. SUMMARY

In this paper we first studied and made observations on the validity of existing log data. Based on our findings, we made recommendations on future data logging. We also studied the visualizations produced from existing log data and made recommendations on how the visualizations could be improved to detect collaboration more effectively. User performance was explored via a comparison to a visualization used by another research team. Ways to utilize currently existing visualizations to detect bad coding habits were also explored. Finally, we examined some problems in measuring learning and knowledge transfer. Additionally, we also pondered on how to detect collaboration that occurred, not during coding but while running the project. We hope that by combining the existing visualization and the suggested improvements, the motives and causes of collaboration could be observed more fully than now, which could be a relevant research theme in the future.

REFERENCES

[1] J.D. Herbsleb, “Global software engineering: the future of socio-technical coordination,” in 2007 Future of Software Engineering, pp. 188-198, IEEE Computer Society, 2007.

[2] K. Crowston, Q. Li, K. Wei, U.Y. Eseryal and J. Howison, “Self-organization of teams for free/libre open source software development,” Information and Software Technology, vol. 49, nr. 6, pp. 564-575, June 2007.

[3] K. Beck and C. Andres, Extreme Programming Explained: Embrace Change, 2nd ed., Addison-Wesley Professional, 2004.

[4] F. Lanubile, C. Ebert, R. Prikladnicki and A. Vizcaino, “Collaboration tools for global software engineering,” IEEE Software, IEEE Computer Society, 2010, vol. 27, nr. 2, pp. 52-55.

[5] P. Louridas, “Using wikis in software development,” Software, IEEE, vol. 23, nr. 2, 2006, pp. 88-91.

[6] L. Singer, F. Filho and MA Storey, “Software engineering at the speed of light: how developers stay current using twitter,” ICSE 2014 Proceedings of the 36th International Conference on Software Engineering, pp. 211-221.

[7] MA Storey, C. Treude, A. van Deursen and LT Cheng, “The impact of social media on software engineering practices and tools,” Proceedings of the FSE/SDP workshop on Future of software engineering research, FoSER ’10, 2010, pp. 359-364, ACM.

[8] Google Docs. Retrieved 3.2.2015, from http://www.google.com/docs/about/

[9] Overleaf: Real-time Collaborative Writing and Publishing Tools with Integrated PDF Preview. Retrieved 3.2.2015 from https://www.overleaf.com/

[10] Cloud9. Retrieved 3.2.2015 from https://c9.io/

[11] Collabode. Retrieved 3.2.2015 from http://groups.csail.mit.edu/uid/collabode/

[12] Saros Eclipse Plugin. Retrieved 3.2.2015 from http://www.saros-project.org/

[13] CoRED. Retrieved 3.2.2015 from http://cored.cs.tut.fi/

[14] Google Web Toolkit. Retrieved 3.2.2015 from http://www.gwtproject.org/

[15] M. Grönroos, Book of Vaadin, Uniprint, 2014

[16] J. Lave and E. Wenger, Situated Learning: Legitimate Peripheral Participation, Cambridge University Press, 1991.

[17] E. Wenger, “Communities of practice and social learning systems,” Organization, vol. 7, nr. 2, pp. 225-246, 2000, Sage Publications.

[18] G. Stahl, T. Koschmann and D. Suthers, Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), Cambridge handbook of the learning sciences, pp. 409-426, Cambridge, UK: Cambridge University Press, 2006.

[19] Y. Wang, P. Wagstrom, E. Duesterwald and D. Redmiles, ”New opportunities for extracting insights from cloud based IDEs,” ICSE Companion 2014 Companion Proceedings of the 36th International Conference on Software Engineering, pp. 408-411.

[20] P. Weissgerber, M. Pohl and M. Burch, ”Visual data mining in software archives to detect how developers work together,” Fourth International Workshop on Mining Software Repositories, 2007.

[21] L. Voinea and A. Telea, ”Visual data mining and analysis of software repositories,” Computers and graphics, vol. 31, pp. 410-428, 2007.

[22] A. Nieminen et al., “Collaborative coding rnvironment on the web: a user study,” Developing cloud software: algorithms, applications, and tools, TUCS General Publication Nr. 60, pp. 275-300, 2013.

[23] J. Lautamäki et al., ”CoRED - browser-based collaborative real-time editor for Java web applications,” Proceedings of The ACM Conference on Computer Supported Cooperative Work, Seattle, ACM, 2012.

[24] T. Kilamo et al., ”Knowledge transfer in collaborative teams: experiences from a two-week code camp,” Companion Proceedings of the 36th International Conference on Software Engineering (ICSE Companion 2014), ACM, New York, USA, pp. 264-271.

[25] B. A. Kitchenham et al., “Preliminary guidelines for empirical research in software engineering,” Software Engineering, IEEE Transactions on, vol 28, nr. 8, pp. 721-734, 2002, IEEE.

[26] V. R. Basili, R. W. Selby and D. H. Hutchens, “Experimentation in software engineering,” Software Engineering, IEEE Transactions on, vol. SE-12, nr. 7, pp. 733-743, 1986, IEEE.

[27] A. Koenig, ”Patterns and antipatterns,” Journal of Object-Oriented Programming 8 (1), pp. 46-48, 1995.

[28] W. Brown, R. Malveau, S. McCormick and T. Mowbray, AntiPatterns: Refactoring Software, Architectures, and Projects in Crisis, John Wiley & Sons, ltd, 1998.

[29] A. J. Riel, Object-Oriented Design Heuristics, Boston: Addison-Wesley, 1996.

Microservices validation: Mjolnirr platform case study

D.I. Savchenko*, G.I. Radchenko* and O. Taipale**
*South Ural State University, Chelyabinsk, Russia
**Lappeenranta University of Technology, Finland
dmitry.savc@gmail.com, radchenkogi@susu.ac.ru, ossi.taipale@lut.fi

Abstract – Microservice architecture is a cloud application design pattern that implies that the application is divided into a number of small independent services, each of which is responsible for implementing of a certain feature. The need for continuous integration of developed and/or modified microservices in the existing system requires a comprehensive validation of individual microservices and their co-operation as an ensemble with other microservices. In this paper, we would provide an analysis of existing methods of cloud applications testing and identify features that are specific to the microservice architecture. Based on this analysis, we will try to propose a validation methodology of the microservice systems.

Keywords - Microservices, Services Oriented Architecture, Cloud computing, PaaS, testing, validation

I. INTRODUCTION

The microservice architecture is a cloud application design pattern that implies that the application is divided into a number of small independent services, each of which is responsible for implementing of a certain feature. Microservices can be considered as meta-processes in a Meta operating system (OS): they are independent, they can communicate with each other using messages and they can be duplicated, suspended or moved to any computational resource and so on.

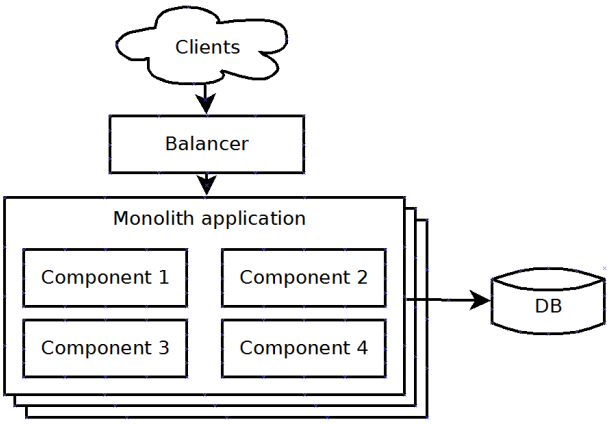


Figure 1. Monolithic system architecture

There are several examples of software systems, which were designed and implemented in accordance with microservice concept. Such companies as Netflix [19] and SoundCloud [3] use the microservice concept in their architectures, while Amazon [20] offers services like S3, which can be considered as microservices.

We can describe microservice architecture by comparing it with the monolithic architecture style, when an application built as a one big single unit (figure 1). The server-side monolithic application will handle HTTP requests, execute domain logic, retrieve and update data from the database, and select and populate HTML views to be sent to the browser. Such a design approach may result in that a change made to a small part of the application, requires the entire monolith to be rebuilt and deployed. Over time, it is often hard to keep a good modular structure, making it harder to keep changes that ought to only affect one module within that system [13].

In a case of microservice architectural style (figure 2), we can divide the application into a several independent components (microservices) and scale them independently. It is useful in a case of high-loaded systems and systems with lots of reusable modules.

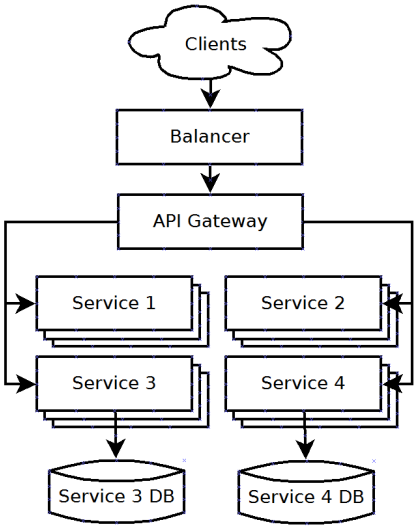


Figure 2. Microservice system architecture

Some developers and researchers believe that the concept of microservices is a specific pattern of implementation of Service Oriented Architecture [11]. In particular, the developers of the Netflix platform, which is considered the embodiment of the microservice architecture, named their approach of a design of software

The reported study was partially supported by RFBR, research project No. 14-07-00420-a and by Grant of the President of the Russian Federation No. MK-7524.2015.9

architecture as "a fine grained Service Oriented Architecture" [12]. However, the microservice pattern defines following specifics of software services development process [13]:

- microservices should use lightweight mechanisms for the communication (often an HTTP resource API);
- microservices should be built around business capabilities;
- microservices should be independently deployable by fully automated deployment machinery;
- there should be a bare minimum of centralized management of these services, architecture should provide decentralized service governance and decentralized data management.

However, the microservice approach is not an ultimate solution to deal with the system complexity. Using microservice approach, we are shifting the accidental complexity [2] from inside of the application out into infrastructure. That is why we should use special approaches to manage that complexity. Programmable infrastructure allows automating the microservices lifecycle management [18]; including automation of service deployment and termination, load balancing, service naming, continuous integration and delivery etc. Such support can be provided by a PaaS cloud platform, in order to scale the system automatically, manage it and provide middleware for the message communication. In addition, there should be implemented monitoring mechanisms to control the flow of the microservice application. It can be implemented as a specialized PaaS [16] solution or integrated in the existing PaaS.

Since the continuous integration is the only option of the implementation of the microservices architecture pattern, the question arises about the support this process on the methodological and software level. One of the most important stages of the process of continuous integration and deployment is validation of the developed software.

Our research goal is to develop the microservice validation methodology and a software support for microservice testing. To reach this goal, during our research we should solve the following tasks:

- to provide a review of existing software validation methods to identify testing features that are specific to microservice architecture;
- to develop a methodology for microservice systems validation, including approaches for component and integration testing;
- to develop a software solution that provides support for microservice systems testing on the basis of the proposed methodology.

In the framework of the current article, we would concentrate on the analysis of the current software validation standards and approaches and their relevance to the microservice architecture pattern. This paper is organized as follows. In Section II we will provide a review of the research and standards, related to the microservices testing and validation. In Section III we would try to identify special aspects of microservice validation, related to the component, integration and system testing. In Section IV we would propose microservice validation techniques for the Mjolnirr cloud platform prototype. In section V we summarize the results of our research and further research directions

II. Related research

Microservices concept is relatively new, so there are few materials in the field of microservice validation and testing right now. Basic information about the microservice approach is provided in the article [13] by James Lewis and Martin Fowler. In this article, they define the concept of microservice as well as the specific features of microservice architecture. Toby Clemson in his article [5] provides strong emphasis on the analysis of possible strategies of microservice systems testing, including certain types of microservice systems architecture; unit, integration, component and contract testing. Neal Ford provides a review of the proposed approaches on this subject is his work [6]. He describes basic ideas about microservice systems development, monitoring and testing. In addition, several commercial companies successfully moved to microservice architectures and published their result. For example in the paper [14] Netflix provide an overview of their PaaS system and mechanisms that they use to provide features of their platform to the components that initially was not designed to be used inside of the Netflix PaaS platform.

In our previous works, we described a prototype of distributed computing system called Mjolnirr [17]. This system works with isolated components, connected using message passing pattern. This system also can be considered as a microservice platform, because each component is isolated, fine-grained and has open standardized interface. In addition, automated scheduling and planning mechanisms have been implemented within the Mjolnirr platform in order to optimize the system load [15].

The ISO/IEC standards, ISO/IEC 25010 – Software Product Quality [7] and ISO/IEC 29119 – Software Testing [8], serve as the starting point of the study of microservice system testing and validation. Generally, we define testing as a combination of verification and validation [8]. In this research, we define microservice testing as microservice validation, because in general it is impossible to apply static testing methods to the compiled microservice.

As we mentioned before, we can consider microservices as a subset of Service Oriented Architecture (SOA). Thus, we can use SOA validation techniques in application to microservices systems. Some of existing approaches to SOA validation use SOA's BPEL business model for constraints-based SOA functional testing [10], other use white-box testing [1] to determine, is service covered with tests enough.

When we are talking about service testing and validation, we should decide, which quality characteristics are important for the product we are going to validate. Quality in use is defined in the ISO/IEC 25011 standard: “Quality in use of a service is the degree to which a

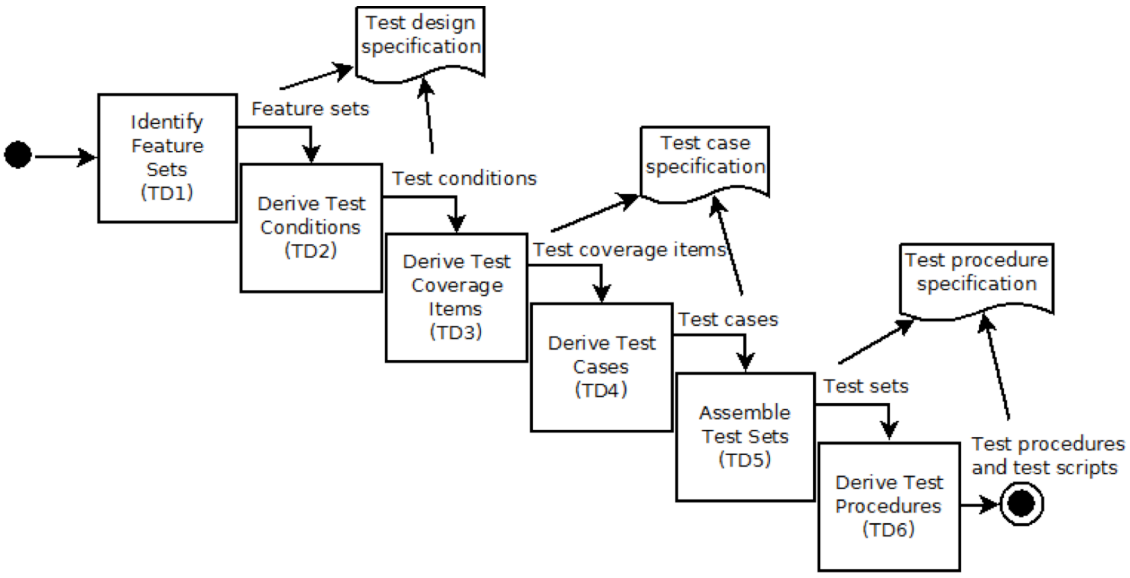


Figure 3. ISO/IEC/IEEE29119-2 Test Design and Implementation Process (modified)

service can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, satisfaction, freedom from risk and SLA coverage" [8]. ISO/IEC 25011 standard defines eight quality characteristics of services: Service Suitability, Service Usability, Service Security, Service Reliability, Service Tangibility, Service Responsiveness, Service Adaptability, and Service Maintainability. All of them, except Service Tangibility, can be applied to microservices. In addition, in a case of microservice system, we do not care about effectiveness and efficiency – these characteristic should be provided and measured by application programmer. However, satisfaction (stability of the whole system, quick response, etc.), freedom from risk and SLA coverage are important when we want to validate microservice system.

According to ISO/IEC 29119-4 standard, test design and implementation process consists of 6 steps – identifying feature sets (TD1), deriving test conditions (TD2), deriving test coverage items (TD3), deriving test cases (TD4), assembling test sets (TD5) and deriving test procedures (TD6) (see Fig. 3) [8]. To describe the microservice validation methodology, we should focus on the TD2, TD3, TD4 and TD6 steps.

III. Special Aspects of Microservice Validation

In this study, we define microservice as a software-oriented entity, which has following features:

- *Isolation* from other microservices as well as from the execution environment based on a virtualized container;
- *Autonomy* – microservices can be deployed, destroyed, moved or duplicated independently. Thus, microservice cannot be bound to any local resource because microservice environment can create more than one instance of the same microservice;
- *Open and standardized interface* that describes all available communication methods (either API or GUI);
- *Microservice is fine-grained* – each microservice should handle its own task.

Let us consider the most significant stages of the validation process, in relation both to the software as a whole and in particular microservice systems and define the features of the validation to be considered for microservice systems. By *unit validation* we understand individual validation of the microservice, by *integration validation* – validation of the interaction between individual microservices and by *system validation* – validation of the whole system (see Table 4).

System validation does not seem to offer features to microservices testing, since validation of the system level does not consider the internal structure of the application. The internal structure of the system can be considered as a black box, so there would be no difference in the system validation process for any application for which network resources access is needed.

A. Microservice unit validation

Functional Unit validation of microservice systems involves independent validation of individual microservices to meet the functional requirements. This

TABLE I. Featured Types and Levels of Testing

	Component	Integration	System
Functional	+	+	–
Load	+	+	–
Security	+	+	–
– - no features for microservice validation + - has features for microservice validation			

type of validation is based on a formal description of the functional requirements (including requirements as input and output data) imposed on each microservice. To provide the functional unit validation, microservice does not require to be deployed in a cloud environment, as this type of validation considers microservice as an isolated component.

Development of a software system in accordance with microservice approach is a process of development of individual independent microservices interacting exclusively based on open protocols. Thus, each microservice represents independent software product that should be developed and tested independently of the other microservices. Functional unit validation of the microservices can be divided into two stages (see Fig. 7):

1. unit, integration and system validation of the microservice source code or/and internal components (basic level);
2. self-validation of the microservice interface.

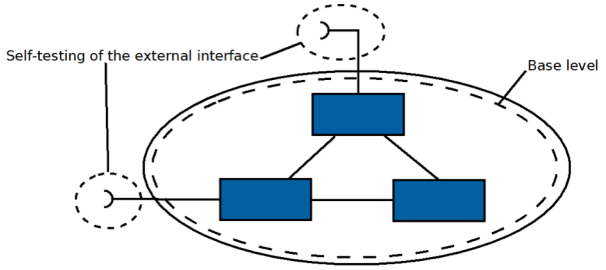


Figure 4. Functional unit validation of the microservice

Each microservice can be a relatively complex software system, which consists of several software components (local storage, web server, etc.) encapsulated in a container. Even in a case of ready-made software solutions, integration of such a component ensemble inside a container must be accompanied by validation the correctness of these components work together.

After that, it is necessary to test the interface provided by the individual microservice on its compliance with the specifications defined in the design. In this case, the entire microservice should be tested in general using its external interface.

Functional unit tests of a microservice should be conducted automatically on each microservice configuration change.

Load unit validation is aimed at the individual microservice validation at a certain load. Such validation may also be performed automatically for each microservice. Microservice load validation allows to identify resources that should be provided to the microservice container to ensure its correct work.

Security unit validation is aimed at the security and isolation validation of the individual microservice. As in the case of functional validation, microservice security does not end with the security of its individual classes. In component security validation we test the security of the whole microservice in the complex (for example, for typical vulnerabilities of web servers, injections, etc.). This stage of validation is performed without the presence of application programmer.

B. *Integration microservice validation*

Integration validation of microservice systems involves validation of all types of communications between the microservices. It includes validation of communication protocols and formats, resolution of deadlocks, shared resource usage and messaging sequence. To provide all of this validation techniques, we need to track all the messages, transmitted between microservices, and build the messaging graph. For this purpose, we need to know the messaging direction, message origin and destination. This information must be provided by microservice standard contract.

Functional integration validation ensures interoperability of the validation individual microservices. In this type of validation, we check for correctness of sending requests from one microservice to another, the sequence of interactions, microservice orchestration and choreography. In this type of validation, we can generate test cases in semi-automated mode – we need to know the communication sequence, and this data should be provided by the developer. Using this data, we can produce communication graph and test the system according to this graph.

Load integration validation is aimed at checking the microservice correctness under automatic deployment, backup and transfer, as well as microservice orchestration and choreography. This type of validation can be done automatically. Unlike functional integration validation, this type of validation includes loading of communication channels to discover the maximum load the microservice can handle before failure.

Integration security validation is aimed at validation of the security of communications between microservices, as well as checking for the interception of messages from outside or inside by anyone, besides the message recipient. Microservice specific validation should also take into account the particular environment in which the microservices are applied.

IV. VALIDATION OF MJOLNIRR-BASED MICROSERVICES

We have developed a prototype of the Mjolnirr microservice platform [17]. This platform supports microservices, implemented using JVM-compatible languages and Java-based containerization. Integration with Mjolnirr will be the first step of our validation framework implementation. This step includes Java-specific methods of software validation, such as jUnit for component validation. Let's describe, how validation implementation steps that we have highlighted in Chapter 2 could be implemented on the basis of Mjolinrr platform.

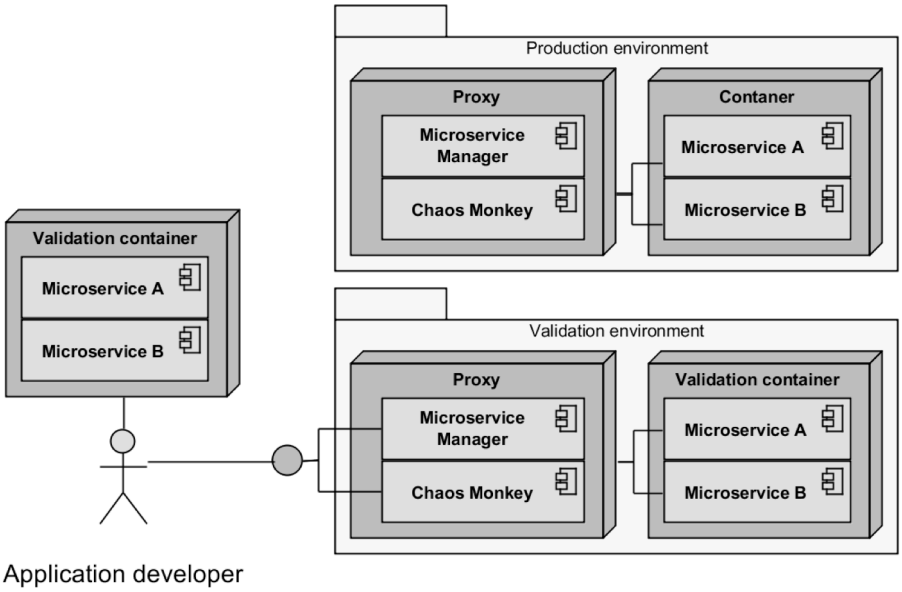


Figure 6. Validation of Mjolnirr-based microservices

TD2 (Derive test conditions) describes the testable aspects of components, in our case - microservices. In Mjolnirr, we can consider microservices and their calls as test conditions. As mentioned in Chapter 3, each microservice is isolated and autonomous and can be tested in a fully isolated JVM. Microservices are communicating using messages. Therefore, conditions, which determine quality of microservice system, are following: full isolation for component level, message communication for integration level and encapsulation for system level.

TD3 (Derive test coverage items) describes the attributes, derived from test conditions. Test conditions are microservices and their communications, therefore, test coverage items would be microservice methods and messages. Each microservice method and communication channel should be covered. Coverage items can be covered automatically or by application developer.

TD4 (Derive test cases) describes the test cases generation process. According to quality characteristics and ISO/IEC 29119-4 standard, we can decide, what test design techniques we should use and find exact technologies to implement those techniques. To perform a component-level testing, we can use only specification-based techniques, such as Boundary Value Analysis, Equivalence partitioning, Use case testing, Random testing and Scenario testing. In a case of component security testing, we can also scan microservice for typical vulnerabilities of the specific middleware, used in its infrastructure.

TD6 (Derive test procedures) describes the testing process. In Mjolnirr we use both automated and user-provided test cases, so we should cover all the coverage items, provided by microservice interface, and call all the test cases, provided by application developer. Test case can modify the microservice environment and state, therefore, microservice validation should be executed in an isolated environment, which should be purged after each set execution. As shown on figure 6, Mjolnirr has three sets of environments – special validation container for isolated component-level validation, validation environment for integration-level validation and production environment for end users. Validation environment and production environment has built-in Chaos Monkey service to ensure system stability. These three steps should be passed by each microservice to go to production.

To use those techniques, we need to have detailed microservice interface description, including input and output variables types, boundaries and syntax. Using this information, we can perform black box testing to each microservice in addition to the simple unit tests. Microservice interface definition described on figure 7.

```
{
  "name": "test_microservice",
  "inputs": [
    {
      "type": "integer",
      "min": 0,
      "max": 10
    },
    {
      "type": "string",
      "syntax": "he.?lo"
    }
  ],
  "outputs": [
    {
      "type": "integer",
      "min": 0,
      "max": 10
    }
  ],
  "input_connections": ["test_microservice2"],
  "output_connections": ["test_microservice3"]
}
```

Figure 7. Microservice interface description

In a case of integration testing we have knowledge about internal structure of the tested system. Therefore, we can use the testing techniques, mentioned before, and structure-based test design techniques, such as data flow testing, branch condition testing, etc. As shown on figure

5, microservice interface also describes input and output connections for each microservice, and we can use this knowledge to produce call graph and validate this structure.

Integration-level validation cannot be implemented on a container level, so this type of validation will be handled by Proxy node. Proxy will store the call graph, compare it with the desired structure, based on component interface. Also, Proxy will use different integration validation techniques, like Chaos Monkey [4] and random messaging, which are useful in a case of high load system.

V. CONLUSION

In this article, we described microservice testing features and validation techniques. In addition, specific test design techniques were mentioned. We have described the basics of microservice approach, provided and overview of methods and standards of software testing and validation. We described microservice validation features and proposed an approach to microservice validation. At last we proposed a brief overview of possible methods of implementation of microservice validation framework based on a Mjolnirr platform.

As a further development of this study, we will implement microservice validation framework, which will be able to perform automated and semi-automated validation to the microservices.

REFERENCES

[1] Bartolini C., Bertolino A., Elbaum S., Marchetti E. Whitening SOA testing. *Proceedings of the the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering.* 161—170.

[2] Brooks F.P. No Silver Bullet. Essence and accidents of software engineering. IEEE computer 20(4); 1987: 10-19.

[3] Calchado P. Building Products at SoundCloud—Part III: Microservices in Scala and Finagle. [Online] Available at: https://developers.soundcloud.com/blog/building-products-at-soundcloud-part-3-microservices-in-scala-and-finagle [accessed 6.02.2015].

[4] Chaos Monkey. [Online] Available at: https://github.com/Netflix/SimianArmy/wiki/Chaos-Monkey [accessed 12.01.2015]

[5] Clemson, T. Testing Strategies in a Microservice Architecture. 2014. [Online] Available at: http://martinfowler.com/articles/microservice-testing/ [Accessed 10.01.2015].

[6] Ford N. Building Microservice Architectures. [Online] Available from: http://nealford.com/downloads/Building_Microservice_Architectures(Neal_Ford).pdf [accessed 13.02.2015]

[7] International Organization for Standardization (2014) 25010. Service Quality Requirements and Evaluation (SquaRE) – System and software Quality Model, 2014, Canada.

[8] International Organization for Standardization (2014) 29119. Software Testing, 2014, Canada.

[9] International Organization for Standardization (2015) 25011. Service Quality Requirements and Evaluation (SquaRE) – Service Quality Model, 2015, China.

[10] Jehan S., Pill I., Wotawa F. Functional SOA Testing Based on Constraints. *Proceedings of the 8th International Workshop on Automation of Software Test.* 33—39.

[11] Jones S. Microservices is SOA, for those who know what SOA is, 2014. [Online]. Available at: http://service-architecture.blogspot.ru/2014/03/microservices-is-soa-for-those-who-know.html. [Accessed: 17.11.2014].

[12] Kant N., Tonse T. Karyon: The nucleus of a Composable Web Service. [Online] Available at: http://techblog.netflix.com/2013/03/karyon-nucleus-of-composable-web-service.html [accessed 17.11.2014].

[13] Lewis J., Fowler M. Microservices. [Online] Available at: http://martinfowler.com/articles/microservices.html [accessed 17.11.2014].

[14] Prana: A Sidecar for your Netflix PaaS based Applications and Services. [Online] Available from: http://techblog.netflix.com/2014/11/prana-sidecar-for-your-netflix-paas.html [accessed 13.02.2015]

[15] Radchenko G., Mikhailov P., Savchenko D., Shamakina A., Sokolinsky L. Component-based development of cloud applications: a case study of the Mjolnirr platform. *Proceedings of the 10th Central and Eastern European Software Engineering Conference in Russia (CEE-SECR '14).* Article 6, 10 pages.

[16] Rhoton J, Haukioja R. Cloud computing architected. [Tunbridge Wells, Kent]: Recursive Press; 2011.

[17] Savchenko D., Radchenko G. Mjolnirr: A Hybrid Approach to Distributed Computing Architecture and Implementation // CLOSER 2014. Proceedings of the 4th International Conference on Cloud Computing and Services Science (Barcelona, Spain 3-5 April, 2014), 2014. P. 445-450.

[18] Thones J. Microservices. Software, IEEE, Volume 24 Issue 3, 2015. P. 116-116.

[19] Tonse S. Microservices at Netflix. [Online] Available from: http://www.slideshare.net/stonse/microservices-at-netflix [accessed 6.02.2015].

[20] Varia J. Cloud architectures. White Paper of Amazon Web Services, 2008. [Online] Available at: https://media.amazonwebservices.com/AWS_Cloud_Architectures.pdf: 16. [accessed 17.11.2014].

Four level provenance support to achieve portable reproducibility of scientific workflows

A. Bánáti[1], P. Kacsuk[2,3] and M. Kozlovszky[1,2]
[1] Óbuda University, John von Neumann Faculty of Informatics, Biotech Lab
Bécsi str. 96/b., H-1034, Budapest, Hungary
[2] MTA SZTAKI, LPDS, Kende str. 13-17, H-1111, Budapest, Hungary
[3] University of Westminster, 115 New Cavendish Street, London W1W 6UW
{banati.anna, kozlovszky.miklos}@nik.uni-obuda.hu,
kacsuk@sztaki.mta.hu

Abstract - In the scientist's community one of the most vital challenges is the issue of reproducibility of workflow execution. In order to reproduce the results of an experiment, on one hand provenance information must be collected and on the other hand the dependencies of the execution need to be eliminated. Concerning the workflow execution environment we have differentiated four levels of provenance: infrastructural, environmental, workflow and data provenance. During the re-execution at all levels the components can change and capturing the data of each levels targets different problems to solve. For example storing the environmental and infrastructural parameters enables the portability of workflows between the different parallel and distributed systems (grid, HPC, cloud). The describers of the workflow model enable tracking the different versions of the workflow and their impacts on the execution. Our goal is to capture the most optimal parameters in number and type as well and reconstruct the way of data production independently from the environment. In this paper we investigate the necessary and satisfactory parameters of workflow reproducibility and give a mathematical formula to determine the rate of reproducibility. These measurements allow the scientist to make a decision about the next steps toward the creation of reproducible workflows.

I. INTRODUCTION

Nowadays in many scientific research fields the experiments are especially compute and data intensive processes. The execution of these experiments in most cases requires parallel and distributed infrastructures.

The successive steps of an experiment are chained to a so called workflow, which can be represented by a directed acyclic graph (DAG). The nodes are so called *job*s, which includes the experimental computations based on the input data accessed through their input ports. In addition these jobs can product output data, which can be forwarded through their output ports to the input port of the next job. The edges of a DAG represent the dataflow between the jobs (Figure 1.).

Different users for different purposes may be interested in reproducing the workflow, for example scientists (author of workflow) in order to prove their results, readers or other scientists in order to reuse results or reviewers in order to verify the correctness of the results [1]. The scientist community has to face two important challenges related on the reproducibility of results: on one hand more and more metadata and provenance data are necessary about the infrastructure, the environment, the changes and the partial results of an execution in order to reconstruct the execution at a later time. The collected data called provenance help to store the actual environments, the partial and final data products and system variables [2], [3]. The saved information can be used in many fields concerning workflows (failure tolerance, optimization, user steering and dynamism [4], [5]). During the processing of scientific workflows four components can change: infrastructure, environment, abstract workflow model and input data. The provenance information collected at the different levels can support different features, services and goals. On the other hand the preparing reproducible workflows require careful and precise design which eliminates the dependencies and includes detailed description about the model and input/output data. Examples of dependencies can be a special hardware requirements, demand of third party or special local services or the usage of random based computing.

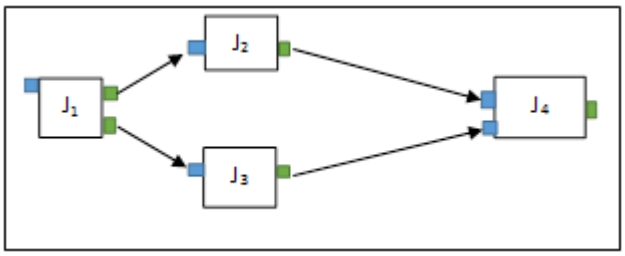


Figure 1. Workflow example with four jobs (J_1, J_2, J_3, J_4)

In this paper we investigate the necessary and satisfactory parameters of reproducibility, we examine the possibility and advantages of a four level provenance. We also give a mathematical model to measure the rate of reproducibility (RoR) to support the scientist in designing and creating reproducible workflow.

The rest of the paper is organized as follows: In the next section we provide a short background and overview about works related to our research. Section 3 presents the necessary and satisfactory requirements of scientific workflow reproducibility including our four level provenance structure. In section 5 we give a mathematical model and define the dependency functions to determine

the rate of reproducibility of a given workflow. Finally we summarize our conclusions and reveal the potential future research directions.

II. RELATED WORKS

Currently the reproducibility of scientific workflows is a burning question which the scientist community has to face with and has to solve. Accordingly in the latter one-two years many researchers investigate this issue. One part of the literature analyzes the requirements of reproducibility and the other part deals with the implementation of such tools or frameworks.

The first group agree on the importance of the careful design [6], [7], [8], [9], [10] which on one hand means the increased robustness of the scientific code, for example with a modular design and detailed description of the workflow, and of the input and output data examples, and consequent annotations [11]. On the other the careful design includes the careful usage of volatile third party or special local services. In these cases two solutions exist, but reproducibility is uninsurable: 1. taking a digital copy of the entire environment using a system virtual machine/hardware virtualization approach 2. capturing and storing metadata about the code and environment that allows it to be recreated later [11].

Zhao et al. [12] in their paper investigate the cause of the so called *workflow decay*, which means that year by year the ability and success of the re-execution of any workflow significantly reduces. They examined 92 Taverna workflows submitted in the period from 2007 to 2012 and found four major causes: 1. Missing volatile third party resources 2. Missing example data 3. Missing execution environment (requirement of special local services) and 4. Insufficient descriptions about workflows. Hettne et al. [13] in their paper list ten best practice to prevent the workflow decay. Grothe et al. [14] analyze the characteristic of applications used by workflows and list the requirements in order to enable the reproducibility of results and determination of provenance. Compared to the former mentioned requirements they assumed the deterministic feature of applications in order to perform appropriate provenance collection.

There exist available tools, VisTrail, ReproZip or PROB [15], [16], [17], which allow the researcher and scientist to create reproducible workflow. With help of VisTrail [15], [18] reproducible paper can be created, which includes not only the description of scientific experiment, but all the links for input data, applications and visualized output which always harmonizes with the actually applied input data, filter or other parameters. ReproZip [16] is another tool, which stitches together the detailed provenance information and the environmental parameters into a self-contained reproducible package.

III. NECESSARY AND SATISFACTORY REQUIREMENTS

The two main difficulties of reproducing a workflow are dependencies of workflow execution and collection of provenance data. The former can be determined as necessary requirements and the latter one gives the satisfactory requirements.

A. *Necessity – Dependencies*

The execution of a workflow may require many resources, such as third party or local services, database services or even special hardware infrastructure. These resources are not constantly available, they can change their location, their access condition or the provided services from time to time. These conditions, which we refer to as dependencies, significantly complicate the chances of reproducibility and repeatability. We have classified the dependencies into three categories: infrastructural dependency, data dependency and job execution dependency as shown in table 1.

TABLE I. CATEGORIES OF DEPENDENCIES OF WORKFLOW EXECUTION

infrastructural	data	job execution
• spec. hardware demand	• changing • TP demand • local spec demand	• deterministic • dependency between jobs • Third Party demand • Local spec demand

By infrastructural dependency we mean special hardware requirements, which are available solely on the local system or not evidently provided by other systems, such a special processing unit (GPU, GPGPU).

In the group of data dependency we listed the cases which does not guarantee the accessibility of the input dataset in another time interval. The causes can be that the data is provided by a third party or special local services. Occasionally the problem origins from the continuously changing and updated database that stores the input data. These changes are impossible to restore from provenance data.

The job execution can also depend on a third party or local services, but the main problem arise when the job execution is not deterministic. The operation of GPU or GPGPU are based on random processes consequently the results of re-executions may differ. More over if the dependency factor is too high between the jobs the reproducibility is harder to guarantee.

These conditions are all necessary to perform reproducibility of workflow execution. In the next section (IV) we give a mathematical formula to determine the rate of reproducibility of a given workflow. With help of this measurement the scientist can see how much part of the workflow can be reproducible with 100 percent at a later period of time. Knowing this information the scientist can decide to apply for example an extra provenance policy with extra resource requirement, which stores the whole third party data or apply virtual machine towards the reproducibility.

B. *Satisfying – Provenance*

We defined four levels of provenance data because during the execution of workflow four components can change that would affect the reproducibility: the infrastructure, the environment, the data and the workflow model.

The first is a system level provenance, which stores the type of infrastructure, the variables of the system and the timing parameters. At this level happens the storing the details of the mapping process as a results we can answer the question of what, where, when and how long has been executed. This information supports the portability of the workflow which is a crucial requirement of reproducibility.

The environmental provenance stores the actual execution details which includes the operating system properties (identity, version, updates, etc), the system calls the used libraries and the code interpreter properties. The execution of a workflow may rely on a particular local execution environment, for example, a local R server or a specific version of workflow execution software, which also has to be captured as provenance data or virtual machine snapshot.

The third category is data provenance. In the literature the provenance often refers to data provenance, which deals with the lineage of a data product or with origin of a result. With this data provenance we can track the way of the results and dependency between the partial results. This information can support the visualization, the deep and complete troubleshooting of the experimental model, the proving of the experiment but first of all the reproducibility. In addition in one of our previous paper [Krakko] we investigated the possibility and the need of user steering. We found that some parameters, filter criteria and input data set need to be modified during execution, which rely on data provenance.

The last provenance level tracks the modifications of the workflow model. The scientist during the workflow lifecycle often performs minor changes, which can be undocumented and later it is difficult to identify or restore. This phenomenon is usually referred as workflow evolution. Provenance data collected at this level can support the workflow versioning.

These structured provenance information of a workflow can support reproducibility at different levels if it meets the requirements of independency discussed before. In addition extra provenance information can be stored in that cases, in which however the workflow contains some dependencies but these dependencies can be eliminated with usage of extra resources.

IV. MEASUREMENT OF REPRODUCIBILITY

A. The model

The *workflow* (Wf) is represented by the set of job, the *job* (*J*) is defined by the series of input and output ports and the *dataflow* is considered between the given output port of a job and input port of another job:

- $WF=\{J_1, J_2, \ldots, J_N\}$,
- $J_i(p_1, p_2, \ldots, p_{Ki}, P_1, P_2, \ldots, P_{Li})$
- $dataflow(J_i, P_l;J_j;p_k)$

where N, K_i, L_i are natural numbers and denote the number of jobs in a given workflow, the number of input ports (p_i) and output ports (P_i) in a given J_i job.

B. Dependency functions

In order to determine the measure of reproducibility of a given workflow we need to define six binary dependency functions. These functions can indicate whether a given job or port depends on any services, input data, random value or each other, in which cases the reproduction of workflow is not possible or especially difficult. We call these criteria together *non-repro* criteria.

$$Dep_{service}(J_i) = \begin{cases} 1, \text{ if the } J_i \text{ job calls a Third party or special local service} \\ 0, \text{ else} \end{cases} \quad (1)$$

$$Dep_{infra}(J_i) = \begin{cases} 1, & \text{if the } J_i \text{ job depends on a special hardware} \\ 0, & \text{else} \end{cases} \quad (2)$$

$$Dep_{input}(J_i, p_j) = \begin{cases} 1, \text{ if the } p_j \text{ input port of the } J_i \text{ job depends on third party, local or changing data} \\ 0, \text{ else} \end{cases} \quad (3)$$

$$Dep_{random}(J_i) = \begin{cases} 1, & \text{if the } J_i \text{ job depends on true random generator} \\ 0, \text{ else} \end{cases} \quad (4)$$

$$Dep_{all}(J_i) = \sum_{i=1}^{N} \left(Dep_{service}(J_i) + Dep_{infra}(J_i) + Dep_{random}(J_i) \right) + \\ + \sum_{i=1}^{N} \sum_{j=1}^{K_i} Dep_{input}(J_i, p_j) \quad (5)$$

$$Dep_{job}(J_i, J_j) = \begin{cases} 1, & \text{if the } J_i \text{ job depends on } J_j \text{ and } Dep_{all}(J_j) \geq 1 \\ 0, & \text{else} \end{cases} \quad (6)$$

The functions (1), (2), (3), decide whether a given job or input port use third party services, local services, or the execution of the job requires special hardware. The function (4) examines the determinism of the results of a given job. A result of a job is deterministic, if during repeated executions the same input data give the same results. For example a computation is non deterministic, if it based on random process and random values or the job requires human intervention. The function (5) returns 1, if at least one of the above mentioned criteria is true. The last function examines the relation between a given job and all other jobs, and it returns 1, if the given job is based on a job which depends on a non-repro criteria (one or more of the functions (1), (2), (3), (4), (5) return 1).

C. Rate of Reproducibility

After determining the dependency functions we can define the *Decay*(J_i) function, which returns 1 if the job is not reproducible.

$$Decay(J_i) = \begin{cases} 1 \ \textit{if} \ Dep_{all}(J_i) + \sum_{j=1}^{N} Dep_{job}(J_i, J_j) \geq 1 \\ 0 \ \textit{else} \end{cases} \quad (7)$$

With help of function (7) we can define the rate of reproducibility, RoR:

$$RoR(Wf) = \frac{\sum_{i=1}^{N} Decay(J_i)}{N} \quad (8)$$

The RoR function says how big part of a given workflow is surely reproducible. At this point, the user – before finally submits his workflow – can think over the model, after viewing the results he can modify the model and eliminates certain dependencies or he can decide to apply extra provenance or virtualization tools to preserve the workflow.

V. CONLUSION

In this paper we investigated the necessary and satisfactory requirements of scientific workflow reproducibility. We approached the problem from two perspectives. On one hand we deal with the design and preparation of a workflow. Many best practices for appropriate designing exist in the literature, but one of the most important issue investigates the dependencies of execution (infrastructural dependency, services dependency or random based compute dependency), which add more difficulties to the reproducibility of a workflow at a later time. To eliminate these dependencies there are the necessary requirements of reproducibility. We gave a mathematical model to describe the execution dependencies and we defined the rate of reproducibility. Our goal is to support the scientists to make their experiment reproducible or to support in using some other tools (extra provenance data collection or virtual machine) to achieve reproducibility.

The second perspective is the possible structures of provenance. We determined a four level provenance, which gives the satisfactory requirements of the reproducibility. The information captured at the different levels can support the different levels of reproducibility, such portability or repeatability.

REFERENCES

[1] D. Koop, E. Santos, P. Mates, T.Vo Huy, P Bonnet, B. Bauer, M.Troyer, D.N. Williams, J.E. Tohline, J. Freire, C.T. Silva, „A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers”, Internatioonal Conference on Computational Science, ICCS 2011. [Online]. Available: http://www.sciencedirect.com.

[2] J. Cheney, A. Finkelstein, B. Ludäscher, és S. Vansummeren, „Principles of provenance (dagstuhl seminar 12091)”, Dagstuhl Reports, köt. 2, sz. 2, 2012

[3] S.M.S da Cruz, Maria Luiza M. Campos, M. Mattoso. „Towards a Taxonomy of Provenance in Scientific Workflow Management Systems”, 259–66. IEEE, 2009. doi:10.1109/SERVICES-I.2009.18.

[4] Bánáti, A., Kacsuk, P., Kozlovszky, M.: Towards Flexible Provenance and Workflow Manipulation in Scientific Workflows. In Proceedings of CGW 14 (2014)

[5] Kail, E., Bánáti, A., Kacsuk, P., Kozlovszky, M.: Provenance based adaptive and dynamic workflows. In: 15th IEEE International Symposium on Computational Intelligence and Informatics, pp 215-219, IEEE Press, Budapest (2014).

[6] P. Missier, S. Woodman, H. Hiden, és P. Watson, „Provenance and data differencing for workflow reproducibility analysis”, Concurrency and Computation: Practice and Experience, 2013

[7] R. D. Peng, „Reproducible Research in Computational Science”, Science, köt. 334, sz. 6060, o. 1226–1227, dec. 2011

[8] J. P. Mesirov, „Accessible Reproducible Research”, *Science*, köt. 327, sz. 5964, o. 415–416, jan. 2010.

[9] D. De Roure, K. Belhajjame, P. Missier, J. M. Gómez-Pérez, R. Palma, J. E. Ruiz, K. Hettne, M. Roos, G. Klyne, C. Goble, és others, „Towards the preservation of scientific workflows”, in Procs. of the 8th International Conference on Preservation of Digital Objects (iPRES 2011). ACM, 2011.

[10] S. Woodman, H. Hiden, P. Watson, és P. Missier, „Achieving reproducibility by combining provenance with service and workflow versioning”, in Proceedings of the 6th workshop on Workflows in support of large-scale science, 2011, o. 127–136.

[11] A. Davison, „Automated Capture of Experiment Context for Easier Reproducibility in Computational Research”, Computing in Science & Engineering, köt. 14, sz. 4, o. 48–56, júl. 2012.

[12] J. Zhao, J. M. Gomez-Perez, K. Belhajjame, G. Klyne, E. Garcia-Cuesta, A. Garrido, K. Hettne, M. Roos, D. De Roure, és C. Goble, „Why workflows break—Understanding and combating decay in Taverna workflows”, in E-Science (e-Science), 2012 IEEE 8th International Conference on, 2012, o. 1–9.

[13] K. M. Hettne, K. Wolstencroft, K. Belhajjame, C. A. Goble, E. Mina, H. Dharuri, D. De Roure, L. Verdes-Montenegro, J. Garrido, és M. Roos, „Best Practices for Workflow Design: How to Prevent Workflow Decay.”, in SWAT4LS, 2012

[14] P. Groth, E. Deelman, G. Juve, G. Mehta, és B. Berriman, „Pipeline-centric provenance model”, in Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, 2009, o. 4.

[15] J. Freire, D. Koop, F. S. Chirigati, és C. T. Silva, „Reproducibility Using VisTrails”, 2014. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download doi:10.1.1.369.9566

[16] F. S. Chirigati, D. Shasha, és J. Freire, „ReproZip: Using Provenance to Support Computational Reproducibility.”, in *TaPP*, 2013.

[17] V. Korolev, A. Joshi, V. Korolev, M. A. Grasso, A. Joshi, M. A. Grasso, D. Dalvi, S. Das, V. Korolev, Y. Yesha, és others, „PROB: A tool for Tracking Provenance and Reproducibility of Big Data Experiments.”, *Reproduce’14. HPCA 2014*, köt. 11, o. 264–286, 2014.

[18] D. Koop, J. Freire, és C. T. Silva, „Enabling Reproducible Science with VisTrails”, *arXiv preprint arXiv:1309.1784*, 2013.

Using the P-TOSCA Model for Energy Efficient Cloud

Bisera Ivanovska, Sasko Ristov, Magdalena Kostoska, Marjan Gusev
Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, Macedonia
Email: ivanovskabisera@yahoo.com, {sashko.ristov, magdalena.kostoska, marjan.gushev}@finki.ukim.mk

***Abstract*—The Topology and Orchestration Specification for Cloud Applications (TOSCA) standard is used to describe a cloud application and cloud architecture in order to allow a portable deployment to other compatible cloud and multi-cloud applications. P-TOSCA is recently proposed model and proved concept, which is an extension of the TOSCA standard to improve the TOSCAs ambiguities and weaknesses. In this paper we use the P-TOSCA model for other issues that are also very important in virtualized datacenters and cloud computing, that is, to enlarge/extend the energy efficient management system. A prototype application that dynamically creates a target virtual machine on utilized physical compute node, ports the application(s) from a virtual machine hosted on an underutilized physical server to the target virtual machine in Eucalyptus cloud is presented, which is specified with P-TOSCA. After migration, the prototype application will shut down the underutilized empty physical node.**

***Keywords*-Cloud computing, green computing, portability.**

I. Introduction

The employment of cloud computing offers new processing models that allows penetrations of applications from scientific, user and business domain. For this purpose virtualization is used, which as first phase is deeply implemented in data centers. The next phase should be directed towards interoperability, portability and energy efficiency.

The amount of energy that servers in cloud/data center environment use altogether with the amount of energy consumed by cooling systems represents a large portion of the overall cost of spent energy. Below are some examples given of how much energy is consumed in order to realize the importance of energy saving and reducing costs.

The US Agency for Protection reported of 100 billion KWh energy spent in 2011, whereas the infrastructure and energy cost was estimated to be 75% from the overall costs in 2014. Worldwide, the amount of energy spent in data centers has increased for 56% for the period from 2005 to 2010 and the proper cost was 1.5% of the total electricity use [1]. In addition, authors in [2] investigated that the average amount of energy spent on one Google site is 180kWatts, if we take into account that at each site there are 40 racks, each with 80 PCs connected. Therefore, reducing the amount of energy is of great importance not only in lowering costs, but also for the reliability, availability and productivity of the cloud environment/data centers. Besides it will contribute to a cleaner environment.

Another challenge is enabling automated portability and interoperability. In this area there are several research directions in creation of standards and complementary solutions like TOSCA, CDMI, OCCI, OCF, jClouds or mOSAIC. An addition of these solutions is P-TOSCA, an extension of the current TOSCA standard.

In [3] we proposed a solution for energy efficient data center, where the administrator manages the physical servers by startting or shutting down, them thus reducing the amount of wasted energy, and cost. On the other hand, in [4], [5] a model of an automated portability of applications is designed that is applied to migrate applications between different cloud providers. In this paper we merge these two models. Moreover, we propose an innovative solution for constant consolidation, integration, migration of VMs based on available resources, network topology and server status, and providing reliable QoS. We are extending the domain of P-TOSCA with energy efficiency module.

The remainder of this paper is organized as follows.

II. Related work

The larger the cloud data center is, the more energy it consumes. Many authors have proposed energy management models and mechanisms, such as the mechanism described in the paper [6]. The goal of this mechanism is to adjust the voltage of the system based on the utilization of the CPU, by transferring tasks from the heavy-server to the server that has more resources available. Another task scheduling model is presented in [7]. Authors have proposed approach for minimizing energy consumption by scheduling tasks to a minimum numbers of servers. The problem is formulated as an integer programming problem which minimizes energy consumption by scheduling tasks to a minimum numbers of servers while keeping the task response time constraints.

A Reinforcement Learning-based Dynamic Consolidation method (RL-DC) is proposed in [8] that aims to minimize active hosts based on current resource management. The servers power mode is determined by an agent that uses reinforcement learning method to learn from past knowledge in order to decide when a host should be switched to proper mode.

In order to minimize the energy consumption and VM migration at same time, authors in [9] focused on dynamic

on-demand resource model. This model is based on network-flow algorithm that switches part of idle servers off. For many researches, the main technique to develop an energy efficient data center is to migrate VMs and power off the unnecessary compute nodes.

Authors in [2] proposed an algorithm that will monitor the servers resources, and then by demand to switch the node off. Although this technique gives good direction for saving energy, it is not dynamic enough touching user specifications and needs. On the contrary, the domain of implementation is not the cloud, indeed in the papers we analyzed that consolidation of VMs between heterogeneous clusters it is not considered.

Similar approach is presented for energy efficient scheduling of Virtual Machines in cloud data centers [10]. A combination of an optimal allocation with a consolidation algorithm of VMs has been proposed, so when resources are leased, a VM migration is performed to adapt placement of those resources. However, as the previous example, this paper as well does not address the cloud environment i.e. is this migration performed within the same cloud data center? Furthermore, is migration between two different cloud platforms considered?

We concluded that in order to spent less energy the focus of the most proposed approaches are shifting tasks between servers or dynamic algorithms are implemented, that on base on available resources VMs are migrated, and therefore the server node is turned off. But, what if specific VM is needed to be created or deleted?

Furthermore, there is not addressed or even designed an energy efficiency and portability model for cloud data center. This gave us an additional motivation to explore and also develop such model that features consolidation, migration and incorporation of VMs on the basis of available resources of the physicals servers, and providing reliable QoS.

III. Background

In this section we will give a brief overview of the P-TOSCA model as well as on the Application for Modern Energy Efficient Data Center.

A. P-TOSCA

Authors in [4], [5] had designed and implemented a model of an automated portability of Platform as a Service hosted applications from one cloud environment to another. This model is called P-TOSCA, it extends TOSCA 1.0 standard [11], thus representing an application specification.

The main idea is to enable a possibility to switch between different cloud providers and vendors. P-TOSCA model [4] is used as an application specification and automated portability of Platform-as-a-Service (PaaS) hosted applications from one cloud environment to another. P-TOSCA is an extension of TOSCA 1.0 standard [12] and defines details on PaaS implementation.

A deployment and migration using TOSCA and P-TOSCA is performed by using Cloud Service Archive (CSAR), artifact that includes a XML description of the application topology

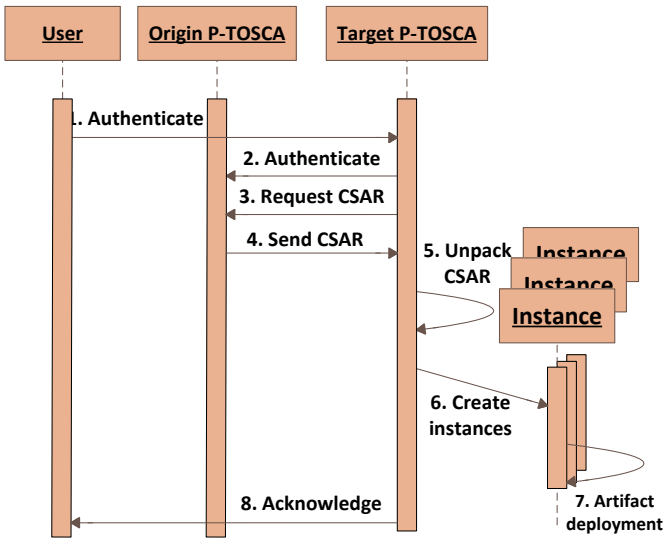


Fig. 1. Activities in P-TOSCA engine application portability

and deployment required artifacts like scripts, war files, zip files, libraries etc. Fig. 1 presents a set of steps performed by the P-TOSCA engine during application migration. In step 1 the user authenticates to the target cloud and specifies the origin cloud and the application. In step 2 authenticates on the origin cloud by using the target P-TOSCA. In steps 3 and the user requests the CSAR for the chosen application which is sent to the target cloud. In steps 5, 6 and 7 the P-TOSCA engine unpacks the CSAR, processes the model, creates the instances and deploys the artifacts. Finally, in step 8, the user is informed of the action result.

B. App for Energy Efficient Data Center

A new approach, named Application for Modern Energy Efficient Data Center was presented in [3]. The purpose is to reduce the amount of wasted energy by remotely managing, i.e. powering on/off the servers in the data center. The data center consists of a pool of servers connected on a switch, thus forming an Ethernet network of one or several VLANs. One server is used as a management station, and the other servers, called clients, must have previously enabled the WOL property on their NICs in the BIOS, so that they can listen to the network for incoming packets. When the administrator selects the proper server to power on/off, the IP and MAC addresses of that server are used to form a magic packet. Then, this packet is send through the Wake-On-LAN protocol as UDP packet to the destination.

IV. The new model

In this section we describe the model we propose. We present the challenge and we propose an extension to the P-TOSCA model which will include energy efficient resource management.

```
<NodeTemplate id="LinuxVm"
              name="LinuxVm for Tomcat"
              type="Server">
       <Properties>
         <ServerProperties>
              <NumCpus>1</NumCpus>
              <Memory>1024</Memory>
              <Disk>10</Disk>
              <EnergyEfficient>
                    true
              </EnergyEfficient>
         </ServerProperties>
       </Properties>
...
</NodeTemplate>
```

Listing 1. Example of extended *Node Template* with *EnergyEfficient* property

A. Challenge

The main challenge is to create energy efficient management system that will operate along with the P-TOSCA engine. The main goal is to create a prototype application that dynamically creates a target virtual machine on utilized physical compute node, ports the application(s) from a virtual machine hosted on an underutilized physical server to the target virtual machine in Eucalyptus cloud. For this goal we use and extent the P-TOSCA. Thus not only that energy efficiency will be achieved, but also a better performance can be obtained.

B. P-TOSCA Extension

In P-TOSCA each virtual node is specified using Node Template, defined and included in the XML description of the application topology. Our main goal is to extend the node description by adding another property to the template which will indicate that the decision for node location will be performed by the application for energy efficient management. For that purpose we extend the ServerProperties element of the Node Template by adding new property: *EnergyEfficient*. This property can have Boolean values (i.e. true or false). Listing 1 illustrates the extended Node Template definition.

C. Architecture

By using this model we can migrate instance from one server to destination, target physical server, and turning off the source server in order to save energy consumption. In Fig. 2 the architecture of the energy efficient portability mode is given followed by an explanation.

When VMs has to be migrated or instantiated, the proper server has to be determined. For this purpose, the Energy Efficient Application utilizes Intelligent Agent, discussed in [13]. If in the cloud data center exists a server that has enough resources needed, then this agent returns the particular server as target server.

In the next step, this target server is used by the P-TOSCA engine to instantiate or migrate the VM on that particular server. (Step 2 from the sequence diagram).

Then, the Intelligent Agent again checks for allocated and/or available resources among the servers in the cloud data center, in order the administrator to be aware of their status. Based on these statuses, the administrator has the privilege to turn off the server that has no resources allocated. (step 5).

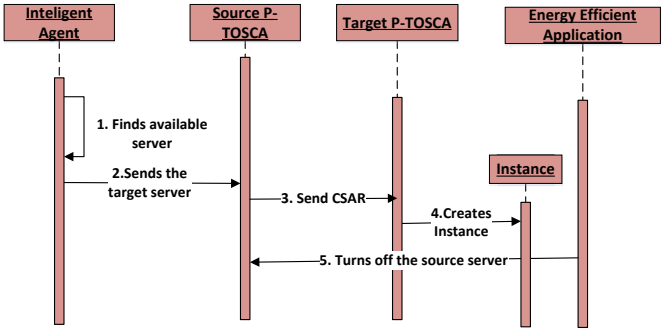


Fig. 2. P-TOSCA energy efficient portability mode

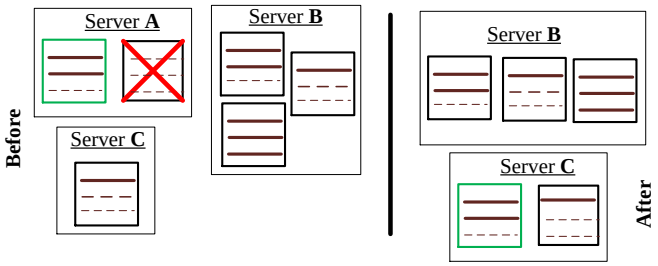


Fig. 3. Case study 1: before and after scenarios

V. Pilot Application

In this section we will describe the energy efficient model with P-TOSCA specification through the following two cases that are usually challenge for cloud:

- Which physical server to migrate particular virtual machine on?
- What physical server to instantiate new virtual machine on?

In both cases energy efficient and QoS is considered.

1) Case 1: Migration using P-TOSCA / Shutdown: This case we tested on four computer nodes, of which three are active. As shown on Fig. 3 servers A and B have two virtual machines instantiated, with corresponding resources, and server C has one. From server B we shut down one virtual machine, leaving one VM active. In order to save energy, we first migrate the virtual machine on some of the active server nodes left, and then power the compute node off. In order to do this, we start by listing all the available and used resources from the controller machine. The question is, on which server to migrate the virtual machine? A resource strategy described in [13] determines/finds the server with enough resources for the new VM. If proper server exists, we set that server as target server in the extended P-TOSCA specification, in order to migrate the virtual machine on that particular server.

2) Case 2: Instantiating new Virtual Machine: The next case that we will consider is when a new virtual machine (with particular resources such as number of cores, memory.)

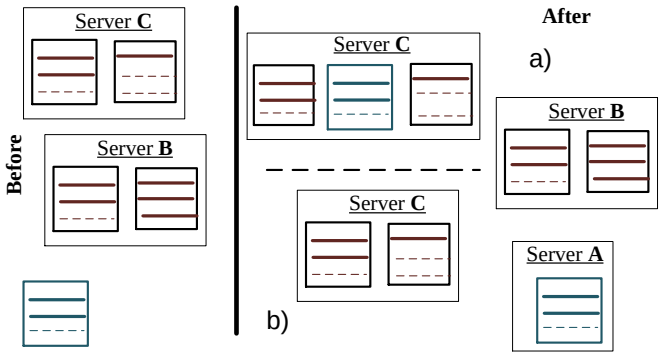


Fig. 4. Case study 2: before and after scenarios

has to be instantiated. The whole scenario is presented on Fig. 4. The testing environment for this case is as same as in case 1, i.e. it differs in number of active servers. Before the new virtual machine is instantiated, the proper server has to be determined/settled, whether it will be an existing or a new one. A brief discussion (of these sub scenarios) follows.

For this scenario we have 2 active server nodes, with 2 and 3 active VMs respectively, and two non-active. In order to determine where to instantiate the new VM, a resource provision strategy is used, based on the greedy algorithm [13]. If the number of required resources of the new VM matches with the number of available resources left from any of the active server nodes, the algorithm returns the particular server as target server. Then, using the P-TOSCA model we instantiate the new VM on that target/particular server.

On the contrary, if none of the active serves matches the required resources, using the application first we click on the server icon to power it on and then mark that server as target server in the P-TOSCA specification. At last, the new VM is instantiated on that server.

VI. Discussion

A. Performance impact

Physical server underutilization will enable better performances. In this paper a diiferent approach with same purpose is presented.

Each physical server consumes $P_0 + L * P_L$ electricity, where:

- P_0 is the energy consumed without utilization,
- L is the overload
- P_L is the energy consumed during overload

As the number of active servers decreases, the total amount of electricity will also decrease.

B. Cooling problems

Our previous paper [3] discussed energy efficient laboratory, that is, usually laboratory with 20 PC-s, without concerning the temperature conditions. Usually, the cloud data system does not consists of a small number of servers, but it consists of several racks and each rack has approximately 20 physicals

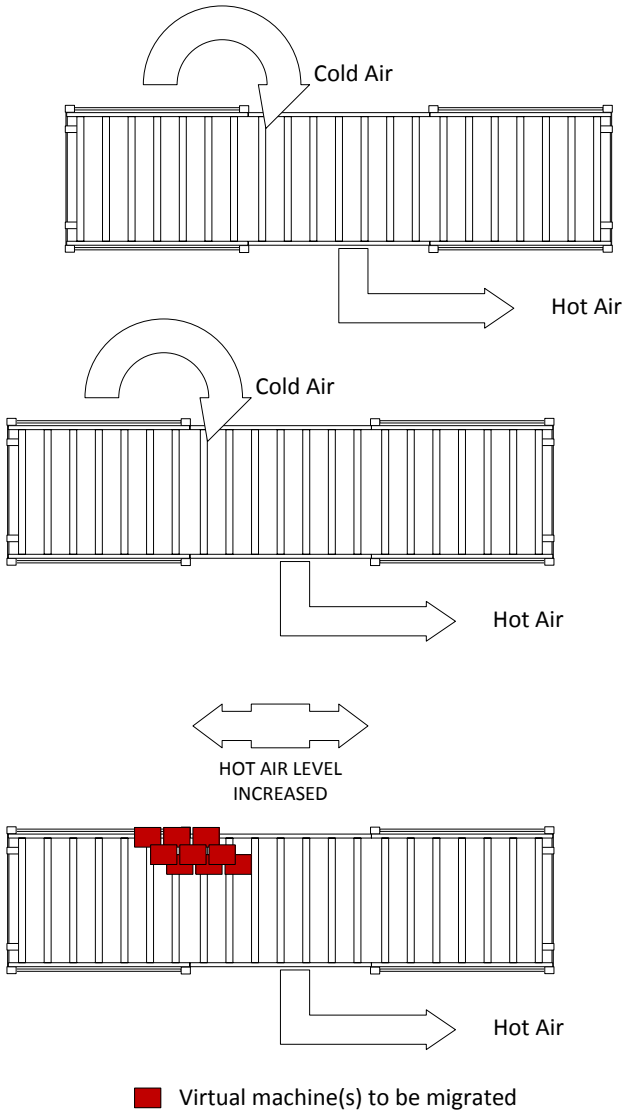


Fig. 5. Cooling consumption in cloud data center

servers, from where multiple virtual machines are instantiated. As analyzed in [14], for every watt of power, server needs 1 to 2 watts of cooling. Therefore cooling takes more than half of the total electricity usage in cloud data centers.

Besides Energy Efficient model, a low-power cooling system in a data center is needed so that costs can be reduced. Therefore, our next challenge is to develop a new solution that will determine which virtual machine from which physical server of a particular rack has to be migrated, so that the physical server is powered off in order to accomplish cooling balancing and spent less electricity.

C. Live migration

As described in [15], the Live migration is performed within seconds whereas, migration with P-TOSCA can take up few minutes. Compared to our solution, live migration is much faster.

On the other hand, the domain of Live migration is limited, while the domain of our proposed solution is much broader. The P-TOSCA model can migrate application between different cloud frameworks (e.g. Eucalyptus and OpenStack) using the native cloud APIs as well as it can be used with different types of applications (e.g. N-Tier [5] and SOA [16]). Besides this, the P-TOSCA model can be applied on application that are on premise hosted and using this model to migrate the application to the cloud, unlike Live migration.

D. Open Virtualization Format (OVF)

Open Virtualization Format (OVF) represents a standard for packaging and distributing virtual appliances [17]. The general goal of the standard is to provide hypervisor agnostic migration of virtual machine software. Every virtual appliance is described using OVF meta-data, which is later used for deployment with pre-built configuration described by the data.

In comparison to other solution, the OVF standard requires a substantial network throughput in order to migrate the data, given the fact that this standard packs and distributes a whole virtual machine. Another drawback is the amount of time required for data migration. Given the fact that most images of virtual machines can be already pre-contextualized with general used software or to use local repository for software installation, the task to move a whole virtual machine can take more time than the other approaches.

VII. Conclusion and future work

This paper proposes a model for energy efficient management of a cloud center. It presents two cases for consideration and offers servers consolidation by migration of virtual nodes from one physical server to other and powering off servers. It also considers the case of shortage of physical resources in the data center and proposes instantiation of new preconfigured physical server in the given cloud as Compute Node.

For this goal we propose an extension of the P-TOSCA model by extending the description of the properties of the Node Template including energy efficiency property. By this we are planning to achieve two goals:

- During the process of new instance creation we can state the physical server to which the application node(s) will be created and the selection of the physical server will be based on intelligent algorithm
- We can migrate applications (with their nodes) from one physical server to other based on intelligent algorithm decision for energy efficiency

In our future work we plan to include metrics related to the cooling process in the intelligent algorithm for energy efficiency management. The algorithm will choose which server from which rack to be powered on given the current temperature conditions. The algorithm will also suggest which server should be powered down based on the temperature conditions in the data center (i.e. which part requires more cooling at the moment).

References

[1] I. S. Dhanoa and S. S. Khurmi, "Energy-efficient virtual machine live migration in cloud data centers," *International Journal of Computer Science and Technology (IJCST)*, vol. 5, no. 1, pp. 43–47, 2014. [Online]. Available: www.ijcst.com/vol5/spl1/inderjit.pdf

[2] E. Pinheiro, R. Bianchini, E. V. Carrera, and T. Heath, "Load balancing and unbalancing for power and performance in cluster-based systems," 2001.

[3] B. Ivanovska, S. Ristov, M. Gusev, and B. Jakimovski, "Application for modern energy efficient data center," in *Telecommunications Forum (TELFOR), 2014 22nd*, Nov 2014, in press.

[4] S. Ristov, M. Kostoska, and M. Gusev, "P-TOSCA portability demo case," in *2014 IEEE 3rd International Conference on Cloud Networking (IEEE CLOUDNET)*, 2014, pp. 269–271.

[5] M. Gusev, M. Kostoska, and S. Ristov, "Cloud P-TOSCA porting of N-tier applications," in *Proceedings of the 22nd International TELFOR Forum*, 2014, pp. 935–938.

[6] L.-T. Lee, K.-Y. Liu, and C.-Y. Tseng, "A dynamic resource management with energy saving mechanism for storing cloud computing," *International Journal of Grid and Distributed Computing*, vol. 6, no. 1, pp. 67–76, Feb 2013. [Online]. Available: http://www.sersc.org/journals/IJGDC/vol6_no1/6.pdf

[7] N. Liu, Z. Dong, and R. Rojas-Cessa, "Task scheduling and server provisioning for energy-efficient cloud-computing data centers," in *Distributed Computing Systems Workshops (ICDCSW), 2013 IEEE 33rd International Conference on*, July 2013, pp. 226–231.

[8] F. Farahnakian, P. Liljeberg, and J. Plosila, "Energy-efficient virtual machines consolidation in cloud data centers using reinforcement learning," in *Parallel, Distributed and Network-Based Processing (PDP), 2014 22nd Euromicro International Conference on*, Feb 2014, pp. 500–507.

[9] J. SiYuan, "A novel energy efficient algorithm for cloud resource management," *International Journal of Knowledge and Language Processing*, vol. 4, no. 2, pp. 12–22, 2013.

[10] C. Ghribi, M. Hadji, and D. Zeghlache, "Energy efficient vm scheduling for cloud data centers: Exact allocation and migration algorithms," in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*, May 2013, pp. 671–678.

[11] Organization for the Advancement of Structured Information Standards. (2013, Mar.) Topology and orchestration specification for cloud applications version 1.0. [Online]. Available: http://docs.oasis-open.org/tosca/TOSCA/v1.0/cs01/TOSCA-v1.0-cs01.html

[12] T. Binz, U. Breitenbücher, O. Kopp, and F. Leymann, "Tosca: Portable automated deployment and management of cloud applications," in *Advanced Web Services*. Springer, 2014, pp. 527–549.

[13] S. Ristov, M. Gusev, and G. Velkoski, "Cloud e-learning and benchmarking platform for the parallel and distributed computing course," in *Global Engineering Education Conference (EDUCON), 2014 IEEE*, Istanbul, Turkey, 2014, p. in press.

[14] Nuventix. (2008) SynJet: Low-Power Green Cooling. [Online]. Available: http://www.nuventix.com/files/uploaded_files/Low%20Power%20Cooling.pdf

[15] C. C. Keir, C. Clark, K. Fraser, S. H, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," in *In Proceedings of the 2nd ACM/USENIX Symposium on Networked Systems Design and Implementation (NSDI*, 2005, pp. 273–286.

[16] M. Gusev, M. Kostoska, and S. Ristov, "P-TOSCA portability of service oriented applications," University Ss Cyril and Methodius, FCSE, Tech. Rep. LiiT:44/2014, 2014.

[17] Distributed Management Task Force. (2010, Jan.) Open virtualization format specification version 1.1.0. [Online]. Available: http://www.dmtf.org/sites/default/files/standards/documents/DSP0243_1.1.0.pdf

Achieving dynamic workflow management system by applying provenance based checkpointing method

E. Kail[1], P. Kacsuk[2,3] and M. Kozlovszky[1,2]
[1] Óbuda University, John von Neumann Faculty of Informatics, Biotech Lab
Bécsi str. 96/b., H-1034, Budapest, Hungary
[2] MTA SZTAKI, LPDS, Kende str. 13-17, H-1111, Budapest, Hungary
[3] University of Westminster, 115 New Cavendish Street, London W1W 6UW
{kail.eszter, kozlovszky.miklos}@nik.uni-obuda.hu,
kacsuk@sztaki.mta.hu

Scientific workflows are data and compute intensive thus may run for days or even for weeks on parallel and distributed infrastructures such as HPC systems and cloud. In HPC environment the number of failures that can arise during scientific workflow enactment can be high so the use of fault tolerance techniques is unavoidable. The most frequently used fault tolerance techniques are job replication and checkpointing. While job replication is based on the assumption that the probability of single failures is much higher than of simultaneous failures, the checkpointing saves certain states and the execution can be restarted from that point later on. The effectiveness of the checkpointing method depends on the checkpointing interval. Common technique is to dynamically adapt the checkpointing interval. In this work we give a brief overview of the different checkpointing techniques and propose a new provenance based dynamic checkpointing method.

I. INTRODUCTION

Scientific workflows being data and compute intensive may require long execution time, which can even last for weeks. During such long intervals it is inevitable to adapt to the dynamically changing environment which can be caused by unwanted input data, crash faults or network problems. In of our earlier works [4] we defined the main requirements of dynamic workflow execution systems as: the ability to react to or to handle unforeseen scenarios raised during the workflow enactment phase, to adapt to new situations, to change the abstract or concrete workflow model or to give faster execution and higher level performance according to the actual environmental conditions and intermediary results. In our other work [2] we have defined the three main areas of dynamism which are optimization of the workflow execution according some criteria, user-steering (user or administrator interaction during execution) [3] and fault tolerance behavior. In this work we investigated fault tolerance behavior.

Fault tolerance is the ability of a system to perform its functions even in the presence of a failure. There are two main groups of failures that could arise during enactment. The first group includes the crash faults or fail-stop faults which may come with faulty system components that result in complete data loss. The other group consists of byzantine faults which result the system components to behave unpredictably and maliciously. Byzantine failures can occur, e.g., due to software bugs, (transitional or permanent) hardware malfunction, or malicious attack. In our work we consider only crash faults where the complete state of the actual task and environment must be restored.

Fault tolerance policy can be reactive and proactive. While the aim of proactive techniques is to avoid situations caused by failures by predicting them and taking the necessary actions, reactive fault tolerance policies reduce the effect of failures on application execution when the failure effectively occurs. There are several solutions in the literature for fault tolerant behavior and other complementary methods in its connected fields [1].

To achieve fault tolerant behavior the most widely adopted methods are:

- Checking and monitoring which is a key factor in failure detection.
- Checkpointing and resubmission where the system state is captured and saved based on predefined parameters (i.e.: time interval, number of instructions) and when the system undergoes some kind of failure the last consistent state is restored and computation is restarted from that point on.
- Replication where critical system components are duplicated using additional hardware or with scientific workflows critical tasks are replicated and executed on more than one processor. We can differentiate active and passive replication. Passive replication means that only one primary processor is invoked in the execution of a task and in the case of a failure the backup ones take over the task processing. In the active form all the replicas are executed at the same time and in the case of a failure the replica can continue the execution without intervention. The idea behind task replication is that replication size r can tolerate r-1 faults while keeping the impact on the execution time minimal. We call r the replication size. While this technique is useful for time-

critical tasks its downsides lies in the large resource consumption, so our attention is focused on mainly checkpointing methods in this work.

We propose a new checkpointing algorithm, that monitors the resources and dynamically adjust the checkpointing interval based on the task's dependency Factor and the already occurred failures.

In the proposed algorithm there is no need to take global checkpoints of the workflow, and therefore there is no need of synchronization of any kind (based on time or based on communication channels between the processors). The parallel threads of the workflow may run on different type of computing infrastructures (for example on virtual machines of different cloud providers) therefore it would be a complex challenge to solve the synchronization between them. The new proposal is based on provenance support. Provenance carries information about the source, origin and processes that are involved in producing data. The main target of collecting provenance support is to provide reusability and reproducibility among a scientist's community but provenance support can provide users, scientists, workflow developers and administrators with wide range of services. For example provenance can also support fault tolerant behavior by providing statistics about historical executions, such as failure rates or distribution and by storing the intermediary results generated by each tasks of the workflow.

Our paper is organized as follows. After the introduction we give a brief overview about existing checkpointing methods, and in chapter III we introduce our checkpointing method. After a brief conclusion the bibliography closes our work.

II. RELATED WORK

Concerning dynamic workflow execution fault tolerance is a very important issue and checkpointing is the most widely used methods to achieve fault tolerant behavior. We investigated the different algorithms in order to give a brief overview of them.

According to the level where the checkpointing occurs we differentiate application level checkpointing, library level checkpointing and system level checkpointing methods. Application level checkpointing means that the application itself contains the checkpointing code. The main advantage of this solution lies in the fact, that it does not depend on auxiliary components however it requires a significant programming effort to be implemented while library level checkpointing is transparent for the programmer. Library level solution requires a special library linked to the application that can perform the checkpoint and restart procedure. System level solution can be implemented by a dedicated service layer that hides the implementation details from the application developers but still give the opportunity to specify and apply the desired level of fault tolerance [5].

From another perspective we can differentiate coordinated and uncoordinated methods. With Coordinated checkpointing (synchronous) the processes will synchronize to take checkpoints in a manner to ensure that the resulting global state is consistent. This solution is considered to be domino-effect free. With uncoordinated checkpointing (independent) the checkpoints at each process are taken independently without any synchronization among the processes. Because of the absence of synchronization there is no guarantee that a set of local checkpoints result in having a consistent set of checkpoints. It may lead to the initial state due to domino-effect.

The frequency of the checkpointing interval also imposes many opportunities in checkpointing algorithms. Young in [6] has already in 1974 defined his formula for the optimum periodic checkpoint interval which is based on the checkpointing cost and the mean time between failures (MTBF) with the assumption that failure intervals follow an exponential distribution.
Sheng et al in [7] has also derived a formula to compute the optimal number of checkpoints for jobs executed in the cloud. His formula is generic in a sense that it does not use any assumption on the failure probability distribution.

The drawback of these solutions lies in the fact that the checkpointing cost can change during the execution if the memory footprint of the job changes, or depending on network reachability issues or when the failure distribution changes. Thus static intervals may not lead to the optimal solution. By dynamically assigning checkpoint frequency we can eliminate unnecessary checkpoints or where the danger of a failure is considered to be severe it can introduce extra state savings.

Meroufel and Belalem [8] proposed an adaptive time-based coordinated checkpointing technique without clock synchronization on cloud infrastructure. Between the different VMs jobs can communicate with each other through a message passing interface. One VM is selected as initiator and based on timing it estimates the possible time interval where orphan and transit messages can be created. There are several solutions to deal with orphan and transit messages, but most of them solve the problem by blocking the communication between the jobs during this time interval. However blocking the communication increases the response time and thus the total execution time of the workflow which can lead to SLA violation. In Meroufel's work they avoid blocking the communication by piggybacking the messages with some extra data so during the estimated time intervals it can be decided when to take checkpoint or logging the messages can resolve the transit messages problem.

The initiator selection is also investigated in Meroufel and Belalem's another work [9] and they found that the impact of initiator choice is significant in term of performance. They also propose a simple and efficient strategy to select the best initiator.

Sheng at al also propose a new adaptive algorithm to optimize the impact of checkpointing regarding the checkpointing or restarting costs in [7].
Theresa et al in their work [10] propose two dynamic checkpoint strategies: Last Failure time based Checkpoint Adaptation (LFCA) and Mean Failure time based Checkpoint Adaptation (MFCA) which takes into account the stability of the system and the probability of failure concerning the individual resources.

To the best of our knowledge there does not exist an adaptive algorithm that takes into account the effect of the failure occurring on a task on the execution time of the whole workflow.

III. PROPOSED MODEL

A. Environmental Conditions

- the system resources are monitored and failures can be detected as soon as possible, therefore the fault detection time (t_f) does not add high latency to the overall makespan of the workflow execution. (t_f=0)
- Task A_j cannot be started before it has received the output from all its predecessors and the results of Task A_i can only be sent to its successor tasks after the task has been finished.
- There is an ideal case so that tasks can be executed as soon as the results from the predecessor tasks are ready and available. The system resources are inexhaustible in number, so the system can allocate the required number of resources to execute all the tasks parallel that are independent from each other.
- The system supports the collection of provenance data, therefore the intermediary results generated by the individual tasks are saved and in case of failure they can be easily retrieved. Thus there is no need to take checkpoints at the end of the tasks, and there is no need to take global checkpoints, since in the case of failure only the effected task should be rolled back.
- The system also support provenance data about failure statistics, so the probability of failures for a certain period of time is available for each resource component taking into account the aging factor as well.

B. General notation

Workflows in general and also scientific workflows are represented as directed acyclic graphs (DAG) $W = (N, E)$, where the nodes (N) represent the computational tasks or jobs and the directed edges (E) represent the dependency between them. The dependency can be data dependency, and control dependency. In the former case the output of a Task A_i gives the input of a Task A_j if there exists an $A_iA_j \in E$ directed edge in the workflow. The control dependency describes the precedence of the tasks: If an A_iA_j directed edge exists in the workflow, then the execution of task A_i must precede the execution of task A_j in time. Here follows a list of the most frequently used symbols in this paper:

T_c the optimal checkpointing interval,
C is the checkpointing cost
X is optimal number of checkpoints during the execution of a task
$T(A_i)$ is the execution time of task A_i
T_f is the mean time between failures (MTBF)
$E(Y)$ is the expected number of failures during the execution of a task
t_l is the loading time, to restore the last saved checkpoint state,
t_f is the fault detection time, the time to detect the failure
A_0 is the first or entry task of the workflow.

C. Algorithm

The primary goal of this algorithm is to minimize the effect of the checkpointing overhead (time, resource) while still keeping to the soft-deadline of the workflow and the performance level at a satisfactory level.

Young [6] and Sheng [7] have already proved that the optimal checkpointing interval can be computed by (1) and (2). In both cases the fault detection time is considered t_f=0. Equation (2) is a more general form of Young's formula, because it does not depend on any probability distribution, unlike Young's (1) formula which needs to assume that failure intervals follow an exponential distribution.

$$T_c = \sqrt{2CT_f} \tag{1}$$

$$X = \sqrt{\frac{T(A_i) \cdot E(Y)}{2 \cdot C}} \tag{2}$$

In our proposed algorithm we use (2) as a starting point to compute the checkpointing intervals. The main idea is that there is a dependency factor between the tasks. Namely if a failure occurs during the execution of a task A_i then it not only has a local effect on the task itself, but has a global effect also on the whole workflow concerning the execution time. Since if a failure occurs during the execution of task A_i then it has to be re-executed from the last checkpoint. It means the execution of the task ends later, so it may cause all of the successor tasks of task A_i to wait for the results. This can result the whole workflow execution to last longer.

We define local cost (3) of a failure on task A_i which is the execution time overhead of a task when during execution one failure occurs.

$$C_{local} = \frac{\overline{ET(A_i)} + \frac{T_c}{2} + t_l}{ET(A_i)} \quad (3)$$

We define global failure cost (4) of a task A_i: which is the execution time overhead of the whole workflow, when one failure occurs during Task A_i

$$C_{global} = \frac{\frac{T_c}{2} + t_l + rank(A_i) + brank(A_i)}{rank(A_0)} \quad (4)$$

where (5) and (6) are classic formulas that are used in tasks scheduling [12] [11]. Basically the rank() function is the critical path from task A_i to the last task, and can be computed recursively backward from the last task. The brank() value is the backward rank value from task A_i backward to the entry task A_0. It is the longest distance from the entry task to task A_i excluding the computation cost of the task itself.

$$rank(A_i) = ET(A_i) + \max_{A_j \in succ(A_j)} rank(A_j) \quad (5)$$

$$brank(A_i) = \max_{A_j \in pred(A_i)} \{brank(A_j) + ET(A_j)\} \quad (6)$$

It can also be calculated recursively downward from task A_0.

Before executing a task A_i, C_{global} can be evaluated. If $C_{global} < 1$ then one failure occurring during execution of task A_i does not add extra latency to the total execution time of the workflow so in that case T_c can be increased until $C_{global} = 1$. From that we get:

$$T_c = 2 \cdot (rank(A_0) - t_l - rank(A_i) - brank(A_i)) \quad (7)$$

Based on the global failure cost and using the assumption that when faults occur during a checkpoint interval (between two consecutive checkpoints) the expected average time loss is half of the checkpointing interval. This is the average time to re-execute the task from the latest checkpoint.

If there occurred already failures during the actual task, or during the predecessors of the actual task, then it may be possible, that default checkpointing intervals cannot be increased, because the cumulative overhead of the occurred faults can negatively affect the whole workflow execution time.

To take earlier faults into account we need information about the realistic execution time of the tasks. With provenance support the real execution time can be obtained and it can be substituted in (7) in place of the brank value. The $C_{global} < 1$ inequality gives the answer whether the checkpointing interval can be increased without lengthening the whole workflow execution time.

IV. CONLUSION

In this paper we investigated the different checkpointing techniques, which are the most widely used proactive fault tolerant methods. We gave a brief overview of the different checkpointing perspectives with special attention on those solutions where the checkpointing intervals are periodic or it changes adaptively during execution. We proposed a provenance based dynamic algorithm that takes into account the global cost of a failure occurring during the execution of a task, and depending on this value can adjust the checkpointing interval in order to eliminate blind checkpoints while still maintaining soft deadlines. In our future work we would like to deeper investigate the various fault tolerant methods for crash faults or even for byzantine faults.

REFERENCES

[1] Bala, Anju, and Inderveer Chana. „Fault tolerance-challenges, techniques and implementation in cloud computing". IJCSI International Journal of Computer Science Issues vol. 9, 2012

[2] Kail, E., Bánáti, A., Kacsuk, P., Kozlovszky, M.: Provenance based adaptive and dynamic workflows. In: 15th IEEE International Symposium on Computational Intelligence and Informatics, pp 215-219, IEEE Press, Budapest, 2014

[3] E. Kail, P., Kacsuk, M. Kozlovszky.: "A Novel Aproach to User-steering in Scientific Workflows" In Proceedings of CGW'14, 2014

[4] E. Kail, A. Bánáti, K. Karóczkai, P. Kacsuk, M. Kozlovszky, Dynamic workflow support in gUSE, MIPRO, 2014 Proceedings of the 37th International Convention

[5] Jhawar, Ravi, Vincenzo Piuri, and Marco Santambrogio. „Fault Tolerance Management in Cloud Computing: A System-Level Perspective". IEEE Systems Journal 7, vol 2, 2013 doi:10.1109/JSYST.2012.2221934

[6] J.W. Young. "A first order approximation to the optimum checkpoint interval" in Comminications ACM, 1974

[7] S. Di, Y. Robert, F. Vivien, D. Kondo, Cho-Li Wang, and F. Cappello. „Optimization of Cloud Task Processing with Checkpoint-Restart Mechanism", 1–12. ACM Press, 2013. doi:10.1145/2503210.2503217.

[8] B. Meroufel, G. Belalem: "Adaptive time-based coordinated checkpointing for cloud computing workfl ows". in Scalable Computing: Practice and Experience, Vol 15, No 2, 2014

[9] B. Meroufel, B. Ghalem: „Policy Driven Initiator in Coordination Checkpointing Strategies". http://www.wseas.us/e-library/conferences/2014/Istanbul/TELEDU/TELEDU-20.pdf.

[10] Antony Lidya Therasa.S, Sumathi.G, Antony Dalya.S.: "Dynamic Adaptation of Checkpoints and Rescheduling in Grid Computing" in International Journal of Computer Applications vol. 3, 2010

[11] Laiping Zhao; Yizhi Ren; Yang Xiang; Sakurai, K., "Fault-tolerant scheduling with dynamic number of replicas in heterogeneous systems," High Performance Computing and Communications (HPCC), 2010 12th IEEE International Conference on , vol., no., pp.434,441, 1-3 Sept. 2010, doi: 10.1109/HPCC.2010.72.

[12] Topcuoglu, H.; Hariri, S.; Min-You Wu, "Performance-effective and low-complexity task scheduling for heterogeneous computing," Parallel and Distributed Systems, IEEE Transactions on , vol.13, no.3, pp.260,274, Mar 2002,

Operating System Impact on CPU and RAM Utilisation when Introducing XML Security

Sasko Ristov and Marjan Gusev
University Ss. Cyril and Methodius, Faculty of Computer Science and Engineering,
1000 Skopje, Macedonia
{sashko.ristov, marjan.gushev}@finki.ukim.mk

***Abstract*—Introducing XML security to a web service increases the message size, which impacts the XML parsing of the greater message and complicates its processing due to complex cryptographic operations. Both tasks impact the web server's CPU and RAM utilisation.**

The performance impact is even more expressed when the number of concurrent messages is rapidly increased. In this paper we analyze the impact of securing a web service message with XML Signature and XML Encryption over the hardware performance, varying the message size and the number of concurrent messages.

The results show that web server that is installed on Linux utilizes the CPU less than the same web server that is installed on Windows. The situation for RAM memory is opposite, that is, the web server installed on Windows operating system occupies less RAM memory.

***Index Terms*—*CPU*; *Operating systems*; *Performance*; *RAM*; *Web service*.**

I. Introduction

Web services are the third generation of the web, after the static web sites and dynamic web applications [1]. They are widely used in today's web applications because they are platform independent and can be deployed on various web servers and operating systems.

The performance of web services is an important factor, both for customers and providers, by different aspects. The former would like to receive the response in a tolerable time, regardless of the period of a day. The latter expect smaller hardware utilisation to predict the capacity of web servers and reduce the power consumption for the equipment and cooling [2]. The important issue for administrators is to predict the web service performance when it is targeted with a particular load (number of messages, their size, type, etc) and hosted on a web server with some amount of hardware resources (CPU, RAM, HDD, I/O).

Web services should be secured if they use sensitive data. Secure Socket Layer / Transport Layer Security (SSL / TLS) can provide a point-to-point security only, but it cannot provide an end-to-end web service's security [3], which can be achieved by WS-Security, but considerably slower [4]. Creating SOAP message and implementing XML Signature increases the message with a constant size, regardless of the message size [5], while introducing XML Encryption additionally linearly increases the message size [6].

Although web services are platform independent, their performance varies on the platform. For example, Ristov and Gusev [7] reported that web services deployed on Linux provide better response time rather than those hosted on Windows, if they are loaded with small number of small concurrent signed or unsigned messages. For all other cases, Windows platform is better.

In this paper we extend this research in order to examine the impact of introducing XML security on web server's hardware resources, hosted with various platforms, i.e. Linux and Windows. Series of experiments are conducted by varying the parameters: message size M and number of concurrent messages N.

The paper is organised in several sections that are described below. Related work in the area of web service performance is presented in Section II. Section III describes the testing methodology for all test cases of series of experiments. The results of the experiments for particular hardware parameter are presented and analysed in Section IV. In Section V we discuss the results, and provide additional analysis of the results for the particular operating system. Finally, the conclusion and plan for future work are given in Section VI.

II. Related work

Many authors propose various techniques to improve the web service performance after detailed analysis. For example, Wichaiwong and Jaruskulchai [8] proposed to use the FTP instead of HTTP protocol to transfer the SOAP messages, thus reducing the data transfer time. Eckert et al. [9] conduct a more comprehensive analysis. They maximized the throughput and minimized the delay by using several worst-case workflow performance optimizations. The SMP (SOAP multicast protocol) reduces the network traffic by grouping the similar SOAP messages into a compact SMP message [10]. Yue-min [11] predicts and maintains the fading performance of the web servers by using a technique based on chaotic theory. Filtering XML data and forwarding to the endpoint web servers such that each web server handles a specific type of XMLs improve the overall performance [12]. Web services composition is optimized by combination of services by Yu et al. [13].

Performance in virtual environment is also analysed. Ristov et al. [14] reported that virtualised environment decreases the web service performance of 30% compared to the same bare-metal compatriot (using the same CPU and memory hardware resources).

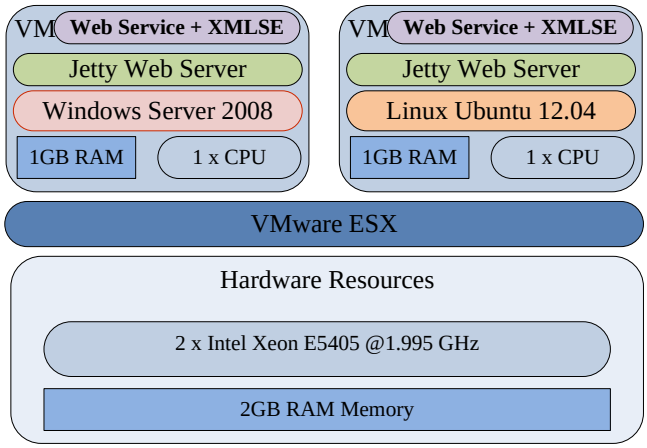


Fig. 1. Virtualised testing environment

Implementing cryptography seriously affect the web service performance. The processor increases its activity and the memory leaks [15]. Chen et al. [16] reported that web service security policies and message sizes impacts a lot on web service performances. Rodrigues et al. [17] achieve a trade-off between security and performance. AES 192 suffers less performance degradation when the number of concurrent messages increases [18]. Tang et al. [19] used the fact that there is a small difference between latencies using Username and X.509, even when the message size grows, and thus they used X.509 keys for XML Signature and XML Encryption. Several optimisations of various WS-Security stages were proposed by Engelen and Zhang [20], [21].

Less analyses were realised on the operating system and hardware impact, which is the goal of this paper. Tan et al. [22] proposed several improvements of hardware bottleneck of different cryptographic algorithms on network processors.

III. Methodology

A series of experiments are conducted to determine the operating systems impact on web server's hardware resources in virtualised environment. This section describes the testing methodology for conducting the experiments.

A. Testing environment

Two experiments with several test cases are realised on a client-server architecture deployed on the same hardware infrastructure and runtime environment, but different operating systems. The environment is virtualised, that is, two virtual machines are instantiated with VMware ESX 4.0, one with Windows Server 2008 and the other with Ubuntu 12.04, as presented in Fig. 1. Both virtual machines are allocated with the same amount of hardware resources, 1GB RAM and 1 CPU Intel Xeon E5405 @1.995 GHz.

Eclipse Jetty [23] is deployed on both virtual machines and test loads are created with SOAPUI [24]. A web service, which receives two strings and returns its concatenation is hosted in web servers in both virtual machines, is used as a demo application. XML Signature and XML Encryption are introduced in web service. That is, the accepted message is encrypted and signed, which the web server decrypts and verifies. Afterwards, the web service generates the concatenation of the input strings into SOAP message, then signs and encrypts it.

Only one virtual machine is active at a time in order to avoid impact on the other [25]. Also, the client and server are in the same LAN to avoid network impact [26].

B. Test cases and experiments

Each test case consists of sending N concurrent messages with M bytes each in a period of 60 seconds. An experiment consists of series of test cases in order to examine the impacts of increasing the parameter M for constant N, and increasing the parameter N for constant M. The parameter M is defined in kilobytes of the original (regular) SOAP messages and has values $M \in \{0.2, 2, 10, 20, 60, 70, 100, 1000\}$, while $N \in \{1, 10, 50, 100, 150, 200\}$. In this range the web server works in normal mode and does not reply any error message.

C. Test data

Average *CPU Utilisations* CPU_L and CPU_W are measured in each test case for both operating systems Linux and Windows in percents (0-100%), where indexes L and W are used for corresponding operating systems - Linux and Windows. Similarly, average *RAM Utilisations* RAM_L and RAM_W are measured in Megabytes.

In order to compare both operating systems for both parameters CPU and RAM, *Relative CPU Utilisation* R_{CPU} and *Relative RAM Utilisation* R_{RAM} are calculated and analysed, which are defined in (1) and (2), correspondingly.

$$R_{CPU} = \frac{CPU_L}{CPU_W} \tag{1}$$

$$R_{RAM} = \frac{RAM_L}{RAM_W} \tag{2}$$

IV. The Results

This section presents a comparative analysis of how the introduction of both XML Signature and XML Encryption (XMLSE) impacts the hardware resources (RAM and CPU) as a function of parameters M and N, on both operating systems.

A. CPU Utilisation Impact

This section elaborates the impact of XMLSE on CPU utilisations CPU_L and CPU_W on a particular operating system for different loads. It also analyses the correlation of different operating system for a constant load.

1) CPU Utilisation on Linux CPU_L: Fig. 2 presents the CPU utilisation on Linux for various values of parameters M and N. CPU_L is almost horizontal for constant N, regardless of the values of parameter M, which means that it does not depend on the message size.

On the other side, two regions are observed for parameter N. Increasing the number of concurrent messages up to $N = 50$ also increases the CPU utilisation linearly up to $CPU_L = 75\%$. After this, the CPU utilisation saturates up to $CPU_L = 80\%$, for both parameters M and N.

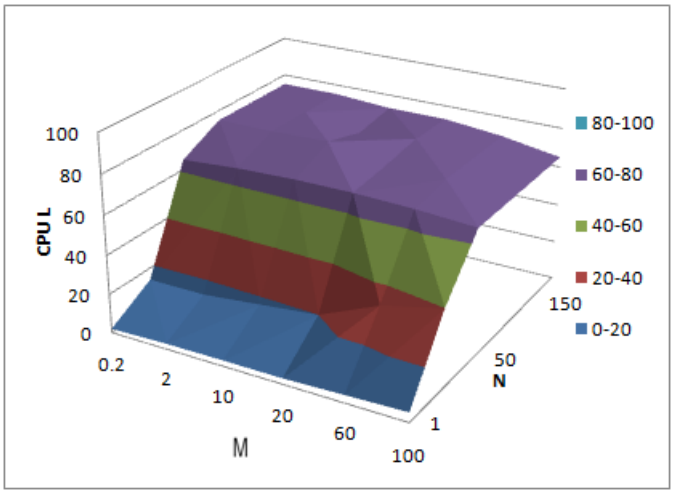


Fig. 2. CPU Utilisation on Linux as a function of both parameters M and N

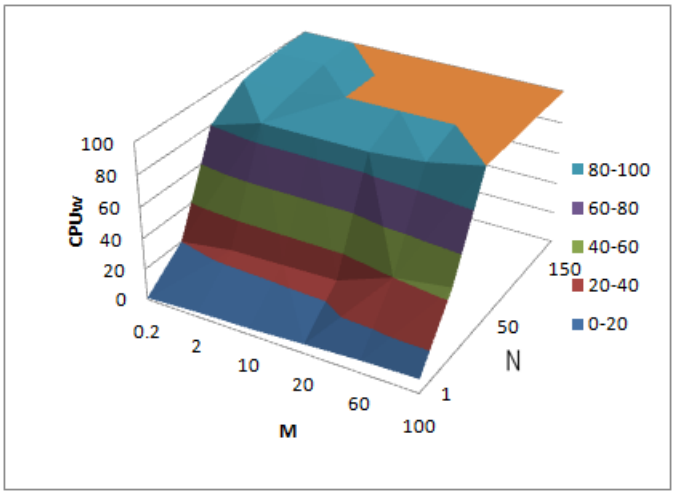


Fig. 3. CPU Utilisation on Windows as a function of both parameters M and N

2) CPU Utilisation on Windows CPU_W: The results of the experiments for CPU utilisation on Windows operating system for various values of parameters M and N are presented in Fig. 3. The CPU_W is also almost horizontal for constant N as CPU_L, regardless of the values of parameter M, i.e. it does not depend on the message size.

The similar linearly increasing region of CPU Utilisation saturation is observed for constant value of M, and changing the parameter N up to $N = 50$. Afterwards, the CPU saturates also, but in this experiment on value $CPU_W = 100\%$ for both parameters M and N.

3) Relative CPU Utilisation R_{CPU}: Comparing figures 2 and 3, one can conclude that the CPU utilisation behaves similarly on both platforms. Also, the impact of both parameters M and N is similar on both operating systems. However, CPU_W's curve is steeper than Linux's for smaller N, and it is higher than Linux's in saturation region.

Better analysis offers the relative CPU utilisation R_{CPU}, which is presented in Fig. 4. The curve is almost horizontal for each value of the parameters M and N.

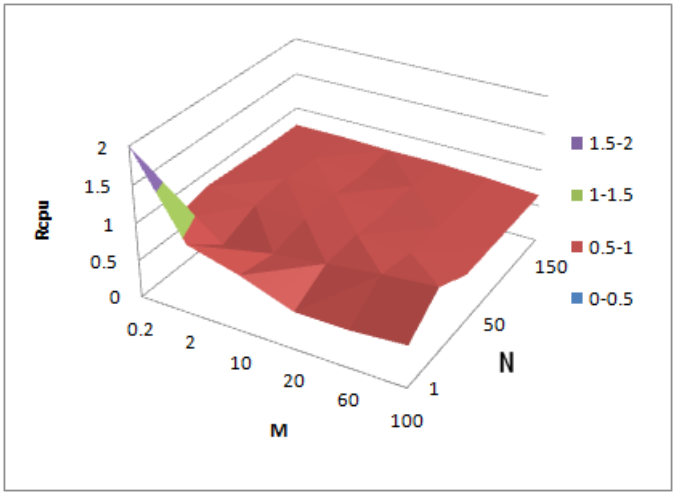


Fig. 4. Relative CPU Utilisation as a function of both parameters M and N

We observe that $0.5 < R_{CPU} < 1$ for almost all test cases, which means that Linux provides smaller CPU utilisation than Windows. Only for small values of parameters M and N, $1 < R_{CPU} < 2$, i.e. Windows provides up to two times smaller CPU utilisation than Linux. But the CPU utilisation in this region is minor and does not impact on the overall web server's performance.

B. RAM Utilisation Impact

Similar to previous subsection, this section elaborates the impact of XMLSE on RAM utilisation for both operating systems RAM_L and RAM_W, and relative utilisation R_{RAM}, for various load.

1) RAM Utilisation on Linux CPU_L: Fig. 5 presents the utilisation of the web server's RAM memory installed with Linux operating system for various values of parameters M and N. The behavior of the RAM_L curve is totally different than the CPU_L's.

RAM_L is almost horizontal (constant) only when the constant parameter (either M or N) is small, but it increases exponentially when the constant parameter (either M or N) is increased. For example, RAM_L increases from 781 up to 1010 for maximum value of constant $N = 150$ and M increases. Also, RAM_L increases from 738 up to 1010 for maximum value of constant $M = 100$ and N increases.

2) RAM Utilisation on Windows CPU_W: The results of the experiments for RAM utilisation on Windows operating system for various values of parameters M and N are presented in Fig. 6. The impact of RAM utilisation is also totally different than the CPU utilisation on the same operating system, but is similar as the RAM utilisation on Linux operating system.

RAM_W is almost horizontal (constant) only when the constant parameter N is small. It increases also as the RAM_L, but starting from lower value and slower. For example, RAM_W increases from 740 up to 930 for maximum value of constant

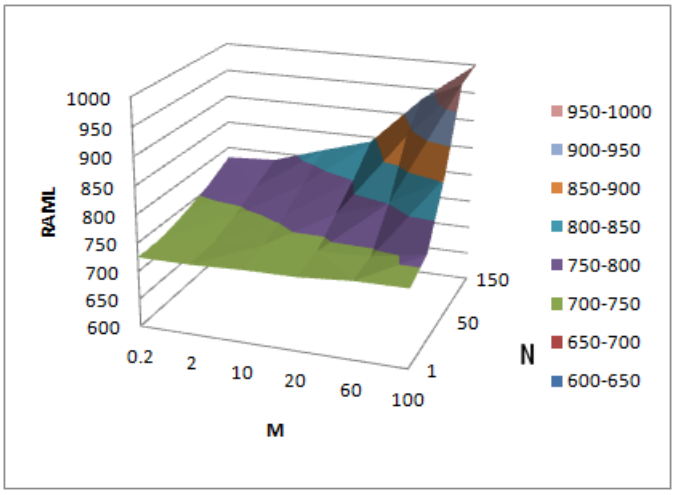


Fig. 5. RAM Utilisation on Linux as a function of both parameters M and N

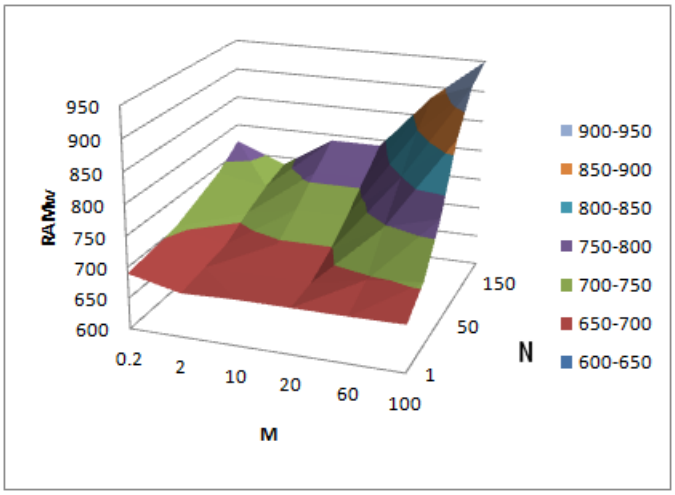


Fig. 6. RAM Utilisation on Windows as a function of both parameters M and N

$N = 150$ and M increases. Also, RAM_W increases from 675 up to 930 for maximum value of constant $M = 100$ and N increases.

3) Relative RAM Utilisation R_{RAM}: By comparing figures 5 and 6, one can conclude that both RAM utilisations RAM_L and RAM_W behave similarly. However, RAM_W's curve starts with lower values for smaller parameters M and N than Linux's and it also finishes with lower values for the greatest M and N. That is, Windows operating system utilises smaller amount of RAM memory than Linux.

More detailed analysis offers the relative RAM utilisation R_{RAM}, which is presented in Fig. 7. We observe that $R_{RAM} > 1$ for all test cases, which leads to the conclusion that Linux operating system utilises more RAM memory than Windows. Even more, this is more emphasized more in the region for the greatest values of parameters M and N (the greates load), where the relative RAM utilisation R_{RAM} has the values in the range of $1.1 < R_{RAM} < 1.2$.

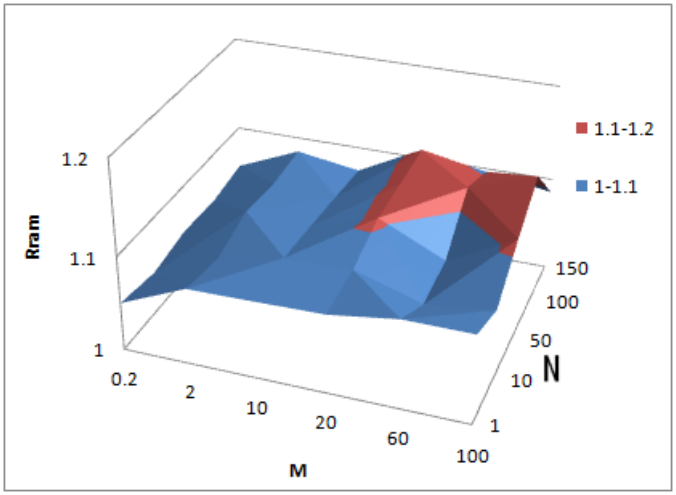


Fig. 7. Relative RAM Utilisation as a function of both parameters M and N

V. Discussion

Previous section presented the results and analysis of the impact of various operating systems (Windows and Linux) on the specific hardware resource: CPU and RAM. The results show that Linux operating system utilises CPU less than Windows, while the Windows operating system utilises less RAM memory than Linux. This is more emphasized for greater load (greater message size and number of concurrent messages).

In this section, we extend the analysis in other direction, that is, to analyse how a specific operating system impacts on both hardware resources CPU and RAM memory, as a function of both parameters M and N. By observing figures 2 and 5, one can conclude that parameters M and N impact different to CPU and RAM on Linux operating system. If a CPU utilisation depends mostly on the number of concurrent messages N, where it saturates after a certain value of N, the RAM utilisation depends on both parameters M and N. This means that CPU suffers more by the number of concurrent messages. This could open other research direction towards grouping several messages into one greater message, that should improve the overall performance.

Very similar conclusion is for Windows operating system, with the opposite behavior of the Linux operating system on RAM and CPU.

VI. Conclusion and Future Work

Several parameters impact the web service performance reported as a response time: the computation that the web service is performing [27]; the CPU power and cores; the number of concurrent messages, the message size and the introduced security [6], even of the resource orchestration in the cloud virtual environment [28] and the platform (operating systems) [7]. This paper analysed the impact of the platform over the other web service performance parameters, i.e. CPU

and RAM utilisation, for two most common web servers' operating systems: Windows and Linux.

The experiments lead to conclusion that web server with Windows operating system should use more powerful CPU than the one installed with Linux, if XMLSE is implemented on a web service, in order to reduce the CPU utilisation, and thus the power consumption. However, this conclusion does not implicitly mean that the web service response time is better if the same web service is hosted on Windows operating system [7].

The conclusion for RAM memory utilisation is opposite. That is, a web server with Linux operating system should have more RAM memory than the one installed with Windwos and the XMLSE is implemented on a web service, in order to improve the overall performance. The gap in the utilised RAM memory is even more emphasized for greater load, regardless of the load type: either the number of concurrent messages or the message size.

The CPU utilisation directly impacts the server's power consumption, which directly impacts to the cost for cooling in the data centers. The results of this paper can be used in consolidating of virtual machines on a specific physical servers to reduce the total costs and improve the performance in the modern data center.

We will continue by analysing the impact on the performance of other types of web services and servers (CPUs, RAM) will be measured in order to generalise the conclusion.

REFERENCES

[1] M. Al-Smadi and C. Gutl, "Soa-based architecture for a generic and flexible e-assessment system," in *Education Engineering (EDUCON), 2010 IEEE*, april 2010, pp. 493 –500.

[2] K. Wang and N. Tian, "Performance modelling of composite web services," in *Circuits, Communications and Systems, 2009. PACCS '09. Pacific-Asia Conference on*, May 2009, pp. 563–566.

[3] E. Bertino, L. Martino, F. Paci, and A. Squicciarini, *Security for Web Services and Service-Oriented Architectures*, 1st ed. Springer Publishing Company, Incorporated, 2009.

[4] B. Alrouh and G. Ghinea, "A performance evaluation of security mechanisms for web services," in *Information Assurance and Security, 2009. IAS '09. Fifth International Conference on*, vol. 2, Aug 2009, pp. 715–718.

[5] M. Novakouski, S. Simanta, G. Peterson, E. J. Morris, and G. Lewis, "Performance analysis of ws-security mechanisms in soap-based web services," Carnegie Mellon Software Engineering Institute, Tech. Rep. CMU/SEI-2010-TR-023, 2010. [Online]. Available: http://www.sei.cmu.edu/library/abstracts/reports/10tr023.cfm

[6] S. Ristov and A. Tentov, "Performance impact correlation of message size vs. concurrent users implementing web service security on linux platform," in *ICT Innovations 2011*, ser. Advances in Intelligent and Soft Computing, vol. 150. Springer Berlin / Heidelberg, 2012, pp. 367–377.

[7] S. Ristov and M. Gusev, "Choosing optimal platform to host web services with xml signature and xml encryption," *J. of Next Generation Information Technology (JNIT)*, vol. 4, no. 5, pp. 110–124, July 2013.

[8] T. Wichaiwong and C. Jaruskulchai, "A simple approach to optimize web services' performance," in *Proc. of the Third Int. Conf. on Next Generation Web Services Practices*, ser. NWESP '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 43–48.

[9] J. Eckert, S. Schulte, M. Niemann, N. Repp, and R. Steinmetz, "Worst-case workflow performance optimization," in *Proc. of the 3rd Int. Conf. on Internet and Web Applications and Services*, ser. ICIW '08. IEEE Computer Society, 2008, pp. 632–637.

[10] K. A. Phan, Z. Tari, and P. Bertok, "Similarity-based soap multicast protocol to reduce bandwith and latency in web services," *IEEE Transactions on Services Computing*, vol. 1, no. 2, pp. 88–103, 2008.

[11] W. Yue-min, "Based on the introduction of chaos theory of web server performance," *Journal of Convergence Information Technology(JCIT)*, vol. 8, no. 6, Mar. 2013.

[12] P. Felber, C.-Y. Chan, M. Garofalakis, and R. Rastogi, "Scalable filtering of xml data for web services," *Internet Computing, IEEE*, vol. 7, no. 1, pp. 49–57, 2003.

[13] X. Yu, Y. Tu, and K. Gao, "Web service performance optimization based on decentralized control mechanisms in cloud computing environment," *International Journal of Digital Content Technology and its Applications(JDCTA)*, vol. 7, no. 9, May 2013.

[14] S. Ristov, G. Velkoski, M. Gusev, and K. Kjiroski, "Compute and memory intensive web service performance in the cloud," in *ICT Innovations 2012*, ser. Advances in Intelligent Systems and Computing, S. Markovski and M. Gusev, Eds. Springer Berlin Heidelberg, 2013, vol. 207, pp. 215–224. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37169-1_21

[15] H. Liu, S. Pallickara, and G. Fox, "Performance of web services security," in *Proceedings of the 13th Annual Mardi Gras Conference*, 2005, pp. 72–78.

[16] S. Chen, J. Zic, K. Tang, and D. Levy, "Performance evaluation and modeling of web services security," in *Web Services, 2007. ICWS 2007. IEEE International Conference on*, july 2007, pp. 431 –438.

[17] D. Rodrigues, D. F. Pigatto, J. C. Estrella, and K. R. L. J. C. Branco, "Comparison and analysis of cryptographic algorithms aiming performance improvement in secure web services," in *Proc. of the 2011 IEEE 13th Int. Symp. on High-Assurance Systems Engineering*, ser. HASE '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 191–192.

[18] ——, "Performance evaluation of security techniques in web services," in *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, ser. iiWAS '11. New York, NY, USA: ACM, 2011, pp. 270–277.

[19] K. Tang, S. Chen, D. Levy, J. Zic, and B. Yan, "A performance evaluation of web services security," in *Enterprise Distributed Object Computing Conf. EDOC '06. 10th IEEE Int.*, oct. 2006, pp. 67 –74.

[20] R. A. van Engelen and W. Zhang, "Identifying opportunities for web services security performance optimizations," in *Proc. of the 2008 IEEE Congress on Services - Part I*, ser. SERVICES '08. USA: IEEE Computer Society, 2008, pp. 209–210.

[21] ——, "An overview and evaluation of web services security performance optimizations," in *Proc. of the 2008 IEEE Int. Conf. on Web Services*, ser. ICWS '08. USA: IEEE Computer Society, 2008, pp. 137–144.

[22] Z. Tan, C. Lin, H. Yin, and B. Li, "Optimization and benchmark of cryptographic algorithms on network processors," *IEEE Micro*, vol. 24, no. 5, pp. 55–69, Sep. 2004.

[23] Eclipse, "Jetty web server," Jan. 2013. [Online]. Available: http://www.eclipse.org/jetty/

[24] SoapUI, "Functional and load testing tool," Jan. 2013. [Online]. Available: http://www.soapui.org/

[25] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, Z. Wen, and C. Pu, "An analysis of performance interference effects in virtual environments," in *Performance Analysis of Systems Software, 2007. ISPASS 2007. IEEE International Symposium on*, april 2007, pp. 200 –209.

[26] M. B. Juric, I. Rozman, B. Brumen, M. Colnaric, and M. Hericko, "Comparison of performance of web services, ws-security, rmi, and rmi-ssl," *J. Syst. Softw.*, vol. 79, no. 5, pp. 689–700, May 2006.

[27] S. Ristov, M. Gusev, and G. Velkoski, "Modeling the speedup for scalable web services," in *ICT Innovations 2014*, ser. Advances in Intelligent Systems and Computing, A. M. Bogdanova and D. Gjorgjevikj, Eds. Springer International Publishing, 2015, vol. 311, pp. 177–186. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-09879-1_18

[28] M. Gusev, S. Ristov, G. Velkoski, and M. Simjanoska, "Optimal resource allocation to host web services in cloud," in *Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing*, ser. CLOUD '13. Santa Clara, CA, USA: IEEE Computer Society, June 2013, pp. 948–949. [Online]. Available: http://dx.doi.org/10.1109/CLOUD.2013.103

Evaluating the impact of the number of processor cores, cache memory and thread during the parallel execution of programs

Flaka Dika*, Valon Xhafa **
* University of Vienna/Faculty of Computer Science, Vienna, Austria
** University of Prishtina/Department of Computer Engineering, Prishtina, Kosovo
valonxhafa@hotmail.com

Abstract - The speed of the computers work depends directly on the number of processor cores used in the parallel execution of the programs. But, there are some other parameters that have impact on the computer speed, like maximum allowed number of threads, size of cache memory inside the processors and its organization. To determine the impact of each of these parameters, we have experimented with different types of computers, measuring the speed of their work during solving a certain problem. To compare impacts of particular parameters on the computers speed, the results we have presented graphically, through joint diagrams for each parameter.

I. INTRODUCTION

Currently, on the market there are different types of computers with multi-core processors. Use of computers with these features has an important impact on their parallel work. Users have difficulties to distinguish which is the best solution for them, knowing clock speed of processor, or the information for the type of processor, like Core i3, Core i5 or Core i7 (in this case often it is assumed that we have to do with the processor that have 3, 5 or 7 cores) [5].

Programs that are written to be executed on a single processor or in a processor core without having the possibility to execute more instructions immediately, are known as serial programs, or sequential programs [1]. On the other side, programs that are written for simultaneous execution in more cores, or more logical processor, are known as parallel programs [2][13]. Parallel execution of programs in a computer is based on the parallel work of processor cores. But, using the Intel's technology that is known as Hyper-Threading it is possible for each core to work as two logical processor [5], respectively it is possible for the processor with 4 cores to work parallel with 8 logical processors.

The parallel work of microprocessors has a great influence on the calculation speed of computers which is important for the massive use of computers in communication through the Internet, also searching on the Web, in the processing of various biometric information, monitoring the climate changes in real time, scientific researches, in complicated three-dimensional games on the computer etc.

Measuring the performance of computers during parallel execution of programs is not an easy task, because in the speed of their work, except clock speed, indicates cores number, model, the support of Hyper-Threading technology, organization of cache memory in separated levels [7] etc.

II. EXPERIMENTAL RESULTS

In order to get relevant conclusions, initially we have experimented with numerous computers that have different processor types of Intel producer. In order to measure the speed of the computer work during experimentation we have used an ordinary algorithm for multiplication of two square matrices with different dimensions, from 100 x 100 to 1000 x 1000 (10 cases). For parallelism of loops are used methods of programming language C# [3][10][12], which are included in Microsoft Visual Studio 2010 package.

We have started the experimentation with computers that are built using previous processor 2 Duo, which have cache memory only on the level L2, then with Core i3, Core i5 and Core i7 processors (short i3, i5, i7) [16], whose memory are on the levels L2 and L3 and processor used for servers, such as Xeon type [15]. Finally from all the computers that are used for experimentation, we have singled out only the results of 15 computers with Intel processor, whose characteristics are given in TABLE I, without differentiating as desktop or notebook computer.

TABLE I. LIST OF USED COMPUTERS

Computer	Model	Clockspeed (GHz)	L2 (KB)	L3 (KB)	Cores	Logical processors
C1	i5	2.90	1024	6144	4	4
C2	i5	3.10	1024	6144	4	4
C3	E3	3.10	1024	8192	4	4
C4	i5	1.70	256	3072	2	4
C5	i3	2.40	256	3072	2	4
C6	i5	3.33	512	4096	2	4
C7	i7	2.20	256	6144	4	8
C8	i3	3.30	512	3072	2	4
C9	2 Duo	2.93	3072	0	2	2
C10	i5	1.60	256	3072	2	4
C11	i7	3.60	1024	8192	4	8
C12	2 Duo	2.66	3072	0	2	2
C13	i5	2.50	256	3072	2	4
C14	i7	2.00	256	4096	2	4
C15	i3	2.40	256	3072	2	4

From the given table we can see that the model of the processor does not automatically determines the number of cores, neither the number of logical processors. Thus,

for example, computer C7 belongs to model i7 and and contains 4 cores and 8 logical processors [8]. But, his memories L2 and L3 are smaller, e.g. than at the model i5 processor of computer C2. On the other side, the computer C14 has model i7 processor, but with two cores and four logical processor [9]. At the same time, processors of earlier generation 2 Duo, which are included in the computers C9 and C12 have relatively large cache memory on the level L2, but don't have the level L3 memory and in parallel work they have shown lower speed than computers with processors of next generation.

A. *Influence of number of cores*

To have a clearer overview in how much does the parallel execution of programs indicates, initially we have experimented with serial execution testing program for matrices multiplication, by activating only one core of the processor. Then, we have measured the execution time of a parallel program when all cores are active, respectively all logical processors. Finally, using the results of measurements, we found the execution speedup and the level of parallelism of the program. Having no chance to present all the results obtained, in the following section we have given only the graphically results for four computers from the list given in TABLE I, who have clock speed approximately the same level (above 3 GHz).

1. *Serial execution time*

Results of serial execution time (Ts) for four specific computers are shown in the diagram given in Fig.1.

The diagram shows that computer C11 is faster, which uses a processor that is ranked in the group of processors that are actually fastest [8], the i7 model, with maximum clock speed of 3.60 GHz, and with 4 cores (short 4-c) and 8 logical processors (short 8-lp). Computer C2, whether it is the i5 model, has competitive speed with computer C11. This is seen in the daily ranking of Intel processors [9], where he has high position in the list. Simultaneously, the computer C6 with i5 processor is slower than computer C8 who have i3 processor. From the given diagram is clear that the curves are ranked based on the clock speed of processors.

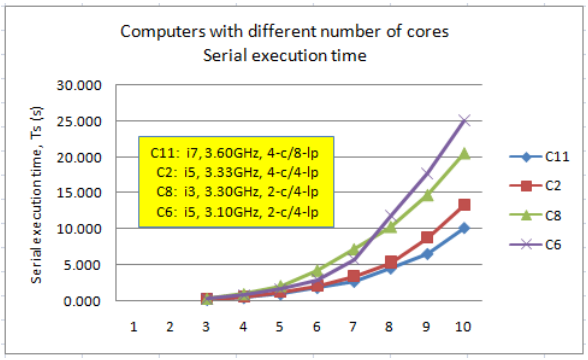


Figure 1. Serial execution time of the program

2. *Parallel execution time*

During the parallel execution of the testing program in separate computers, parallel execution time (Tp) have big difference from serial execution time (Ts), which can be seen from the diagram given in Fig.2.

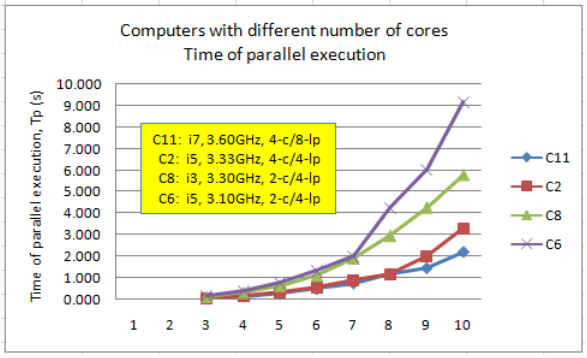


Figure 2. Parallel execution time of testing program

Here it is seen that computer C11 has higher speed, and computer C6 is slower, which is consistent with their speed in serial work. But, in this case execution times are significantly shorter. Thus, for example, computer C11 multiplication of matrices with dimensions 1000 x 1000 performs for more than 2 seconds, and during serial execution is needed more than 10 seconds.

Parallel execution time in computers C11 and C2 do not differ much between them, despite being different models of processors (i7 and i5). During the parallel work of computer C11 are 8 logical processor active, and in computer C2 - 4 logical processors. On the other side, as it is shown in the diagram, with increased dimensions of matrices, the speed of the computer C6 drops dramatically, despite having the i5 processor and its clock speed of 3.10 GHz. From what was said and by experimentation that we had with other computers, we can conclude that the speed of parallel execution of the program can't be determined based on the model of the processor (i3, i5 or i7), or number of cores, because it is also influenced by the clock speed of processor.

3. *Execution speedup*

If it's known the serial execution time (Ts) and parallel execution time (Tp) of the testing program, can be calculated execution speedup (S) through the expression [6]

$$S = \frac{Ts}{Tp} \qquad (1)$$

Using the experimental results mentioned above, execution speedup of 4 computers taken into consideration, it will look like is given in Fig. 3.

Even the execution speedup maintain the sequence of curves for specific computers as well as in diagrams for execution time. But curves have variations, with increases and decreases. At the end of their, obvious that only the computer C11 continues accelerating growth (finally reaches the value of 4.5 times), while the three other computers values stay the same, or decrease. As it is seen from the diagram given, and even from experimentation that we had with other computers too, the execution speedup is greater if the number of operations that are executed (e.g. multiplication of matrices with large dimensions) increases up to a certain thresholds. During

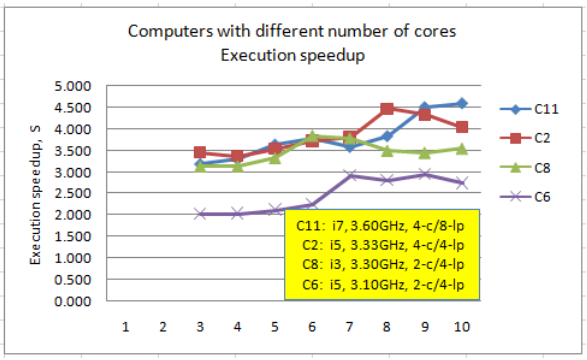


Figure 3. Execution speedup of the execution process

this there is no continual growth execution speedup, which shows that parallelism is not perfect.

B. *The impact of cache*

In the speed of the computer work it is important even the size of the cache memory. Models of the last generations of processors contain cache memory on the three levels. Initially we have experimented with computers of prior generations, like 2 Duo processor cores. In TABLE I we've included computers C9 and C12, whose processors have only memory L2 level and we have noticed that compared with other computers in parallel work are extremely slow, despite having relatively higher clock speeds. Then we experimented with other 6 computers also included in the TABLE I and the results obtained for them are presented in the diagram given in Fig. 4.

From the given diagram it is shown that computer C7 with i7 processor is faster and has large cache memory on the third level whether that it has the lowest clock speed. Computers C2 and C1 have approximately the same speed, with processor of model i5, clock speed much greater than computer C7 and have a maximum cache memory at two levels. In the overall ranking of

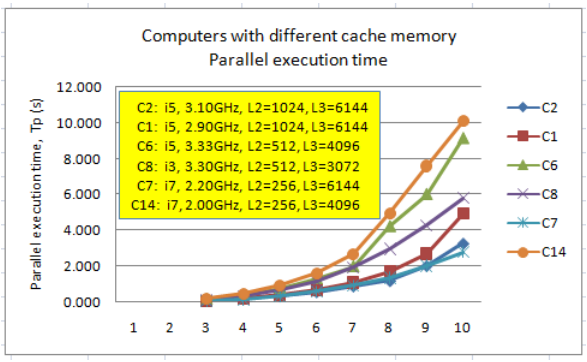


Figure 4. Parallel execution time of computers with processors that have different cache memory

processors, processor of computer C1 have higher rank than the computer C2 [9].

Speed of computers C8 and C6, despite that they have processors i3 and i5, in their speed ranking have higher speed clock and cache memory size. The Computer C14, with i7 processor, have the highest execution time.

C. *The level of parallelism*

The maximum execution speedup of program can be determined by Amdahl's Law [4][6]

$$S = \frac{1}{(1-P)+\frac{P}{N}}. \qquad (2)$$

Here, N is the number of processors, and P represents the relative percentage of the program parallelism. Starting from the given Law, through mathematical operations can be derived expression for the calculation of the level of parallelism of a program:

$$P = \frac{N(S-1)}{S(N-1)}. \qquad (3)$$

Since in the measurement of execution speedup values are calculated using (1), and knowing the number of cores, i.e. the number of logical processors (N), the level of parallelism can be calculated through (3). Thus, for computers with i7 processor, the level of parallelism will appear as is given in Fig. 5.

From the diagram provided and experimentation that we have done with other computers, it is shown that the relative level of parallelism of the program execution is over 0.65 and just a bit under 1.00, which is directly linked with the speed of parallel execution of the program.

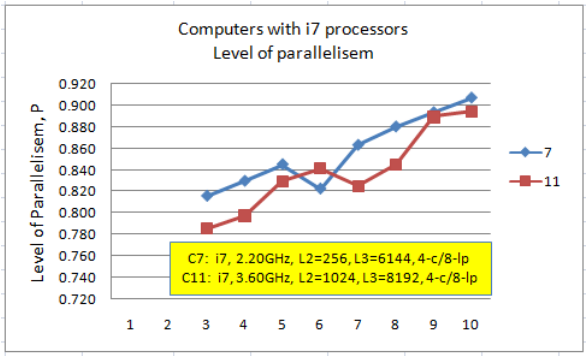


Figure 5. Parallel execution time of computers with processors that have different cache memory

D. *The maximum rate of parallelism*

During the parallel work of computer, the operating system determines the number of threads that are activated [10], or being allocated as needed certain number of threads to programs that run in parallel [11]. However, this can be done by the user too, by limiting the number of threads. If parallel execution relates only to a program, computer work can be optimized if the number of threads deal as far as the number of cores, i.e. the number of logical processors. To show this we experimented with several computers. But in Fig. 6 have just given diagram of parallel execution time of testing computer C4, limiting the number of threads between the values 4 and 20.

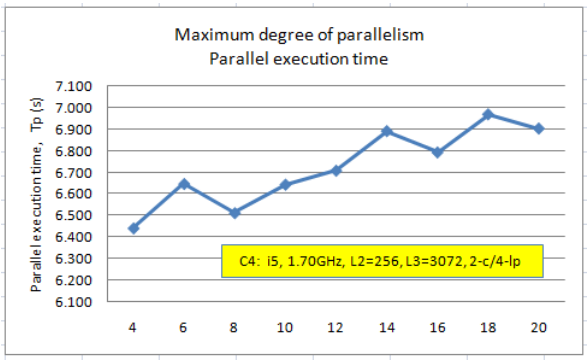


Figure 6. The rate of parallelism

The given diagram clearly shows the execution time is minimal if the maximum degree of parallelism obtained 4, which corresponds to the number of processor cores of the computer where the program is executed. Execution time increases if it's allowed a higher degree of parallelism, which is observed in most of the computers with which we experimented. From this we can conclude that the higher speed of parallel program execution is usually achieved when the maximum degree of parallelism is taken as far as the number of processor cores.

III. CONLUSION

The speed of the serial execution of the program for different computers mainly is indicated directly from the processor clock speed. But the speed of parallel execution of the program can't be determined based on the model of the processor, or the number of cores, because the cache memory size and clock speed have an important impact. Execution speedup of the computer, depends on the number of processor cores, as well as the level of parallelism of the program being executed. But the level of a parallelism of a program execution depends directly on the speed of the parallel work of the computer. In computers with which we experimented, level of parallelism is over 0.65 and just under 1. In order to optimize the parallel speed of computer we can manipulate with the allowed number of threads. Experimental results show that if a single program is executed, the speed of computer work is usually better if the number of threads deal as far as the number of processor cores.

REFERENCES

[1] Blaise Barney. Intoduction to Parallel Computing. Lawrence Livermore National Laboratory. https://computing.llnl.gov/tutorials/parallel_comp/ [Accessed on 10.9.2014]

[2] Ananth Grama, Anshul Gupta, George Karypis, Vipin Kumar. Introduction to Parallel Computing, Second Edition. Addison Wesley, January 16, 2003

[3] Parallel Programming with .Net. http://blogs.msdn.com/b/pfxteam/ . [Accessed on 17.8.2014]

[4] Frank Willmore. Parallel Programming. The University of Texas at Austin https://portal.tacc.utexas.edu/c/document_library/get_file?uuid=d7f742a0-42c8-4407-b3c6-9fe1cbda14f1&groupId=13601 [Accessed on 25.11.2014]

[5] http://www.pcworld.idg.com.au/article/386100/what_difference_between_an_intel_core_i3_i5_i7_/ [Accessed on 18.12.2014]

[6] Jan Zapletal. Amdahl's and Gustafson's laws. VSB - Technical University of Ostrava, 2009

[7] Shared Memory Multiprocessors. http://www.inf.ed.ac.uk/teaching/courses/ pa/Notes/lecture04-multi.pdf [Accessed on 23.12.2013]

[8] http://www.pcworld.idg.com.au/article/386100/what_difference_between_an_intel_core_i3_i5_i7_/ [Accessed on 16.12.2014]

[9] http://www.cpubenchmark.net [Accessed on 18.1.2015]

[10] Optimal number of threads per core. http://stackoverflow.com/questions/1718465/optimal-number-of-threads-per-core [Accessed on 12.3.2013]

[11] https://msdn.microsoft.com/en-us/library/system.threading.tasks.paralleloptions.maxdegreeofparallelism%28v=vs.110%29.aspx [Accessed on 23.1.2014]

[12] Gaston C. Hillar. Proffesional Parallel Programming with C#, Wiley Publishing, Inc., Indianapolis, Indiana, 2011

[13] Peter S. Pacheco. An Introduction to Parallel Programming, University of San Francisco, Morgan Kaufmann Publishers, 2011

[14] Hans-Wolfgang Loidl. Parallel Programing in C#. Heriot-Watt University, Edinburgh, 2012

[15] http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e7-family.html [Accessed on 18.1.2015]

[16] http://www.intel.com/content/www/us/en/processors/core/5th-gen-core-processor-family.html [Accessed on 15.7.2014]

Performance Analysis of a New Cloud e-Business Solution

Kristina Kolic
Haybrooke Associates
10 Pelham Street, Leicester, LE2 4DJ, United Kingdom
Email: kristina.kolic@haybrooke.com

Marjan Gusev and Sasko Ristov
Ss. Cyril and Methodius University, FCSE
1000 Skopje, Macedonia
Email: marjan.gushev, sashko.ristov@finki.ukim.mk

***Abstract*—In this paper we explore if the applications with scaled user demands, such as the number of HTTP transactions achieve proportionally scaled performance. In addition, we explore which configuration is most convenient and achieves the best performance as a final conclusion about performances when applying the scaling of resources or demands. The experiments are performed for an e-Business solution hosted as a SaaS application on Windows Azure. Finding an optimal cloud infrastructure is not a trivial problem, since there are a lot of possibilities to organize multiple virtual machines in a cloud. The results obtained from the experiments show that the optimal configuration is based on use of a lower number of bigger VMs. In addition to this, the number of operations has an impact on the performance proportionally.**

***Keywords-e-Business solutions, SaaS, cloud computing, scaling, performance, Windows Azure, load testing, jMeter*.**

I. Introduction

Lately the cloud becomes one of the most popular platforms for hosting a solution. The power that the cloud represents in terms of the scalability and the pay-as-you-drink paying model attracts more and more companies to use it for hosting their software solutions.

The cloud is considered as industry platform when a solution becomes open to more than one company. The Software as a Service model (SaaS) went that far beyond the expectations of the clients development platforms. The offered web services or APIs of the SaaS applications make a system 'open, but not open' or 'closed, but not closed'. The SaaS model on the cloud represents an efficient way to use the hardware resources by multiple customers [1].

There are a few cloud computing platforms on the market, such as Amazon AWS, Google Cloud and Windows Azure. The latter provides neutral platform regarding the technologies, and at the same time allows the developers to extend the range of programming languages they use. On the other hand, Microsoft managed to implement their services and products in the Azure platform (Windows Live, Office Live, SQL Server, .NET and Sharepoint services, Visual studio) [1].

The cloud use has large impact on saving the budget that a company can afford for hosting a software solution. In this paper we aim to find out what is the right resource configuration from which a company can benefit? Does it depend on the software solution architecture and also the cloud infrastructure used?

To obtain relevant conclusions one has to provide measurement of a real-world user traffic and demo of a real application deployed on the cloud. We use an approach of performing load testing or stress testing. The test execution takes place in the cloud environment. The main reason why the cloud testing is important is to leverage the cloud resources for the price paid. It also helps in the domain of scalability of the solution in the cloud [2].

The model of the e-business system, which will be used in this research represents the processes by the means of a ticket management system. The benefit includes familiarity with monitoring an issue state (for example, an order was put through the system as new, than it was accepted and processed by the salesman, after which the invoicing and payment are being processed by the commercialist)[3].

We suppose that the system may behave differently under different conditions: when heavier operations are executed, the demand for more powerful server resources is requested, and other way around, when lighter operations are executed, the system will require lower server power. On top of this is the user load, considered to be in a range between small to medium (0-1000 users).

The cloud solutions are hosted on a set of virtual machines (VMs) and the main goal in this research is to find the optimal organization of VMs to host a real e-Business application, by analyzing the performance of various configurations for providing sufficient processing speed and quality of service. For this purpose we set the following hypotheses:

Hypothesis 1: An application that requires more processing activities is executed proportionally in more processing time.

Hypothesis 2: A system with a configuration that consists of multiple smaller VMs, such as A1 type VMs achieves better performance than a system with a configuration that uses the same total number of cores but uses large VMs, such as A3 type VMs.

II. Related work

The research conducted by Gushev et al. [4] shows the optimal cloud configuration in which a ticket management system reaches the best performance with small or medium user load. The configuration that enables the best performance in Microsoft Azure clouds is the one that uses several virtual machines with low number of cores. How Amazon EC2

environment behaved with an application that is data-intensive was examined by Schad et al. [5]. Similar parameters were considered as in the previous research: CPUs, memory speed, disk I/O, network bandwidth and they were examined on instances of different size (small, large etc). From this paper it is concluded that there is a large difference in the performance of different sized instances.

Another research conducted by Gusev et al. [6] suggests that the optimal configuration of compute intensive algorithms is better in configurations that use lower number of more powerful VMs, each with more cores (for example, at least 4 cores). This case is recommended when the company wants to scale the current cloud resources.

Xiong et al. [7] carried out a research that examines the percentile of the response time of a cloud solution. A solution that uses a queueing network model is used for the research purposes. Using the calculation of the probability distribution function of the response time (cumulative distribution functions), they developed a method to predict the service performance of the cloud computing software.

Iosup et al. [8] studied the behavior of a scientific software of the many-tasks nature. Their results showed that the cloud environments are still not as powerful as grids or parallel production infrastructures, but on the other hand, it gives great value for the money when the scientists need resources immediately and for short time.

Khanghahi et al. [9] summarized the way the cloud computing performance evaluation can be categorized. The evaluation can be based on the criteria and the characteristics of the solution or it can be based on simulation (that uses data centers locations, number of users etc. In our research, we will use simulation that is based on different cloud configuration and the number of users at specific moment.

III. Methodology

The demo application with an e-Business solution is hosted on the Microsoft Windows Azure cloud platform. There are two ways of consuming the platform resources: Windows Azure sites or virtual machines. The first approach is used for smaller web applications, with small user load and transactions, while the second is used for business applications or resources demanding applications. In this particular case, we consider the load of topmost thousand users, and our e-Business solution will be hosted on VMs.

For the research purposes, we will use three types of the standard set of VMs available on the Windows Azure, with capacities presented in Table I, which shows that the prices increase proportionally depending on the number of used CPU cores. It suggests that the price per core will be the same regardless of the VM. For example, renting one big (A3) VM will cost the same as four small (4 x A1) [10].

The architecture of the testing environment is presented in Fig. 1. The database of the system is hosted on a dedicated single VM. Since the database needs more server power, it is recommend to use the A3 type VM. The application used for the load testing is jMeter in its desktop version. It is installed on the same VM as the database, presented in Fig. 1.

TABLE I
Azure VM resources configuration and prices

Type	Cores	RAM	Price
A1	1	1.75GB	$ 0.09/hr (approx. $ 67/mo)
A2	2	3.5GB	$ 0.18/hr (approx. $ 134/mo)
A3	4	7GB	$ 0.36/hr (approx. $ 268/mo)

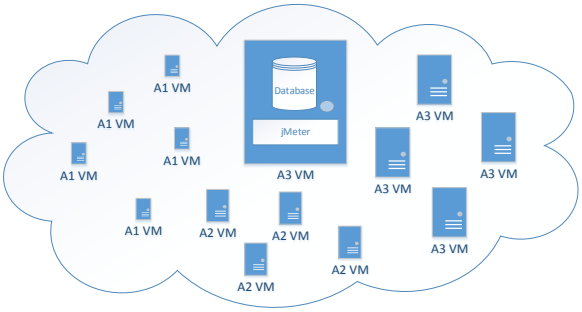


Fig. 1. Architecture of the testing environment and demo application

TABLE II
Testing configurations

cores	organization	id
1 core	1 A1 type VM	1xA1
2 cores	2 A1 type VMs	2xA1
2 cores	1 A2 type VM	1xA2
3 cores	3 A1 type VMs	3xA1
4 cores	4 A1 type VMs	4xA1
4 cores	2 A2 type VMs	2xA2
4 cores	1 A3 type VM	1xA3
5 cores	5 A1 type VMs	5xA1
6 cores	6 A1 type VMs	6xA1
6 cores	3 A2 type VMs	3xA2
8 cores	8 A1 type VMs	8xA1
8 cores	4 A2 type VMs	4xA2
8 cores	2 A3 type VMs	2xA3
10 cores	10 A1 type VMs	10xA1
10 cores	5 A2 type VMs	5xA2
12 cores	6 A2 type VMs	6xA2
12 cores	3 A3 type VMs	3xA3
16 cores	8 A2 type VMs	8xA2
16 cores	4 A3 type VMs	4xA3

The application itself will be hosted on several VMs of the same type and configuration. For example, a set of 1,2, 3 or 4 A1 VMs to be used along with the database VM. For our testing purposes, the number of VMs is limited to 20 and the test cases are defined with less than 20 cores (since the target is SME). Table II presents the VM resource organizations with various number of cores.

In our experiments, all VMs run in the same Azure datacentre placed in South Central US.

Demo used in this paper is the SaaS solution [11], as a hybrid between the ordinary e-Business system and the ticket management system. The idea is to use states for the processes in the e-Business, so each document can be tracked by the users. It is a realization of a transactional web application that processes documents related to the e-Business ordering,

invoicing and financial clearing processes.

The target users of this system are the SMEs (Small and medium enterprises). It allows the SMEs to be aware of every process that happens in their company, and at the same time it saves the budget of the company reserved for software solutions.

System features include management of: products, product catalogs and pricelists, offers for the customers, customers orders, invoices and payments according to the bank statements. Each of these documents and corresponding processes is described by several states of its lifecycle [11].

The system is developed in ASP.NET framework as a multi-tier web application. The database tier is realized by the Microsoft SQL Server database management system.

Two test cases are defined for the load test experiment.

TC1: Selection of all the orders from the system - This test case consists of a few test steps. In the first step, a user from the manager role logs in to the system. In the second step, the user navigates through the system to find the page where all the orders are listed. The final step is to open the page with the listed orders. Overall, the test case makes 3 HTTP requests.

TC2: Insertion of a new offer in the system - This test case is more complex than the previous one. It is made of 6 HTTP requests. First of all, a user from the manager role logs in to the system, and navigates to get to the page where all a history of the offers is presented. By clicking a button for adding a new offer on this page, a new page is presented with several information input fields. The required fields of this form are to be populated and the form submitted to save the offer. After this, the user is redirected to the list of all offers.

Both test cases are loaded with various users considered as small or medium users loads: starting from 1, 5, 10, 50, 100, and increasing by 50 up to 1000 users.

The basic test data for each test case and configuration is the response time T_R for executing the HTTP requests of the test case under a certain user load.

Performance P is equivalent to the processing speed (number of executed requests), derived from the response time for each configuration, calculated by (1).

$$P = N/T_R \tag{1}$$

where N is the number of HTTP requests of the test case and T_R is the response time in seconds. The unit which expresses the speed is *requests per second*.

The other performance indicators are based on evaluation of the performance (processing speed). The Relative speedup R is calculated for two differently structured configurations with the same number of cores, for instance 4xA1 and 1xA3 both having 4 cores. The relative speedup is always calculated within one test case (TC1 or TC2). The relative speedup between two configurations marked with x and y is calculated by (2) where P_x is the speed of the configuration x and P_y is the speed of the configuration y.

$$R_{xy} = P_x/P_y \tag{2}$$

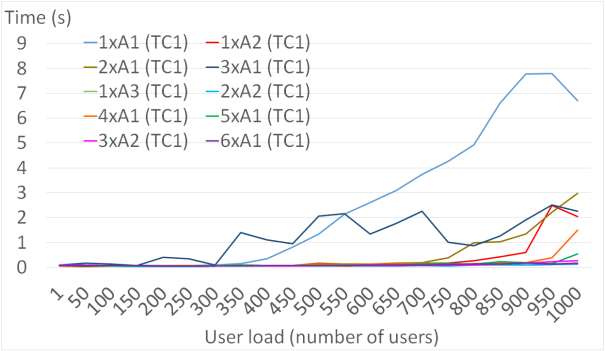


Fig. 2. Response time for TC1 on configurations with <8 cores

Let's analyze how the evaluation will be performed. To prove Hypothesis 2 we have to compare two configurations x and y that have the same number of cores. Let the configuration x use configurations with smaller number of more powerful VMs, such as A3, and y those with greater number of less powerful VMs, such as A1 or A2. If the relative speedup satisfies the relation $R_{xy} \leq 1$ it means that the x configuration achieves less performance and the Hypothesis 2 is valid, meaning that the configurations with smaller VMs achieve better performance and are optimal.

The experiment to prove the Hypothesis 1 compares two test cases realized as systems x and y, where the system y denotes the test case TC2 and x the test case TC1 for the same resource configuration. It means that system y needs more processing in comparison to x and one expects that $R_{xy} \geq 1$ to prove the Hypothesis 1. It shows that the processing activities take proportionally more time to be finished.

IV. Analysis of Results

In the first test case (TC1) we experiment how the system behaves when a list of the orders have to be selected under a certain user load. The test case contains 3 HTTP requests which is the total number of request per each user.

The second test case (TC2) shows how the system behaves when a user inserts a new offer under a certain user load. The test case contains 6 HTTP requests which doubles the amount of requests per user of the previous test case.

The parameters that are taken to compare the configurations are the response time of each configuration for user load and processing speed for each configuration.

The results for obtained response times are presented by charts, where the X-axis corresponds to the user load and Y-axis to the response time measured in seconds.

The results of the load tests for TC1 on cloud configurations using VMs with less than 8 cores (1, 2, 3, 4, 5 or 6 cores) are presented in Fig. 2. If we compare the configurations with 2 cores, we can see from the chart that both configurations have nearly the same time up to 750 users, and afterwards the configuration with A2 machine has better response time. When 4 cores are used, the chart represents that there is no such a difference in the response time of the configurations

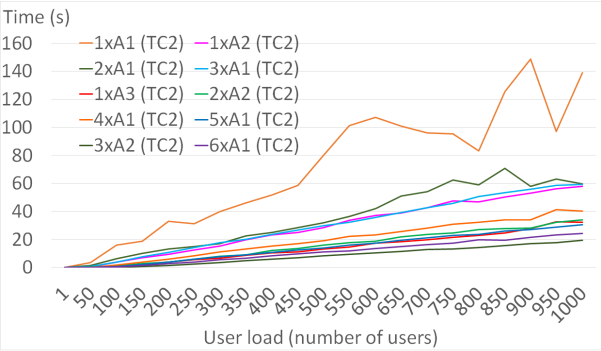


Fig. 3. Response time for TC2 on configurations with <8 cores

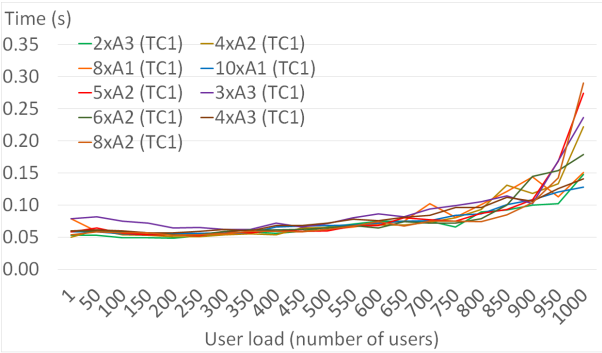


Fig. 4. Response time for TC1 on configurations with ≥8 cores

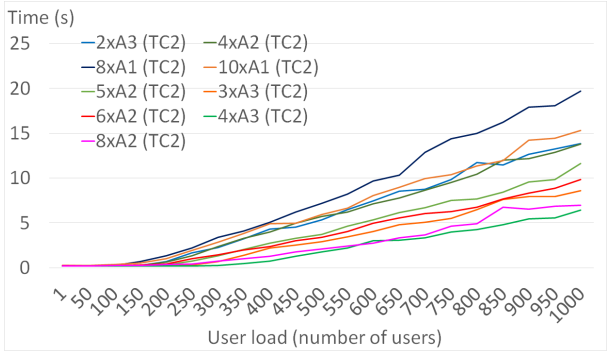


Fig. 5. Response time for TC2 on configurations with ≥8 cores

with lower number of VMs and higher number of cores (A2, A3) while the usage of A1 type of VMs has close response time to the previous configurations up to 900 users. When configuration of 6 cores is used for TC1, we obtained slightly better response time when the type A1 is used. Overall, one might easily conclude that the configurations with higher number of resources result with smaller response times.

The results of the load tests for TC2 on cloud resources using VMs with less than 8 cores are presented in Fig. 3. This test case contains more requests which results in a higher response time than TC1 at all times. When 2 cores configurations are used for TC2, it is clearly the configuration with VM of A2 type has better response time. For the set of 4 cores, the best response time in TC2 has the configuration of VM of A3 type regardless of the user load. In the case of 6 cores configuration in the TC2, we have the opposite result than TC1. For this configuration, at all the times, the configuration of A2 type VMs has better response time at all the times. The best case is when there are 6 cores and 3 VMs of A2 types, meaning VMs with multiple cores.

Figures 4 and 5 show the results of tests on configurations with a total of 8 or more cores (8, 10, 12 and 16 cores) respectively for the load test of TC1 and TC2. All the cores in one configuration are part of the same VM type (A1, A2 and A3).

In the case of using 8 cores in TC1, the configuration of less VMs shows better response time. The other two with bigger number of VMs get closer in the response time up to the load of 700 users. In the 10 cores configuration in TC1, better response time is obtained when higher number of VMs of A1 type is used. The other configuration of A2 type VMs provides response time closer to the previous configuration up to load of 900 users. Similarly to the previous configuration, when we have configuration of 12 cores in TC1, a better response time is given by the configuration with higher number of VMs of A2 type this time. The results for the configuration of 16 cores used in TC1 show that the set of higher number of VMs of type A2 has better response time up to the load of 900 users. Above this load, the configuration of lower number of VMs of A3 type has better response time. From this, we can sublime that we have the optimal results in the case when there are 8 cores in a configuration 2 VMs of type A3. For TC2 of this core number range, that is one of the configurations with the lowest number of cores.

The configurations that have A2 and A3 type of VMs in 8 cores set give almost the same response time in the terms of TC2. Both of the response times are better than the configuration that has A1 type of VMs. For 10 cores configuration set, the configuration with better response time is the one that uses A2 type of VMs in every user load. When 12 cores are used for TC2, the configuration with lower number of VMs of A3 type provides a better response time for each user load. Using 16 cores, the configuration that has better response time in almost all user loads is the configuration of A3 type of VMs and that is the optimal response time obtained for all the configurations in this case.

Based on the measurements of the response time for each test case and each configuration, we are able to calculate the processing speed dependent on the number of requests by (1). The results about the performance (processing speed) are presented in Figures 6 and 7 respectively for for TC1 and TC2. X-axis represents the user load and Y-axis the speed measured in requests per second.

We observe that at certain load, the cloud configuration gets saturated. After these points the speed gets even lower, due to queues and incapability to process all the requests. This is more evident for more processing demands, such as TC2.

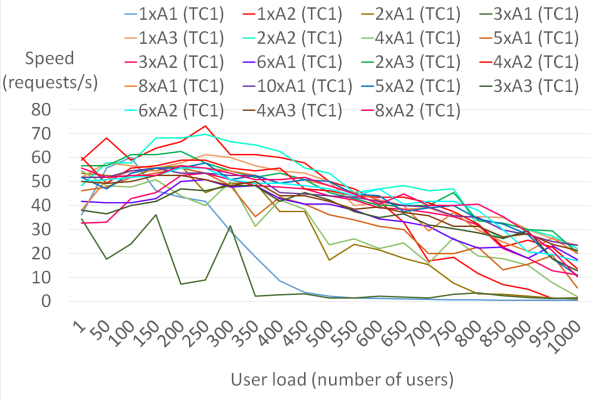


Fig. 6. Processing speed for TC1

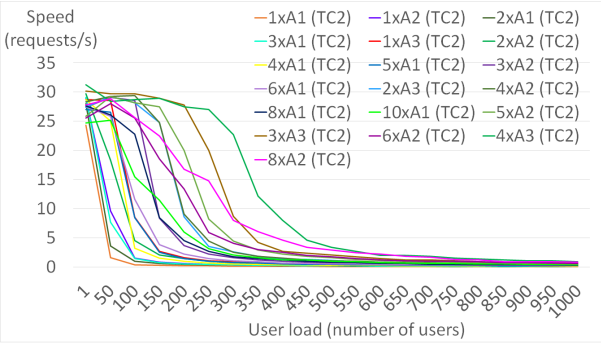


Fig. 7. Processing speed for TC2

When analyzing TC1, saturation appears mainly for the configurations that use A1 type of VMs (1xA1, 2xA1, 3xA1) and there is one case when A2 type VM is used, as presented in Fig. 6. At the worst case, the saturation starts from the user load of 500 users for 1xA1 configuration, while the best case when saturation appears is 2xA2 on 950 users at the user load. The chart also shows that the configuration 10xA1 has the optimal speed for the maximum user load of 1000 users, which is 23.4 requests/second.

Every configuration enters the saturation state at some point of the user load scale, as shown in Fig. 7. The worst case happens again for the 1xA1 configuration, which enters saturation when the load is only 100 users. The best case is the 4xA4 configuration which is saturated at 600 users of user load. At the user load of 1000 users, the configuration 4xA3 has the best speed, 0.4 requests/second.

In this section we compare the speeds of all the configurations for TC1 and TC2, where the processing demands are doubled. According to (2) we expect that TC2 will achieve at most half processing speed of TC1.

Fig. 8 presents the relative performance between TC1 and TC2 for each configuration. TC2 demands more processing and consumes more resources, at least double than TC1. The measurements show that the speedup value for some user loads and configurations goes up to 264 times. This approves the

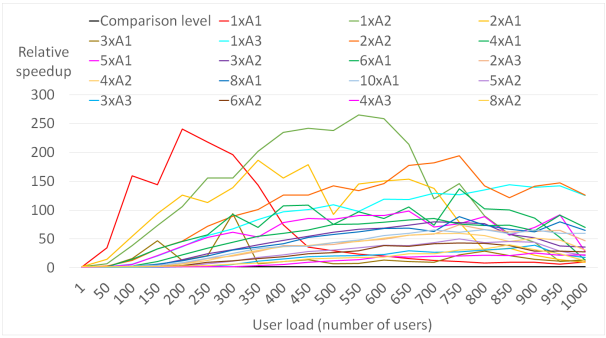


Fig. 8. Relative comparison for each configuration in TC1 and TC2

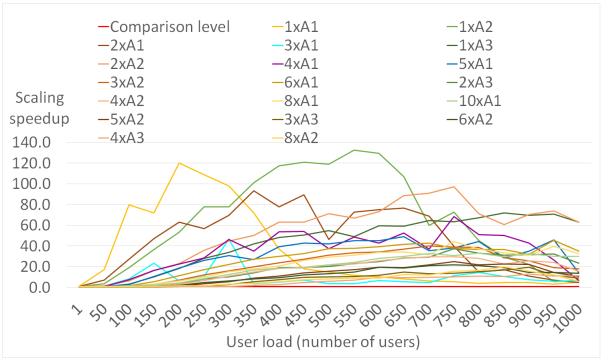


Fig. 9. TC1 and TC2 scaling speedup

Hypothesis 1.

More precise indicator is the scaling speedup. Since TC2 uses at least twice as much HTTP requests than TC1, the scaling factor is 2. Fig. 9 shows the speedup values of the same configurations for TC1 and TC2. We can see that all the configurations get scaling speedup greater than the linear, which means, the scaling factor is at least 2.

The optimal configuration is analyzed by the relative speedup of the configurations with the same number of cores. We perform experiments with 2, 4, 6, 8, 10, 12, and 16 cores, for both TC1 and TC2.

Fig. 10 shows the relative speedup of the configurations for TC1. There are only four cases that achieved results below the linear speedup, which means the better configuration is the one with more VMs of smaller type (it is A2 type in all the cases). The rest of the cases show different values for the relative speedup above the linear, which means the configuration with less larger VMs is better, and the optimal configuration can be chosen depending on the user load.

The relative speedup for the selected configurations for TC2 is shown in Fig. 11. In most of the configurations, the configurations using larger VMs perform better and can be chosen as optimal. Yet the optimal configuration for this test case is 3xA3 that contains 12 cores.

Table III shows the optimal configurations for number of cores for TC1 and TC2. From here, we can sublime that in most of the cases the optimal configuration is using less large

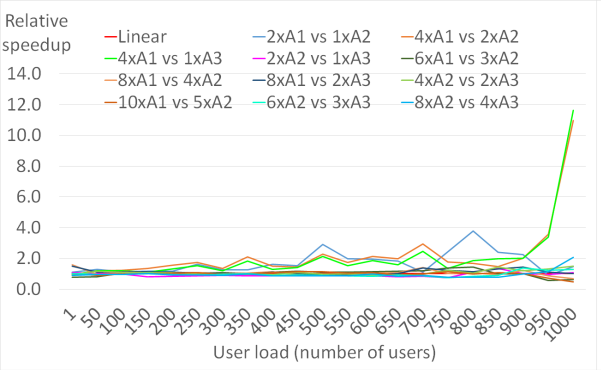


Fig. 10. Relative performance for TC1

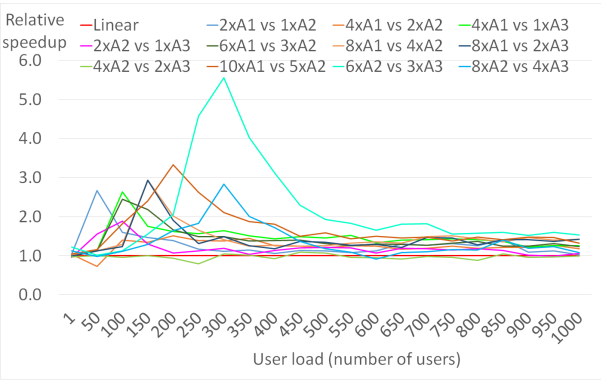


Fig. 11. Relative performance for TC2

TABLE III
OPTIMAL CONFIGURATIONS FOR TC1 AND TC2

Configurations	Cores	TC1	TC2
2xA1 vs 1xA2	2	1xA2	1xA2
4xA1 vs 2xA2	4	2xA2	2xA2
4xA1 vs 1xA3	4	1xA3	1xA3
2xA2 vs 1xA3	4	2xA2	1xA3
6xA1 vs 3xA2	6	3xA2	3xA2
8xA1 vs 4xA2	8	4xA2	4xA2
8xA1 vs 2xA3	8	2xA3	2xA3
4xA2 vs 2xA3	8	4xA2	4xA2
10xA1 vs 5xA2	10	5xA2	5xA2
6xA2 vs 3xA3	12	6xA2	3xA3
8xA2 vs 4xA3	16	8xA2	4xA3

VMs with higher number of cores (or type A2 or A3).

It means that the Hypothesis 2 fails, and the system performs better when using a configuration with small number of large VMs.

V. CONCLUSION

The results showed the optimal solutions for the e-Business system depends on the operations that it might have, 'heavy' and 'light' requests that also include the database transactions.

Our first test case in the realized e-Business solution is a 'light' operation, since it contains operations, such as changing the document status or listing the documents. These operations are quite common and massively used in the e-Business system since they do not involve a lot of database transactions. However, the e-Business system can be considered as a system with a lot of 'light' operations, but there are even more 'heavy' operations that include numerous insertions of new data, such as offers, orders, invoices or bank statements which requires a lot of input fields, various calculations of the account balance, and also processing the data for the purposes of reporting.

Moreover, the frequency of the heavy operations depends on the company business. If it is a business that has a huge number of orders everyday, the operations will be very often. Once the order is created and processed, it triggers the creating of an invoice, which then partially guarantee that there will be a bank statement which matches the invoice. Each process will trigger an operation which is considered to be heavy, and will eventually make the system work better on more powerful server resources under a heavy user load.

Our experiments on the performance of the e-Business SaaS solution has showed that:

- Hypothesis 1 holds when comparing the results of the relative performance of two test cases, one of which has double the amount of HTTP requests.
- Hypothesis 2 fails, which means that the optimal configuration is the one with lower number of large VMs.

We conclude that for a real transactional based e-Business application, it is better to choose configurations that use lower number of larger VMs (A3 type). When scaling, the software provider must have in mind that the speedup will be equal or higher than the number of cores. An application, such this, that requires more processing (for HTTP requests, database transactions etc.) demands proportional processing time to the requests.

REFERENCES

[1] M. Cusumano, "Cloud computing and saas as new computing platforms," *Communications of the ACM*, vol. 53, no. 4, pp. 27–29, 2010.
[2] J. Gao, X. Bai, and W.-T. Tsai, "Cloud testing-issues, challenges, needs and practice," *Software Engineering: An International Journal*, vol. 1, no. 1, pp. 9–23, 2011.
[3] K. Kolic, S. Ristov, and M. Gusev, "A model of SaaS e-business solution," in *Proceedings of the 22nd International TELFOR Forum, IEEE Conference Publications*, 2014, pp. 1146–1149.
[4] P. Gushev, S. Ristov, and M. Gusev, "Performance analysis of SaaS ticket management systems," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conf. on*. IEEE, Sept 2014, pp. 753–760.
[5] J. Schad, J. Dittrich, and J.-A. Quiané-Ruiz, "Runtime measurements in the cloud: observing, analyzing, and reducing variance," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 460–471, 2010.
[6] M. Gusev, S. Ristov, B. Koteska, and G. Velkoski, "Windows Azure: Resource Organization Performance Analysis," in *Service-Oriented and Cloud Computing*. Springer, 2014, pp. 17–31.
[7] K. Xiong and H. Perros, "Service performance and analysis in cloud computing," in *Services-I, 2009 World Conf. on*, 2009, pp. 693–700.
[8] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, and D. H. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 6, pp. 931–945, 2011.
[9] N. Khanghahi and R. Ravanmehr, "Cloud computing performance evaluation: Issues and challenges," *computing*, vol. 1, p. 5, 2013.
[10] Microsoft. (2014). [Online]. Available: http://azure.microsoft.com/en-us/pricing/details/virtual-machines/
[11] K. Kolic, M. Gusev, and S. Ristov, "A new e-business solution with ticket management," in *Proceedings of i-Society 2014, IEEE Conference Publications*, 2014, pp. 203–208.

Interoperable Job Execution and Data Access through UNICORE and the Global Federated File System

Shahbaz Memon*‡, Morris Riedel*‡, Chris Koeritz† and Andrew Grimshaw†
*Juelich Supercomputing Center, Forschungszentrum Juelich GmbH
Juelich, Germany, m.memon@fz-juelich.de
†Department of Computer Science,University of Virginia
Charlottesville, USA
‡School of Engineering and Natural Sciences, University of Iceland,
Reykjavik, Iceland

***Abstract*—Computing middlewares play a vital role for abstracting complexities of backend resources by providing a seamless access to heterogeneous execution management services. Scientific communities are taking advantage of such technologies to focus on science rather than dealing with technical intricacies of accessing resources. Multi-disciplinary communities often bring dynamic requirements which are not trivial to realize. Specifically, to attain massivley parallel data processing on supercomputing resources which require an access to large data sets from widely distributed and dynamic sources located across organizational boundaries. In order to support this abstract scenario, we bring a combination that integrates UNICORE middleware and the Global Federated File System. Furthermore, the paper gives architectural and implementation perspective of UNICORE extension and its interaction with Global Federated File System space through computing, data and security standards.**

I. Introduction

Distributed computing middleware platforms are expected to provide an abstract and convenient interface through which scientific communities can manage complex simulations and workflows. As we have seen a recent wave of big data which covers data intensive applications requiring unified data access and bringing large data sets to execution services representing high performance computing resources. The execution service which is a part of compute middleware, aid and orchestrate these large workflows and aggregates data from multiple sources for pre-processing and analysis tasks. These sources might be managed by different kinds of remote data management services. The data management interfaces for reading and writing data are significantly different from each other. Thus, it is imperative to let execution service know the interface of each data management service instance. In order to enable an execution service for multiple data services may require adaptors, which in the worst case proportional to the number of intended data management interfaces, therefore it is not practically feasible to develop and maintain all of them. With this, we deduce two following requirements from the execution service's perspective. Firstly, a federated data management system should be in place that supports interoperable data sharing through a unified interface. Secondly, the execution service which exposes high performance computing resource, is capable of interacting with data management systems in a secure and reliable manner.

For the sake of gaining more insight on requirements and architectural implementation we consider XSEDE infrastructure as a case study. XSEDE has a diversified set of user communities which are running jobs on high performance computing (HPC) and high throughput computing (HTC) resources. In XSEDE the UNICORE [1] middleware provides execution service capabilities and abstraction over multiple batch systems deployed on high performance computing clusters. For some applications the use of resources can span both HPC and HTC type of resources, for example in case of a complex workflow. In our scenario the federated data access across multiple sites is realized through the Global Federated File System (GFFS) [2]. It is an open standards based data management service for accessing a myriad of file systems in a consistent heirarchical representation. The GFFS is an integral part of GenesisII middleware [3], which is a services suite for realizing high throughput computing scenarios. The scope of our work is to extend UNICORE for the scientific communities who are using the GFFS and the GenesisII middleware. This will be useful to the scientific communities who are already using the UNICORE services over XSEDE infrastructure and may want to access data from GFFS space. Another impact will be on a group of users who are intending to run complex workflows spanning the simultaneous use of both high performance (through UNICORE) and throughput (GenesisII) jobs that use and share data through the common GFFS space. To enable UNICORE managed jobs to interact with the GFFS, the execution service of UNICORE should understand the interfaces through which the GFFS data can be managed and transported. In this paper we describe which part of the UNICORE's execution tier is extended for the GFFS, and how the GFFS server and client is modified to recognize UNICORE services.

The rest of the paper is conformed to the following structure. Section II gives a background on technologies and standards involved. Section III elaborates the security model, while the technical aspects of middleware extensions are described in Section IV. Section V describes the related work.

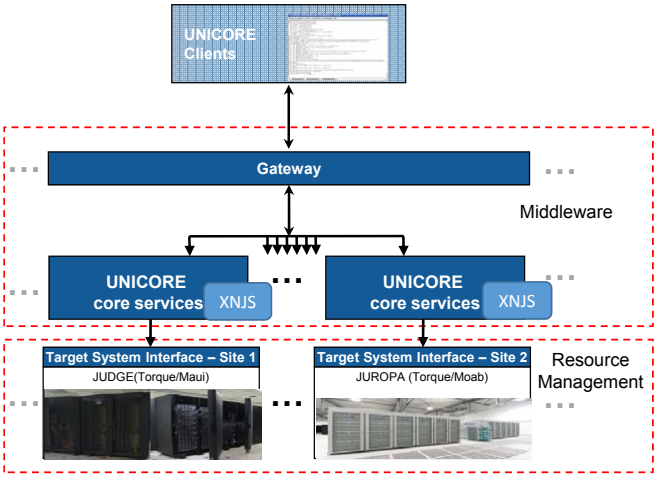


Fig. 1: An overview of UNICORE Architecture.

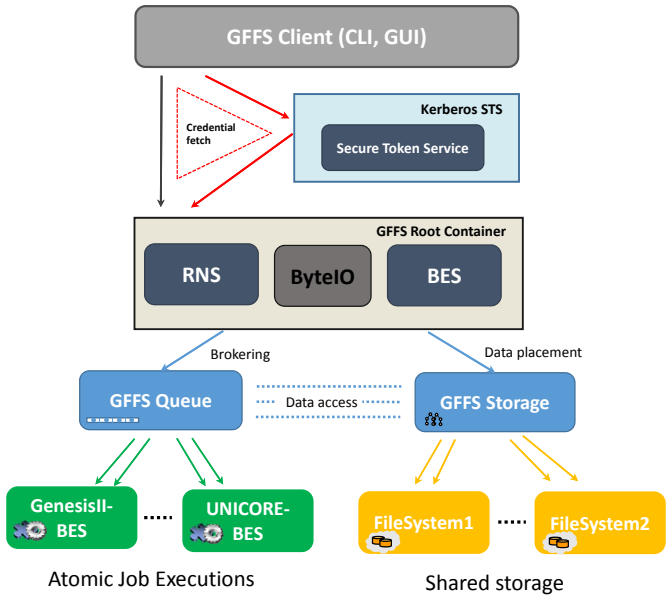


Fig. 2: A holisitic view of the Global Federated Filesystem architecture.

We conclude the paper in Section VI.

II. Background

This section provides an overview of the technologies which are involved in the implementation.

A. UNICORE

UNICORE is a service oriented middleware which provides a set of compute and data services to access high performance computing resources. The UNICORE architecture is composed of three layers as shown in Figure 1.

The client layer is responsible for providing user interface, which includes preparing and sending client requests to the remote deployed UNICORE web services. The server side is founded upon the principles of service oriented architecture. It encapsulates a SOAP based web services environment capable of hosting WS-I (Web Services - interoperability) and WSRF (Web Services Resource Framework) compliant service endpoints. The job execution and data management capabilities are exposed through web service endpoints. Once the client submits a job, it is forwarded to XNJS (Extended Network Supervisor), which processes a job request by first validating the submission request and then after a successful validation it is communicated to the Target System tier. At this layer the Target System Interface (TSI) component is deployed on a login node of a batch system. It directly interacts with the low level batch system, and has a main responsibility of translating job requests into batch system specific command scripts. It also monitor job statuses, and manage underlying running file transfers during the job's execution life cycle.

B. The Global Federated File System

The Genesis II Global Federated File System (GFFS) is at its root a networked file system that provides researchers with tools for securely managing and sharing their computational data. The GFFS also provides a means of managing jobs that use that data via the GFFS-Queue. The GFFS implements a web services architecture that uses standards-based protocols for storing and accessing data over a network. To support federation across different organizational entities, the GFFS provides a rooted file system structure with standard namespace locations for storing users (via Secure Token Services), groups, home directories, and computational resources such as queues and Basic Execution Service endpoints (BES) for processing jobs. Figure 2 shows a simplistic view of the integrated architecture encompassing GenesisII and UNICORE Basic Execution Service (BES) endpoints interaction with the GFFS's root container. The UNICORE's BES implementation is explained in the next section.

Several mechanisms are implemented for storing and retrieving data in the GFFS. The GFFS implements many of the standard Unix commands (such as "cp" and "mv") in a console-mode grid shell. There is also a graphical user interface view of the GFFS, which supports drag and drop file management. The GFFS also provides a "FUSE" file system that allows users to mount the GFFS on a Unix directory and operate on files in the GFFS as if they were simply part of the local file system. Lastly, the GFFS also provide an exports feature that allows users to make part of their own file system visible within the GFFS. The GFFS Queue supports submission of multiple jobs for subsequent distribution to the execution service endpoints connected to the queue. The GFFS Queue provides researchers with a mechanism for managing and controlling their computations via a graphical user interface as well as with familiar command line tools such as "qstat" and "qkill". Jobs will be distributed to BES resources automatically by the GFFS Queue, but can also be rescheduled as needed. The XSEDE project benefits from the GFFS by giving researchers a way to securely share data with their colleagues and by providing a high level view of the computational resources available at XSEDE in the GFFS Queue.

```
<Application>
  <ApplicationName>SCIAPP</ApplicationName>
</Application>
<DataStaging>
  <FileName>sciapp.in</FileName>
  <Source>
    <URI>rns:/work/app/sciapp.in</URI>
  </Source>
</DataStaging>
<DataStaging>
  <FileName>sciapp.out</FileName>
  <Target>
    <URI>rns:/work/app/sciapp.out</URI>
  </Target>
</DataStaging>
```

Fig. 3: JSDL XML instance showing GFFS data staging.

C. Related Standards

In order to carry out file system operations such as mkdir or cp, the RNS (Resource Namespace Service) [4] protocol is used. The RNS specification provides client applications to have human readable mappings with WS-Addressing based endpoints. Considering the WSRF based endpoints wherein web service resources can have a nested heirarchical structure, the RNS can be very helpful in this case by providing a fault and location transparent access to resources in a simple and human readable representation. The GFFS implements the RNS and also its Web Services Resource Framework (WSRF) [5] rendering to access web service resource endpoints. As far as the data transfer is concerned the GFFS primarily uses the ByteIO [6] standard. ByteIO provides a set of interfaces to interact with bulk data sources and sinks.

Job Submission Description Language (JSDL) [7] is XML based data model that describe computational job's requirements in terms of application, resources, that will run in distributed computing environments. UNICORE and GenesisII clients specify job requirements in the JSDL format. UNICORE server side implements most of the JSDL, and its related profiles and extensions.

OGSA-Basic Execution Service (BES) [8] provides a web services based interface for managing and monitoring computational jobs in high performance and high throughput computing environments. For a job submission use case BES interface accepts a JSDL instance as a parameter and then runs it on backend resource. In our scenario UNICORE server exposes its computing capabilities via BES, and GenesisII client use this interface to invoke remote calls on the UNICORE endpoint.

III. Security Model

The GFFS' security model is based on the SAML [9] standard, and specifically adopts the UNICORE's SAML profile [10] and implementation. The GFFS implementation of SAML extends the UNICORE model to understand the "delegation chains" of the Genesis II security model. A delegation chain encodes that a User has delegated some amount of trust to a resource in the GFFS in order to achieve a task, such as processing a job. Most delegation operations can be considered to start with three "entities" involved. These entities are (1) a grid user (for our example, called U, (2) a TLS connection identified by an X509 certificate (called C), and (3) a grid resource (called R). Longer delegation chains usually contain all three of these types of entities as the first links in the chain.

In the GFFS, the first entity U is always a user defined as a grid STS (Secure Token Service) object. This entity is the prime mover for any operations that are performed in the GFFS. The user's rights within the grid's access control list permission system dictates what that user can and cannot do with regards to every grid resource.

The client software must authenticate as the grid user U to obtain services from a GFFS container. This is where the second security entity C comes in; it is the connection made by the client software at the behest of the user. Initially, the client credentials only contain C, as one makes the connection before STS (Secure Token Service) authentication occurs. In the XSEDE Login process, this connection is always based on X509 credentials obtained from the XSEDE MyProxy server, and thus can be a well-known identity within both Genesis (via a Kerberos-based STS) and UNICORE (via the grid-mapfile). In the example, the client software first authenticates to MyProxy by using user's name and password to obtain the certificate C. Then, the client authenticates to the Kerberos STS in the GFFS to obtain grid user U.

After the user authenticates, the client software's "credential wallet" will contain the first delegation of trust, U -> C. This states that the grid user U trusts the TLS connection C to act on its behalf. Afterward, all of the actions taken by the TLS connection C are understood to be U's actions. Any access that provide U with permissions will also be granted to C. That may include submitting a job to a BES (Basic Execution Service) named R. This is the second point where trust is delegated; the certificate for TLS connection C signs a new trust delegation that says "C trusts R" to perform a job execution. This extends the length of the delegation chain by one, so that it now has all three entities involved in two trust delegation objects. This can be represented with delegation arrows such as:

First delegation: U -> C (The grid user trusts the TLS connection) Second delegation: C -> R (The TLS connection trusts the resource) Full Chain: U -> C -> R (The user U trusts the connection C which trusts R to run a job)

This new delegation chain can be presented by the resource R when it needs to do further actions on the grid user's behalf. Actions might include storing file staging results back to RNS space (U -> C -> R -> D, where D is a Data folder) or possibly submitting the job to another BES for final processing. The important point is that entity D is just another resource to which trust can be delegated, and the chain can be continued in that manner for as long as its delegation depth limit allows.

The prior implementation of Genesis II security, was not based on SAML, used nested, serialized Java objects to define a tightly coupled chain, with each nesting level signed by the issuer at that level. This implementation was considered to be a problem for interoperability because it is tightly-coupled with the GFFS Java classes and cannot in any sense be considered a standard (or standardizable). The new SAML implementation relies instead on UNICORE's SAML assertions, which represent the delegation chain as isolated X.509 based assertions that are tied together with unique identifiers. Each SAML assertion is also signed by the issuer, but there is no nesting and no reliance on serialized GFFS Java objects.

One challenge to interoperating between the GFFS and UNICORE implementations arose due to a difference in interpretation of the SAML assertions. The delegation chains in the GFFS are tightly-coupled, and do not allow mixing and

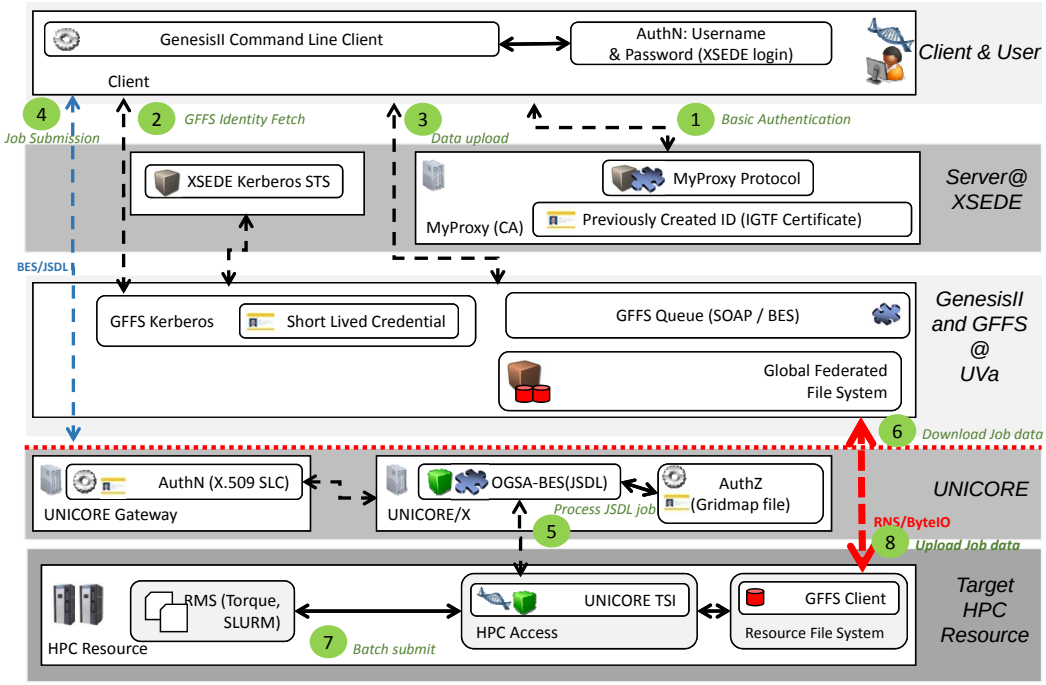


Fig. 4: Integrated architecture showing GenesisII client interacting with UNICORE and the GFFS

matching of individual entities. This is not directly provided by the UNICORE's SAML implementation, which permits the receiver to mix and match any delegations provided in a message (U -> C -> R is considered to be two separable delegations U -> C and C -> R). In the GFFS model, a delegation chain must be used in its entirety or not at all.

To address this difference, the GFFS implementation of SAML adds a unique identifier to each SAML assertion. A chain such as U -> C -> R is built by embedding the identifier of the U -> C assertion in the C -> R assertion. To make this cryptographically secure, the signature of the U -> C assertion is also embedded in the C -> R assertion before C -> R itself is signed. This enforces the connections between the GFFS delegation chains while still leveraging the UNICORE's Security Assertion Markup Language (SAML) implementation. Upon reception, the chains are reassembled and any assertions that are referenced by a longer delegation chain are removed from the pool of available assertions.

IV. Middleware Enhancements

UNICORE's server tier is extended to accept the incoming requests from GenesisII remote clients. The remote clients here implies the GFFS's GFFS-Queue component which is an entry point for a user to submit job. The GFFS-Queue acts as a meta scheduler that brokers a job based on its resource requirements on available set of living Basic Execution Service (BES) based computing endpoints. While considering UNICORE as one of the BES endpoints we need to understand how to interpret, authenticate, and authorize the incoming GFFS-Queue requests. A separate UNICORE server extension is implemented that is invoked when server finds a security token containing GFFS related information in the incoming client request. The extension validates the SAML chain by looking into every element of the chain. These elements are entities which contain every stake holder including end user or service through which the request was passed. The details of the entities are described in the previous section.

Before a GenesisII client is able to send jobs to a UNICORE's Basic Execution Service (BES) endpoint, a Genesis II container should recognize and link the UNICORE's Basic Execution Service (UNICORE-BES) instance into the RNS space. The linking is achieved through Endpoint Reference Minting process. An EPR (Endpoint Reference) is the basic component of the Resource Namespace Service (RNS). Every location in the GFFS namespace has an EPR that identifies (1) where the resource lives and (2) the X.509 certificate that represents the resource. Minting an EPR is the process of creating a new EPR as an XML document that represents an external resource, such as a UNICORE-BES. The process of minting an EPR combines the URI where the resource is located with the X.509 certificate expected as the resource's identity (which it would report over a TLS connection). Once an EPR is minted, the EPR's XML document can be stored locally as a file or added as a new link to the grid namespace. When linked into the grid, a user with appropriate credentials can see the entry in the GFFS files ystem and can use it to obtain whatever services the resource provides.

The user's XSEDE identity is extracted from the delegation chain, and the retrieved identity is validated against the authorization store of the UNICORE server deployment. If the user is found under the authorization store then the required user context is created for carrying out GFFS data staging call outs. After the context creation phase, the server extension releases its control and job moves to the next phase of execution. In the beginning of the execution phase the request is processed further to carry out the GFFS based data stagings of jobs. Figure 3 shows the job request encoded in JSDL containing application requirements and data staging elements pointing

to the GFFS space. Note that the job will be executed on the UNICORE site, therefore GenesisII-BESes are not involved in this sequence.

The GFFS based data transfer is realized through an XNJS extension. The GFFS download component prepares a command to pull data from the GFFS. The command which looks like "grid cp remote-source job-working-directory", will be forwarded to the Target System Interface (TSI) that invokes the GenesisII's "grid" command gracefully and downloads the data to the job's working directory. In a likewise manner the GFFS upload takes care of uploading output files to the remote GFFS location. Any failure will make UNICORE fail the job and abandon any further processing related to it. For every job which is sent by GenesisII client UNICORE server extension maintains an additional folder in each of the job's working directory that contain user's contextual information.

The pictorial representation of a simple job submission sequence is shown in Figure 4. In the first step GenesisII GUI client asks a user to login through an XSEDE provided credentials. After the authentication phase, she uploads a data to the GFFS folder. The user then creates a job request with application details, and location of the data that is mentioned just before. In a similar manner, output data paths are also specified. The user then selects a UNICORE endpoint and run the job. The cient continuously monitors the job until it reaches to a terminal state. Once the job is finished successfully the outputs are fetched back from the remote job working directory which is located on cluster's file system to the GFFS space. For the sake of brevity a sequence of major steps are being highlighted.

V. Related Work

In this section we present computing middlewares integrated with wide area file systems. GridFTP [11] is one of the major data transfer protocols used in today's scientific and commercial data infrastructures. Specifically, GlobusOnline data transfer service is mainly using this protocol to move data across widely distributed endpoints. UNICORE's GridFTP extension [12] help scientists to submit job executions on UNICORE and using GridFTP based data endpoints for data stagings. From the implementation perspective, both the GridFTP and GFFS extensions are hooked in at the same architectural level, that is in XNJS. In the case of GridFTP integration, the clients requiring UNICORE server to perform data staging on user's behalf need to send at least X.509 proxy certificate chain along with the job submission request. For the GFFS the entities communicating the job request to the server must send the complex chain of SAML assertions.

ARC [13] is a middleware suite used by communities which mostly use high throughput computing capabilities. ARC's integration with [14] and DDM (Distributed Data Management) [15] solutions are mostly used by particle physics community, ATLAS [15] at the Large Hadron Collider. dCache and DDM are distributed data management platforms providing storage and retrieval of huge amounts of data. dCache and the GFFS share mostly the same set of scenarios, but the major difference is, the GFFS expose its interface via RNS and ByteIO, whereas dCache is accessed through the SRM [16] interface.

WS-PGRADE / gUSE (grid and cloud user support environment) [17] is an open source scientific gateway framework that allows access to hetrogeneous grid and cloud resources. gUSE provides a client extension in the form of DCI bridge [18] to the GFFS by invoking GenesisII clients. It is much similar to the way UNICORE integrates the GFFS. The framework provides access to UNICORE and ARC job submission services through the OGSA-BES interface.

VI. Conclusion

UNICORE's OGSA-Basic Execution Service (BES) based execution and the RNS- and ByteIO-enabled Global Federated File System allow standards based combination for sharing and accessing resources across infrastructural boundaries. The representation and accessibility of resources such as files, directories, jobs, compute endpoints, and other web services can be accessed through usable hierarchical structure which is common to physical file systems. Presently, the proposed work has passed rigorous testing by the XSEDE's Software Development & Integration, and XSEDE operations team according to functional, security, and performance use cases. Now the implementation is accepted for production deployment, for which some sites are in transition of updating UNICORE endpoints for the GFFS access. In the current version, UNICORE explicitly invokes the GenesisII command line client to access the GFFS space. This usage of the GFFS client requires that client be installed on each resource management system where UNICORE resides. In the next iteration of UNICORE, the dependency of the client will be removed and instead the GFFS' programmatic interface will be adopted.

Enabling UNICORE to access the wide-area distributed GFFS can help scientific communities perform effective processing of large data sets on high performance computing platforms. This work has generated an interest from the science gateway communities across XSEDE who are intending to run data and compute intensive jobs that leverage the GenesisII based high throughput features and UNICORE supported high performance computing resources.

Acknowledgment

This work is partially funded by the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] A. Streit and et al., "UNICORE 6-Recent and Future Advancements," *Annals of telecommunications*, vol. 65, pp. 757 – 762, 2010. [Online]. Available: http://juser.fz-juelich.de/record/11372

[2] A. Grimshaw, M. Morgan, and A. Kalyanaraman, "GFFS - The XSEDE Global Federated File System," *Parallel Processing Letters*, vol. 23, no. 02, p. 1340005, 2013.

[3] M. Morgan and A. Grimshaw, "Genesis ii - standards based grid computing," in *Seventh IEEE International Symposium on Cluster Computing and the Grid, 2007.*, 2007, pp. 611–618.

[4] M. Morgan, A. Grimshaw, and O. Tatebe, "RNS Specification 1.1," OGF, December 2010.

[5] M. Morgan and O. Tatebe, “RNS 1.1 OGSA WSRF Basic Profile Rendering 1.0,” OGF, December 2010.

[6] M. Morgan, “ByteIO Specification 1.0,” OGF, Oct 2006.

[7] A. Anjomshoaa and et al., “Job Submission Description Language (JSDL), Version 1.0,” OGF, July 2008.

[8] I. Foster and et al., “OGSA Basic Execution Service (BES), Version 1.0,” OGF, Nov 2008.

[9] S. Cantor and et al., “Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0,” OASIS, March 2005.

[10] K. Benedyczak, P. Bała, S. van den Berghe, R. Menday, and B. Schuller, “Key aspects of the UNICORE 6 security model,” *Future Generation Computer Systems*, vol. 27, no. 2, pp. 195 – 201, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X10001524

[11] A. Sim and et al., “GridFTP v2 Protocol Description,” OGF, May 2005.

[12] “UNICORE/X Manual,” http://tinyurl.com/olp25fe, [Online; accessed 11-Sept-2014].

[13] M. Ellert, M. Grønager, A. Konstantinov, B. Kónya, J. Lindemann, I. Livenson, J. L. Nielsen, M. Niinimäki, O. Smirnova, and A. Wäänänen, “Advanced resource connector middleware for lightweight computational grids,” *Future Gener. Comput. Syst.*, vol. 23, no. 2, pp. 219–240, Feb. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.cam.2006.05.008

[14] P. Fuhrmann and V. Gülzow, “dCache, Storage System for the Future,” in *Euro-Par 2006 Parallel Processing*, ser. Lecture Notes in Computer Science, W. Nagel, W. Walter, and W. Lehner, Eds. Springer Berlin Heidelberg, 2006, vol. 4128, pp. 1106–1113. [Online]. Available: http://dx.doi.org/10.1007/11823285_116

[15] G. Behrmann, D. Cameron, M. Ellert, J. Kleist, and A. Taga, “Atlas ddm integration in arc,” *Journal of Physics: Conference Series*, vol. 119, no. 6, p. 062015, 2008. [Online]. Available: http://stacks.iop.org/1742-6596/119/i=6/a=062015

[16] A. Sim and et al., “The Storage Resource Manager Interface Specification, Version 2.2,” OGF, May 2008.

[17] P. Kacsuk, Z. Farkas, M. Kozlovszky, G. Hermann, K. Balasko, Akos an d Karoczkai, and I. Marton, “WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities,” *Journal of Grid Computing*, vol. 10, no. 4, pp. 601–630, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10723-012-9240-5

[18] “DCI Bridge Manual - v3.6.7,” http://sourceforge.net/projects/guse/files/3.6.7/Documentation/, [Online; accessed 11-Sept-2014].

Multi-stream 2D frequency table computation on dataflow architecture

P. Škoda*, V. Sruk**, and B. Medved Rogina*
* Ruđer Bošković Institute, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
pskoda@irb.hr

Abstract – Frequency table computation is a common procedure used in variety of machine learning algorithms. In this paper we present a parallelized kernel for computing frequency tables. The kernel is targeted for dataflow architecture implemented on field programmable gate array (FPGA). Its performance was evaluated against a parallelized software implementation running on a 6-core CPU. The kernel with six concurrent input data streams running on 300 MHz achieved speedup of up to 6.26×, compared to 6 threaded software implementation running on 3.2 GHz CPU.

Keywords – frequency table, dataflow, field programmable gate arrays, decision tree learning

I. INTRODUCTION

The 2D frequency table is a two-dimensional extension of histogram. In machine learning context, the frequency table is often used to provide information on how frequent specific combinations of attribute and class values are in a given dataset. This specific use is a key part of the attribute selection process in decision tree learning algorithms [1]. Applications of decision tree learning, as well as the field of data mining in general, are faced with continuing increase in dataset sizes. This drives efforts to develop new faster and more efficient algorithms, as well as developing new hardware platforms that provide high computational power [2].

One hardware platforms that is gaining more ground in computing is the field programmable gate array (FPGA). FPGA. FPGAs are digital integrated circuits designed to be user-configured after manufacturing [3], [4]. They enable fast development and deployment of custom digital hardware, and thus allow implementation of custom computational architectures that can be reconfigured on demand.

The most suitable computational mode for FPGA is the dataflow model. Most computer systems implement a control-flow model, in which the computation is described by a sequence of operations that are executed in specified order. In dataflow model, computation is described by chain of transformations applied to a stream of data [5]. The dataflow is described in form of a graph. Nodes of the graph represent operations, and edges represent data links between the nodes. The dataflow graph maps naturally to the hardware implementation. Nodes/operations are translated to functional hardware blocks, while edges are translated to data buses and signals. By dividing the graph into stages separated by registers, it's transformed into a pipelined structure suitable for implementation on FPGA.

In this paper we present a novel implementation of 2D frequency table computation targeted for dataflow architecture realized on FPGA. The kernel is developed for integration with C4.5 decision tree learning program [6], [7]. Its result is used to calculate dataset split quality, using an entropy based measure, which is then used for attribute test selection.

II. RELATED WORK

2D frequency table computation implemented on FPGA has been previously published as a part of efforts in implementing decision tree learning algorithms. FPGAs have been proven to be suitable of implementing data mining algorithms, though so far little attention has been directed to decision tree learning algorithms. Only a few attempts of using FPGAs to accelerate execution of decision tree learning have been reported [8], [9].

Chrysos et al. [9] implement frequency table computation on FPGA as part of HC-CART system for learning decision trees. Frequency table computation is implemented in Frequency Counting module, which receives attribute-class label pairs. The pairs are used to generate addresses for frequency table locations which contents are then incremented by one. In their implementation, all attribute-class label pairs received by the kernel have to be prepared by program running on CPU, and then preloaded to memory attached to FPGA. Input data is transferred from CPU to FPGA memory many times in the course of program execution.

In our previous work [10], we implemented a 2D frequency table computation using FPGAs. The *ComputeFreq* kernel receives two input streams – one for address and one for class labels – and counts occurrences of pairs in FPGA's internal memory. The dataset is held in FPGA attached SDRAM. The dataset is transferred to the FPGA's SDRAM only once, and only data transfer between the kernel and the host system is the transfer of computed frequency tables. We build upon this work by expanding it to multiple attribute streams which allow computation of several tables simultaneously, and by implementing efficient data transfer for the computed tables.

TABLE I. BASIC INFORMATION ON MAXELER VECTIS-LITE FPGA BOARD

FPGA	Xilinx Virtex-6 XCVSX475T
Off-chip RAM (LMem)	6 GiB (6×1 GiB) DDR3-800 SDRAM
On-chip RAM (FMem)	~ 4 MiB Block RAM
CPU ↔ FPGA bandwidth	2 GB/s
LMem ↔ FPGA bandwidth	38.4 GB/s max.

III. DATAFLOW ENGINE ARCHITECTURE

The accelerator is implemented in form of a dataflow engine (DFE) which consists of at least one kernel and a manager. The computation is implemented by kernels, while the role of manager is to organize data movement within the DFE. Kernels are instantiated within the manager, and data links between the kernels, external RAM, and CPU are defined.

The frequency table computation was implemented using the Maxeler platform [11]. The platform includes an FPGA board, drivers, and API. The DFE is coded in Java, using the API which provides objects and methods that are translated to hardware units by the compiler. The compiler translates the Java code into VHDL, and generates software interface for CPU code [12]. The DFE was implemented on Maxeler Vectis-Lite board, which contains a single FPGA, and on-board SDRAM. The board is connected to the host PC workstation via PCIe bus. Basic information on the board is shown in Table I.

A. Kernel architecture

Architecture of the *ComputeFreq-MS* kernel is shown in Fig. 1. The kernel has two scalar inputs for parameters: one for number of items to process (*items*), and one for stream length in items (*strmLen*). Stream length sets the number of items to read from the on-board SDRAM, which requires all accesses to be made in 96 byte blocks, i.e. 24 item blocks (4 bytes per item). The kernel processes the first *items* elements from the stream, while the rest are read from memory and ignored.

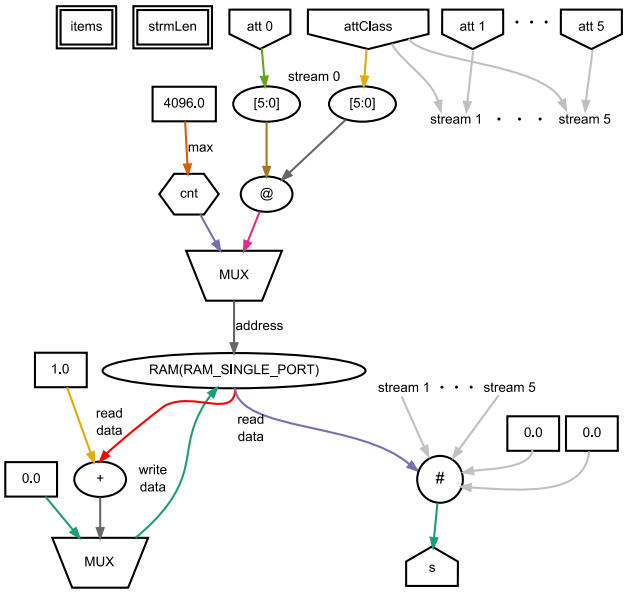


Figure 1. ComputeFreq kernel architecture for up to 64 unique attribute and class values (N_A = 6, N_C = 6).

The kernel consists of six identical frequency counter structures. Each frequency counter receives two input streams – one for attribute and one for class values. Both streams carry 32 bit unsigned integers. The input streams are sliced so the low N_A bits retained from the attribute, and N_C bits are retained from the class values. The retained bits are concatenated to form the frequency table address. In the table address, the attribute value is the high, and the class value is the low part of the address word.

The kernel has a total of seven stream inputs, six for attribute (att_0 – att_5), and one for class (*class*) value streams. Each attribute value stream is connected to a single frequency counter. The class value stream fans out to all frequency counters, so that they all receive identical class value streams.

The frequency tables are stored in on-chip memory – block RAM (BRAM). Each frequency counter has one BRAM which is addressed by the address formed from the low bits of attribute and class streams. The addressed location is read, its value incremented by one, and written back to the same location. The BRAM is configured as single port RAM with read-first synchronization. Due to access latencies of the BRAM itself, and the registers inserted by the compiler, there are latencies inside the loop. To ensure correct counting, the kernel's input streams are throttled to allow input of one element each *m* clock cycles, where *m* is the internal latency of the loop.

After all elements of interest are streamed in, the contents of the BRAMs are streamed out to the host main memory through stream output *s*. The output stream *s* is made by joining output streams from all frequency counters (s_0 – s_5) into a one vector stream. The output vector is padded to 8 elements, i.e. two additional dummy elements are added to it. The padding enables efficient conversion of the output stream to word width of 128 bits which is used in communication over PCIe bus.

As BRAMs are read, they are at the same time reset to zero. During the read phase, the BRAMs are addressed by a counter, and their write ports are switched from increment-by-one values to constant zero. In this case there are no loop dependencies, and the output stream is not throttled – it outputs one element every clock cycle. At the same time, any remaining elements in the input stream are read un-throttled (one per clock) and ignored.

B. Manager

A single *ComputeFreq-MS* kernel is instantiated in manager. Input streams *attClass* and att_0 – att_5, are linked to on-board SDRAM. All input streams use linear memory access pattern. Output stream *s* is linked to

TABLE II. FPGA RESOURCE USAGE BY THE DFE

Resource	Used	Total available	Utilization
LUTs	34,724	297,600	11.67 %
Flip-flops	50,064	297,600	16.82 %
DSP blocks	0	2,016	0.00 %
Block RAM	226	2,128	10.62 %

computer main memory. Source addresses for input streams att_0 – att_5, *attClass*, values for scalar inputs *items* and *strmLen*, and destination address for output stream *s* are defined through the generated DFE interface.

The kernel parameters were set for up to 64 unique attribute and class values (N_A = 64, N_C = 64). Each frequency table holds 4096 32-bit words. Due to the frequency counter's loop latency, it can process one item (attribute - class pair) every five clock cycles. Kernel clock frequency is set to 300 MHz. With six attribute streams, the DFE can process up to 360×10^6 elements per second. FPGA resource usage for the DFE is given in Table II.

IV. Experimental Results

A. Test environment

The kernel was benchmarked using code form C4.5 Release 8 decision tree learning program [13]. Parts needed to load the datasets, and computing the frequency tables were extracted from the program and used for benchmarking. The frequency table computation was parallelized using OpenMP [14] by distributing attributes of the dataset over the threads. This included replicating the frequency table data structure to accommodate multi-threaded execution. The original *ComputeFrequencies* function was modified to use the replicated data structure. Since the attributes were distributed over threads, multiple invocations of *ComputeFrequencies* function were necessary if number of attributes exceeded the number of threads.

Functions for transforming the dataset to the appropriate format, and loading it to the DFE were added to the benchmark program. For execution on DFE, a new function was created by modifying the original *ComputeFrequencies* function. Modifications involve removing the computation loop and replacing it with function calls to the DFE. Received results are transformed into the frequency table data structure used by the parallelized software implementation. Since the DFE processes six attributes in parallel, the *ComputeFrequencies* function is invoked once for every group of six attributes.

Time measurement was added to the *ComputeFrequencies* function. Entire run time of the function is measured. The function's throughput is calculated from the measured times using the formula:

$$T_{a,i} = \frac{a\,i}{t_{a,i}} \tag{1}$$

where a is number of attributes, i is number of items, $t_{a,i}$ is function execution time, and $T_{a,i}$ is the calculated function throughput for the dataset with a attributes, and i items.

A set of 102 synthetic datasets were used in the benchmark. Number of attributes ranged from 6 to 1,536, in a geometric sequence 6×2^i, with i = 0, 2... 7. Number of items ranged from 2048 to 4×2^{20}, in a geometric sequence 2^i, with i = 11, 12... 22. The dataset size was limited to 3 GiB, i.e. to 768×2^{20} elements. Datasets were generated randomly, with uniformly distributed values in range of 1-63 for attribute values, and 0-64 for class values.

Benchmark program was compiled using gcc compiler, version 4.4.7. Benchmarks were run on Intel Xeon E5-1600 workstation, with 16 GiB DDR3-1600 RAM, under Centos 6.5 Linux OS.

B. CPU benchmark results

For the baseline CPU performance, the measurements were conducted on a six-threaded parallelized implementation of *ComputeFrequencies* function. Measured execution times are given in Table III, and are shown in Fig. 2. For datasets with over 128×2^{10} items, execution time scales approximately linearly with number of items. There is a larger increase in execution in area of 16×2^{10} – 32×2^{10} items, for datasets with 96 or more attributes, and 32×2^{10} – 256×2^{10} for datasets with less than 96 attributes.

This increase in execution is more obviously visible in function throughput, shown in Fig. 3, as a sharp drop of the throughput. From the figure it's visible that for datasets with fewer items, the throughput increases with increasing number of items. Once the dataset exceeds a certain number of items (dependent on number of attributes), throughput drops, and quickly stabilizes to an approximately constant value. As number of attributes increases, function throughput decreases, assuming equal

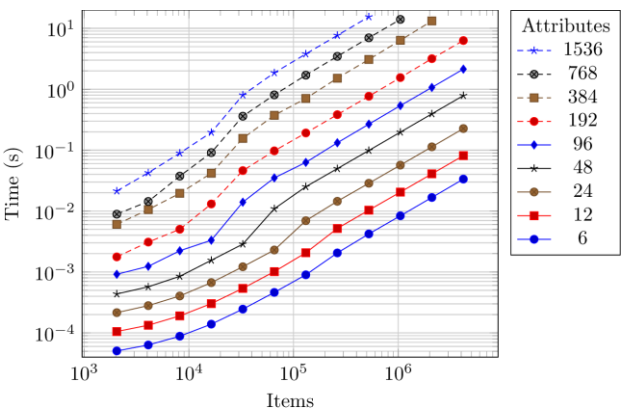


Figure 2. Execution time on CPU as function of number of items, measured for datasets with 6 to 1,536 attributes

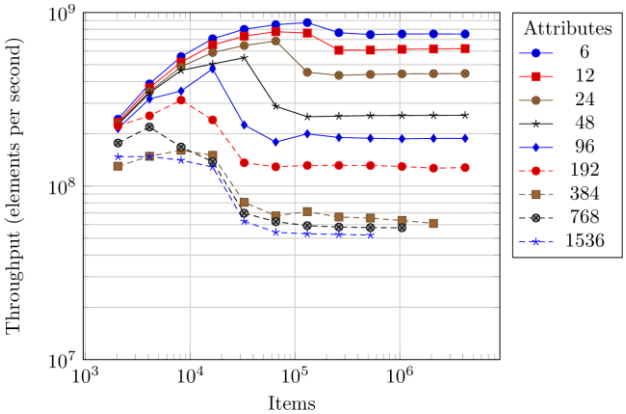


Figure 3. Throughput on CPU as function of number of items, measured for datasets with 6 to 1,536 attributes

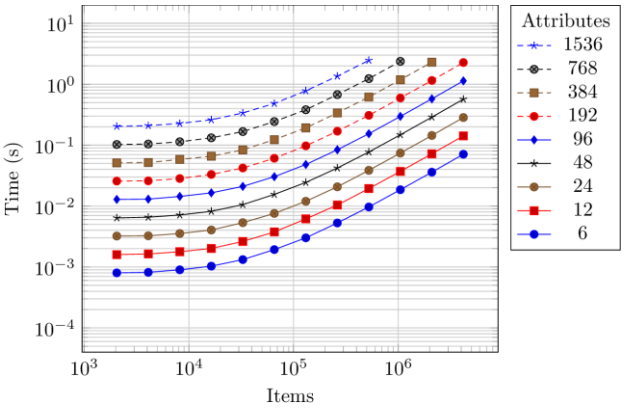


Figure 5. Execution time on DFE as function of number of items, measured for datasets with 6 to 1,536 attributes

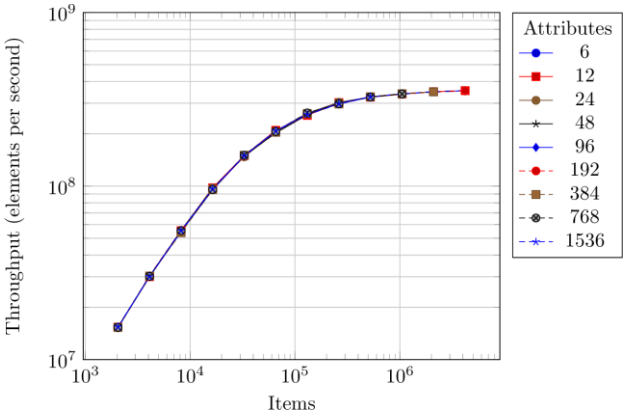


Figure 4. Throughput on DFE as function of number of items, measured for datasets with 6 to 1,536 attributes

number of items. The number of items at which peak throughput is recorded, also decreases with increasing number of attributes, as well as the peak throughput value itself. For datasets with 384 or more attributes, there is no increase of throughput. Throughput just drops from one approximately constant value to another.

Maximum measured constant (invariant to number of items) throughput is 617×10^6 elements/s, achieved on dataset with 12 attributes. Minimal measured constant throughput is 52.2×10^6 elements/s, measured on dataset with 1,536 attributes.

The drop in function throughput is most likely a consequence of interaction of CPUs cache system and main SDRAM. Smaller datasets have a smaller likelihood of cache miss, which translates into higher throughput for smaller datasets. Another important factor is the data structure for storing ht dataset. The dataset is stored in attribute major order, which leads to memory accesses with stride. With more attributes, the stride is larger, and consequentially, so is the likelihood of cache miss.

C. *DFE benchmark results*

DFE was benchmarked under the same condition as the six-threaded software implementation. Execution times measured on DFE are given in Table IV, and shown in Fig 4. Execution time on DFE scales approximately linearly with number of items in the dataset, on datasets with 512×2^{10} or more items. On datasets with fewer items,

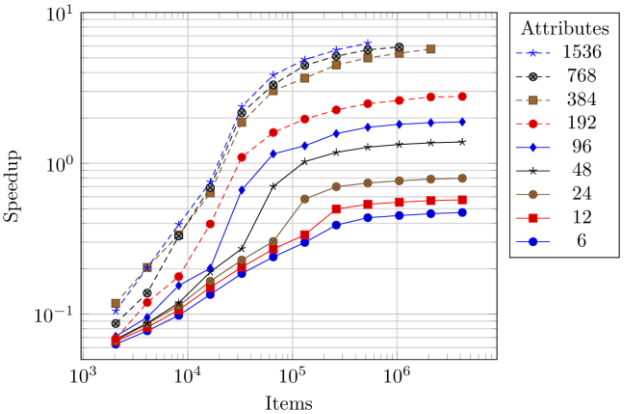


Figure 6. Speedup, measured for datasets with 6 to 1,536 attributes

the execution times asymptotically approach a certain minimal value, dependent on number of attributes. For a fixed number of items, the execution times scale linearly with number of attributes.

The kernel throughput graph, shown in Fig 5, shows that curves for all number of attributes form a tight group. The curves for different number of attributed cannot be distinguished one from another. On datasets with 512×2^{10} items, the throughput is approximately constant at 340×10^6 elements/s. The measured throughput is close to the theoretical maximum of 360×10^6 elements/s.

The DFE performs better with datasets with larger number of items, close to theoretical maximum when number of items is 512×2^{10} or more. This is a consequence of communication and control overheads between the DFE and the host computer. The overhead is independent from number of items, and its influence diminishes as the number of items increases. As can be seen from Table IV, execution times are approximately constant for datasets with 8,192 or less items. The minimum DFE execution time was calculated from these measurements, amounting to 801.1 µs.

D. *Comparison of the results*

DFE and CPU results were compared by computing speedup, i.e. ratio of execution times on CPU and on DFE. The speedup is shown in Fig 6. For most of the test dataset sizes, execution on DFE is slower than on CPU. For datasets with fewer than 48 attributes, the DFE is slower no matter what the dataset size is. For 48 attributes the speedup exceeds unity value for numbers of items 128×2^{10} or more. The maximum speedup is achieved on datasets with 384 and more attributes. Speedups on these datasets are close in value, suggesting that by further increasing number of attributes will result in negligible increase in speedup. These values can then be used as approximation of the upper bound on speedup. The maximum speedup measured is 6.26×, achieved on dataset with 1,546 attributes and 512×2^{10} items.

The total performance of the DFE was estimated by calculating the weighted average of speedups, weighted by number of elements in the dataset:

TABLE III. MEASURED EXECUTION TIMES ON CPU

Number of items	Number of attributes / Execution time								
	6	**12**	**24**	**48**	**96**	**192**	**384**	**768**	**1,536**
2,048	50.59 µs	105.3 µs	215.0 µs	434.9 µs	916.9 µs	1.769 ms	6.049 ms	8.887 ms	21.32 ms
4,096	63.31 µs	133.5 µs	280.4 µs	571.0 µs	1.237 ms	3.096 ms	10.59 ms	14.37 ms	42.51 ms
8,192	88.21 µs	190.1 µs	403.6 µs	847.8 µs	2.228 ms	5.027 ms	19.56 ms	37.46 ms	89.34 ms
16,384	139.2 µs	302.8 µs	668.8 µs	1.557 ms	3.321 ms	13.08 ms	41.85 ms	91.15 ms	195.6 ms
32,768	245.6 µs	538.0 µs	1.219 ms	2.867 ms	13.99 ms	46.18 ms	156.3 ms	360.6 ms	803.1 ms
65,536	462.1 µs	1.014 ms	2.296 ms	10.92 ms	35.06 ms	97.45 ms	373.2 ms	808.1 ms	1.860 s
131,072	899.0 µs	2.066 ms	6.949 ms	25.08 ms	62.94 ms	191.5 ms	706.2 ms	1.702 s	3.795 s
262,144	2.056 ms	5.184 ms	14.48 ms	49.73 ms	132.1 ms	382.5 ms	1.518 s	3.467 s	7.658 s
524,288	4.215 ms	10.35 ms	28.61 ms	98.80 ms	267.1 ms	766.7 ms	3.079 s	6.987 s	15.44 s
1,048,576	8.365 ms	20.48 ms	56.80 ms	197.3 ms	537.8 ms	1.553 s	6.359 s	14.01 s	–
2,097,152	16.70 ms	40.80 ms	113.3 ms	394.1 ms	1.071 s	3.176 s	13.21 s	–	–
4,194,304	33.49 ms	81.27 ms	226.3 ms	787.0 ms	2.139 s	6.303 s	–	–	–

$$S = \frac{\sum_a \sum_n a\, n\, s_{a,n}}{\sum_a \sum_n a\, n} \quad (1)$$

where S is the average speedup, and $s_{a,n}$ is the speedup on dataset with a attributes and n items. The average speedup is 4.1.

For further comparison, execution efficiency was computed for CPU and DFE. The efficiency is defined as:

$$E_{a,n} = \frac{m\, f}{F_{a,n}} \quad (1)$$

where $E_{a,n}$ is efficiency and $F_{a,n}$ is throughput for dataset with a attributes and n items, f is clock frequency, and m is number of threads or number of streams on CPU and DFE respectively. Peak throughput values were used to calculate efficiency.

The software implementation (CPU baseline) was executed on six threads, on CPU clocked at 3.2 GHz. Peak throughput was 875×10^6 elements/s. The DFE uses six streams, and was clocked at 300 MHz. Peak throughput on DFE was 355×10^6 elements/s. Execution efficiency on CPU is 21.9 clock cycles per element, and on DFE it's 5.08 clock cycles per element. Of the 5.08 clocks, 4 are consequence of the frequency counter's internal loop latency.

V. CONLUSION

In this paper we presented a multi-streamed compute architecture for 2D frequency matrix computation, implemented on FPGA platform. This multi-streamed architecture is an advancement on our previous work [10].

Benchmark results reveal that CPU outperforms DFE for smaller datasets. This is a consequence of control and communication latencies between the FPGA board and the host computer. Minimal time required to execute any action on DFE is approximately 800 µs, regardless of dataset size. Small datasets are easily processed by the CPU in less time. The DFE outperforms CPU for larger datasets, that have at least 48 attributes, 32×2^{10} items, and 6×2^{20} elements. Best speedup achieved by DFE is 6.26×. The DFE is more efficient in processing data, requiring only 5.08 clock cycles per dataset element, while CPU requires 21.9, when performing at peak efficiency. Of the DFE's 5.08 clock cycles, 4 cycles are consequence of frequency counter's internal loop latencies.

The kernel can be further improved by vectorizing inputs stream in the same manner its output was vectorized. This will allow efficient processing of more

TABLE IV. MEASURED EXECUTION TIMES ON DFE

Number of items	Number of attributes / Execution time								
	6	**12**	**24**	**48**	**96**	**192**	**384**	**768**	**1,536**
2,048	800.6 µs	1.597 ms	3.220 ms	6.385 ms	12.86 ms	25.69 ms	51.29 ms	102.5 ms	203.4 ms
4,096	817.1 µs	1.634 ms	3.248 ms	6.559 ms	13.00 ms	25.90 ms	51.89 ms	104.2 ms	209.2 ms
8,192	898.9 µs	1.780 ms	3.544 ms	7.133 ms	14.39 ms	28.33 ms	58.38 ms	113.6 ms	227.3 ms
16,384	1.031 ms	2.014 ms	4.043 ms	8.187 ms	16.47 ms	33.04 ms	65.85 ms	131.9 ms	260.4 ms
32,768	1.322 ms	2.626 ms	5.338 ms	10.55 ms	21.03 ms	42.13 ms	83.68 ms	167.0 ms	337.5 ms
65,536	1.929 ms	3.758 ms	7.576 ms	15.54 ms	30.35 ms	60.91 ms	123.2 ms	243.9 ms	483.4 ms
131,072	3.009 ms	6.153 ms	11.98 ms	24.43 ms	48.17 ms	97.42 ms	192.3 ms	381.3 ms	779.6 ms
262,144	5.277 ms	10.41 ms	20.70 ms	42.10 ms	83.98 ms	169.3 ms	338.4 ms	675.5 ms	1.356 s
524,288	9.664 ms	19.35 ms	38.62 ms	77.18 ms	154.0 ms	308.5 ms	616.5 ms	1.236 s	2.467 s
1,048,576	18.55 ms	37.08 ms	74.19 ms	147.9 ms	296.4 ms	594.5 ms	1.184 s	2.374 s	–
2,097,152	36.03 ms	72.09 ms	144.2 ms	288.3 ms	576.1 ms	1.154 s	2.307 s	–	–
4,194,304	70.96 ms	142.0 ms	284.1 ms	567.9 ms	1.136 s	2.272 s	–	–	–

elements in parallel, and better utilization of available memory bandwidth. Another improvement would be reducing or eliminating the internal loop latencies. This can potentially quadruple its performance assuming that operating frequency of the kernel does not drop due to higher design complexity.

In future work, in addition to the stated improvements, this kernel will be integrated in the C4.5 program as accelerator unit. However, to achieve significant performance gains the communication latency between host computer and FPGA will have to be compensated for. This will most likely involve adding support for batch processing of a series of small dataset, and require some modifications of the original algorithm. Overall, with additional improvements the kernel is expected to outperform the CPU for a wider range of dataset sizes.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7), under grant agreement no. 317882 – E2LP Embedded Computer Engineering Learning Platform.

REFERENCES

[1] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theroy and Applications*. River Edge, NJ, USA: World Scientific Publishing, 2008.

[2] A. N. Choudhary, D. Honbo, P. Kumar, B. Ozisikyilmaz, S. Misra, and G. Memik, "Accelerating data mining workloads: current approaches and future challenges in system architecture design," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 41–54, Jan. 2011.

[3] I. Kuon, R. Tessier, and J. Rose, "FPGA Architecture," *Found. Trends Electron. Des. Autom.*, vol. 2, no. 2, pp. 153–253, 2008.

[4] Mark L. Chang, "Device Architecture," in *Reconfigurable Computing: The Theory and Practice of FPGA-based Computation*, S. Hauck and A. DeHon, Eds. Morgan Kaufmann, 2008, pp. 3–27.

[5] J. B. Dennis, "Data Flow Supercomputers," *Computer (Long. Beach. Calif).*, vol. 13, no. 11, pp. 48–56, Nov. 1980.

[6] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993, p. 302.

[7] S. Ruggieri, "Efficient C4.5," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 2, pp. 438–444, 2002.

[8] P. Škoda, B. Medved Rogina, and V. Sruk, "FPGA implementations of data mining algorithms," in *MIPRO, 2012 Proceedings of the 35th International Convention*, 2012, pp. 362–367.

[9] G. Chrysos, P. Dagritzikos, I. Papaefstathiou, and A. Dollas, "HC-CART: A parallel system implementation of data mining classification and regression tree (CART) algorithm on a multi-FPGA system," *ACM Trans. Archit. Code Optim.*, vol. 9, no. 4, pp. 1–25, Jan. 2013.

[10] P. Škoda, V. Sruk, and B. Medved Rogina, "Frequency Table Computation on Dataflow Architecture," in *MIPRO, 2014 Proceedings of the 37th International Convention*, 2014, pp. 357–361.

[11] "Maxeler Technologies," 2014. [Online]. Available: http://www.maxeler.com/. [Accessed: 10-Feb-2014].

[12] O. Pell and V. Averbukh, "Maximum Performance Computing with Dataflow Engines," *Comput. Sci. Eng.*, vol. 14, no. 4, pp. 98–103, Jul. 2012.

[13] J. R. Quinlan, "C4.5 Release 8," 1993. [Online]. Available: http://rulequest.com/Personal/c4.5r8.tar.gz. [Accessed: 23-Jun-2013].

[14] OpenMP Architecture Review Board, "OpenMP Application Program Interface," 2011. [Online]. Available: http://www.openmp.org/mp-documents/OpenMP3.1.pdf. [Accessed: 05-Feb-2014].

Porting an N-tier Application on Cloud using P-TOSCA: A Case Study

Magdalena Kostoska, Aleksandar Donevski, Marjan Gusev, and Sasko Ristov
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Skopje, Macedonia
magdalena.kostoska@finki.ukim.mk, donevski.aleksandar@students.finki.ukim.mk,
{marjan.gushev, sashko.ristov}@finki.ukim.mk

***Abstract*—Although many companies and universities migrated their services, applications and data on the cloud, still significant part did not move yet, due to variety of reasons. Migration to other platform can cause the client many problems, and several directions have emerged to automate it as much as possible.**

P-TOSCA is a recently proposed Platform as a Service (PaaS) extension for TOSCA for automated application portability. In this paper, we demonstrate the migration of an application with most common N-tier architecture from on-premise to cloud, as well as from one cloud to another, by using P-TOSCA portability model, the development of application topology and execution plan. Although the presented demo was tested to migrate and transfer an N-tier application to Eucalyptus and OpenStack open source clouds, this model and procedure can be used for automated migration and transfer among any open standard that supports PaaS and P-TOSCA application specification.

***Index Terms*—*Automated management*; *Cloud*; *Portability*.**

I. Introduction

Several years after Amazon promoted the first cloud, IT managers had a dilemma whether to migrate their IT systems to a cloud. It was a new emerging concept with many challenges, such as security, trustworthiness, how to choose an appropriate cloud service provider, vendor lock-in or migration cost [1] and problems. Some challenges are mitigated. For example, almost all most common cloud service providers implemented information security management system [2], defined strict SLAs with several "nines" for their availability, free trial accounts, thus improving the trustworthiness with their customers.

Raising new cloud service providers with new and improved cloud services, transformed the dilemma *whether to migrate* into *when to migrate* to a cloud. Unfortunately, some challenges remained, such as automated portability. This means to migrate an application to a cloud neither with changing the application, nor server configuration, nor the new operating and cloud environment [3]. The automated portability will mitigate the risk of vendor lock-in [4] for the applications that are built on open standards. This will allow a customer to easily migrate to the cloud, transfer from one cloud to another, or even migrate back to its premise, without concerning the cloud platforms, infrastructures, APIs etc.

Recently, Kostoska et al. [5] identified several weaknesses and ambiguities of TOSCA, and promoted the P-TOSCA model as an extension of the TOSCA 1.0 standard [6], which can be used for portable cloud services [7]. P-TOSCA is used to describe a PaaS application specification and for automatic porting it to cloud environment. In this paper, we demonstrate how an N-tier application analyze the PaaS applications hosting portability regardless of their deployment model, whether they are public or private, as long as they integrate the open standards or use standardized (Application Programming Interfaces) APIs.

The rest of the paper is organized as follows. Related work in the area of cloud application portability is elaborated in Section II. Section III describes the P-TOSCA and definition and procedures of porting the N-tier application. Evaluation of results and discussion are given in Section IV, while the contribution of our work and plan for our future work in Section V.

II. Related Work

In the area of cloud portability standards there are few active working areas. Two standard oriented researches represent Open Virtualization Format (OVF) and Topology and Orchestration Services for Applications (TOSCA). The Distributed Management Task Force (DMTF) suggests the OVF [8], which specifies transport mechanism for moving virtual machines from one hosted platform to another and between different hypervisors. TOSCA defines a portable and manageable specification of services and applications that can be deployed on any cloud provider.

The TOSCA standard is research interest in many directions. OpenTOSCA, approach defined by Binz et al. [9], represents execution environment that offers imperative processing of the TOSCA. Other directions include modeling tools for TOSCA like Vino4TOSCA and Winery [10], [11], as well as TOSCA topology specification [12], [13].

Other efforts in the area of standardization are delivered by IEEE Cloud Computing Initiative (CCI), where two new standards are in progress: Cloud Portability and Interoperability Profiles (CPIP) [14] and Standard for Intercloud Interoperability & Federation (SIIF) [15].

Other researches propose introduction of abstraction-driven approach. In this area Ranabahu et al. [16] use an abstraction-driven approach to specify solutions. This approach is used in two research projects using Domain Specific Language (DSL). Hill and Humphrey [17] develop a storage abstraction layer

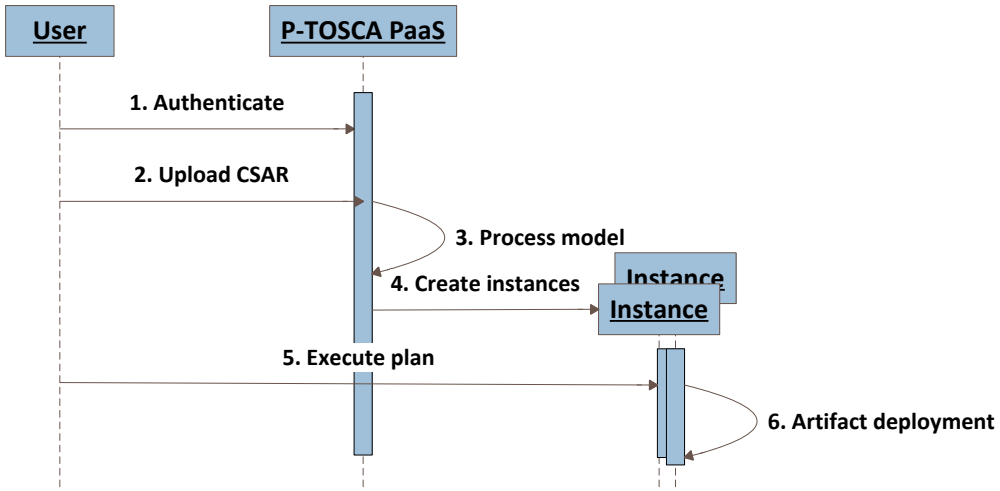


Fig. 1. Migration procedure using P-TOSCA PaaS deployment

(CSAL) that provides Blob, Table, and Queue abstractions. This approach is centered on the cloud storage and can be used in addition to another approach that solves application portability.

Usage of semantic technology is also popular approach. Loutas et al. [18] focus on this technology to enable deployment, migration and interoperability of PaaS hosted applications. Cretella and Di Martino [19] propose an approach that aims at automatic analysis of cloud vendor APIs. Petcu et al. [20] exploits the this technology to create a new set of APIs and to enable functional portability through the vendor-agnostic API and data portability through the use of drivers for the same type of data

In our previous research we demonstrated porting of a simple SOA application by using the P-TOSCA model and P-TOSCA engine [21]. In this paper, we present how to describe an N-tier application and migrate from on-premise to two different cloud environments, as well as to transfer the application from one cloud to another.

III. Cloud application portability based on P-TOSCA

This section describes the N-tier application and the P-TOSCA specification for its porting.

A. About P-TOSCA

The P-TOSCA model [5] enables:

- better description of application topology and
- definition of specific cloud properties,

in addition to the original TOSCA specification.

The application topology, which contains the types, templates and artifacts, are described by XML file. Additional files (artifacts) represent files needed for The information about the application deployment and configurations (scripts, libraries, wars or zips) are stored in other files called artifacts. The specification and the required artifacts are packed in a Cloud Service Archive (CSAR).

1) Application migration: The detailed procedure of migration is presented in Fig. 1. The first step is the client authentication, and then upload the CSAR archive in the step 2. The P-TOSCA PaaS platform processes the model and contacts the cloud controller to create instances (it is an N-tier application and need at least two instances). After creation of the instances, the user selects the plan for execution, followed by artifact deployment.

2) Application transfer from one cloud to another: This procedure differs of the migration at the first steps. After the authentication to the target P-TOSCA cloud, the user does not upload the CSAR, but he/she selects the hosted application and specifies the source and destination clouds. The CSAR is requested by the source and sent by the destination automatically between P-TOSCA. After unpacking the CSAR, the destination P-TOSCA creates instances, deploys the artifacts and acknowledges the user for finished transfer of the application.

B. N-Tier application

The application is a Java MVC student records management application. The selected application is a 2-Tier application (database and application server), but the model can be generalized for 3- or 4-Tier application. The first tier is a database tier and it is installed with MySQL DBMS to host a databases. The second tier consists of the presentation and logic tiers. It is installed with Tomcat web application server, to host the application.

C. P-TOSCA topology description for the N-Tier application

We have generated the topology description for the N-Tier application.

The generated *Node Types* (base, specific and custom) for this application are shown in Fig. 2. It is specified by three main categories:

- *Base type* - requires basic components (operating system, web server, web application, database and DBMS);

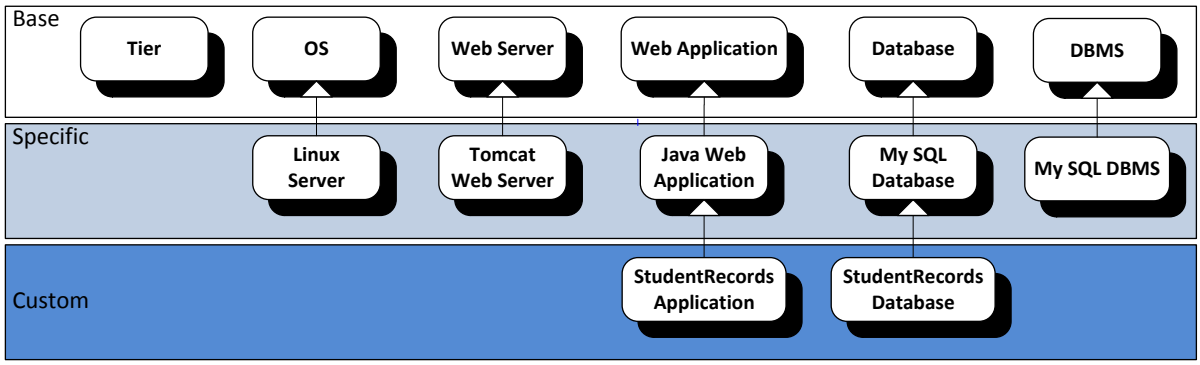


Fig. 2. *Node Types* and their hierarchy for Student Records (N-tier) application

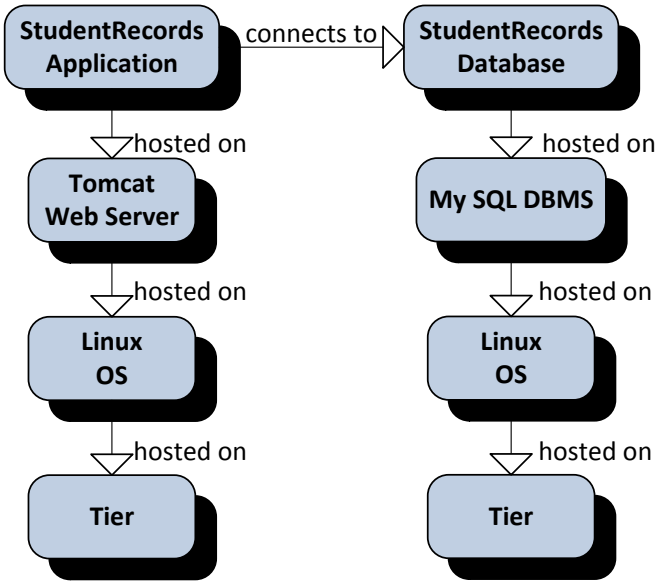


Fig. 3. Topology template for Student Records (N-tier) application

- *Specific type(s)* - requires application topology (Linux operating system), Tomcat web server, Java Web Application, MySQL database and MySQL DBMS); and
- *Custom node types* - defines specifically developed application components.

The *Topology Template* defines the topology elements (nodes) that the application is consisted of. Each node is specified by node template and is defined (i.e. type of) given node type. The *Topology Template* presented in Fig. 3, which shows the topology of the application.

Each node of the topology is specified by a node template, and the relationships between the nodes are specified using relationship templates. Same as nodes, the relationship types are initially set in the specification and each relationship template specifies the type of relationship.

The information about the infrastructure (required instance for application and security protocols and port settings) is defined in Table I. Listing 1 presents the XML Tomcat Tier node template definition.

TABLE I
TIER NODE TEMPLATE DEFINITION

Server property	value
Number of CPUs	2
Memory size	2 GB
Disk size	10 GB
Initial instances	1
Security rule protocol	TCP
Security rule ports	80, 443

TABLE II
LISTENING PORTS OF THE TOMCAT SERVER

Server property	value
HTTP port	80
HTTPS port	443

Table II defines http and https ports where the Tomcat server listens for request, while Listing 2 defines the TomcatWebServer Node template definition.

Values for artifact template for configuration of Student Records (N-tier) application are given in Listing 3. The configuration script requires input parameters for location (IP address) of the database server, as well as username and password required for the connection string to the database.

The CSAR is a zip archive that packs:

- all node and template definitions,
- Tomcat installation and deployment artifacts,
- MySQL installation and deployment artifacts,
- application deployment artifact, and
- database deployment artifacts.

IV. EVALUATION OF THE RESULTS AND DISCUSSION

This section presents the testing environment and testing data and evaluates and discusses the results of the migration from on-premise to both cloud environments and transfer from one to the other cloud environment.

A. Testing environment

The testing environment consists of two open source cloud frameworks:

```
<NodeTemplate id="VmTomcat" name="VM for Tomcat"
    type="Server">
      <Properties>
       <ServerProperties>
            <NumCpus>2</NumCpus>
            <Memory>2048</Memory>
            <Disk>10</Disk>
            <InitialNumInstances>
                  1
            </InitialNumInstances>
            <ServerSecurityProperties>
           <ServerSecurityProperty>
            <protocol>
                        TCP
                  </protocol>
            <port>
                        80
            </port>
           </ServerSecurityProperty>
           <ServerSecurityProperty>
            <protocol>
                        TCP
            </protocol>
            <port>
                        443
                  </port>
           </ServerSecurityProperty>
          </ServerSecurityProperties>
       </ServerProperties>
      </Properties>
      <Requirements>
        ...
      </Requirements>
      <Capabilities>
        ...
      </Capabilities>
</NodeTemplate>
```

Listing 1. Tomcat Tier node template definition

```
<NodeTemplate id="TomcatWebServer" name="Tomcat Web
    Server" type="TomcatWebServer">
      <Properties>
            <httpport>80</httpport>
            <httpsport>443</httpsport>
            <username>x</username>
            <password>x</password>
      </Properties>
      <Requirements>
        ...
      </Requirements>
      <Capabilities>
        ...
      </Capabilities>
</NodeTemplate>
```

Listing 2. TomcatWebServer Node template definition

- OpenStack cloud, which is installed with Ubuntu 12.04.3 LTS on one server with both cloud and node controller), and
- Eucalyptus cloud, which is installed on three servers, each with CentOS 6.5: one for the node tier - node controller, and two for cloud tiers - cloud and cluster controllers).

Fig. 4 depicts the cloud environments. More servers are used for Eucalyptus since it is more scalable than other open source clouds [22]. All physical servers of both clouds are the same, each containing 4 cores CPU and 4GB RAM. Internal networks are isolated from external. Public (float) networks of both clouds are isolated in separate VLANs [23], and since they are private IP addresses, they are NAT-ed to Internet through router.

Each experiment is conducted five times and average values are presented.

```
<ArtifactTemplate
    id="StudentRecordsAppConfigureScript"
    type="ScriptArtifact">
   <Properties>
    <ScriptArtifactProperties>
     <ScriptLanguage>sh</ScriptLanguage>
     <PrimaryScript>
                scripts/StudentRecords/configure.sh
          </PrimaryScript>
    </ScriptArtifactProperties>
   </Properties>
   <ArtifactReferences>
    <ArtifactReference
        reference="scripts/StudentRecords">
     <Include pattern="configure.sh">
                <InputParameters>
                      <InputParameter
                          nodeTemplateId="VmMySql"
                          property="ServerIPAddress"/>
                      <InputParameter
                          nodeTemplateId="VmMySql"
                          property="DBuser"/>
                      <InputParameter
                          nodeTemplateId="VmMySql"
                          property="DBpass"/>
                </InputParameters>
          </Include>
    </ArtifactReference>
   </ArtifactReferences>
</ArtifactTemplate>
```

Listing 3. Artifact template definition for Student Records application configuration

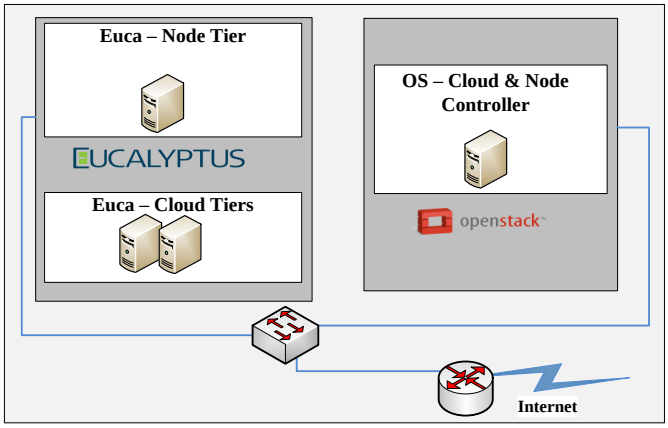


Fig. 4. Cloud environments for OpenStack and Eucalyptus

B. Evaluation and discussion

The evaluation of the migration and transfer using P-TOSCA model is presented in Table III.

First of all, all four experiments were successful, that is, migration from on-premise to OpenStack and Eucalyptus

TABLE III
TIME EVALUATION OF PORTING THE N-TIER APPLICATION

Source	Destination	Value (sec)
Eucalyptus cloud	OpenStack cloud	269
OpenStack cloud	Eucalyptus cloud	237
On-Premise	Eucalyptus cloud	1136
On-Premise	OpenStack cloud	1202

clouds, as well as transfer of the N-Tier application from OpenStack to Eucalyptus and vise-versa were successful on each cloud environment.

Migration to OpenStack takes almost 6% more time than to Eucalyptus, due to better Eucalyptus' architecture. Similar to migration, transfer of the application is also faster for 13.5% from OpenStack to Eucalyptus, rather than from Eucalyptus on OpenStack.

Finally, we observe that migration is much slower than the transfer between the clouds, which is normally since the former requires upload of CSAR from outside network.

V. CONCLUSION AND FUTURE WORK

This paper demonstrates the proof-of-concept and a procedure for porting a typical N-tier application neither with changing the application nor source nor target environment. A P-TOSCA model of a typical N-tier application is defined.

Two demo scenarios are presented: migration from on-premise to two different clouds Eucalyptus and OpenStack, and transfer the application from one cloud to another. Not only that the porting is proved, but it is conducted in a reasonable time, even with definition of the P-TOSCA specification. That is, after preparing the P-TOSCA specification of an application, it can be used for efficient porting to a cloud. This portability demonstration can be generalized if an application is portable and uses open standards.

This is only a starting point for automated porting of an application of any type. Our further research will be to define and conduct the porting of other type of applications, such as software oriented architecture, and on other open source clouds. Special focus will be on improving the performance and definition of security rules.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University", Skopje, Macedonia through the project XaaSinCloud.

The iKnow project has been developed as a part of the Tempus JPGR 511342 Project.

REFERENCES

[1] O. Rana, "The costs of cloud migration," *Cloud Computing, IEEE*, vol. 1, no. 1, pp. 62–65, May 2014.

[2] S. Ristov, M. Gusev, and M. Kostoska, "A new methodology for security evaluation in cloud computing," in *MIPRO, 2012 Proceedings of the 35th International Convention*, 2012, pp. 1484–1489.

[3] M. Hajjat, X. Sun, Y.-W. E. Sung, D. Maltz, S. Rao, K. Sripanidkulchai, and M. Tawarmalani, "Cloudward bound: Planning for beneficial migration of enterprise applications to the cloud," *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 243–254, Aug. 2010.

[4] B. Satzger, W. Hummer, C. Inzinger, P. Leitner, and S. Dustdar, "Winds of change: From vendor lock-in to the meta cloud," *Internet Computing, IEEE*, vol. 17, no. 1, pp. 69–73, Jan 2013.

[5] M. Kostoska, M. Gusev, and S. Ristov, "P-TOSCA portability model for paas hosted applications," University Ss Cyril and Methodius, FCSE, Tech. Rep. LiiT:22/2014, 2014.

[6] OASIS. (2013, Mar.) Topology and orchestration specification for cloud applications version 1.0. [Online]. Available: http://docs.oasis-open.org/tosca/TOSCA/v1.0/cs01/TOSCA-v1.0-cs01.html

[7] T. Binz, G. Breiter, F. Leyman, and T. Spatzier, "Portable cloud services using tosca," *Internet Computing, IEEE*, vol. 16, no. 3, pp. 80–85, May 2012.

[8] Distributed Management Task Force. (2010, Jan.) Open virtualization format specification version 1.1.0. [Online]. Available: http://www.dmtf.org/sites/default/files/standards/documents/DSP0243_1.1.0.pdf

[9] T. Binz, U. Breitenbücher, F. Haupt, O. Kopp, F. Leymann, A. Nowak, and S. Wagner, "OpenTOSCA – a runtime for TOSCA-based cloud applications," in *11th Int. Conf. on Service-Oriented Computing*, ser. LNCS. Springer, 2013, pp. 692–695.

[10] O. Kopp, T. Binz, U. Breitenbücher, and F. Leymann, "Winery – modeling tool for TOSCA-based cloud applications," in *11th International Conference on Service-Oriented Computing*, ser. LNCS. Springer, 2013, pp. 700–704.

[11] U. Breitenbücher, T. Binz, O. Kopp, F. Leymann, and D. Schumm, "Vino4TOSCA: A visual notation for application topologies based on TOSCA," in *OTM 2012, Part I*, ser. LNCS, vol. 7565. Springer-Verlag, 2012, pp. 416–424. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33606-5_25

[12] F. Li, M. Vogler, M. Claessens, and S. Dustdar, "Towards automated IoT application deployment by a cloud-based approach," in *Service-Oriented Computing and Applications (SOCA), 2013 IEEE 6th Int. Conf. on*, Dec 2013, pp. 61–68.

[13] M. Kostoska, I. Chorbev, and M. Gusev, "Creating portable TOSCA archive for iKnow university management system," in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2014, pp. 710–715.

[14] IEEE Cloud Computing. (2014, Jul.) Cloud portability and interoperability profiles (CPIP). [Online]. Available: http://cloudcomputing.ieee.org/standards/standards-guidance-p2301

[15] ——. (2014, Jul.) Standard for intercloud interoperability and federation (SIIF). [Online]. Available: http://cloudcomputing.ieee.org/standards/standards-guidance-p2302

[16] A. Ranabahu, E. Maximilien, A. Sheth, and K. Thirunarayan, "Application portability in cloud computing: An abstraction driven perspective," *IEEE Transactions on Services Computing*, vol. PP, no. 99, p. 1, 2013.

[17] Z. Hill and M. Humphrey, "CSAL: A cloud storage abstraction layer to enable portable cloud applications," in *IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom)*, 2010, pp. 504 – 511.

[18] N. Loutas, E. Kamateri, and K. Tarabanis, "A semantic interoperability framework for cloud platform as a service," in *IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom)*, 2011, pp. 280 – 287.

[19] G. Cretella and B. Di Martino, "Towards automatic analysis of cloud vendors APIs for supporting cloud application portability," in *Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, 2012, pp. 61 – 67.

[20] D. Petcu, G. Macariu, S. Panica, and C. Crăciun, "Portable cloud applicationsfrom theory to practice," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1417–1430, 2013.

[21] S. Ristov, M. Kostoska, and M. Gusev, "P-TOSCA portability demo case," in *2014 IEEE 3rd International Conference on Cloud Networking (IEEE CloudNet'14)*, Luxembourg, Oct. 2014, pp. 269 – 271.

[22] D. C. Marinescu, *Cloud Computing Theory and Practice*. Elsevier, 2013.

[23] A. Donevski, S. Ristov, and M. Gusev, "Security assessment of virtual machines in open source clouds," in *MIPRO, 2013 Proceedings of the 35th International Convention, IEEE Conference Publications*, Opatija, Croatia, 2013, pp. 1375–1380.

Cloud Portability Standardization Overview

Elena Markoska, Ivan Chorbev, Sasko Ristov, Marjan Gusev
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Skopje, Macedonia
Email: elena.markoska@live.com, {ivan.chorbev, sashko.ristov, marjan.gushev}@finki.ukim.mk

***Abstract*—With the development of various cloud technologies and their increased usage, the newly arisen challenge is the migration of a cloud-based application and data among various cloud environments and vendors. In the absence of a standardised mean by which full portability can be achieved, various tools exist in the form of APIs, standards, and protocols. In this paper we conduct an overview of the state of the art regarding the topic of migration and portability of cloud-based applications, in order to determine the possible challenges that today's technologies might have, as well as to find the most appropriate automated way to port an application from one to another cloud, or to migrate an on-premise application to the cloud.**

***Keywords*-Cloud Computing, services, standards.**

I. Introduction

The increase of usage of cloud computing technologies also introduces the existence of various clouds framework softwares that operate separately. Portability implies the ability to use software components across multiple different hardware and software platforms and environments. In this paper, portability between clouds, is defined as the ability to smoothly and fluently transfer a software component, application, service or data into another cloud environment, ensuring undisturbed continuity and elasticity.

As the current trend implies the usage of cloud services, the cloud consumers should be offered the possibility to use services across different clouds, whether it is to simply migrate to another cloud or seek an on-demand scaling of resources by expanding their system onto external resources, outside the native cloud where their application or service resides [1].

Various different approaches have been taken to explore the possibilities of theoretically and practically achieving some level of portability of an application between clouds. These efforts range from scientific papers which offer ideas and overviews, to proof-of-concept implementations or models, to even commercial products. Most groups stress the importance of such a study, given the fact that portability is still one of the greatest and most important obstacles that hinder the increase of adoption of cloud technologies [2].

The aim of this paper is to produce an analysis of the state-of-the-art efforts and their approaches in cloud portability, including their structure and a study of their respective APIs. Furthermore, the paper presents a study of the structure of web and cloud based applications. Lastly, conclusions are drawn regarding the current efforts which identify the weak points in the approaches attempted today.

This paper is separated into several sections which explain the following. Section II elaborates the need for cloud portability and clearly outlines the difference between portability and interoperability. In addition, it outlines the architecture of web applications, as the most common type of application in the cloud computing service market, and compares it to the architecture of the cloud applications with the intention to understand the transition process during the migration process. Section III introduces an overview of the state-of-the-art efforts that attempt to achieve cloud portability, elaborating upon TOSCA, mOSAIC, AWS CloudFormation as well as Heat. Lastly, Section IV draws the conclusions, along with our plans for future work.

II. Motivation and Background

The cloud is a rapidly developing technology that bases its greatest advantages upon the ubiquitous network access to a shared and virtualised set of hardware and software resources. These resources can be numerous and of various nature, most of which including processing, network, storage, message transport and queuing, all of which can be provisioned. The existence of various resources that make up the cloud computing model leads to a different cloud structure depending on the implementation of each of the aforementioned resources. The current software stacks of most cloud services are very heterogeneous and therefore the features each of them provide offer varying levels of incompatibility between each other. This kind of a heterogeneity is a great obstacle in the face of the demands of the cloud consumer as well as the vendor, and can be both seen as an issue with portability and interoperability [3].

A. Difference between cloud portability and interoperability

Cloud portability and cloud interoperability are very frequently seen as synonyms that share the same concept, mainly because numerous solutions solve both, even though originally have been intended for one of them. Still, these two concepts have their differences which will be elaborated below.

Cloud interoperability ensures the simultaneous and mutual work of diverse cloud services, data and application. Primarily it is a concern among cloud service providers, who must offer fluent functionality of the services they offer with the external world, be it public clouds or other cloud providers. This kind of a concept implies a previously specified knowledge of both systems, including agreements upon which the systems will interoperate. While the cloud service providers work to maintain

the interoperability between the systems, the cloud customers are usually entirely unaware of these added functionalities.

On the other hand, cloud portability is a concept that also concerns both the cloud customer and provider. By definition, it deals with the minimization of the human efforts when migrating an application, data or service from one cloud to another, which in turn lessens the need for re-designing and re-writing parts of the application. This way, cloud consumers are able to not only run instances of their application on multiple clouds, but also use external resources [4].

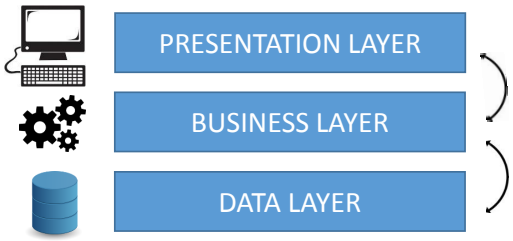


Fig. 1. Three layer architecture.

B. Motivation

1) Economic reasons: Portability within the cloud is necessary from an economic aspect for several reasons. On one hand, the customers of the cloud would gain protection over their investments in their application developments. Due to the heterogeneity of the cloud, the migration of a software application, along with its data, from one cloud to another often requires the rewriting of large parts of both the services and the application, all in order to comply with the new environment to which the application is migrating. On the other hand, the cloud service providers would be interested in enhancing the opportunity for application portability in order to promote their own attractiveness within a highly dynamic and demanding market. In addition, the need for portability also allows the development of third party organisations which would be able to mediate between the cloud customer as well as multiple cloud providers. This kind of arrangement would enable deployments depending on the customer's requirements, and all kinds of necessary adaptations would be provided by the third party.

2) Technical reasons: From a technical point of view, cloud portability is of immense importance in order to exploit cloud elasticity to a maximum extent, as well as to ensure the continuity in a given application along with its functionality of a service. Throughout a lifetime of an application, a point might come at which external resources are required from a public cloud, and so it needs to be redeployed from a private cloud. This kind of a porting process is triggered by the cloud customer. Furthermore, in order to achieve continuity, precaution measures such as regular back-ups must be made, not only of data, but of application instances as well. This implies that an application must be able to use various cloud services which in turn will aid the replication of the application as well as its data. This kind of a porting process is most commonly in accordance to the provider's agreements, and is usually considered in the design phase.

3) Legal reasons: The laws within a country may change rather abruptly, triggering the need for software to be deployed from one cloud system to another. A cloud service provider may also provide faulty service which in turn would call for immediate redeloyment of applications on another cloud platform. In this sense, efficient portability mechanisms are essential for fast recovery. Situations like the aforementioned one are also attempted to be avoided by creating contracts which guarantee a sufficient contingency plans if the cloud provider fails. These can only be provided if portability between various cloud services exists. In addition, contracts between parties may also demand the usage of certain metrics which measure the quality of the cloud's performance, including its portability, which in turn reveals the demand for a metric system that can assess the level of portability within a cloud [5].

C. Components of cloud applications

While the applications in the cloud computing service market range in nature, the most successful ones are the web applications. Understanding the differences between web applications and Cloud applications is essential for the development of a solid transition process from web applications to cloud applications.

The structure of a traditional web 2.0 application heavily relies upon a three layer architecture which allows numerous benefits, as well as drawbacks in some instances. For example, the separation of the layers into presentation, logic and data outlines the application into parts which interact with each other, but are essentially disjunct Fig. 1 [6]. If a certain application is made up of several components, then this kind of a separation in layers greatly aids the connectivity between the components, allowing layers to interact with their respective layer. On the other hand, this architecture runs into a problem when the need for scalability and elasticity escalates. The design patterns that are used in its development are most commonly adapted for a limited number of users. The data in these applications is also often used in databases which have difficulties scaling or replacing in the events of failure, and any change in their schema requires the disruption of the continuity of the application life.

A cloud application is expected to have an unpredictable load and a much larger user base that would exploit the elasticity of a system. The transformation of a web application into a cloud application requires a general architecture of a cloud application. Petcu et al. [7] propose an architecture form presented in Fig. 2. There are numerous commercially available APIs that aid the developers with the building of applications. Some of which attempt to hold to the outline of the aforementioned multi-layer cloud application architecture. TOSCA and mOSAIC specifically offer APIs which facilitate the development of portable and interoperable cloud applications.

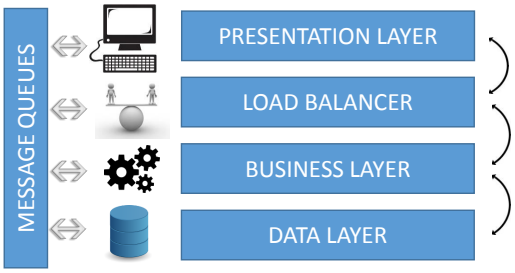


Fig. 2. Cloud application architecture.

III. Overview of current standards

There have been numerous efforts at researching as well as implementing some of the standards that are already out on the market, most of which free to use and require no royalty compensation if applied. Most of these are independent projects developed by either organisations or companies and have not yet been widely adopted [8] [9] [10].

In this section several of these solutions are elaborated, describing their structure and methodology, as well as the basic outline of their APIs and how they can be used. Some of them are presented in greater detail with the aim to portray the contact points with the developer, explaining in detail how their architecture aids the portability and interoperability of cloud applications.

A. CAMP

One of the very few steps that need to be taken towards an enhanced interoperability is the unification of management interfaces across several cloud infrastructures. OASIS proposes the Cloud Application Management for Platforms (CAMP) standard, which is a specification designed to ease the management of applications across PaaS. CAMP specifies a RESTful self-service application and a platform management API, which is neutral in several aspects, including the language, framework as well as platform. Its expectations primarily lie in the creation of an ecosystem filled with common tools, plugins, libraries and framework, which then offer vendors increased performance when it comes to interoperability between clouds [11].

The aforementioned CAMP API is made up of a resource model and a protocol that is used to manipulate the resources remotely. The resources contained in the resource model are generally separated into a Platform, Assemblies, Platform Components and Application Components. Excluding the Platform, all of these resources can exist in an instantiated and in a deployed form. The platform type of resource describes a PaaS as a whole and references all applications on the platform, allowing the discovery of the other existing Platform Components. These platform components can be database services, web services etc. An Application Component is an artefact, such as source code or metadata which can use other Application Components, depending on its respective Platform Component. Finally, an Assembly is a resource that represents an application by referencing the components it is composed of, allowing management at runtime [12].

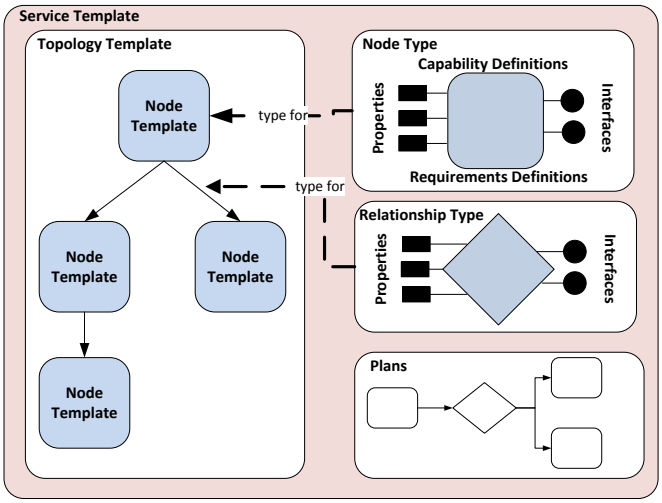


Fig. 3. Tosca service template.

B. TOSCA

TOSCA, Topology and Orchestration Specification for Cloud Applications, is created with the aim to enhance the portability specifically for cloud applications and services. TOSCA enables the interoperable description of applications and the infrastructure cloud services, the relationship between various components of the service, as well as the operational behaviour of these services. All of the aforementioned are independent of the suppler that creates the service, including any particular cloud provider or hosting technology. OASIS ensures that increasing the service and application portability within a vendor-neutral ecosystem, TOSCA enables a portable deployment to any compliant cloud, easier and more intuitive migration of existing applications to the cloud as well as offering dynamic, multi-cloud provider applications.

As opposed to CAMP's limited functionalities that deal with portability and orchestration, TOSCA is superior as it proposes a more comprehensive approach. This industrially-endorsed specification allows the software components of an application along with their respective relationship and management procedures to orchestrate basic operational behavior (deployment, patching, shut down), using descriptions based on the specified XML-based language which ensures a high level of universality and interoperability [13].

TOSCA introduces Service templates which are a concept used to specify the topology and orchestration of IT services. This mainly refers to the structure and invocation of management behavior within the service. The general structure of a Service template can be seen in Fig. 3, encapsulating a Topology Template which describes the structure of an application. In addition, this Topology template is a directed graph consisted of a set of Node Templates and Relationship Templates.

A node in this graph is represented by a Node Template which specifies the occurrence of a Node Type as a component of a service. A Node Type defines the properties of these components and the operations available that have the opportunity to manipulate the component.

A Relationship Template, on the other hand, specifies the occurrence of a relationship between the nodes in the Topology template. Each Relationship Template contains a Relationship Type which defines the nature of the relationship as well as any properties that belong to it. These are most commonly defined separately so they may later be reused [14].

The first available open source tool was first published in 2013 through the OpenTOSCA initiative. A modelling tool called Winery [15] was published, which allows the definition of a cloud application topology using a web GUI. An initial version of the TOSCA container infrastructure was also released. Kostoska et al. [16] pointed the benefits and drawbacks of TOSCA specification. The P-TOSCA was introduced as an extension of TOSCA to overcome several TOSCA's drawbacks. Using P-TOSCA, porting of a small SOA [17], [18] and N-tier application [19] was conducted.

Though a relatively new standard, TOSCA has been offered for implementation on a free basis, and various companies, non-profit organisation, academic institutions, individuals including governments have already attempted to implement their applications using this specification. Numerous companies which are world leaders in the IT industry, such as ASG, Fujitsu, HP, IBM, SAP and others, congratulate this specification as an advancing of higher levels of interoperable and portable cloud solutions [20].

C. mOSAIC

mOSAIC is an open source API and platform that ensures the portability and interoperability between multiple clouds. While the API is the key element that ensure smooth portability, the platform offers numerous features, which aid the portability of an application.

In short, mOSAIC's intention is to aid the cloud application developer build an application that is easily portable between different clouds, as well as offer guidance as to which are the proper resources required to run an application. Moreover, mOSAIC also offers a software platform including a set of application tools, a semantic engine, a cloud agency as well as a service discoverer. The main issue that mOSAIC attempts to resolve is the one of the developer of the cloud application. Instead of worrying about implementing the application for a specification of a particular cloud, the developer can only focus on the development of the application using mOSAIC's API, and later on deploy the application on any cloud the developer wishes. Therefore, the application's portability and elasticity are ensured and the developer has a choice of a programming language of implementation.

mOSAIC offers several layers of abstraction which in turn ease the development of cloud applications. Separating the abstraction into various layers, the developers are able to assemble an application based on several components with

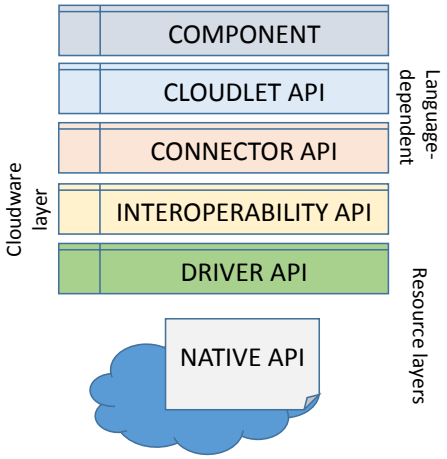


Fig. 4. mOSAIC's API layers.

defined properties and relationships, previously described by an application descriptor which is used in the lower ends of the deployment process. An Eclipse plug-in, if the developer uses Java, for instance, will define the graphical interface of the application, generating its own code which is then passed onto higher levels of development. In addition, a layer of abstraction is also designed for the owners of existing applications who intend to deploy the application onto other clouds. First parts of the application must be rewritten to comply with mOSAIC's requirements, which in turn allows for the smooth deployment onto multiple clouds. Lastly, two separate layers of abstraction deal with debugging, testing and manual control of the resources, as well as billing purposes [10].

Fig. 4 presents the layers of the mOSAIC APIs intended to support the cloud components' development and communication. Applications that use mOSAIC are made up of configurable Cloud Building Blocks which have the abilty to communicate between each other. These building blocks can be either a Cloud Resource or a Cloud Component. Cloud components are controlled by the application developer and follow the implemented behaviour which defines its functionalities.

As shown in Fig. 4, Cloudlet APIs are language dependent and provide standard programming components to the developer. The Connector API provides an abstraction for the cloud resources. Depending on the chosen programming language the developer of the application may interact with this API, or not at all. This API ensures the uniformity of the programming paradigms, assuming that the implementation of the connector API in an object oriented programming language generally have similar hierarchies and patterns. The Interoperability API as a member of the Cloudware layer aims to provide interoperability on programming language level, enforcing protocol syntax as well as semantics. Most often it enables Connectors to invoke operations that are enabled in the Driver API, without any knowledge of the technology

it supports. The Driver API, also known as a wrapper, has the role to wrap the native API and provide the initial level of uniformity. All cloud resources of the same type are processed and exported with the same interface. The Native resource API is most often a library provided by the cloud vendor and is usually language-dependent. This layer provides a very low level of uniformity.

In order to develop a fully mOSAIC compliant application, the developer is required to use the Cloudlet API which would ensure the scalability of the application, offer improved fault tolerance, and provide autonomy, in the sense that the components will be able to run in a cloud environment independent of other components. A Cloudlet Container is the host of a single Cloudlet, however it is possible to host several running instances of the same Cloudlet within the same Container. Each of these containers has a unique identifier which enables its identification during runtime. Therefore, it is not possible to distinguish between the different instances running in the same Container. Still, as a general rule, mOSAIC ensures that the functional behaviour of the Cloudlet must be independent from the number of instances or Containers [21].

D. AWS CloudFormation

AWS CloudFormation is an orchestration approach developed by Amazon, and as such it acts as a proprietary to the Amazon infrastructure. Using JSON templates, it describes a set of AWS resources, forming a stack. This stack, once created, can be managed as an entity. The AWS CloudFormation is able to automatically resolve dependencies. In addition, its primary aim is to orchestrate the creation of AWS resources which have previously been declared by the JSON templates, with the ability to connect these resources afterwards. This way, there is no need to specify management or deployment processes [22].

The AWS CloudFormation implies the work with templates and stacks. Templates are used to describe the user's AWS resources as well as their properties. This is most commonly done by saving the template with any extension compliant with the JSON standard (.json, .template, .txt), which can then be read by the CloudFormation and considered a guideline for building the specified AWS resource.

A stack, on the other hand, unifies and therefore manages related resources. The user can create, update and delete a set of resources by creating stacks and organising their resources within them. By modifying the stack or applying any changes to it, the resources that belong to the collection assigned to the respective stack are also changed [14].

E. HEAT

As one of the most popular cloud middleware software, OpenStack suggests its program, Heat, aimed to deal with orchestration. Heat is portable between OpenStack clouds, and it enables declarative infrastructure provisioning. It implements a cross-compatible AWS CloudFormation for OpenStack clouds, using multiple languages for describing templates. It provides a template based orchestration that describes cloud applications by executing appropriate OpenStack API calls which in turn generate cloud applications [23].

Like the AWS CloudFormation, Heat also introduces the usage of stacks and templates. These templates allow the creation of most OpenStack resource types, as well as some more advanced functionality such as instance autoscaling or nested stacks.

Heat generally uses two templates: HOT (Heat Orchestration Template) and CFN (short for AWS CloudFormatioN). The HOT template can only be used with OpenStack and is not backwards compatible with AWS CloudFormation templates. HOT templates are typically required to be expressed in YAML. CFN, on the other hand, are templates normally expressed using JSON [24].

F. IEEE P2301

The IEEE P2301 - Guide for Cloud Portability and Interoperabilty Profiles (CPIP) standard is a guideline written by the IEEE Computer Society. Its purpose is to assist cloud computing vendors as well as users in the processes of producing cloud systems with increased portability and interoperability based on standardized means and services. The guideline identifies various elements that Cloud Computing Systems possess and provides an enumerated group of portability, management, interoperability interfaces, file formats and conventions used for each of these elements. The guideline groups the suggested enumerations into logical entities which in turn increases the level of portability and commonality [25].

G. Unified Cloud Interface

Recognizing the high level of diversity among the numerous approached towards a standardized mean that increases Cloud portability, the Unified Cloud Interface Project (UCIP) has the goal of producing an API that uses all other API's that the Cloud systems provide as external interfaces. Since independent cloud systems provide both similarities as well as difference in the way fundamental and additional functionalities are implemented, the UCIP attempts to arive at a certain level of abstraction that's beyond the API of the Cloud API. The resolution of the heterogeneities between the cloud APIs and the existing protocols and standards is achieved through the usage of various ontologies as well as the Web Ontology Language (OWL) [26].

H. Open Cloud Computing Interface

With the establishment of the several cloud offerings, spanning the stack, ranging within the Information as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The Open Cloud Computing Interface (OCCI) comprises a large set of open specifications, offering a Protocol and an API for different types of management tasks. Initially the scope of the project were simply IaaS model based Serices, however it has developed into a rather flexible API with a focus on interoperability, innovation, integration and portability, while still holding scalability in a high regard.

Currently, all service models are covered by the API. Due to its high level of flexibility, many academic and industry giants currently support the OCCI.

IV. Conclusion and future work

Cloud portability is a well-recognised problem ever since the beginning of the cloud computing technologies. Different approaches for portability across multiple cloud platforms already exist, some of which in the form of standards, while others as platforms or APIs, proof-of-concept attempts etc.

Web applications are the most common and successful type of applications. Their structure heavily relies upon a three layer architecture that consists of a Presentation layer, a Business layer and a Data layer. The cloud application architecture assumes the existence of a Presentation Layer, a Load Balancer, a Business Layer, a Data Layer as well as Message Queues. The Load balancing layer acts as a distributor of requests through instances of the application in order to avoid overload, while the Message queues serve as communication pathways between each of the layers.

None of the existing approaches are widely adopted for several reasons. Firstly, most of them focus on a specific topic and attempt to excel achieving completeness in certain aspects of interoperability and portability. Secondly, the majority of cloud providers are reluctant to share their work as it provides a crucial advantage in their companies on the market. Lastly, despite the numerous attempts and approaches, there is no solution yet that envelops all aspects of portability and migration due to the great versatility of applications, platforms, services and languages.

Several additional approaches can be relevant for the future advancement in the field of cloud portability and interoperability. For instance, the introduction of global standards will enhance the number of implementations compliant with these standards, therefore increasing the likelihood of a unified portability solution. In addition, these standards can be made up to establish metrics, monitoring and security. TOSCA and CAMP are newly emerged standards which still have an increase of users, while ISO/IEC 27017 is a standard for the information security management for cloud systems and is currently in preparation [27]. Furthermore, the introduction of a limitation in the sphere of an architectural design for these portability solutions would help immensely with reducing the heterogeneity between each other. Limiting the architecture to several fixed levels of abstraction will provide greater room for automated and flawless portability, while reducing the room to implement exotic features which would hinder the smoothness of the solution.

References

[1] F. Gonidis, I. Paraskakis, and D. Kourtesis, "Addressing the challenge of application portability in cloud platforms," in *7th South East European Doctoral Student Conference*, 2012, pp. 565–576.

[2] D. Petcu and A. V. Vasilakos, "Portability in clouds: approaches and research opportunities," *Scalable Computing: Practice and Experience*, vol. 15, no. 3, pp. 251 – 270, 2014.

[3] P. Mell and T. Grance, "The nist definition of cloud computing sp 800-145," 2011. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[4] CSA, "Security guidance for critical areas of focus in cloud computing," 2011. [Online]. Available: https://cloudsecurityalliance.org/research/security-guidance/

[5] J. Bozman, "Cloud computing: The need for portability and interoperability," *IDC Analyze the Future, Sponsored by Red Hat, Inc*, 2010.

[6] "Three-layered services application." [Online]. Available: https://msdn.microsoft.com/en-us/library/ff648105.aspx

[7] D. Petcu, G. Macariu, S. Panica, and C. Crăciun, "Portable cloud applications – from theory to practice," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1417–1430, 2013.

[8] J. Wettinger, M. Behrendt, T. Binz, U. Breitenbücher, G. Breiter, F. Leymann, S. Moser, I. Schwertle, and T. Spatzier, "Integrating configuration management with model-driven cloud management based on TOSCA." in *CLOSER*, 2013, pp. 437–446.

[9] D. Michelino, "Implementation and testing of openStack Heat," CERN, 2013. [Online]. Available: http://zenodo.org/record/7571/files/CERN_openlab_report_Michelino.pdf

[10] F. Moscato, R. Aversa, B. Di Martino, T. Fortis, and V. Munteanu, "An analysis of mosaic ontology for cloud resources annotation," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*. IEEE, 2011, pp. 973–980.

[11] M. Carlson, M. Chapman, A. Heneveld, S. Hinkelman, D. Johnston-Watt, A. Karmarkar, T. Kunze, A. Malhotra, J. Mischkinsky, A. Otto *et al.* (2012) Cloud application management for platforms. [Online]. Available: https://www.oasis-open.org/committees/download.php/47496/CAMP-24-2012-11-15.pdf

[12] G. Katsaros, M. Menzel, A. Lenk, J. R. Revelant, R. Skipp, and J. Eberhardt, "Cloud application portability with TOSCA, Chef and OpenStack," in *Cloud Engineering (IC2E), 2014 IEEE International Conference*. IEEE, March 2014, pp. 295–302.

[13] T. Binz, G. Breiter, F. Leyman, and T. Spatzier, "Portable cloud services using TOSCA," *IEEE Internet Computing*, vol. 16, no. 3, pp. 80–85, May 2012.

[14] "TOSCA," OASIS, 2015. [Online]. Available: http://docs.oasis-open.org

[15] O. Kopp, T. Binz, U. Breitenbücher, and F. Leymann, "Winery – modeling tool for TOSCA-based cloud applications," in *11th International Conference on Service-Oriented Computing*, ser. LNCS, vol. 8274. Springer, 2013, pp. 700–704.

[16] M. Kostoska, I. Chorbev, and M. Gusev, "Creating portable TOSCA archive for iKnow university management system," in *Federated Conf. Comp. Science and Information Systems (FedCSIS)*, 2014, pp. 767–774.

[17] S. Ristov, M. Kostoska, and M. Gusev, "P-TOSCA portability demo case," in *2014 IEEE 3rd International Conference on Cloud Networking (IEEE CloudNet'14)*, Luxembourg, Oct. 2014, pp. 269 – 271.

[18] M. Gusev, M. Kostoska, S. Ristov, and A. Donevski, "A survey on approaches for interoperability and portability of cloud computing services," in *Proceedings of 5th International Conference on Cloud Computing and Service Science (CLOSER)*, 2015, in press.

[19] M. Gusev, M. Kostoska, and S. Ristov, "Cloud P-TOSCA porting of N-tier applications," in *Proceedings of the 22nd International TELFOR Forum*, 2014, pp. 935–938.

[20] "OASIS approves TOSCA standard." [Online]. Available: https://www.oasis-open.org/news/pr/oasis-approves-tosca-standard-for-deployment-and-operational-management-of-applications-acro

[21] "mOSAIC java API - programming guide," mOSAIC. [Online]. Available: http://developers.mosaic-cloud.eu/confluence/display/MOSAIC/mOSAIC+Java+API+-+ProgrammingGuide

[22] S. Mathew and J. Varia, "Overview of amazon web services," 2013, amazon Whitepapers. [Online]. Available: http://www.ajsnetworking.com/wp-content/uploads/2015/01/aws-overview.pdf

[23] "heat-specs design specifications," OpenStack. [Online]. Available: http://specs.openstack.org/openstack/heat-specs/

[24] R. Kumar and B. B. Parashar, "Dynamic resource allocation and management using openstack," *Nova*, vol. 1, p. 21, 2010.

[25] K. Stravoskoufos, A. Preventis, S. Sotiriadis, and E. G. Petrakis, "A survey on approaches for interoperability and portability of cloud computing services," in *Proceedings of 4th International Conference on Cloud Computing and Service Science (CLOSER)*, 2014, pp. 95–102.

[26] S. Kolb and G. Wirtz, "Towards application portability in platform as a service," in *Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on*. IEEE, 2014, pp. 218–229.

[27] "ISO/IEC DIS 27017." [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43757

In-Situ Anonymization of Big Data

Tomislav Križan, Marko Brakus, Davorin Vukelić
Poslovna Inteligencija d.o.o., Zagreb, Croatia
{tomislav.krizan, marko.brakus, davorin.vukelic}@inteligencija.com

Abstract - With organizations storing and even openly publishing their data for further processing, privacy becomes an issue. Such open data should retain its original structure while protecting sensitive personal data. Our aim was to develop fast and secure software for offline anonymization of (distributed) big data. Herein, we describe speed and security requirements for anonymization systems, popular techniques of anonymization and de-anonymization attacks. We give a detailed description of our software for in-situ anonymization of big data distributed in a cluster tested on a real Telco customer data record (CDR) dataset (dataset size is around 500 GB).

I. INTRODUCTION

Vast amount of data is being generated from versatile sources and organizations are primarily interested in analyzing it for further information and improved decision making. Data analysis surely leverages data value for the organization and opening data publicly benefits wider community and fosters further scientific research. However, analysis means disclosing data to other employees, third party organizations or even publishing it on the internet for everybody to use. Stored data is often secret, proprietary or protected by data protection and privacy laws and organization can suffer significant financial loss if that data is misused. Huge cost of data theft and misuse urged organizations to improve security measures and protect data.

Ideally, data would be void of any sensitive information which is protected by law while retaining all other information and structure valuable for further analysis. These sensitive parts of dataset need to be changed in a way to make it impossible or at least infeasible to trace the changed values back to the original ones. Such a transformation would ensure that confidential data, even when outsourced, published or stolen, could not be misused. *Anonymization* (de-identification, depersonalization, masking, obfuscation, etc.) is a process that changes or removes sensitive data while preserving its structure. It is considered successful if it is impossible or computationally infeasible to track original sensitive data from the anonymized dataset.

Anonymization is traditionally done on organizations' files, databases or centralized warehouses. Data is usually extracted from different sources, validated, transformed and loaded into databases. Due to nonstandard data demands, organizations employ newer *Big Data* technologies characterized by increased data volume (great data size), velocity (real-time and near real-time systems to meet high-speed data streams) and variety (lack of data structure). Such high volume of data should be processed preferably at velocity close to the one at which it was generated so that it can possibly be included in a data pipeline. Distributed environments and traditional centralized processing assume costly data movements (CPU time, network bandwidth and energy consumption) and additional vertical scaling on the place of processing. Therefore, data should preferably be processed at the same place where it was generated to avoid costly data movement. *In-situ* methods are inspired by such demands, where inter-node communication should be reduced. Thus, data movement should be minimized by doing computing locally where data was created. One of the most common applications of the in-situ processing is in visualization by doing rendering on the nodes where data was generated [1]. Also, in-situ method is appropriate for monitoring distributed environment [2] generating real time streams of data (network equipment, sensors, etc.).

In this work, we give a description of anonymization process which encompasses defining organizations' sensitive data, setting requirements for anonymization system, choosing appropriate anonymization techniques and considering their vulnerabilities. We then present our implemented solution for anonymization of Big Data inspired by in-situ data processing. It is exhibited how various aspects of Big Data are met with in-memory object store for high data volumes and fast messaging for inter-node communication. It is shown how the system is designed for anonymization of data at the nodes where data is

created by minimum client-server communication. Lastly, we present system performance measurements for anonymizing an example telco mobile customer data record (CDR) dataset in a distributed environment.

A. Types of Sensitive Information

Personal data [3] is defined as "any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity." Some examples of such data are: address, credit card number, bank statements, criminal record.

Sensitive data attributes/fields/columns/identifiers can be classified into several types regarding their sensitivity. *Direct (explicit) identifiers* contain values that can uniquely identify persons or organizations (i.e. sensitive entities). Examples are name, telephone number, email address, credit card number, various IDs, etc. *Indirect (implicit, quasi) identifiers* can identify the entity but only indirectly. Examples include geographical location, gender, marital status, profession, date of birth, etc.

B. Data Anonymization System Requirements

Data anonymization is meant to make restricted and confidential data harder to misuse and available to unauthorized people, both inside and outside of the organization.

Even (database column-level) encryption using SQL scripts is a valid approach for data anonymization. If the databases are different, inconsistency may pose a problem, i.e. the same customer would appear differently in each database. There could also be inconsistency between departments using different versions of the same technology and "referential integrity" thus cannot be maintained. Also, such scripts are not flexible and hard to maintain. Therefore, a specialized data masking tool would be more appropriate for any enterprise-wide data anonymization initiative.

A system for anonymization should adhere to several requirements, primarily security and speed.

Security should be strong enough to make any tracing of masked (anonymized) values back to original sensitive information hard enough or impossible. This can not only be achieved by strong encryption, where used keys also need to be protected properly, but also by other masking methods. Letting only authenticated users run the masking would also improve system's security.

Speed is measured in number of records masked in unit of time. This capability determine whether the system is capable of working online (real-time mode, data in motion) or only offline (batch mode, data at rest). The process should in any case be finished inside reasonable time.

Techniques. Supported algorithms should be able to anonymize/generate/format all the different value types of the dataset (credit card numbers, zip codes, telephone numbers, etc.). Also, different ways of masking should be supported, e.g. partial masking (masking only part of the value).

Connectivity. With data commonly being stored in databases, the system could offer connecting to various databases, creating queries and preserving physical structure, indices, primary keys, triggers and referential integrity across tables. For input from file, various file formats (fixed width records, delimiters and encodings) should be supported. Output can also be of different types. Integration into a larger data process such as ETL (Extract-Transform-Load) is also desirable.

Resilience. The system should skip/report invalid data such as improper or missing values in a record/row. Checkpointing (marking the last successful record) would enable restarting the masking from before the record of last failure. Also, examining system's logs should reveal the source of possible failure (improper data or system error).

C. Data Anonymization Techniques

The term anonymity has a different meaning in different countries. In some countries it corresponds to *computational anonymity* (difficulty to computationally trace original values from the masked ones) and in other countries to *perfect anonymity* (theoretically impossible to trace original values from masked ones) [4].

There are two main approaches to anonymization: one based on *randomization* and other on *generalization* [4]. Randomization comprises any modification of the original values either by removing them (nullifying), replacing them with randomly generated or chosen values, adding noise or swapping. Generalization techniques convert values into more general ones and is therefore not suitable when we want to retain data structure and information.

Techniques can be further described with several characteristics. *Deterministic* (*repeatable*) - same original value always maps into the same masked value. *Uniqueness* – two different original values cannot map into the same masked value. *Conditional sensitive* - different masking for different given conditions/parameters. *Partial* – masking only part of the value (e.g. only several last digits of the telephone number). *Realistic* – masked output looks meaningful. *Length preservation* - masked value is of the same length as the original. *Reversibility* – it is possible to get original value from the masked one (e.g. provided the encryption key). *Information loss* – how much information is lost during the masking.

Substitution technique replaces the original value with some other fake masked value. This fake value can be either randomly chosen from the given list (*pseudonymization*) of fake values or generated randomly. The original value can also be changed by *noise addition* where random noise is added to the original value. The variant when the masked value is chosen from the given

list (usually external file) is called *lookup substitution.* While it replaces original values with meaningful masked fake values its list can easily be exhausted for a large number of different original values.

Shuffling or data swapping randomly rearranges values inside one dataset column while preserving the order in other columns. To make masked data more realistic, *group shuffling* can be used to rearrange several columns together. It is safer to do shuffling on larger dataset where tracing the original values is harder.

Blurify technique changes the original value by a given variance. Usually financial data such as salaries are randomly increased or decreased for the specified variance percentage, e.g. -/+ 5% (*Numeric variance*). Minimum and maximum limits can be specified along with the variance to keep the masked value within meaningful boundaries. The technique can also be applied to temporal data such as birth dates (*Date variance*).

Nulling out simply deletes the original value. Normally, the technique cannot be used on non-nullable columns of the database.

Character masking technique is similar to nulling out and replaces the original value with a specified character constant. E.g. first several digits of a credit card number are masked with a character constant.

Encryption technique encrypts the original value into anonymized one by using the provided key. Although the confidentiality is provided, the output value is not realistic. Also, the key needs to be managed and if disclosed, data can be compromised. It is suitable when the anonymized data later needs to be de-anonymized back into original form.

Hashing is a form of encryption technique where no key is used and data of arbitrary length is irreversibly masked into fixed length message digest output. Same as with encryption, hashing is wrongly equated with anonymization, and although it may put itself as a first intuitive choice for obfuscating data, it is not advised [4].

Aggregation aggregates several clustered records and replaces their values with generalizations. Numerals can be averaged, dates can be generalized into months or years, exact salaries into ranges, etc.

K-anonymity is a requirement that pseudo-identifiers of each record must match at least *k* other records in the anonymized dataset. Optimal k-anonymization is an NP-hard problem and there are various approaches to meet this requirement. *L-diversity* and *t-closeness* models are refinements of *k-anonymity.*

D. Data De-Anonymization Risks

Anonymization techniques are as secure as difficult it is to trace the original value from the masked or infer any details about the sensitive information. There is no generally accepted mature method to evaluate this difficulty of de-anonymization [4]. Therefore, choosing the most resistant anonymization technique is nontrivial and dependent on the use case.

According to [4], there are three main approaches to de-anonymization. *Singling out* corresponds to the possibility to isolate record(s) which identify an individual. *Linking attack* tries to link records from different datasets containing common indirect identifiers.

The widespread misbelief about pseudonymization, substitution, encryption or hashing is that if the original sensitive data is replaced by random values or encrypted with state of the art encryption with long enough a key or even completely removed, it is perfectly anonymized [4]. Hashed data can be reversed back to original by brute force attacks or reversing the algorithm. While state of the art encryption does offer high degree of protection, weak custody of the key could make it possible to identify the data. Furthermore, abundant data on the internet can be used to form auxiliary datasets which are then combined with the anonymized ones to help re-identify sensitive information. De-anonymization of Netflix prize is a popular case where data, ostensibly anonymized by randomization, was revealed with the help of auxiliary dataset obtained from the internet [5]. Generally, a dataset with not only perfectly obfuscated but even removed direct identifiers remains vulnerable to singling out and linking to other datasets.

Generalization techniques, while resistant to singling out, require sophisticated tuning against linkability and inference attacks. The main flaw of the k-anonymity is the inference attack, which is improved in l-diversity and t-closeness techniques. Although these refinements are imperfect, they make inference attacks much less successful.

II. RELATED WORK

Growing volume of data and privacy concerns spawned work both in newer anonymization techniques and their merging with popular large scale distributed computing systems. These better anonymization techniques strive to be more resilient to background knowledge and linking attacks done by using increasingly available online datasets. Incorporating the anonymization and privacy preservation into distributed frameworks adds special requirements for secured network protocols, prevention of data disclosure to other nodes in the environment and restricting types of data queries.

In [6] a solution for k-anonymizing vertically distributed data is presented. Vertically distributed datasets are first anonymized locally and then joined. A special secured protocol is constructed that achieves global k-anonymization between two separate nodes. [7] describes anonymization of data horizontally distributed among several cluster nodes. It argues that anonymizing data separately on each node fails to meet referential integrity and that centralizing in on a third party secured server is infeasible. Therefore, a special distributed anonymization protocol is implemented that queries distributed data sources and guaranties that such virtual union of datasets is k-anonymous. Similarly, in [8] an

encrypted global view of the horizontally or vertically distributed dataset is constructed by means of a special secured protocol.

Cloud computing and Big Data offer organizations cost effective computing resources. Yet, organizations are hesitant or legally forbid to use them because of the possible security risks. There are several proposed solutions for privacy protection of Big Data in the cloud. [9] presents data anonymization framework which upgrades MapReduce computing with privacy protection. MapReduce query jobs are constrained by enforced value ranges, preventing unwanted disclosure of original data. The framework enables large-scale computation on data that originates from different sources and belong to different owners. It concentrates on preserving differential privacy by adding noise to the output of the computation. However, data stored in the cloud is not actually anonymized, not all query types are possible and the query results contain added noise. Other two systems described in [10] and [11] provide a privacy-aware cloud computing based on partitioning MapReduce jobs to process sensitive data on private clouds and insensitive data on public clouds. The systems transparently upgrade MapReduce framework and not much adjustments to the existing jobs are needed. It is argued that traditional techniques fail to anonymize large scale data and that it is more efficient to modify or restrict the running MapReduce jobs that access data. Thus, stored data itself remains unanoymized.

III. ANONYMIZATION SYSTEM OVERVIEW

The anonymization system described herein was developed for the purposes of European FP7 FERARI (Flexible Event pRocessing for big dAta aRchItectures) project [12]. The project's goal is to develop novel in-situ methods for detection of events in distributed environments producing Big Data. The system is originally used to anonymize real world telco Customer Data Records (CDR) dataset, which is then used in the project primarily for the research purposes of developing methods of Deep Packet Investigation (DPI) for Mobile Fraud Detection.

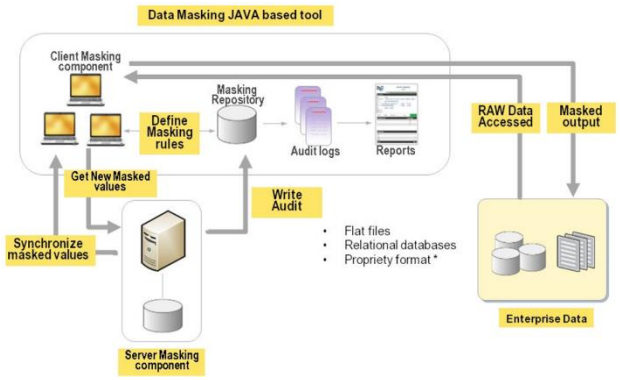


Figure 1 : Architecture of in-situ data anonymization system

The intention was to develop a generalized Java solution that would work both for data residing on a single node and for dana scattered across a distributed environment as exhibited in Figure 1. In addition to anonymization of stationary data, the system attempts to meet demands of high velocity dynamic data. Therefore, multiple inputs are supported, client server communication is minimized and data structures are adjusted for high volume data.

A. Data Connectivity

Connection to standard stationary data repositories such as databases and flat files is supported. Connection to database table is done by JDBC connection and executing the table retrieval query given in the client's configuration. Currently, two main flat file types are supported: delimited text files and fixed width text files. Description of file type and columns is given in the client's configuration. Files are read and written by buffering with *java.io.BufferedReader* and *java.io.BufferedWriter*, which showed to be the simplest and fastest way of reading and writing text files. Both flat files and table rows retrieved by the query should match the columns description specified in the client's configuration. Input of data in motion for real time (dynamic) masking is implemented by supporting JMS messaging queue through *javax.jms.QueueReceiver*. Specifically, ActiveMQ [13] high-performance asynchronous messaging queue library was used for testing. Similarly as for input, data output is provided into database table, delimited or fixed width text file or JMS message queue.

B. Anonymization Techniques

The system currently implements several anonymization techniques of randomizing type. Namely, substitution by pseudonymization, partial substitution, key generator, nullify, blurify and random.

The implemented substitution masking method replaces each value with a fake masked one. The list of possible masked values is given in a file in masker client configuration. Also, the technique can be made deterministic and unique. Similarly, partial substitution masking method replaces only part of the value with a randomly generated characters. The part to be masked is specified by starting position on the left and maximum masked length starting from the right.

The implemented key generator method generates random alphanumeric or only numeric strings of specific length. It can also be made deterministic and unique. Similarly, the random method works on the client side by replacing the value with a random string of specific length and can not be deterministic or unique.

The nullify method simply erases the value by returning an empty string.

To preserve referential integrity among different data sources possibly on different nodes, some mapping data used for anonymization has to be centralized on a server node accessible from all other client nodes. In case of substitution methods, mappings of keys into masked values are kept in a map data structure, which is accessible on the server node.

C. Mapping Data Structure

Pseudonymization masking techniques to be deterministic and preserve referential integrity store mappings of keys into masked values. We implemented hash table of key to masked value mappings for retrieving past mappings. Also, a Bloom filter was implemented for checking if some masked value already exists in the hash table. When processing higher volumes of data, the heap available to Java program can easily become a constraint. The increase of allocated heap also increases the time that the garbage collector needs to collect unused memory objects. Contrary to the popular belief, it is the number of used (live) objects that slows down garbage collection. Garbage collection can therefore greatly and unpredictably increase response time by creating long pauses. Thus, offloading data structures to off-heap native direct memory successfully scales vertically improving performance [14] and adjusting the system for higher volume and higher velocity of Big Data. Access to off-heap direct memory is provided by *java.nio.ByteBuffer* and *java.nio.MappedByteBuffer*. Although the former is faster, we employed the later to map the off-heap hash table data structure to a disk file and thus ensure persistence. Such memory mapped file is guaranteed by the OS to be stored to the actual file on disk in the case the program brakes. However, one single *java.nio.MappedByteBuffer* instance can refer to no more than approximately 2 GB of memory as it is accessed by 32-bit integer index. To overcome this limitation, several overlapping buffers are used to refer to larger memory area. Thus, the complete mapping data structure is located in an array structure inside multiple *java.nio.MappedByteBuffer* objects pointing to one large memory area.

The map data structure located on the server node is accessed from multiple client nodes (Client Masking Component and Server Masking Component in Figure 1) and is therefore made concurrent. Concurrency is implemented with multiple locks similar as in *java.util.concurrent.ConcurrentHashMap*, where each lock *guards* one part of the table. The maps used on clients do not need concurrency and are not run with locks to avoid unnecessary overhead.

D. Internode Communication

The system is divided into multiple clients and one server to support anonymization of datasets scattered or generated on different nodes across the network or cluster. The server node contains shared data needed to perform unique masking across multiple nodes. To comply with the in-situ idea, when masking a value, only if the specific masking technique does not find necessary data locally, it requests it from the server node. For example, when pseudonymizing string value by using the substitution technique, if the client node does not find the mapping in its local map, the server node is queried for that mapping. The server node looks up the key in its map and returns it if it is found, otherwise the mapping is generated. The client node stores the received mapping in its local map to avoid the future query to the server. In case no maps specified by the configuration are found locally when the node is run, the client requests the maps from the server. Map files are quickly transferred as data streams.

Finagle [15] asynchronous Remote Procedure Call (RPC) system is used for client server communication. Finagle server exposes callable RPC services *com.twitter.finagle.Service* for each masker. The client handles asynchronous requests to the server by *com.twitter.util.Future* objects, which invoke registered callbacks when the result from the server arrives. Thus, the masking blocks for rows of data, not for each field.

E. Integration into Apache Storm

We integrated anonymization system into Storm [16] - a distributed, fault-tolerant system for processing streams of data. Its topology comprises two main types of nodes – *spouts* and *bolts*, which are scattered across the network cluster and perform data reading/writing and data processing, respectively.

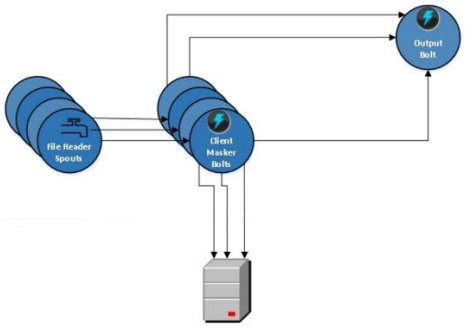


Figure 2 Masking clients inside multiple Storm bolts

As depicted in Figure 2, input into the topology is done by *spout* node type, which is subscribed to Kafka [17] Topic to get lines of text to anonymize. Kafka is a messaging queue distributed on a cluster, accompanying Storm with a messaging system. The input is then parsed in *LineSplitterBolt* node and sent down the topology into multiple *ClientDataMaskingBolt* nodes in form of tuples. Tuples contain sets of values to anonymize. Each such node functions as a client in the masking process and contacts the *ServerDataMaskingBolt* if the masking technique requires so. *ServerDataMaskingBolt node*

provides access to one or more maps, same as in the non-Storm masking cluster. Each *ClientDataMaskingBolt* is provided with the details of masking server location before the topology is submitted. The Storm topology is defined in a java program and automatically submitted to the cluster from any cluster node. Anonymized data is passed from each running *ClientDataMaskingBolt* bolt to *KafkaSinkBolt*, which writes the output.

F. Configuration

Configuration for client part of the process comprises description of input (file format, database table connection, messaging queue parameters) and specification of dataset columns to anonymize (column index, type of anonymization, masker id and masker seed). For the server part of the process it contains description of the server-side maskers (masker id, masker seed and masker specific parameters). When running the system on a single node, configurations are passed as XML files from the command line. In case of both in-situ processing or processing on Storm, client and server anonymization nodes obtain configuration from Zookeeper [18] - a centralized service for maintaining configuration information and distributed synchronization. The configuration from the XML file is stored into Zookeeper by passing path to XML configuration files to Zookeeper's store similar to file system by '*zkcli.sh –cmd upconfig –zkhost 127.0.0.1:2181 –d /path/to/config/files –n anonymization'* script. Uploaded configuration files are retrievable from the running Zookeeper service like '/configs/anonymization/client_config.xml'.

G. Dataset and Measurements

We used real telco *Call Data Record* (CDR) dataset for testing the performance of the anonymization system. CDR data is the fundamental subject of analysis such as churn prevention and user segmentation in telecoms. It is generated by telecoms' equipment to record the details of the phone call that is made through that specific device. As it is generated in different places of a distributed environment, it is suitable for in-situ data processing. Also, analysis of customer data raises concerns about privacy and demands anonymization.

The sample CDR dataset used for testing comprises approximately 465 GB of post-paid and 73 GB of pre-paid customer data records for a period of 6 months. The dataset is stored in multiple CSV text files and comprises attributes like id, call start and end time, geographic location, duration, cost, event type, tariff, etc. For post-paid data, we chose to mask fields BILLED_MSISDN, CALLING_NUMBER, CALLED_NUMBER and OTHER_PARTY_TEL_NUMBER with partial substitution masking technique using the same server-side masker. Similarly, we masked fields MSISDN_NUMBER and B_NUMBER of the pre-paid dataset with partial substitution technique. Thus, the same partial substitution masker was used to mask the chosen fields and provide consistency across two independent dataset.

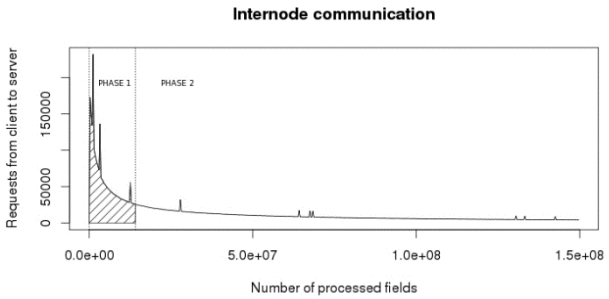


Figure 3 Frequency of masking requests *from client to server*

The effect of in-situ processing is depicted in Figure 3, where it is shown that clients' requests to server exponentialy decrease as data fields are anonymized. In *Phase 1*, phone numbers (in reality generated at different telecom base stations) are unknown and requests to the server are frequent. Later in *Phase 2*, no new phone numbers appear and there is no need for requests to the server because the client masking nodes (in reallity located on base stations) contain all the mapping data.

The spikes visible in the Figure 3 are due to the fact that the provided CDR dataset is in combined form with calls not strictly ordered by date but clustered based on their arival from appropriate base station. Also, spikes within *Phase 2* correspond with seasonality of summer holidays and additional new numbers entering into system (by roaming users).

In everyday use of distributed anonymization system, impact of *Phase 1* communication can be minimized by using pre built memory maps, or utilizing historical dataset on server side with only gathering distinct values to speed-up memory map buildup. After that, in-situ processing on all distributed nodes can be started and behaviour of whole system from that point will be only aligned to *Phase 2* of internode communication.

After anonymizing more than 3 billion CDR records, the map contained 21,308,495 distinct phone number mappings. Original dataset didn't have standardized phone numbers, thus same number could appear multiple times within masked value set depending on formatting of that specific number (e.g. with country prefix, with city prefix, shortcode, etc.). Standardization of phone numbers was out of scope for this paper.

Actual test results compared towards existing (and non-distributed) commercial Enterprise Data Masking solutions were omitted from this paper because we didn't get necessary approvals in due time.

IV. CONCLUSION

Anonymization is becoming more important with growing volume of generated data and privacy protection concerns. Newer anonymization techniques are required to resist the background knowledge and linking attacks that are possible by versatile datasets available online. Also, the techniques need to be built into processing systems capable of handling high volume and high velocity of Big Data. Our work presents a system suitable for in-situ anonymization of data in places where it was generated. It is appropriate for high volume of data with masking techniques' appropriate mapping data structures. Moreover, latencies and unexpected pauses are minimized to adhere to the requirements of high velocity real time data. Also, internode communication in a distributed environment is minimized. Beside from the in-situ processing architecture, the system is also used in Storm distributed realtime computation system. In further work, datasets from other domains will be anonymized and distributed environments of different sizes and data generation speeds will be tested.

[1] Moreland, Kenneth, et al. "Examples of in transit visualization." Proceedings of the 2nd international workshop on Petascal data analytics: challenges and opportunities. ACM, 2011.

[2] Mert Akdere, U˘ur Çetintemel, and Nesime Tatbul. Plan-based complex event detection across distributed sources. Proc. VLDB Endow., 1(1):66–77, August 2008. ISSN 2150-8097. doi: 10.14778/1453856.1453869

[3] Directive 95/46/EC of The European Parliament and of the Council, 24 October 1995

[4] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, adopted on 10 April 2014

[5] A. Narayanan,V. Shmatikov. "Robust de-anonymization of large sparse datasets." Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, 2008.

[6] JIANG, Wei; CLIFTON, Chris. A secure distributed framework for achieving k-anonymity. The VLDB Journal - The International Journal on Very Large Data Bases, 2006, 15.4: 316-333.

[7] LEFEVRE, Kristen; DEWITT, David J.; RAMAKRISHNAN, Raghu. Workload-aware anonymization techniques for large-scale datasets. ACM Transactions on Database Systems (TODS), 2008, 33.3: 17.

[8] KOHLMAYER, Florian, et al. A flexible approach to distributed data anonymization. Journal of biomedical informatics, 2014, 50: 62-76

[9] ROY, Indrajit, et al. Airavat: Security and Privacy for MapReduce. In: NSDI. 2010. p. 297-312.

[10] KO, Steven Y.; JEON, Kyungho; MORALES, Ramses. The HybrEx model for confidentiality and privacy in cloud computing. Proc. of HotCloud, 2011.

[11] ZHANG, Kehuan, et al. Sedic: privacy-aware data intensive computing on hybrid clouds. In: Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011. p. 515-526.

[12] http://www.ferari-project.eu

[13] http://activemq.apache.org/

[14] Harris, Steven T., Christopher Dennis, and Saravanan Subbiah. "Off-heap direct-memory data stores, methods of creating and/or managing off-heap direct-memory data stores, and/or systems including off-heap direct-memory data store." U.S. Patent No. 8,832,674. 9 Sep. 2014.

[15] http://twitter.github.io/finagle/guide/

[16] https://storm.apache.org/

[17] http://kafka.apache.org/

[18] http://zookeeper.apache.org/

VISUALIZATION SYSTEMS

A GPU-based Simultaneous Real-Time EEG Processing and Visualization System for Brain Imaging Applications

Z. Juhasz and Gy. Kozmann
University of Pannonia/Department of Electrical Engineering and Information Systems, Veszprem, Hungary
{juhasz,kozmann}@virt.uni-pannon.hu

Abstract – A data-driven prototype software is presented for EEG processing and visualization. The system relies on the GPU architecture for providing simultaneous processing and visualization of the EEG data. Two example brain imaging algorithms, the surface Laplacian and the spherical forward solution are used for illustrating the effective use of the massively parallel GPU hardware in speeding up computations. The paper describes the architecture of our system, the key design decisions, and the performance optimization of the parallel implementation. Using the CUDA-OpenGL interoperability, the computing subsystem can directly modify potential data in the OpenGL vertex memory, avoiding unnecessary GPU-Host data transfers. The system and our parallel implementations demonstrate that real-time processing and visualization is possible for a range of algorithms during EEG processing. We are confident that these results can pave the way for supercomputing-class implementations and open up new opportunities in the clinical practice and neuroscience research.

I. INTRODUCTION

High-resolution EEG imaging is an increasingly important tool in neuroscience [1]. Unlike CT or MRI that only provide structural information, EEG can measure functional brain activity. Its time resolution typically is in the millisecond range which is vastly superior to other functional imaging methods, such as PET or fMRI. Consequently, EEG is indispensable in areas where high temporal resolution brain activity mapping is required, e.g. in epilepsy diagnosis in the clinical setting [2], [3] or in cognitive experiments based on evoked or event-related potentials [4].

The neuroscience community relies on commercial as well as open source software products for EEG processing. The commercial systems are more common in the clinical area, whereas open source packages, such as the EEGLab [5], Fieldtrip [6] or CSD [7] Matlab toolboxes, dominate in the research labs. While the open source approach reduces development time, facilitates sharing of methods and data as well as serves as a verification infrastructure, by being a de-facto standard, halts innovation in the area of EEG processing software. We argue that innovation is required as the processing speed achieved during routine evaluations is intolerably low.

Since the state-of-the-art EEG systems can employ up to 256 electrodes and operate at as high as 16 kHz sampling frequency, even a few second measurement in such a system can generate hundreds of MBytes of data. The evaluation of these experiments results in long execution times (varying from minutes to hours depending on methods, data size and control parameters). This can be somewhat tolerated in single-shot experiments but it clearly presents a serious obstacle in large multi-patient studies or if EEG is ever to be considered as a diagnostic tool in the daily clinical routine.

This paper presents a novel parallel computing approach to EEG data processing and describes an experimental software environment developed at the University of Pannonia. We have developed a GPU-based massively parallel system that demonstrates that near real-time processing speed is achievable in key brain imaging algorithms, such as the spherical surface Laplacian [8] or the forward solution for source localization [9]. Although the program is created for EEG brain imaging, it can be used after little modifications for ECG-based multi-channel body-surface mapping applications [10], [11], [12].

At the center of our implementation is the GPU architecture that acts as a computing and visualization engine. The most advanced cards today can achieve up to 5 teraflops (10^{12} floating point operations per second) computational performance on a single chip using approximately 100 W power consumption. This performance-per-watt value cannot be matched in any other CPU-based parallel system. We use OpenGL [13] for visualization and the NVIDIA CUDA [14] technology for the parallel implementation on the GPU.

The structure of our paper is the following. Section II describes two illustrative imaging method, the surface Laplacian mapping as well as the forward algorithm indispensable in solving the inverse problem. Both approaches aim at detecting the spatial and temporal patterns for cortical brain activity. Section III describes the rationale behind using graphical processors for increasing processing speed as well as providing data visualization capabilities in our program. We highlight peculiarities of these systems from the algorithm design and optimization point of view. Section IV explains the core architecture of our software and the key design concepts, while Section V discusses our results to date.

II. EEG IMAGING METHODS

The electric potential measured on the scalp is the combined result of different activity sources in the brain. The potential is rarely usable as recorded due to the low signal-to-noise ratio and the presence of unwanted artifacts, such as the effect of eye movements, hence require filtering and other pre-processing steps. Nonetheless, the filtered potential data is seldom used on its own for activity mapping. A fundamental problem in the interpretation of scalp potential maps and in the solution of the forward/inverse problem is the so-called smearing effect of the skull. As the electric field propagates within the head, regions with the different conductivity spread the effect of the original sources, making their size grow on the scalp.

A. The Surface Laplacian

Laplace imaging can be used to 'sharpen' the scalp potential images, i.e. to reduce the smearing effect. The surface Laplacian [15], [16], [17] – the second derivative of the surface potential – computes the current source density (CSD) of the brain that is strictly related to the activity level of cortical areas. Several methods are known that apply the Laplace operator to focus the potential image before solving the inverse problem [18], [19], [20], [17].

The advantage of the Laplacian is that it is independent of the reference electrode choice, relatively insensitive to eye-movement artifacts and generates a relative topographic image that is well suited to activity pattern recognition. The Laplacian of the potential filed in the three-dimensional Cartesian coordinate system is expressed as

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

Where f is the potential function represented by the discrete potential values measured at the electrodes. One of the most widely used algorithm for computing the surface Laplacian on a sphere was proposed by Perrin [8], [21], which uses spherical splines

For a point *E* on the sphere, the Laplacian (a.k.a. the current source density), *C*, is given as

$$C(E) = \sum_{i=1}^{n} c_i h\big(cos(E, E_i)\big),$$

where E_i is the i^{th} electrode on the surface of the sphere, and the c_i s are the solution of

$$GC + Tc_0 = Z$$
$$T'C = 0$$

with

$$T' = (1,1,\dots,1),\ C' = (c_1, c_2, \dots, c_n),$$

$$Z' = (z_1, z_2, \dots, z_n),\ G = (g_{ij}) = \Big(g\left(cos(E_i, E_j)\right)\Big),$$

where z_i is the potential measured at electrode E_i. The function $h(x)$ is defined as the sum of the following series

$$h(x) = \frac{1}{4\pi} \sum_{n=1}^{\infty} \frac{(2n+1)}{n^{m-1}(n+1)^{m-1}} P_n(x)$$

where $P_n(x)$ is the n^{th} degree Legendre polynomial and m is the stiffness parameter of the spline.

A typical application of the surface Laplacian is in event-related experiments, in which normally multi-trial average of a 5-10 second time window of brain activity is analysed. In the widely used Matlab-based CSD Toolbox [7] program, the calculation is carried out in two steps; first, an H matrix is created by solving $h(x)$, than the Laplacian is calculated for the entire data window. The Matlab execution times are given in Tables I and II measured on an Intel Core i7-3820 2.70 GHz processor. Table I shows the calculation time of the *H* matrix for a different number of electrodes and summation limits for $h(x)$. Table II shows the overall execution time including the H matrix and the window calculations for $N = 15$. Even at this summation limit, it is evident that one window can take 10s of seconds to calculate. For $N = 50$, the execution time moves to the order of minutes.

TABLE I. CALCULATION TIME [SEC] OF THE H MATRIX (CSD)

N	Number of electrodes		
	64	128	256
15	5.74	22.84	92.50
50	34.00	137.51	552.12

TABLE II. TOTAL SURFACE LAPLACIAN CALCULATION TIME [SEC], N=15

Number of samples	Number of electrodes		
	64	128	256
1	5.74	22.84	92.51
512	5.81	23.03	92.89
1024	5.89	23.15	93.28
2048	6.04	23.47	94.07
5120	6.50	24.40	96.33
7160	6.80	25.03	97.94

B. The forward solution

While the surface Laplacian can help in focusing the potential image obtained on the scalp, it does not contain information about the cortical placement of the activity sources. Source localization methods can help in extracting this information, whose aim is to identify the location and magnitude of activity sources (dipoles) based on the measured potentials. The difficulty lies in the problem that any number of dipole combinations can generate identical scalp potential maps. This is the bioelectrical inverse problem which has no unique solution except when a priori constraints can be applied.

The basis of the inverse solution is that one or more hypothetical dipoles are placed in the brain within the head model and the generated potentials are computed – this is called the *forward problem* – then compared to the measured values. During the process, the position, orientation and magnitude of the dipoles are varied until an acceptable solution is found. Since there is no theoretically unique solution (except in the case of a single equivalent dipole), several solution methods have been proposed. Reviews and comparisons of these and other

frequently used inverse solution methods can be found in [22], [23], [24], [25].

The inverse solution requires the computation of hypothetical dipole generated potentials, known as the bioelectrical forward problem. This is based on the electrical field computation in the head model. The simplest head model is the spherical models that can contain up to five layers. Spherical models have closed form solutions, therefore they are computationally less demanding. Various solutions are developed for the multi-layer spherical head model forward calculation [26], [9]. The main advantage of the spherical head models is computational efficiency. Due to the closed form solution, the computational cost is relatively modest. This is crucial in localization where a large number of dipoles need to be tested for.

The widely used Fieldtrip Matlab toolbox [6] contains the implementation of the 4-layer spherical forward solution algorithm proposed by Cuffin and Cohen [9]. The computation of the generated potential is based on the following equation that calculates the effect of the x component of a source dipole vector. Similar equations are used for the y and z components.

$$V = \frac{P_x \cos\varphi}{4\pi\sigma_4 R^2} \sum_{n=1}^{\infty} \frac{(2n+1)^4 f^{n-1} (c*d)^{2n+1} P_n^1(\cos\theta)}{n\Gamma}$$

where

P_n^1 is the associated Legendre polynomial,

$$\Gamma = d^{2n+1}\{b^{2n+1} n(k_1-1)(k_2-1)(n+1) + c^{2n+1}(k_1 n + n + 1)(k_2 n + n + 1)\}\{(k_3 n + n + 1) + (n+1)(k_3-1)d^{2n+1}\} + (n+1)c^{2n+1}\{b^{2n+1}(k_1-1)(k_2 n + k_2 + n) + c^{2n+1}(k_1 n + n + 1)(k_2 - 1)\}\{\mathrm{n}(k_3-1)(k_3 n + k_3 + n)d^{2n+1}\},$$

R is the radius of the scalp sphere, f is the coefficient for the dipole position and

$k_1 = \frac{\sigma_1}{\sigma_2}, k_2 = \frac{\sigma_2}{\sigma_3}, k_3 = \frac{\sigma_3}{\sigma_4}$, and b, c, d are radius coefficients for each layer given as

$$b = \frac{r_1}{r_4}, c = \frac{r_2}{r_4}, d = \frac{r_3}{r_4}, R = r_4.$$

C. Computational cost

The forward solution is a critical operation since it has to be calculated many times during the solution of the inverse problem. In the most general case, the location, orientation and magnitude parameters of each dipole are adjusted until the best fit is found. Table III contains illustrative execution time figures obtained using the Fieldtrip Toolbox using 64 and 128 electrodes and varying number of dipoles. Execution time is measured on an Intel Core i7-3820 2.70 GHz processor.

III. EMBRACING GPU TECHNOLOGY

The traditional approach to speeding up the illustrated algorithms is to implement the algorithms in C/C++ and ideally write it in a parallel manner so it could scale over a range of CPU cores. Unfortunately, the achievable parallelism is limited in traditional computers as the number of cores is limited to 8 on Core i7 systems and 16 on state-of-the-art Xeon processors. We ignore the distributed-memory systems using multiple CPUs as they are not typical in the medical field.

TABLE III. SEQUENTIAL EXECUTION TIMES [SECONDS] OF THE SPHERICAL FORWARD ALGORITHM IN THE FIELDTRIP MATLAB TOOLBOX

Number of dipoles	Number of electrodes	
	64	128
1	0.25	0.32
10	0.52	0.95
100	3.34	7.22
500	15.93	35.13
1000	31.90	70.03
10000	314.86	700.02

A. GPU overview

Graphical processors, due their origin in 3D graphics, are designed for massively parallel execution of pixel graphics instructions. Today's GPUs are generally programmable and can be used the general computational tasks. State-of-the-art chips can contain over 2000 cores and their computational performance can exceed 4 teraflops. The key to this performance is the very efficient and coordinated execution of thread groups. GPU programs are ideally data-parallel, executing the same instruction sequence for a large amount of data. Threads can be arranged into 1D, 2D or 3D grids and efficiently mapped onto physical cores.

GPUs run the program from their own global memory that resides on the GPU card. They also use faster memories, such as the shared memory that can be used by a group of threads or registers for the exclusive use of a single thread. Global memory access is very expensive, thousands of instructions can be executed with a single read/write cycle. Even more expensive is the data transfer between the host and the GPU, shown in Fig. 1. While the GPU is treated by many as a co-processor, the efficient use requires that the GPU can execute at its own speed without the tight control of the CPU program.

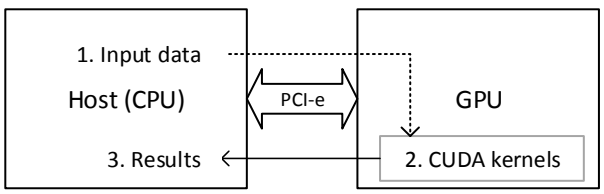


Figure 1. Data transfer between the Host and the GPU during operation.

B. Parallel execution strategy

Our primary aim was to explore various GPU parallelization strategies and to compare different implementations in order to *a)* speed up computation as

much as possible as well as to *b)* investigate the effect of design decisions on performance.

Both algorithms fall into the data-parallel class of problems. The first design decision is how to partition the problem in a scalable way, the second is how to use the memory the most efficient way and the third is how to minimize the number of cycles not spent with computation.

1) The GPU surface Laplacian

In contrast to the original surface Laplacian algorithm, in which the Laplacian is only computed for the electrode coordinates, in our system we are interested in the values at every position of the head. Our implementation computes interpolated values for a larger set or points. The exact number of points is determined by the resolution of the 3D sphere mesh used to build the head model.

The problem is partitioned by the surface points. The Laplacian at each point is computed independently from the other points. Since the spline interpolation is a global operation, all the electrodes are shared by the parallel threads (kernels).

2) The GPU forward solver

Our parallel forward solver implementation is based on the work reported in [27]. The underlying structure of the algorithm is a dipoles × channels thread grid, in which each thread calculates the effect of one dipole on one scalp electrode. We have added several performance optimizations to this algorithm. Using an optimized grid allocation strategy, restructuring code to optmize memory access and exploit on-chip instruction level parallelism better, we were able to almost double the performance of the original algorithm.

Another approach for improving the overall execution time was to use a more efficient base algorithm. The algorithm proposed by Sun [28] uses an interpolated function to avoid the lengthy computation of the Legendre sums in the original Cuffin algorithm. This version brings another ten-fold increase in performance.

IV. Software Architecture

Traditional EEG processing systems follow the traditional batch processing style of computation; data is input from a file, a set of algorithms are executed on the data and the results are saved in output files. They are also CPU-centric, i.e. the CPU controls the entire execution flow as well as performs the heavy mathematical computations. This type of software architecture is the result of the type of traditional computer architecture used in the past few decades that are often commonly referred to as "Denial Architecture" [29]. They pretend that computers operate sequentially and that the memory architecture is flat. GPU architecture explicitly exposes the underlying parallel hardware components and encourage the user to take advantage of locality in the hierarchical memory organization offering new opportunities for increasing performance and optimizing execution.

Our software architecture follows the explicit parallelism design principle and is based on the stream computing paradigm. The basic underlying model is of the data pipeline. Data is pumped from the measurement instrument or fetched from the data file, as required, and passed on through FIFO channels to processing nodes. The nodes output data to further processing or visualization nodes. The execution of the algorithms are triggered by the arrival of data and is carried out with maximum level of hardware parallelism. Other key design decisions were to remove any unnecessary data movement (keep as much data as possible on the GPU) and overlap CPU-GPU data transfer with computation to hide data latency.

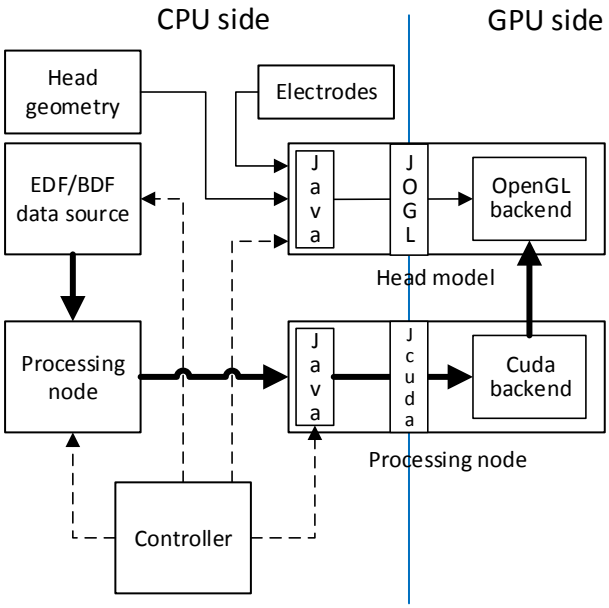


Figure 2. The high-level view of the GPU stream architecture. Solid arrows indicate data movement, thick arrow depicts live data update path (stream), whereas dashed arrows indicate control port connections. Data copied to the GPU memory are available to both the OpenGL and the CUDA backend. Processing node may or may not have GPU implementation parts. Their number and order are not limited.

A. Visualization

The visualization subsystem is implemented in OpenGL. We have avoided the use of any open-source visualization toolkit to retain maximum design freedom. Since the primary visualization aim is the display of time-varying potential (or potential-derivative) data on a static 2D or 3D geometry, head and electrode cap geometry data is cached into the GPU memory on initialization. Only potential data is updated continuously. A thin Java object wrapper layer is created to structure the view into higher-level objects (head model, electrodes, etc.) and to control the underlying OpenGL implementation.

B. Computation

Potential data is processed and transformed by the GPU using CUDA kernels. The largest potential performance penalty in a GPU-based computational and visualization system is if the GPU-computed results are copied back to the host in order to be displayed by the visualization system. The CUDA-OpenGL interoperability architecture allows us to avoid this by directly modifying OpenGL data structures from CUDA kernels.

Our algorithms are designed for scalability. As new GPU processors appear with more compute cores, the implementation automatically scales. If needed, the system can run with several GPU cards installed, in which case computation and visualization can be partitioned to separated GPU devices.

The CUDA threads are called via the JCuda wrapper library. To better manage the GPU code sections, the kernels are embedded in Java processor objects that manifest high-level EEG operations, e.g. `MovingAverageFilter`, `SphericalLaplace`, etc. These objects provide a clean programming interface to the rest of the system.

V. Results

Our prototype EEG processing and visualization system is operational. The current system consists of an Intel Core-i5 hosted NVIDIA Quadro K4200 card (1344 cores, 2.1 teraflops) and four 46" Samsung edge-lit display panels arranged as a 2×2 grid. The surface area and the 4K resolution of the display wall is sufficient to simultaneously display the EEG measurement plots, the 2D and 3D potential and Laplacian maps as well as RMI images and (if available) video recordings of epileptic patients.

The EEG plot (Fig. 3) allows for amplitude and time axis zoom, selection of measurement sample with marker. The marker can be moved or the measurement can be replayed at selectable speed. Fig. 4 illustrates the result of the surface Laplacian computation and its visualization on a 3D realistic head model. The focusing effect the of the Laplacian is clearly visible.

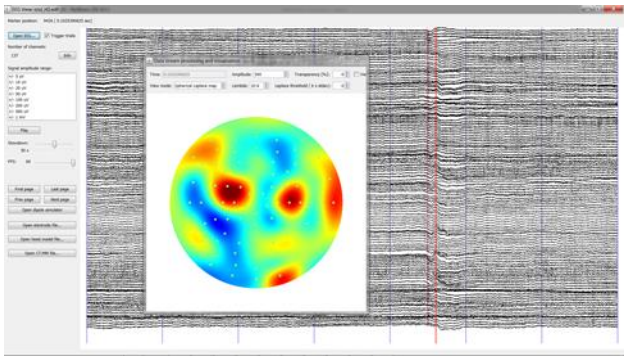


Figure 3. Screenshot of the prototype showing the multi-channel EEG plot and the 2D Laplacian map computed from the potentials.

A. Achieved computational performance

Our system has achieved considerable increase in computational performance. As shown in Table IV, using 128 electrodes and up to 2K (N=50) or 8K (N=15) surface points, the system can compute and visualize the Laplacian map real-time up to 2 kHz sampling frequency. The execution times shown in Tables IV and V are measured on an NVIDIA GTX 980 GPU card (2048 cores).

The performance results for the forward problem are similarly impressive. Using several code optimization steps we were able to improve the performance of the GPU-version of the Cuffin-Cohen algorithm [27]. Using Sun's algorithm could further increase the performance and reach msec range execution time for 128 channels and 5000 dipoles. The results are displayed in Table V.

TABLE IV. GPU-based Surface Laplacian Performance (128 electrodes)

Number of surface points	Execution time [msec]	
	N=15	***N=50***
1024	0.144	0.497
2048	0.152	0.578
4096	0.293	1.086
8192	0.444	1.761
16384	0.752	3.001
32768	1.386	5.642

TABLE V. GPU-based Spherical Forward Solution Performance

Number of dipoles	Execution time [msec]			
	Cuffin & Cohen		***Sun-Stok***	
	E=64	***E=128***	***N=64***	***N=128***
1	0.341	0.720	0.144	0.497
10	0.402	0.891	0.152	0.578
100	0.626	1.337	0.293	1.086
1000	1.721	3.374	0.444	1.761
10000	8.852	17.002	0.752	3.001

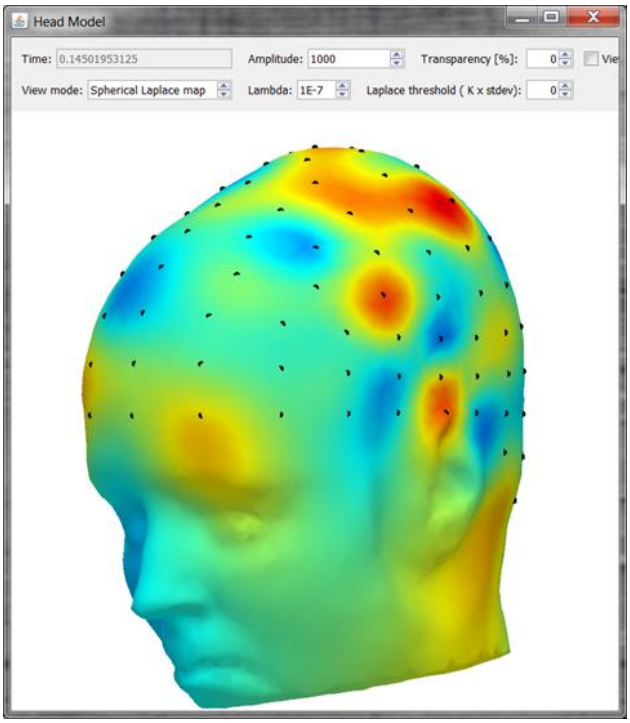


Figure 4. 3D visualization of the GPU-based surface Laplacian map generated from the measured scalp potentials. The realistic head model is segmented from real patient MRI data using the Freesurfer program package. The skull and the cortex layers are not shown.

VI. Conclusion

This paper presented a novel, GPU-based stream-oriented EEG processing software system. We demonstrated that the GPU architecture is well suited to simultaneous EEG visualization and processing. The computational performance offered by current GPUs make it possible to carry out simultaneous computation and visualization in real-time, or online if live measurement is performed. We were also interested in exploring new architectural and performance optimization methods that could lead towards the development of EEG software systems higher flexibility and with several orders of magnitude larger performance than that available today.

We have demonstrated with two important imaging algorithms how a GPU-based stream-oriented system can be created, and showed that computational performance up to 3 orders of magnitude higher than Matlab can be achieved while providing simultaneous interactive 2D/3D visualization. Our forward solver implementation is faster than any other GPU-based methods known to us in the literature. Our fast implementation of the surface Laplacian can serve as a basis for high-resolution cognitive studies and BCI applications. The system is integrated into a 4K resolution 2×2 display wall system creating a very high-performance, yet cost-effective brain imaging system. Future development plans include the integration and coupling of other imaging modalities such as MRI and fMRI, implementing fast realistic head-model based imaging algorithms as well as adapting the system for use in body-surface potential mapping applications.

Acknowledgment

The authors wish to thank their students A.G. Nagy (MSc) and D. Kenyeres (BSc) for their invaluable help in the development of the prototype system described in this paper.

References

[1] C. M. Michel and M. M. Murray, “Towards the utilization of EEG as a brain imaging tool.,” *Neuroimage*, vol. 61, no. 2, pp. 371–85, Jun. 2012.

[2] L. Ding, G. A. Worrell, T. D. Lagerlund, and B. He, “3D source localization of interictal spikes in epilepsy patients with MRI lesions.,” *Phys. Med. Biol.*, vol. 51, no. 16, pp. 4047–62, Aug. 2006.

[3] B. Alkonyi, C. Juhász, O. Muzik, E. Asano, A. Saporta, A. Shah, and H. T. Chugani, “Quantitative brain surface mapping of an electrophysiologic/metabolic mismatch in human neocortical epilepsy.,” *Epilepsy Res.*, vol. 87, no. 1, pp. 77–87, Nov. 2009.

[4] N. Srinivasan, “Cognitive neuroscience of creativity: EEG based approaches,” *Methods*, vol. 42, no. 1, pp. 109–116, May 2007.

[5] Z. A. Acar and S. Makeig, “Neuroelectromagnetic forward head modeling toolbox.,” *J. Neurosci. Methods*, vol. 190, no. 2, pp. 258–70, Jul. 2010.

[6] R. Oostenveld, P. Fries, E. Maris, and J. Schoffelen, “FieldTrip : Open Source Software for Advanced Analysis of MEG , EEG , and Invasive Electrophysiological Data,” vol. 2011, 2011.

[7] J. Kayser and C. E. Tenke, “Principal components analysis of Laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks.,” *Clin. Neurophysiol.*, vol. 117, no. 2, pp. 348–68, Feb. 2006.

[8] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, “Spherical splines for scalp potential and current density mapping,” 1989.

[9] N. Cuffin and D. Cohen, “Comparison of the Magnetoenchephalogram and Electroencephalogram,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 47, pp. 132–146, 1979.

[10] R. Trobec, “Computer analysis of multichannel ECG,” *Comput. Biol. Med.*, vol. 33, no. 3, pp. 215–226, May 2003.

[11] B. He, G. Li, and J. Lian, “A spline Laplacian ECG estimator in a realistic geometry volume conductor.,” *IEEE Trans. Biomed. Eng.*, vol. 49, no. 2, pp. 110–7, Feb. 2002.

[12] G. Li, J. Lian, and B. He, “Spatial resolution of body surface potential and Laplacian pace mapping.,” *Pacing Clin. Electrophysiol.*, vol. 25, no. 4 Pt 1, pp. 420–9, Apr. 2002.

[13] M. J. Kilgard, “Modern OpenGL Usage : Using Vertex Buffer Objects Well Mark J . Kilgard 1 Introduction 2 Maximizing Vertex Processing Rates 3 Start with Immediate Mode,” pp. 1–19, 2008.

[14] N. Cuda, “Programming Guide,” 2009.

[15] P. L. Nunez and R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG, 2nd Edition*. Oxford University Press, USA, 2005.

[16] P. L. Nunez and K. L. Pilgreen, “The spline-Laplacian in clinical neurophysiology: a method to improve EEG spatial resolution.,” *J. Clin. Neurophysiol.*, vol. 8, no. 4, pp. 397–413, Oct. 1991.

[17] P. L. Nunez and A. F. Westdorp, “The surface laplacian, high resolution EEG and controversies,” *Brain Topogr.*, vol. 6, no. 3, pp. 221–226, Mar. 1994.

[18] F. Babiloni, C. Babiloni, F. Carducci, L. Fattorini, P. Onorati, and A. Urbano, “Spline Laplacian estimate of EEG potentials over a realistic magnetic resonance-constructed scalp surface model,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 98, no. 4, pp. 363–373, Apr. 1996.

[19] J. Le, V. Menon, and A. Gevins, “Local estimate of surface Laplacian derivation on a realistically shaped scalp surface and its performance on noisy data,” *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.*, vol. 92, no. 5, pp. 433–441, 1994.

[20] F. Babiloni, C. Babiloni, F. Carducci, L. Fattorini, C. Anello, P. Onorati, and A. Urbano, “High resolution EEG: a new model-dependent spatial deblurring method using a realistically-shaped MR-constructed subject’s head model,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 102, no. 2, pp. 69–80, Feb. 1997.

[21] “Corrigenda,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 76, no. 6, pp. 565–566, Dec. 1990.

[22] R. Grech, T. Cassar, J. Muscat, K. P. Camilleri, S. G. Fabri, M. Zervakis, P. Xanthopoulos, V. Sakkalis, and B. Vanrumste, “Review on solving the inverse problem in EEG source analysis.,” *J. Neuroeng. Rehabil.*, vol. 5, p. 25, Jan. 2008.

[23] C. M. Michel, M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. Grave de Peralta, “EEG source imaging.,” *Clin. Neurophysiol.*, vol. 115, no. 10, pp. 2195–222, Oct. 2004.

[24] H. Hallez, B. Vanrumste, R. Grech, J. Muscat, W. De Clercq, A. Vergult, Y. D’Asseler, K. P. Camilleri, S. G. Fabri, S. Van Huffel, and I. Lemahieu, “Review on solving the forward problem in EEG source analysis.,” *J. Neuroeng. Rehabil.*, vol. 4, p. 46, Jan. 2007.

[25] O. Hauk, “Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data.,” *Neuroimage*, vol. 21, no. 4, pp. 1612–21, Apr. 2004.

[26] P. Berg and M. Scherg, “A fast method for forward computation of multiple-shell spherical head models,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 90, no. 1, pp. 58–64, Jan. 1994.

[27] N. B. Bangera and D. Lewine, “Accelerated large scale spherical model forward solutions for the EEG / MEG using CUDA,” in *GPU Technology Conference*, 2010, vol. 23, no. 11, p. Poster.

[28] M. Sun, “An efficient algorithm for computing multishell spherical volume conductor models in EEG dipole source localization.,” *IEEE Trans. Biomed. Eng.*, vol. 44, no. 12, pp. 1243–52, Dec. 1997.

[29] B. Dally, “The End of Denial Architecture and The Rise of Throughput Computing,” *Electronics*, 2009.

Web Portals for Visualizing and Searching Spatial Data

Lejla Bećirspahić*, Almir Karabegović *
* University of Sarajevo, Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina

Abstract - Spatial data have proven to be extremely useful in many areas. Accordingly, there are more and more web GIS solutions. These solutions allow the visualization of spatial data to all Internet users. A large number of available online resource introduced the problem of efficient retrieval of relevant information. The aim of this paper is to describe the way in which web portals can be used to search and presentation of spatial data.

The paper first analyzes the area of the web portal. It provides an overview of the web GIS technology, with particular emphasis in its advantages and disadvantages and how the client/server architecture can be applied to web GIS solutions. Then, it describes how the service oriented architecture can be used to resolve interoperability problems, and gives an overview of OGC standards: WMS, WFS, and WCS. In the end, it explains the need for introducing a geoportal, as a special type of a web portal.

In the practical part of this paper is implemented interactive map of Bosnia and Herzegovina. An overview of the technologies was given: PostGIS, Boundless SDK, GeoServer and OpenLayers, which are an integral part of the OpenGeo Suite. The paper also describes how to integrate local data sources with public web services: MapQuest, Bing Maps and OpenStreetMap.

I. INTRODUCTION

A geographic information system (GIS) is a computer system that allows capturing, modeling, manipulation, retrieving, analysis and presentation of spatial data. A majority of the early GIS users were professional users. Expansion of the Internet has made GIS available to casual users. The most important web GIS functionality is web mapping or visualization, which is the process of creating and using online maps. Modern web GIS solutions let users interact with maps.

A single source is almost never sufficient for satisfying information needs of today's users. A real potential of web GIS can be achieved through the integration of spatial data from multiple sources. WMS, WFS and WCS are three web services that are suggested in order to support the interoperability.

Web portal centralizes access to information from different sources and enables searching large amounts of data in a fast and efficient way. A geoportal is a type of web portal used to find and access spatial data and services via the Internet.

II. WEB PORTALS

There is no unique definition of a web portal. Depending on their own standpoints, different authors have proposed different definitions. One group of authors suggests that a web portal is a specific web page, which allows access to other web pages. When interpreting previous definition, one has to keep in mind that a web portal is more than a simple web page containing hyperlinks to other related pages. By its design, web portal should facilitate finding of relevant information in today's informational chaos. Information originating from different sources are consolidated and represented in a web portal, in a way that enables more efficient and fast search. In a broad sense, the use of web portal is not limited to accessing web pages, but also to accessing various network resources [1]. Other group of authors emphasizes the possibility of a portal's customization, so that user needs can be satisfied [2]. User customization in business environment implies adaptation to user role in an organization, so that user does not need to worry about inessential data. This enables user to fulfill tasks in a quality manner. Web portal also supports decision making process, because it provides a single point of access to useful information [3].

There are several ways to classify web portals. Mainly, web portals are classified as horizontal or vertical. Horizontal web portal is designed as a starting point for browsing the web [4]. Vertical web portals or vortals can be divided into three subclasses: corporate portals, used by employees to get access to organizational data and services, industry portals, used for information sharing and business transactions between business partners, and customer portals, used for customer support. Another common type of web portals are informational web portals. Informational portals are designed to satisfy information needs of specific groups in community. These groups are formed based on different criteria: age, gender, ethnic, national or racial affiliation [2].

Web portal can be compared to the front page of a newspaper [1]. No one is going to buy a newspaper with poorly designed front page and irrelevant news. The same rule applies to web portal – its success depends on how

fast required information is delivered to user, but also on capability to represent this information in accessible manner.

III. WEB GIS

Desktop GIS solutions are intended for professional users, who have specific knowledge and skills. In addition, desktop GIS application requires an installation of some necessary software. This approach cannot expand the full potential of GIS. The web has introduced a new era in GIS development, that is characterized by web GIS. The main idea of web GIS is online publishing, searching, processing, analysis and presentation of spatial data.

In simple terms, web GIS is any GIS that uses web technologies. This means that a web browser is all that is needed for accessing a web GIS application and it is simpler and more accessible than desktop GIS solutions. The combination of the Web and GIS has expanded the range of users, who can access the GIS functionality regardless of their current location. Wider availability has led to new GIS application, such as decision-making support, planning and problem solving. Areas of application are diverse: government, economy, health, research and scientific programs. Spatial data have proven useful in various fields, wherefore the number of users is further expected to increase [5].

Web GIS provides following features [5] [6]:

- Visualization or web mapping – the most used functionality. This includes a presentation of spatial data in a form of map.
- Collection of geospatial information – professionals and amateurs voluntary collect spatial data within collaborative projects like Wikimapia and OpenStreetMap. Data collected in this way is termed as VGI (volunteered geographic information).
- Dissemination of geospatial information – although VGI is useful, reliability of data can be doubtful. When making an important decisions, data sources must be accurate and reliable. Web GIS can integrate spatial data from multiple sources and present them on a single map.
- Personalization – web GIS can support personalization option. This allows user to create custom maps, add locations, descriptions, images, videos or links, and share that with other users.
- Geospatial analysis – web GIS can support analytical functionalities.
- Printing maps – user can select an area and layers for printing.

Web GIS has following advantages [6]:

- A global reach – after publishing to the Internet, Web GIS application is accessible to the whole world.
- A large number of users – unlike desktop GIS applications, which can be accessed only from a computer with appropriate software installed, web GIS can be used concurrently by hundreds of users.
- Cross-platform compatibility – web browsers support different operating systems. This means that web GIS is platform independent.
- Lower cost – instead of buying a software, individual users can access web GIS functionalities from a web browser.
- Ease of use – web GIS is intended for a wide range of users, who might not have knowledge about GIS.
- Automatic update – most web GIS clients are automatically updated.
- Diverse applications – desktop GIS applications are intended for professionals, and have limited use. On the other hand, web GIS is available to casual users.

Web GIS has following disadvantages [7]:

- Internet can be unreliable – availability and accuracy of external sources cannot be guaranteed.
- A high bandwidth is required – users with slow connection may have difficulty loading the map, which can be frustrating.
- Spatial data are expensive – spatial data are publicly available and free of charge in the USA. However, spatial data can be very expensive in other parts of the world.
- Copyright infringement – use of spatial data without owner's permission can cause copyright violations and legal consequences.
- Limited screen space – devices with small screens are less suitable for viewing web maps.

A. Web Mapping

Web mapping or visualization is the process of designing, implementing, generating and delivering maps on the Web [7]. Web mapping is extremely important functionality, since it allows visualization of complex spatial data on a map. When shown on the map, spatial data are suitable for different analysis. In this way, it is easier to detect relationships among them. Web mapping has introduced maps to a wide range of users. It is sufficient that a user opens a web browser and utilize numerous opportunities offered by web GIS.

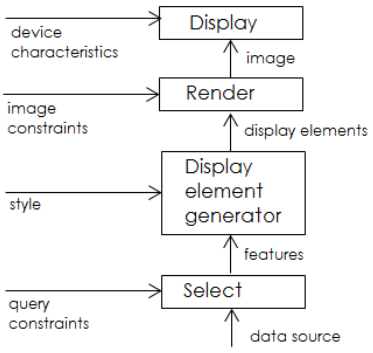


Figure 1. OpenGIS model for interactive web mapping [8]

Interactive web mapping involves creation of dynamic web maps on user demand. OpenGIS Consortium has developed a model of interactive web mapping, shown in figure 1.
The model of interactive mapping has four tiers [8]:

- Selection retrieves spatial data from a data source according to the constraints set in the query. Connection between the map server and the spatial database is achieved at this stage.
- The display element generator turns the selected spatial data into a sequence of display elements. It also applies appropriate styles.
- Render generates a map from input display elements.
- The rendered map is displayed on the output device.

In interactive web mapping, map server generates a new map on each user request. Generating maps from raw data is time-consuming operation. Response time of the server is determined by the bandwidth between a client computer and a remote computer. If a response time is slow, users will wait longer for the server's response. Large load of a map server can lead to worse performance. This means that user will wait longer to load the map, which is unacceptable for web applications.

Tile-based maps represent a significant improvement. This technique divides the map into discrete set of zoom levels, where each zoom level is divided into tiles of the same size. Described zoom levels form a hierarchy that can be visualized as a pyramid. The biggest advantage is in accessing certain tiles, instead of the entire map. Therefore, only the requested tiles are rendered. Tile-based maps enhance the performance and achieve savings of computing power and memory space. For this reason, they are used frequently in today's GIS solutions [9].

B. Web GIS Architecture

Web GIS architecture is similar to the client/server three-tier architecture [8]. A client is usually a web browser, which can be accessed on the local computer. Web browser interacts with a remote computer – server. A client sends a request to the server, while the server processes this request and sends a response to the client.

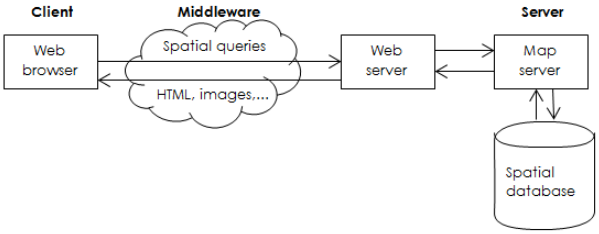


Figure 2. Client/server architecture of web GIS application [8]

Figure 2 show main components in client/server architecture of web GIS application: web browser, web server, map server and spatial database.

GIS components on the server side [7]:

- Web server – handles http requests. In the simplest case, web server returns static documents (HTML pages, raster images). Also, if there is a need for an interaction, web server sends requests to CGI applications or server-side scripting languages. The most popular web servers are Apache and Microsoft Internet Information Server.
- Map server – handles spatial data and generates maps (usually in PNG, JPG or SVG format). The most popular open-source map servers are UMN MapServer, GeoServer and Mapnik. Commercial alternatives are ESRI ArcIMS, Intergraph GeoMedia WebMap and Autodesk MapGuide.
- CGI applications – implement Common Gateway Interface standard and enable communication between the web server and map server. A CGI application is usually executed on the web server. It processes the request sent by client and forwards this request to the map server, then receives server's response and forwards this response to the client.
- Scripting languages (PHP, Perl, ASP, Python, Ruby) – are built into the web server and can be used to achieve communication between the web server and map server.
- Spatial databases – are relational databases with spatial data types and methods. Examples of spatial databases are PostGIS, Oracle Spatial, Microsoft SQL Server, IBM DB2 and MySQL.

GIS components on the client side [7]:

Web browser is sufficient for viewing static pages. Most of today's GIS solutions require additional plugins:

- ECMAScript – is standardized version of JavaScript. It is necessary for the implementation of the interaction on the client side.

- Event support – events can trigger script execution. Examples of events are mouse moves, mouse clicks, keyboard presses, etc.
- Network requests – are necessary to load additional data on the website.
- The Document Object Model (DOM) – is a standard for access and manipulate HTML and XML documents.
- Scalable Vector Graphics (SVG) – is a format that integrates vector graphics, raster graphics and text.
- Java applet – is a small application that can be included in a web page.
- Web browser plugins – Adobe Acrobat, Adobe Flash, Apple QuickTime, Adobe SVG viewer, Sun Java plugin.

C. *GIS Web Services*

A real potential of web GIS can be achieved through the integration of spatial data from multiple sources, so that a picture of the overall situation is created. On this way, information is exchanged within and outside the organization. Data integration raises the question of interoperability between spatial data and services [10]. Interoperability is defined as the ability to communicate, program execution and data exchange between different functional units. The user knows little or nothing about characteristics of these units. To make this possible, data is exchanged over common procedures, with use of agreed formats and protocols. As a result, interoperability contributes to overcoming problems in communication between components of different manufacturers, reduces redundancy and encourages knowledge exchange [11].

A single map cannot satisfy information needs of today's users. In these situations, it is inevitable to access different data sources. The most acceptable solution is to combine data from multiple sources into one composite map. However, standard client/server architecture offers little interoperability. Figure 3 shows situation where user, in order to access data from three different sources, needs to run three different web applications. The server will generate a map for each request.

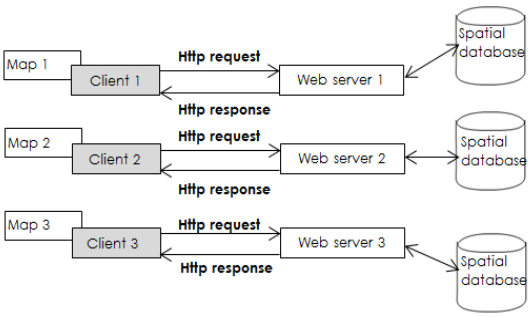


Figure 3. Lack of interoperability in the case of a client/server architecture [10]

In these solutions is lacking transparency, because each source is accessed through a specific server [10].

Interoperability can be achieved by introducing web services [11]:

- The Web Map Service (WMS) is a standard for visualizing spatial data on the Internet. It provides uniform access to maps rendered by map server. Maps can be generated from the raw data, as well as vector and raster data. Client, which implements WMS, can access any server that implements WMS. In this way is achieved the integration of data from different sources.
- The Web Feature Service (WFS) is a standard specified by OGC that is used for requesting/sending spatial objects on the Web. Spatial objects can be viewed as the source code behind a map. The WMS returns a map in the form of an image, on which are not supported any modifications or analysis. The WFS provides operations for manipulation of spatial objects.
- The Web Coverage Service (WCS) is a standard specified by OGC that is used for requesting/sending spatial data as coverages. A coverage is the digital spatial information that represents an object in some geographic area. While the WMS returns static images, the WCS gives more information, including valuable metadata. The WCS is WFS for raster data.

Service oriented architecture (SOA) is an approach to building software applications based on the use of web services. In this architecture, web services are the only way to access network resources. Figure 4 shows that the user can access data from three different databases from a single client. All the required information is displayed on a single map [10].

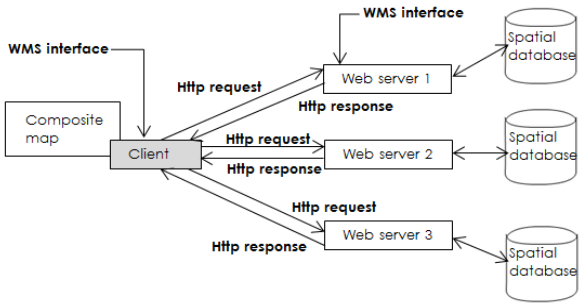


Figure 4. Interoperability of a SOA [10]

IV. GEOPORTAL

Organizations from different domains collect and publish spatial data. It is common that each organization focuses on their issues. This means that one organization will have data about roads, the other will have data about

woods, etc. A user needs a complete picture of some area. For him, this organization of spatial data does not have usability. The same problem can arise inside an organization, if each department collects processes and analyses data in distinct way. An efficient solution for the described problem is the development of geoportal. A geoportal is a type of web portal used to find and access spatial data and services via the Internet [12]. A geoportal can also be defined as a central point of access to spatial information, regardless of their location, format and structure [13].

V. INTERACTIVE MAP OF BOSNIA AND HERZEGOVINA

A. *Used Technology*

OpenGeo Suite is used for implementation of the interactive map of Bosnia and Herzegovina. OpenGeo Suite is a complete platform for managing spatial data and building maps for web browsers, desktop and mobile devices. OpenGeo Suite includes following open-source components [14]:

- PostGIS – a spatial extension to PostgreSQL.
- GeoServer – an open-source server written in Java, which allows retrieving, sharing and editing spatial data.
- GeoWebCache – a tile caching server, which accelerate the display of the maps.
- Boundless SDK – provides tools for building web GIS applications. It uses OpenLayers - a JavaScript library for viewing and interacting with spatial data in the web browser

B. *The Application Architecture*

Figure 5 shows the architecture, used for the realization of interactive map of Bosnia and Herzegovina. The bottom layer is a spatial database – PostGIS, in the middle is application server – GeoServer and GeoWebCache, while at the top is a user interface with OpenLayers library. This is client/server architecture.

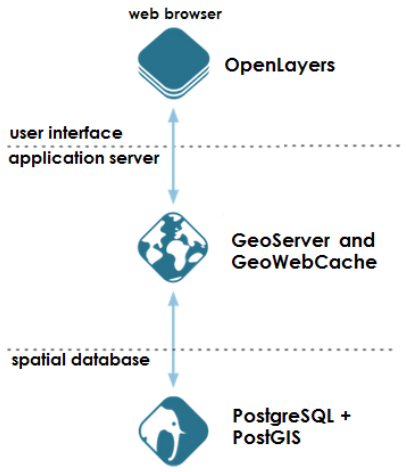


Figure 5. The Application Architecture [14]

C. *Web Services Integration*

OpenLayers 3 (OL3) allows easy integration with public web services. Restriction is that these maps have to be background (base) layer. For this reason, every layer that was created in this way will be grouped in background group.

MapQuest is a free web mapping service. A creation of a layer, that will use a MapQuest server as a source, is very simple and not much different from creating a layer form a local source. OL3 includes a class ol.source.MapQuest that allows the integration of a MapQuest map in the project. Figure 6 shows a MapQuest map.

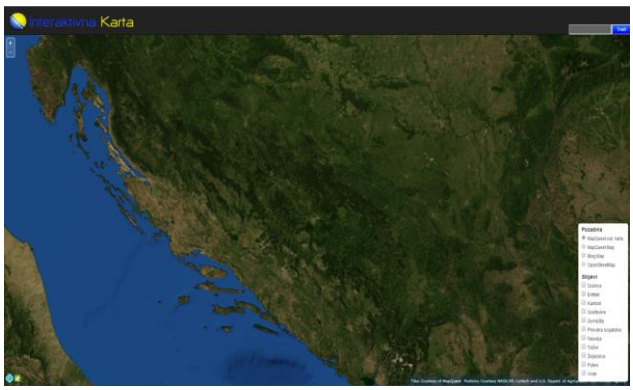


Figure 6. MapQuest satellite map

Bing Maps, originally known as Virtual Earth, is web mapping service developed by Microsoft. OL3 includes a class ol.source.BingMaps that allows the integration of a Bing map in the project. To gain an access to the Bing Map service, it is necessary to obtain appropriate API key. Figure 7 shows a Bing map.

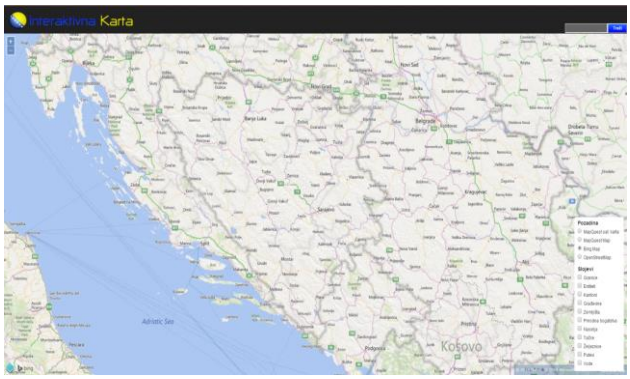


Figure 7. Bing map

OpenStreetMap is a collaborative project that aims to create a free map of the world. OL3 includes a class ol.source.OSM that allows the integration of an OpenStreetMap map in the project. Figure 8 shows an OpenStreetMap map.

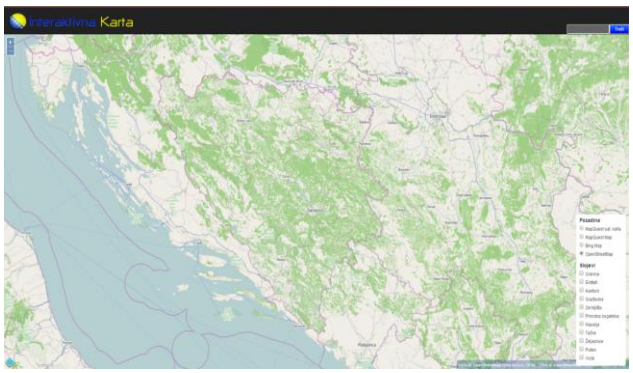


Figure 8. OpenStreetMap map

D. Viewing Layers with information about Bosnia and Herzegovina

Available layers are shown in the lower right corner of the screen. Layers are grouped into two groups: background and layers. This control enables user to select the layers, which will be shown on the map located in the central part of the screen. Only one background can be active at a time. On the other hand, multiple layers can be displayed on a single background. Figure 9 shows how to combine data about the borders of Bosnia and Herzegovina with an external source – a MapQuest satellite map.

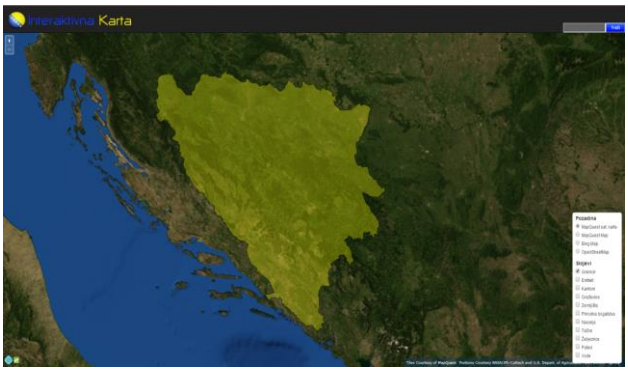


Figure 9. Layer with borders of Bosnia and Herzegovina

Zoom in/out controls are included by default. They are shown in the upper left corner of the screen. Figure 10 shows how to zoom in the layer in order to see the building located in the streets of Sarajevo.

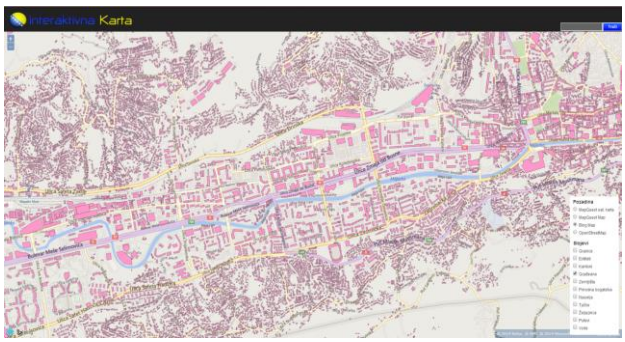


Figure 10. Layer with building in Bosnia and Herzegovina – zoomed view

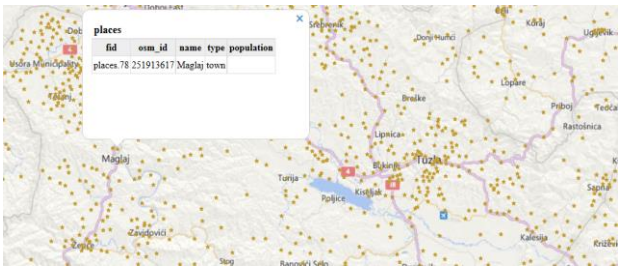


Figure 11. Layer with places in Bosnia and Herzegovina – a pop-up window with details about the place

When a user clicks on some point on the map, a pop-up window will appear, containing information about the place positioned at this location.

E. The Place Search on the Map

Interactive map has a place search functionality that allows locating nearly any place in Bosnia and Herzegovina. The user enters the name of the required place in the search box, which is located in the top right corner of the screen, and then click search. If that place exists, a red marker will appear on map.

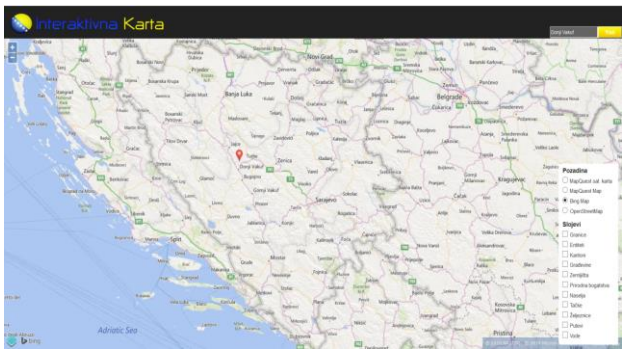


Figure 12. Layer with places in Bosnia and Herzegovina – a pop-up window with details about the place

VI. IMPLEMENTATION STEPS

1) Creating a database
   a. Finding adequate sources of vector data is the first step in the making of a map. List of used .shp files is shown in table 1.
   b. Next step is to design a logical data model.
   c. PostgreSQL is selected for a DBMS. PostGIS, that is a PostgreSQL extension, is installed for a spatial support.
   d. Separate table is created for each .shp file in the table. These files are loaded into the database using pgShapeLoader. A physical data model is created.

1) SPATIAL DATA ABOUT BOSNIA AND HERZEGOVINA

Name	*Description*
Borders	Data about the borders
Entities	Data about the entities
Cantons	Data about the cantons
Buildings	Data about museums, universities, bus stations, hotels, hospitals, etc.
Terrains	Data about the terrains
Natural resources	Data about the natural resources (woods, parks, etc.)
Places	Data about the populated places
Points	Data on a variety of objects
Railways	Data about the railways
Roads	Data about the roads
Water	Data about the water

2) Creating an infrastructure for an web application
   a. GeoServer 2.5 is chosen for a server that will generate maps and manipulate with spatial data.
   b. Eclipse Jetty is a web server that will contain web application.
   c. OpenLayers 3 is chosen for the front end development.
3) Creating layers in GeoServer
   a. To create layers of the map, the user interface of the GeoServer is used for adding a new source of vector data.
   b. These data are published as a new layer by GeoServer.
4) Integrating with external sources
   a. Local layers will be integrated with the following external maps: MapQuest, Bing Maps and OpenStreetMap.
5) WebGIS application development
   a. A user interface is designed.
   b. The code for creating map is written. A map of Bosnia and Herzegovina will be rendered in the central part of the user interface.
   c. Then, a control for layer selection is added.
   d. To implement the popup window, a listener for a single click event must be defined.
   e. Final step is the implementation of search functionality.

VII. CONCLUSION

The primary goal of building a geoportal is to provide a single source of spatial data. Namely, organizations from different domains are collecting and publishing spatial data. Such data sources introduce the problem of inconsistency and redundancy. In such environment, finding the relevant information is time-consuming. Geoportals integrate data from different sources, and provide an efficient mechanism for their search. Because of this, it is considered that a geoportal can satisfy information needs of today's users.

Besides searches, geoportals are useful for spatial data visualization on the web map. Spatial data are complex. They need to be presented on a map to have some value for a typical user. Visualization facilitates the identification of the relationship between spatial data and performing an analysis.

OpenGeo Suite is used for implementation of the interactive map of Bosnia and Herzegovina. An experience of working with this tool has confirmed that the emergence of web GIS has made spatial data available to ordinary users: interactive map was implemented without a wider knowledge of geography.

REFERENCES

[1] A. Tatnall, "Portals, Portals Everywhere," *Web Portals: The New Gateways to Internet Information and Services*, 2005.s

[2] T. Martins, V. Carvalho, F. Soares, "Web Portal: Total Challenge," *International Journal of Web Portals*, vol. 4, pp. 54-65, 2012.

[3] D. Jalal, M. M. Al-Debei, "Portals and Task Innovation: A Theoretical Framework Founded on Business Intelligence Thinking," *The Eleventh Annual International Conference on Business Intelligence and Knowledge Economy*, vol. April, pp. 1-20, 2012.

[4] H. Strauss, "All About Web Portals: A Home Page Doth Not a Portal Make," Web *Portals and Higher Education - Technologies to Make IT Personal*, 2002.

[5] Z. A. Khan, M. Adnan, *Usability Evaluation of Web-based GIS Application - A Comparative Study of Google Maps and MapQuest (Master Thesis)*, Ronneby: School of Computing, Blekinge Institution of Technology, 2010.

[6] P. Fu, J. Sun, "GIS in the Web Era," *Web GIS: Principles and Applications*, Redlands, Esri Press, 2010, pp. 1-24.

[7] S. Shekhar, H. Xiong, Encyclopedia of GIS, New York: Springer, 2008.

[8] A. A. Alesheikh, H. Helali, H. Behroz, "Web GIS: Technologies and Its Applications," *Symposium on Geospatial Theory, Processing and Applications*, Ottawa, 2002.

[9] R. Garcccia, J. Pablo de Castro, E. Verdu, "Web Map Tile Services for Spatial Data Infrastructures: Management and Optimization," *Cartography - A Tool for Spatial Analysis*, Infotech, 2012, pp. 26-48.

[10] K. Sahin, U. Gumusay, "Service Oriented Architecture (SOA) Based Web Services for Geographic Information Systems," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 27, pp. 625-630, 2008.

[11] R. Stopper, I. I. Enescu, "Open Geospatial Consortium (OGC) and Web Services (WMS, WFS, http://www.e-cartouche.ch/content_reg/cartouche/webservice/en/text/webservice.pdf.

[12] A. Karabegović, M. Ponjavac, "Geoportals as Interface for Data Warehouse and Business Intelligence Information System," *Advances in Intelligent Systems and Computing*, vol. 257, pp. 27-40, 2014.

[13] H. Akinci, C. Comert, "Geoportals and Their Role in Spatial Data Infrastructures," *5th International Conference on Geographic Information Systems*, Istanbul, 2008.

[14] "OpenGeo Suite," Boundless, http://boundlessgeo.com/solutions/opengeo-suite/

Prototyping of visualization styles of 3D scalar fields using POV-Ray rendering engine

J. Opiła*
*AGH University of Science and Technology
Department of Applied Computer Science, Faculty of Management, Cracow, Poland
jmo@agh.edu.pl

Abstract – There is a persistent quest for novel methods of visualization in order to get insight into complex phenomena in scientific domains as various as physics, biomedicine or economics. Research teams involved achieved excellent results, however some problems with elaboration of novel visualization styles connected with flexibility of the software used and quality of the final images still persist.

In the paper results of inspection of four visualization styles of 3D static scalar field employing POVRay ray-tracing engine are discussed, i.e. equipotential surface method using direct implementation of isosurface{} object, cellular trilinear interpolation approach, application of texture and eventually pseudo-particles design. All styles presented have been tested for hybrid visualizations and compared concerning computing time, informativeness and general appearance. It is shown in the work that Scene Description Language (SDL), domain specific language implemented in POV-Ray is flexible enough to use it as a tool for fast prototyping of novel visualization techniques. Visualizations discussed in the paper were computed using selected components of API of ScPovPlot3D, i.e. templates written in the SDL language.

Keywords - POVRay, pseudo-particles, scalar field visualization, ScPovPlot3D, visual data analysis

I. INTRODUCTION

In the recent years technologic development enabled elaboration of numerous outstanding data gathering methods, which produces a large amounts of data including numeric and raster image formats. Amongst most productive disciplines one can mention astronomy, engineering, economics, econometrics, medical sciences including medical imaging and this short list is hardly complete. Often collected or computed data are in the form of 3D scalar field, static or dynamic, for example spatial temperature distribution in the sleeve and block of piston engine or pressure distribution in the rocket engine combustion chamber. Scalar fields can be described by differential equations which sometimes can be reduced to a simple formula e.g. for gravitational, electrostatic or (approximate) molecular Lenard-Jones potential. However in numerous cases they can be obtained by measurements only for example distribution of greenhouse gases in Earth's atmosphere [1] or distribution of mineral resources, principally ore deposits, such as coal or oil. Reliable analysis of such a data more often than not requires intensive usage of computer data processing and subsequent visualization.

Though there is an agreement on visualization styles of many data arrangements, visualization of 3D scalar fields still may be refined according to characteristic of a specific case, e.g. implementation of hybrid designs. Thus visualization software or library should allow not only for minor adjustments like color or shape of data points but for experimenting with new styles as well. In order to meet this requirement visualization software has to be sufficiently flexible. Unfortunately, ready-made, even state of the art software, no matter if proprietary (i.e. closed source), or Free and Open Source (FLOSS), always offers only closed list of variants of charts and graph layouts [11], as all of them have to be explicitly compiled into the machine code of the final application or library. As a result, efficient and fast prototyping of novel designs is usually difficult, even if using low level visual toolkits, and sometimes impossible without writing many lines of a new code. This weakness can be overcome by choice of efficient renderer (in order to obtain high quality images) equipped with scripting high level language (to avoid compilation overhead) and published as FLOSS software (to be free from license restrictions).

II. VIRTUAL LABORATORY

In order to inspect problems set above, one needs to conduct some experimental research. However, all "experiments" to be done requires computer only, thus name "numerical experiments" fits well. Because numerical experiments were performed in "virtual laboratory", below is explained how such "laboratory" has been designed.

A. The choice of rendering engine.

In search for the suitable software platform several packages have been examined: BioBlender [23], Blender [14], Diderot [18], gnuplot [10], ParaView [25] RasMol [24], Shadie [20],[21], [22], VTK [17] but only POV-Ray [2] met the all, drafted above, requirements. Some of them just cannot produce visualization of 3D scalar field but molecules only [24], [23] or have disappointing output [11], the over requires tedious low level coding in Python [14] or C++ [17] Below is a short resume of this analysis, software is listed in alphabetic order, except for the POV-Ray which is discussed as the last item.

The work was supported by AGH University of Science and Technology under grant nr. 11.11.200 (job .7)

BioBlender [23] – It is a very interesting add-on to Blender, however is limited to plain visualization of molecules; may use native Blender, Cycles or POV-Ray engine for rendering output images.

Blender [14] – Blender is a 3D animation program, equipped with a choice of efficient raytracing modules (Blender native or Cycles) but capable of employing POV-Ray for raytracing as well. Blender offers user space API written in Python, which can be used for preparing custom charts. While this is really flexible solution for experts, for end user it would be too complex. Thus one of the main objectives of this work, which is possibility of easy building new types of 3D charts, could not be reached because every new kind of plot ought to be separately programmed using complex triangular meshes.

Diderot [18] – excellent, but limited to medical imaging and because of usage of rarely used computer language SML/NJ [19], has very sloppy learning curve,

gnuplot [10] – is a simple but powerful as well as multiplatform chart plotting program, licensed as Free Open Source. However it offers only simple line-art level graphics and limited number of charts and styles. Moreover, plotting algorithms cannot be extended by user written shaders.

RasMol [24] – This toolkit has been designed to model molecules. However it seems to be discontinued, latest release available is from 2009,

Shadie [20] – Python based toolkit and domain specific language, able to use GPU CUDA engine in order to do volume visualization using CT or X-Ray data. Originally designed for medical purposes may be used for other fields of science as well, however this application is restricted, because Shadie requires as input data grayscale bitmaps. Another problem is that the project seems to be discontinued; the latest source files are dated on 2011 year.

VTK [17] – excellent and extended toolkit, but requires laborious coding phase, in order to built case specific final program. While adding novel designs is possible by extending base classes it requires excellent programming skills.

POV-Ray [2] is the well known, raytracing program producing stunning photorealistic images. Unfortunately it is controlled by text files containing sequence of statements written in specific to POV-Ray domain specific language dubbed by developers "Scene Description Language" – SDL. SDL as domain specific language is built into POV-Ray, and its full documentation is available on POV-Ray's home page and dedicated Forum. SDL implements a couple of predefined graphics primitives (i.e. readymade "shapes"), simple [1] and complex [2] ones, all of them defined as mathematical entities. Objects inserted into scene description file are thus parameterized by a set of attributes and can also be scaled, rotated, translated and skewed. Because POV-Ray is founded on mathematical modeling it offers to the end user a number of valuable numerical methods like root solvers (including Sturmian solver), spline functions, fractals and vector analysis functions (ex. gradient finding). While SDL is good enough for programming of even very complex scenes, it lacks some features characteristic for modern languages like constants, structures or classes. Another problem is connected with coordinate system employed by the program, which not only is left-handed but rotated too. Because of these weaknesses only those skilled in programming and 3D aware users are able utilize its full potential and power. Luckily, SDL implements many constructs sufficient for coding of efficient Application Programming Interface (API), like macros, inline functions, function and macro pointers, conditionals, loops and arrays. Furthermore, POV-Ray is equipped with comfortable Integrated Development Interface [3] (IDE), featuring a.o. syntax coloring, word completion and simple debugger.

B. What ScPovPlot3D is and why has been used.

In the recent years set of templates for POV-Ray, named ScPovPlot3D (**S**cientific **POV**-Ray **Plot**s in **3D**) has been developed [3]. Beginning with simple histogram chart some years ago it can now help prepare a range of 3D charts including raw or smoothed 3D surfaces, wireframe "surfaces" and surfaces based on irregularly dispersed data [7],[8],[9] (employing simple kriging method) available in the stable branch. Experimental modules include procedures for a.o. fuzzy numbers [5], hybrid charts [6], vector field and potential visualization. Moreover helper modules delivers easy to use items like smart cameras ("Cameras.inc" library), coordinate systems as well as text with enhanced formatting.

Main achievement of the toolkit is a reasonable balance between ease of use, flexibility and quality of produced images. As a result, less experienced users can prepare custom charts in minutes, just by modifying included examples, while expert users may tune code of the scene to obtain novel designs. In the following chapters prototyping of visualization styles of 3D static scalar fields using the POV-Ray, complemented with ScPovPlot3D templates is discussed.

C. Choice of exemplary scalar field

To compare different visualization styles the same scalar field produced by selected molecular structure should be used for all tested styles. The structure chosen should obey at least two, contradictory, constraints. Firstly, it should be complex enough to reveal unknown properties, for example at boundary or in case of structural defects (e.g. lack of one of atoms). Secondly, it has to be as simple as possible to be computable in reasonable time on regular workstation. Based on these assumptions Lennard-Jones potential (often referred as L-J potential) around single layer hexagonal structure (thus referred in the paper as "graphene like") composed from 40 neutral molecules, has been chosen. It should be mentioned, that electrostatic potential around real graphene flake is generated by electric charge distributed continuously due to the principles of the quantum mechanics and should be

[1] ex.: plane, cylinder, torus.
[2] 3D Bezier patch, isosurface, fractals,

[3] Windows™ version only

computed by integration over space using Coulomb's Law.

Although in the paper electrostatic potential is taken into account conclusions remain the same for any static scalar field, e.g. tissue density restored from NMR or X-Ray DICOM images.

D. Lennard-Jones potential

Visualization of molecules is a real challenge. First of all, world in nanoscale is governed by quantum mechanics so phenomena like hybridization of orbitals [15] or elusive Casimir effect [16] are often observed. In particular nanoscale objects, due to presence of continuous electric charge distribution, does not sport a well defined (i.e. rigid) surface like a billiard or golf ball so a special convention should be adopted, e.g. specific value of potential or ionic radius resulting from measurements, to be enable drawing "surface" of the molecule. It should be noted, that ionic radius depends on the electronic structure of an element in a molecule. Another problem is connected with numerical calculations, namely, in equations very small values (ε, r_0, r_m, r) are placed one in denominator and the other in nominator what is a serious source of round-off errors. This requires special attention at designing and writing source code.

For a single neutral atom or spherical molecule a relatively simple formula for electrostatic potential was developed by Lennard-Jones (1) [12]:

$$V_{LJ}(\vec{r}) = \varepsilon\left[\left(\frac{r_m}{|\vec{r}-\vec{r}_0|}\right)^{12} - 2\left(\frac{r_m}{|\vec{r}-\vec{r}_0|}\right)^{6}\right], \qquad (1)$$

where $\vec{r}_0$ is position of the source molecule, $\vec{r}$ points at observation point, $\boldsymbol{\varepsilon}$ is a depth of the potential well and r_m is distance at which value of the potential is minimum. As results from (1) L-J potential is spherically symmetric.

For a set of molecules summation over all $\boldsymbol{N}$ sources have to be done (2). For the case presented in this paper $\boldsymbol{N}$=40:

$$V_{LJ}(\vec{r}) = \sum_{i=1}^{N} \varepsilon\left[\left(\frac{r_m}{|\vec{r}-\vec{r}_{0,i}|}\right)^{12} - 2\left(\frac{r_m}{|\vec{r}-\vec{r}_{0,i}|}\right)^{6}\right]. \qquad (2)$$

While alternative formulations of the L-J potential are possible but this one has been implemented in ScPovPlot3D API. Prototypes presented in the next part of the paper require passing value of the scalar field in the form of POV-Ray "function". However, functions in POV-Ray have many limitations and resembles inline functions of C++ rather than fully featured subroutines. Although formula (2) seems to be simple, coding it in the form of POV-Ray function is not. Because loops and temporary variables are not allowed inside POV-Ray "functions" whole definition of the function, including header and statements, has to be explicitly injected into the code of the scene at runtime. The solution is to design dedicated function factory using #macro{} syntax completing this task. The code of function factory for L-J potential is presented in the Fig. (1). Although code of the function factory is short, after expansion may inflate to hundreds of lines for larger structures.

Code snippets discussed in next sections are reduced to the most important lines. Full source code, required to re-compute included graphics is available on-line [4]. POV-Ray documentation is available from within IDE.

E. Computer

Most calculations were done on ASUS 1215N, which is rather slow, if compared to high-end workstations. Computation time on multicore computer, in comparison to obtained times, may be over a dozen times shorter.

In the next chapters, prototyping and search for optimal 3D visualization style of Lennard-Jones potential produced by "graphene like" flake is presented.

III. ISOSURFACE

Spatial distribution of the scalar field **V**, given by a function **V**=V(x,y,z) (not necessarily continuous) is being customarily visualized using "same value" surfaces, called "isosurface" or "equipotential" if the potential field (for example gravitational or electrostatic) is under

```
#macro CreateVLJ(_PTs)  // creates Lenard-Jones potential function from network  of neutral molecules given by structure table _PTs[][]:
    // using variant of L-J formula employing radius of minimum of the potential well r_m
    #if (_PTs[0][0]>0)      // need one charge at least!
       function(x,y,z){
          _PTs[1][6]*( pow( _PTs[1][5] *_PTs[1][5]/(pow(x-_PTs[1][1],2) + pow(y-_PTs[1][2],2) + pow(z-_PTs[1][3],2)), 6)
                    -2*pow( _PTs[1][5] *_PTs[1][5]/(pow(x-_PTs[1][1],2) + pow(y-_PTs[1][2],2) + pow(z-_PTs[1][3],2)), 3))
          #for(i, 2, _PTs[0][0]) // loop  #for...#end  is part of macro not function
             +
             _PTs[i][6]*(pow( _PTs[i][5] *_PTs[i][5]/(pow(x-_PTs[i][1],2) + pow(y-_PTs[i][2],2) + pow(z-_PTs[i][3],2)), 6)-
                       2*pow( _PTs[i][5] *_PTs[i][5]/(pow(x-_PTs[i][1],2) + pow(y-_PTs[i][2],2) + pow(z-_PTs[i][3],2)), 3))
          #end
          }
    #else
         function(x,y,z){0} // just for safety
    #end
#end // CreateVLJ() end definition
```

Figure 1. Function factory for L-J potential defined as macro injecting raw code into scene file at runtime. Relevant data are defined in the global table _PTs[][]. First index, i=1…N (first row for i=0 has special meaning), enumerates all molecules in the structure, and the second has following interpretation (i-index of the molecule): [i][0] charge, [i][1,2,3] x, y and z of observation point, [i] [4] L-J ionic radius R [i][4], L-J radius of minimum of potential well [i][5], potential well ε [i][6].

consideration. POV-Ray sports dedicated object for this purpose called *isosurface*, definition is presented below. Notation adheres to common convention: key identifiers are in bold, optional parts are in square brackets, options are separated by '|' character [2]:

```
isosurface {
  function { FUNCTION_ITEMS }
  [contained_by { SPHERE | BOX }]
  [threshold FLOAT_VALUE]
  [accuracy FLOAT_VALUE]
  [max_gradient FLOAT_VALUE]
  [evaluate P0, P1, P2]
  [open]
  [max_trace INTEGER] | [all_intersections]
  [OBJECT_MODIFIERS...]
}
```

Only one item is obligatory, namely 3D scalar function. The most important optional item is a float *threshold*, which determines value constant over the isosurface object. Object modifiers describe look and finish of the surface. However, if omitted, default values are used. Detailed description of other attributes is given in documentation of POV-Ray.

Isosurface visualization style may be simple or hybrid. In the first case only shape of equipotential surface is revealed and colorful texture may be used to differentiate between surfaces with different threshold values. Thus a few surfaces may be shown in a single picture.

In the second case texture may be calculated using, for example, value of the gradient of the field over the surface. That way two different functions may be visualized at the same time as shown in Fig. 2.

However two hard to solve problems arise, both related to POV-Ray limitations. The first one is connected with computation time, the other with syntax.

Computation of the perfect looking isosurface requires use of very precise root finding algorithm, which in turn makes a lot of calls to potential function. Moreover, the potential function, due to limitations of POV-Ray syntax, has to be defined in rather inefficient way as a very long formula. Furthermore, algorithm is forced to search for the roots everywhere around the particle. Reducing searched space using **contained_by** modifier may solve this problem to some extent.

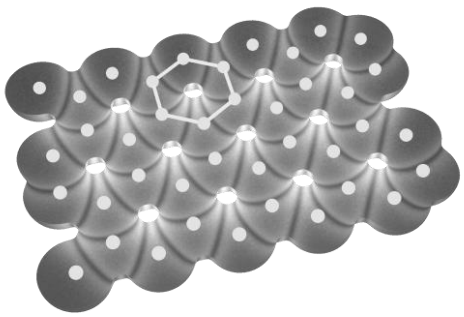


Figure 2. Hybrid representation of electrostatic L-J field around "graphene like" flake. Light gray circles marks "carbon" atoms, bonds of one hexagon are shown. The shape of the surface represents surface of constant potential and tint represents gradient of the potential: dark – low gradient, white-high gradient.
Computation time was 1h 07m. Source: own.

Limitations of syntax of function makes programming of user-function governed textures very difficult. In general, user defined potential function has to be passed as an argument to the gradient function building macro. Then the gradient aware texture may be defined:

```
#local gMin =    75; // assumed gradient minimum
#local gMax = 150; // assumed gradient maximum
SetGradientAccuracy(.0001) // const. for fn_Gradient()
// fn_Gradient(VLJ) makes from passed VLJ function
// gradient length of VLJ as a function.
#local grV0 = fn_Gradient(VLJ);
// definition of clipped function grV, useful for
// color_map{} texture
#local grV = function{
   clip(grV0(x,y,z)-gMin, 0, gMax-gMin)/(gMax-gMin)
};
SetIsoTexture(  // setting gradient governed texture..
  texture{
    pigment{
      function{ grV(x,y,z) } // gradient aware texture
      color_map{ _ColorMap }
    }
    finish{ Dull }
  }
)
// preparing of the isosurface,
#declare ISO3 = MakeEquiPlane( VVV, 5, Min, Max )
// and displaying it:
object{ISO3}
```

The aim of cellular trilinear interpolation method, presented in the next chapter is to overcome inefficiency of function coding and to improve efficiency of root searching algorithm.

IV. CELLULAR TRILINEAR INTERPOLATION

In case of very large structures counting hundreds of atoms potential function produced by function factory becomes extremely inefficient. Additionally finding isosurfaces over extended areas abruptly increases computation time. The first problem may be solved by simplification of function code through some kind of multi-linear interpolation the other by splitting whole search space into cells, then rejecting from analysis those cells where no piece of isosurface can be found.

Trilinear interpolation function $f:\Re^8 \rightarrow \Re$ implemented in ScPovPlot3D is defined in recursive manner (3) [26], [27], relevant geometry and symbols are displayed in the Fig.3. First, by linear interpolation of "Q" along edges of the cell, $Q_{ij} = \lambda Q_j + (1-\lambda) Q_i$ (where Q is scalar defined in vertices of the given cell - Q may be interpreted as charge, potential value, etc.) edge intermediate values denoted as e.g. Q_{12} are computed (3a). Then edge intermediate values are combined across common face (3b) producing two face intermediate values, in Fig. 3 denoted as Q_D and Q_U. Finally face intermediate values produce approximated value of Q denoted as Q_A (3c). It can be easily shown, that

interpolated value Q_A does not depend on specific sequence of calculation of edge and face intermediate values. Sequence presented in the Fig.3 may be described by [yxz] notation and the same values for Q_A will be obtained if sequence [xyz] or [zxy] would be used.

$$\lambda_x = \frac{x}{L_x}; \quad \lambda_y = \frac{y}{L_y}; \quad \lambda_z = \frac{z}{L_z};$$

$$\begin{cases} Q_{12} = \lambda_y Q_2 + (1-\lambda_y) Q_1 \\ Q_{34} = \lambda_y Q_4 + (1-\lambda_y) Q_3 \\ Q_{56} = \lambda_y Q_6 + (1-\lambda_y) Q_5 \\ Q_{78} = \lambda_y Q_8 + (1-\lambda_y) Q_7 \end{cases} \quad a) \qquad (3)$$

$$\begin{cases} Q_D = \lambda_x Q_{12} + (1-\lambda_x) Q_{34} \\ Q_U = \lambda_x Q_{56} + (1-\lambda_x) Q_{78} \end{cases} \quad b)$$

$$Q_A = \lambda_z Q_U + (1-\lambda_z) Q_D \quad c)$$

Because (3) uses only simple arithmetics, i.e. additions and multiplications, it can be computed relatively fast, however this does not mean, that whole computation time is short. Test hybrid image (isosurface plus gradient) presented in the Fig. 4, has been rendered using 300×150×60 cell grid, hence potential value in 2'772'511 nodes had to be calculated. After that every cell was tested, if isosurface threshold may be found inside, if yes, isosurface object was injected into scene file, if no, processing was passed to the next cell. Computing of nodal values and preparing isosurface pieces for rendering took approximately 90% of scene computing time. Because algorithm is approximated, some circular artifacts for gradient may be observed on the calculated surface. As isosurface algorithm avoids crawling through empty space, computation of surface itself is fast.

The overall efficiency may be improved using external command line driven program written in C language as numerical calculations using machine level instructions

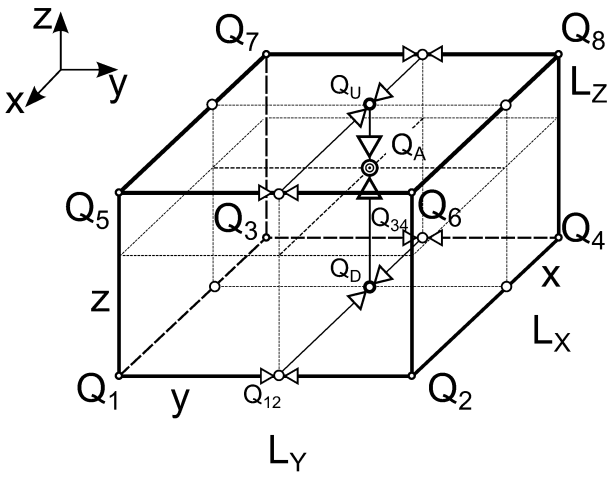


Figure 3. Illustration of multilinear interpolation algorithm. Approximated value of potential $V(Q_A)$ in point $Q_A(x,y,z)$ marked with a triple circle is given by (3). White arrows shows sequence of calculations. In the upper left corner symbolic reference to coordinate system is shown. Source: own.

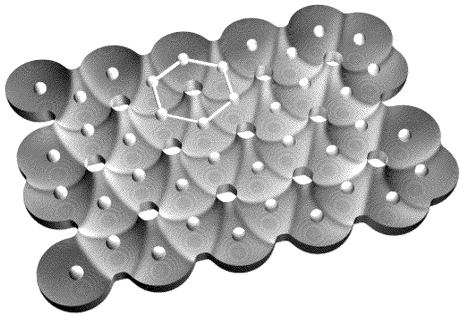


Figure 4. Hybrid representation of LJ potential and related gradient computed using multilinear interpolation over 300×150×60 grid of cells. Computation time: 3h 18m. Source: own.

(as is common in "compiled" languages) are very fast. Furthermore, as cells are mutually independent, this algorithm may be easily implemented as parallel employing GPU CUDA technology. Another improvement toward reducing number of cells and enabling smooth surface transitions between them, requires implementation of more sophisticated interpolation algorithm, based for example on spline paradigm.

V. PSEUDOPARTICLES APPROACH

Another possible visualization style is based on pseudoparticle idea. "Particles" in 3D computer graphics, implemented as "media" in POV-Ray, simulate presence of 0-sized objects scattering, absorbing or emitting light. These objects may get specific color due to programmer's demand. Similar paradigm has been implemented in very old program HPOV [15], designed for visualization of hybridized electron orbitals.

Because media calculations in POV-Ray are very slow a plenty of very small spheres may be used instead. Hence visualization is reduced to point calculations now, limitations of syntax are lifted. As a result this style is very flexible, up to possibility of including representation of third scalar function as e.g. shape variation.

In presented implementation all pseudo-particles were obtained by the drawing of x,y,z coordinates from three independent pseudo-random number generators (PRNG), then comparing potential for the point drawn with desired threshold value. If difference is smaller than assumed accuracy, position of this point is refined using gradient of the potential. Eventually a little sphere is injected into the scene, with texture related to gradient of the potential. Exemplary image obtained using this technique is presented in the Fig. 5.

Although this style has some advantages it has shortcomings as well. First of all, PRNG implemented in POV-Ray is very slow and most of "shots" are missed. However positions of pseudoparticles may be computed by external program and then read in from text file.

VI. COLOR MAPS

"Color map" is feature of POV-Ray allowing for assignment of color (texture or pigment as well) to the float value given by space dependant "pattern function". Thus any scalar function including potential function and

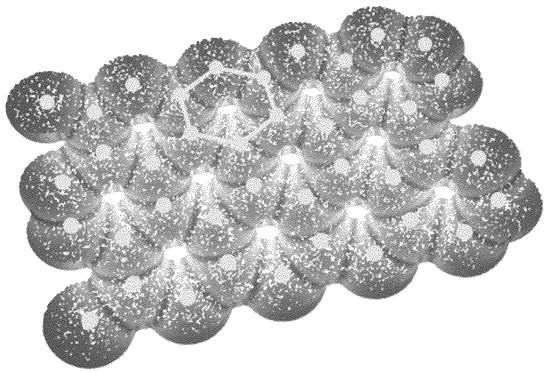


Figure 5. Hybrid representation of LJ potential and gradient using pseudoparticles approach employing 10^5 spheres. Due to small efficiency of drawing algorithm computation time was 4h 8m. Despite large number of spheres drawn,still same artifacts are visible. Source: own.

its variants (gradient length, divergence …) may be passed to texture definition as pattern function. As a probe object a plane or another suitable for research aim object may be used. Amongst advantages of this approach one can mention very short computation time and high flexibility.

The "Potential.inc" module from ScPovPlot3D library [3] offers relevant API functions. Exemplary image rendered that way is presented in Fig. 6. As input, decimal logarithm of LJ potential was used. Black and white stripes are painted every tenfold increase of potential on three perpendicularly oriented planes, revealing potential holes as possible two-state molecular switches (single particle may be trapped up or below the plane).

VII. Conlusion

High quality 3D render program equipped with

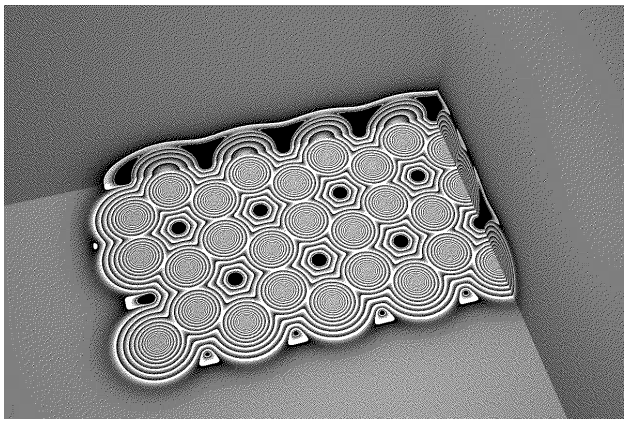


Figure 6. Color map representation of decimal logarithm of LJ potential painted on three perpendicular probe planes. BW stripes marks 10 times increase of potential value. Computation time, 0h 2m. Source: own

efficient scripting language and supplementary API, like POV-Ray, may be used for fast prototyping of complex visualization styles.

Because of numerical efficiency problems in some cases external program should be used.

Due to relatively short computation time color map style is very well suited for fast introductory analysis of new structures.

Acknowledgment

I thank to Mr. Prof. J.T. Duda for encouragement and personal support.

References

[1] NASA Earth Observatory, *Global Patterns of Carbon Dioxide*, http://earthobservatory.nasa.gov/IOTD/view.php? id=82142, accessed 2014-12-28,

[2] POV-Ray, http://povray.org, accessed 2014-12-28,

[3] J. Opiła, *ScPovPlot3D POVRay templates*, http://scpovplot3d.sf.net, accesed 2014-12-28,

[4] http://sourceforge.net/projects/scpovplot3d/files/Dev/4MiPro38.zip/download, accessed 2015-03-21,

[5] I. Skalna, J. Opiła, J. Duda, T. Pełech-Pilichowski, B. Gawel, et al., *Advances in Fuzzy Decision Making*, unpublished,

[6] J. Opiła, "*Business data visualization using ScPovPlot3D, POVRay based toolkit*", in press,

[7] J. Opiła, *Some remarks on visualisation of irregular or dispersed data* (pol. *Kilka uwag o wizualizacji danych nieregularnych lub rozproszonych*) in: Diagnozowanie stanu środowiska, red. J. K. Garbacz. Bydgoszcz: BTN, 2012, ISBN: 978-83-60775-34-9,

[8] J.Opiła, I. Skalna, T. Pełech-Pilichowski, *Interval arithmetic for irregularly distributed data visualisation by kriging* in: Enterprise Management-Theory and Practice, XIV International Science. Conference.: 22–23 Nov. 2012, Kraków,

[9] J.Opiła, T. Pełech-Pilichowski, *Visualisation of dispersed or irregular data*, Zeszyty Naukowe WSEI, Kraków, nr.8., 2012, pp. 151-178, ISSN 1734-5391

[10] gnuplot, http://www.gnuplot.info, accessed 2015-02-09,

[11] http://www.gnuplot.info/demo, gnuplot gallery for gnuplot v. 5.0, accessed 2014-12-28,

[12] J. E. Lennard-Jones,, *On the Determination of Molecular Fields*, Proc. R. Soc. Lond 1924, A 106 (738): 463–477,

[13] Inkscape, v.0.91.1, www.inkscape.org., accessed 2015-03-21

[14] Blender, v.2.73a, www.blender.org, accessed 2015-02-09

[15] Paltanavi-Źius A., *Hybridized Orbital Preview*, HOPV.exe, 2006, http://www.adomas.org/hopv/, accessed 2015-02-09,

[16] *Quantum Levitation*, http://www.st-andrews.ac.uk/~ulf/levitation.html, accessed 2015-02-09,

[17] VTK, *Visualization Toolkit*, http://www.vtk.org/, accessed 2014-06-27,

[18] *Diderot* http://diderot-language.cs.uchicago.edu, accessed 2015-01-27

[19] compiler for the Standard ML '97 programming language, http://smlnj.org/, accessed 2014-06-27

[20] Shadie project, https://code.google.com/p/shadie/ (foll. from Diderot home page) accessed 2015-02-04,.

[21] M. Hašan, *Short but Comprehensive Introduction to Shadie*, http://miloshasan.net/Shadie/shadie.pdf, accessed 2015-02-04,

[22] M. Hašan, *Volume Rendering of Dosimetric Distribution and Biological Response from 3D/4D Treatment and Delivery*: http://miloshasan.net/Shadie/aapm-talk.zip, accessed 2015-02-04

[23] BioBlender, http://www.bioblender.eu/, accessed 2014-06-27,

[24] RasMol, http://www.openrasmol.org/, accessed 2014-06-27,

[25] ParaView, http://www.paraview.org/, accessed 2015-02-09

[26] P. Bourke, http://paulbourke.net/miscellaneous/interpolation, accessed 2015-03-23,

[27] R. Wagner, *Multi-Linear Interpolation*, http:\\bmia.bmt.tue.nl\people\BRomeny\Courses\8C080\Interpolation.pdf, accessed 2015-03-23

Freeform spatial modelling using depth-sensing camera

Mario Volarević, Petar Mrazović, Željka Mihajlović
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
mario.volarevic@gmail.com, petar.mrazovic@gmail.com, zeljka.mihajlovic@fer.hr

Abstract - This paper proposes a novel 3D direct interaction modelling system for manipulating voxel based objects using currently available 3D motion sensing input devices such as Microsoft Kinect. The guiding principle while developing the system was to imitate natural human modelling behaviour and provide a real life experience of 3D object manipulation, inspired by the techniques used in modelling clay. Properties of a functional prototype application are presented and software architecture of the created tool is analysed. Descriptions of newly developed algorithms are included, grouped into several categories. Visualization algorithms are used for defining properties and creating usable modelling mass from volumetric models. On the other hand, algorithms related to object recognition and human computer interaction include various techniques for depth segmentation, contour detection, finger recognition and virtual control with gestures.

I. INTRODUCTION

Computers of various sizes and forms surround us in everyday life and it is hard to imagine life without them, but in some cases interaction is still somewhat unintuitive because methods of indirect control are used. Usually, we first have to learn how to use the controller and only then can we perform the desired task. Notable examples include using a mouse to control a pointer and activate onscreen actions or using gamepads or joysticks to control the environment with different buttons and sticks. Rapid development of new and interesting technologies related to human-computer interactions encouraged us to research and develop a novel interaction model which could improve our experience when interacting with 3D objects in a computer generated world. The key motivation behind technologies which use direct manipulation is not only to naturally represent human actions, but also to promote the notion of people performing tasks themselves, and not through some intermediary, like a computer. Problem with unnatural controls is especially apparent when doing 3D modelling since usually 2D controllers are used for manipulating objects in 3D world. One of the possible solutions for performing these tasks in a more natural way will be presented in this paper. Final goal is to have a functional prototype which would be more like a real life experience, for instance, as you would do clay modelling. Therefore, suitable hardware for direct 3D control has to be found and accompanying software support has to be developed.

In the last couple of years situation with direct control in consumer devices is getting better because smartphones and tablets became widely available and the main method of interaction is via touchscreen surfaces. Touch interfaces allowed users to use movements from real life (like translations) to be directly mapped to corresponding actions for onscreen elements. There are some programs like Autodesk 123D Sculpt on iOS which use this kind of interaction but unfortunately, with current technology, there is no depth information so operations like extruding or making holes in the model feel unnatural and don't provide good feedback. Reconstruction of depth information can be somewhat alleviated by adding gyroscopes and accelerometers which provide limited 3D movement recognition. One of the best examples of such controller is Nintendo Wiimote because of its high precision, wide availability, user friendliness and popularity. Drawback is that the controller has to be held in a user's hand which means that fingers can't be used for modelling with a fine level of detail [1, 2]. Also, our goal is to provide a truly natural user interface without any additional devices that would be a burden to a user.

An alternative approach is offered by Microsoft with Kinect where a user's own body acts as a controller. The advantage in this case is the ability to use the full 3D range of movement "without" constraints[1] and have free hands to interact with the virtual world in a natural way, and that was one of the initial prerequisites. Another device with similar characteristics, which could be also suited for the 3D modelling methods described here, is Leap Motion Controller. Since the mode of operation and detection is similar compared to Kinect, it should be relatively easy to adapt the developed prototype for use with Leap.

This paper is divided into six sections. Section 2 investigates related work where relationship with our prototype is based on similar ideas or a similar controller was used. Section 3 consists of several subsections where technical implementation details of used algorithms are described. In section 4, both performance and functional results are presented and discussed. Section 5 provides insight into possible future development of the prototype and finally, section 6 concludes the paper.

II. RELATED WORK

During the development of our prototype, we focused our research on a combination of computer modelling and

[1] There are still some constraints like usable sensor range or speed with which the user moves the hands.

motion gesture interfaces. Our proposed 3D direct interaction modelling system was significantly inspired by the work of Cho et al. [3] presented in the paper under the title "Turn: A Virtual Pottery by Real Spinning Wheel" and later expanded journal article [4]. Cho proposes a digital pottery system that successfully reflects real-world gestures to their virtual counterparts. The study focuses on the specific aspect of clay modelling where the freedom of movement and possible creations are limited by a spinning-wheel modelling technique. On the other hand, the interaction model proposed in our paper tries to implement virtual freeform modelling and object manipulation without spatial restrictions or any predefined movements.

In the paper "Gesture Interface for 3D CAD Modelling using Kinect", Dave et al. [5] successfully bring the intuitiveness of motion gestures to a CAD environment using Kinect in combination with constructive solid geometry and deformations, which produces results similar to traditional clay modelling. However, the restriction of the proposed creative solution is that Kinect could be only used to control tools which are already present in standard CAD programs. Using their prototype as an inspiration, we tried to remove any kind of restriction that could be a potential obstacle in providing a real life experience of freeform modelling in its true sense. Oliver Kreylos [1,2] also created modelling prototypes with motion input devices, but these prototypes were using motion controllers like Razer Hydra and Wiimote. Oliver's presented work did not try to achieve direct interaction via natural user interfaces, but instead, it utilized benefits of motion controllers to develop highly accurate and usable modelling system. The evaluated results are very impressive, and therefore present great challenges to be overcome, compared to our modelling system with less accurate, object recognition based, methods, i.e. without motion controllers.

Finger recognition techniques used in this paper were partially based on the algorithms described by Daniel James Ryan [6] in his thesis. Although it was possible to directly use the algorithms as they were described in his thesis, we decided to adapt and optimize recognition algorithms to better accommodate our requirements where main reasons for modifications were strict performance requirements.

III. PROPOSED SOLUTION

Our solution can be divided into several distinct components which, when combined, achieve the final goal of having a real-time system which allows completely freeform 3D modelling that also imitates real life clay modelling behaviour. Basic components of the proposed solution are representation of the virtual clay, the finger recognition system and the visualization and interaction framework.

A. Virtual clay

The first component of the proposed solution defines the properties of the virtual modelling clay compound. An approach based on voxels was chosen because while analysing different alternatives for mesh representation this seemed like the most natural solution when dealing with multiple simultaneous extrusions and intrusions. The marching cubes algorithm [7] was used to render the volumetric model, but in a GPU accelerated variant, to make the prototype sufficiently fast to be executed in real-time. GPU acceleration could have been achieved using some GPGPU APIs, like OpenCL or CUDA, but since our goal was also to provide support for older computers, a more compatible approach using shaders was chosen. The method used for GPU acceleration of marching cubes was based on the guidelines from the NVidia's GPU Gems article [8]. The main idea is to load the model in a 3D texture and then use GPU shaders to process all cubes simultaneously. A version with multiple shader passes was used, which enabled moving as much computations as possible from the geometry shader to the vertex shader. Reducing the workload in the geometry shader was required because that shader is extremely slow compared to other shaders.

Two passes were used, where coordinates of the corners of the voxel were the initial input and rendered polygons were the final output. The first pass in the vertex shader calculates positions of the voxels and their polygon configurations. Voxel positions and polygon configurations are then sent to the geometry shader where empty voxel configurations are rejected and simple markers for relative positions of triangle vertices are generated. Up to five markers are generated since five is the maximum number of polygons per voxel. Markers are then used in the vertex shader of the second pass where real 3D world space polygons coordinates are calculated, as well as normals. All of these previous steps led to an optimization which improved performance because it reduced the workload on geometry shader. The shader just had to stream out newly generated vertices and did not have to calculate anything. Finally, pixel shader is used to calculate lightning and render the model with proper colours.

B. Finger recognition

The second component in the developed system is related to finger recognition and it is based on Microsoft Kinect combined with algorithms for detecting and tracking body and various joints. The official Kinect API provides support for obtaining 3D coordinates of major joints so hand positions could be easily found, but for coordinates of individual fingers custom segmentation and recognition algorithms had to be developed and this process can be divided into several stages. In the first stage Kinect API is used to roughly detect the centre of the hand, and then, we propose algorithm based on the coordinates of the detected centre hand points. The image around the detected points is segmented by depth and hand contours are extracted. Segmentation was based on the assumption that hands are always in front of the body so it could be performed more easily.

Since the obtained contour is saved in a 2D field, it has to be converted to a 1D aligned array so it could be used with a k-curvature algorithm that is used to detect fingertips. Conversion was done with a custom algorithm which collects neighbouring contour points using circular search pattern where starting search position for the next

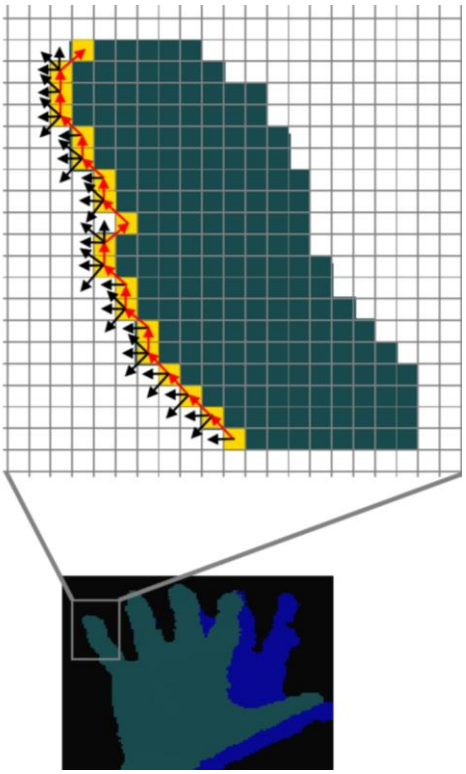


Figure 1. Execution of the contour detecting algorithm. Arrows show how the algorithm advanced while searching for the contour. Black arrows point to rejected background pixels and red arrows show the direction of contour pixels.

point depends on the positions of the current and previously found points. Figure 1 shows a simplified example of the contour-finding algorithm. In the extracted contour, k-curvature calculates the fingertip positions by taking three points from the array separated by factor k. Curvature and centroid calculations performed on the selected three points are used as features for detection of fingertips, as seen on Figure 2. The proposed algorithm works with the assumption that hands are mostly curved smoothly compared to the fingers, and centroid calculation is used to distinguish fingertips from valleys between fingers because these are the only places with sharp angles.

After finger positions are calculated, they have to be converted to Kinect's projection space coordinates so they are comparable with real world dimensions. Coordinate conversion was done using methods from Kinect API. After the final feature positions have been obtained, they were transferred to the interaction system where they were processed and converted into gestures.

C. Interaction system

Finally, there is a visualization step and interaction system that connects the other two, previously described, components and provides an input/output interface for the user. It is implemented using SharpDX Toolkit API, which is based on DirectX 11 and can be considered as an unofficial successor to Microsoft XNA Framework. SharpDX is currently the fastest managed DirectX APIs and is just 1.5x times slower than native code. It also forms a basis for many graphics engines, of which MonoGame is the most famous one.

Primary output is visual and is represented by a window where volumetric model with its bounding box is displayed. In the main window, there is also an abstract model which uses simple spheres to represent hands and fingers. Also, to ease the positioning in front of Kinect, because of the limited range and viewport, an additional sub-window is displayed, which shows depth and finger recognition data.

Hand gestures are used as primary input and were designed to be as intuitive as possible, but it still takes some time to get used to controls since there is no force feedback. Gestures are defined in the following way: if hands are not touching the model, it is possible to adjust the view by closing them (like grabbing an object) and then move, rotate and zoom it, and if they are inside the model or touching it, virtual clay is either added or removed on coordinates calculated based on the position of the fingertips. Some possible manipulations while using the program can be seen in the images in Figure 3. Additionally, keyboard controls can also be used for view manipulation, as a backup for cases when Kinect isn't connected or stops functioning properly due to bad lighting.

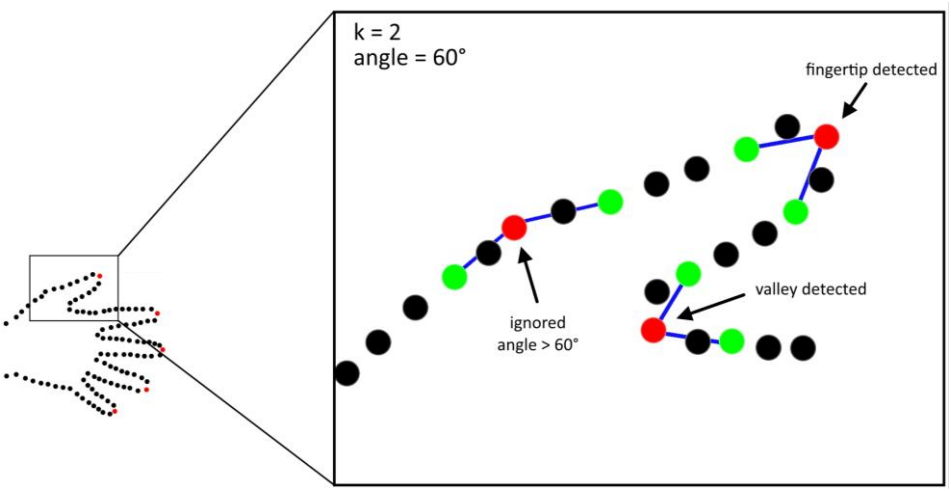


Figure 2. Simplified representation of the k-curvature algorithm. Fingertips are detected based on the angle between three points separated by factor k. Valleys as false positives are ignored because their centroid isn't located inside a finger

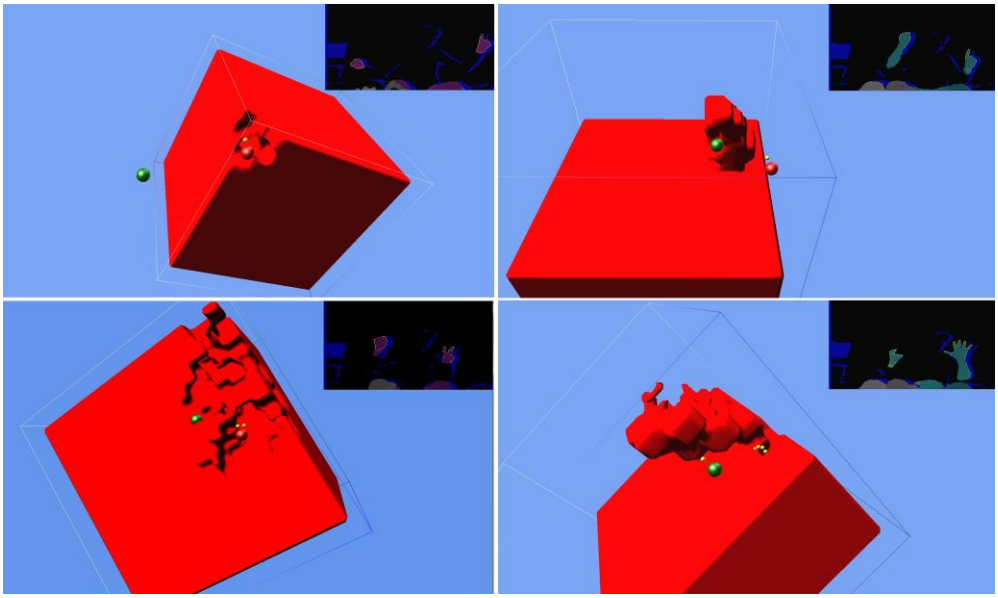


Figure 3. Several images of a working program where various intrusions, extrusions and view manipulations on a model can be observed

IV. RESULTS

The developed prototype application showed that results are promising, both considering performance and functionality. Even on relatively underpowered hardware, represented by Configuration 1, compared to other listed configurations, it is possible to track the fingers and render the model with reasonable quality in real-time with all, initially planned, features enabled. There were several hardware configurations on which this program was tested to show its scalability and also good compatibility, considering the variety of the used components. Table 1 lists relevant parts of the specifications for each of the configurations. Since the host part of the program is single-threaded and the main computation, which is multi-threaded, is executed on the graphics card, turbo frequencies will be listed for the CPUs, and the number of cores with their frequencies for the GPUs, so scaling can be more easily evaluated.

Performance results were measured on a cubic model with Kinect connected and finger recognition turned on. An example of the model and the pose used for evaluation are shown in Figure 4. Samples were taken for several resolutions to show which cases are usable in practice for different configurations and how the algorithm scales depending on the complexity. A numerical representation of the results and used resolutions can be seen in Table 2. In Figure 5, which shows graphical representation of the results, it can be seen that drawing speed, measured in frames per second, scales as expected depending on the GPU architecture and considering the small differences between CPU performances it can be concluded that this prototype is GPU bound.

While measuring performance results where Kinect was active were not much different than results with Kinect inactive. There was a slight performance hit, but it was observable only on smaller resolutions, which means that image processing for Kinect took at most 5-10 ms of GPU time. Memory usage was also observed in both cases and results showed that when Kinect is not connected, each time when model resolution is changed, it steadily increases from 50 MB to 320 MB, but when garbage collection activates, after all assets have been loaded, transferred to GPU and not needed on CPU side anymore, it drops to around 60 MB. With Kinect connected, memory usage is constant and floats around 200 MB, even after garbage collection has activated.

Functionality-wise, gestures and model manipulations are working well while Kinect detects user joints correctly, but quite often it does not. One of the cases when Kinect does not detect joints at all is when users leave the Kinect detection range. The detection range is limited to distances between 0.8–3.5 m. This usually happens when moving hands closer to Kinect to change farther parts of the model and it could be avoided if users would not stand too close to Kinect, but in that case another problem arises because Kinect's resolution is not high enough and the algorithm which detects fingers does not work well anymore. In our case, experiments showed

TABLE I. HARDWARE CONFIGURATIONS USED FOR TESTING

	Configuration 1	Configuration 2	Configuration 3
CPU	Core2Duo T7500, 2.4 GHz	Core i5-430M, 2.53 GHz	Core i7-3630QM, 3.4 GHz
GPU	Radeon HD 3650m, 120x600 MHz	Radeon HD 5650m, 400x550 MHz	Quadro K2000M, 384x750 MHz
RAM	4 GB DDR2, 800 MHz	4 GB DDR3, 1066 MHz	8 GB DDR3, 1600 MHz
OS	Win8.1 x64	Win8.1 x64	Win7 x64

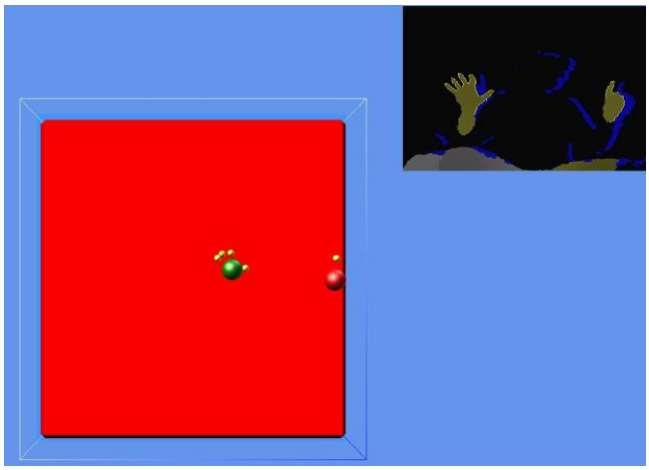


Figure 4. Example of the used benchmark scenario where simple cube was loaded while finger recognition was active

that the best distance is if the user is positioned at around 1.2m. Another case is when Kinect confuses other parts of environment or other joints as hands (usually when a user moves too fast) in which case wrong image fragment is used when trying to detect fingers. In general, most problems happen because Kinect does not have high enough temporal and spatial resolution and all of these problems should disappear in future implementations since use of other controllers is already planned.

Since there are no directly comparable products on the market which offer gesture based modelling, we subjectively evaluated the usability and satisfaction of users after using our prototype application. Although our proposed solution, in its current state, lacks the precision of the professional 3D modelling tools, feedback collected from the users proves that the concept of intuitive, computer based, 3D modelling is feasible. Simplicity and ease of use of the developed prototype application became apparent when users without any modelling experience managed to quickly figure out the gestures and started modifying the preloaded model.

TABLE II. PERFORMANCE DEPENDING ON THE MODEL SIZE

Cube length	Configuration 1	Configuration 2	Configuration 3
(N vertices)		*(FPS)*	
50	127	358	670
100	24	57	124
150	7	20	42
200	4	8	18
250	2	4	10

V. FUTURE WORK

As an interesting future research topic and potential improvement, we consider changing the controller. Since Kinect was designed to recognize larger body parts, competitive devices with finer recognition capabilities would be a better choice for this application, such as Leap Motion Controller or the new Kinect v2. Another interesting feature, which would improve immersion and usability, would be an interactive perspective projection transformation based on the user's head position which is easily obtainable through Kinect API. The idea is that the virtual view responds to the user's head movement, and we believe this feature would positively affect the experience of the virtual 3D world. There is also high probability that the mentioned feature would be also implemented with a virtual headset, such as Oculus Rift, since it has higher availability compared to other similar devices.

There is also some room for improvement regarding the application's performance. One of the optimizations is related to eliminating the redundant vertices created by the marching cubes algorithm since the edges on the neighbouring cubes share the same vertices. The idea would be to create only a single instance of each vertex and then use indexing to describe where each vertex is located. Another useful optimization would be partial dynamic model loading, especially when dealing with high resolution volume models. Since the complexity of the algorithm is high, this feature could considerably

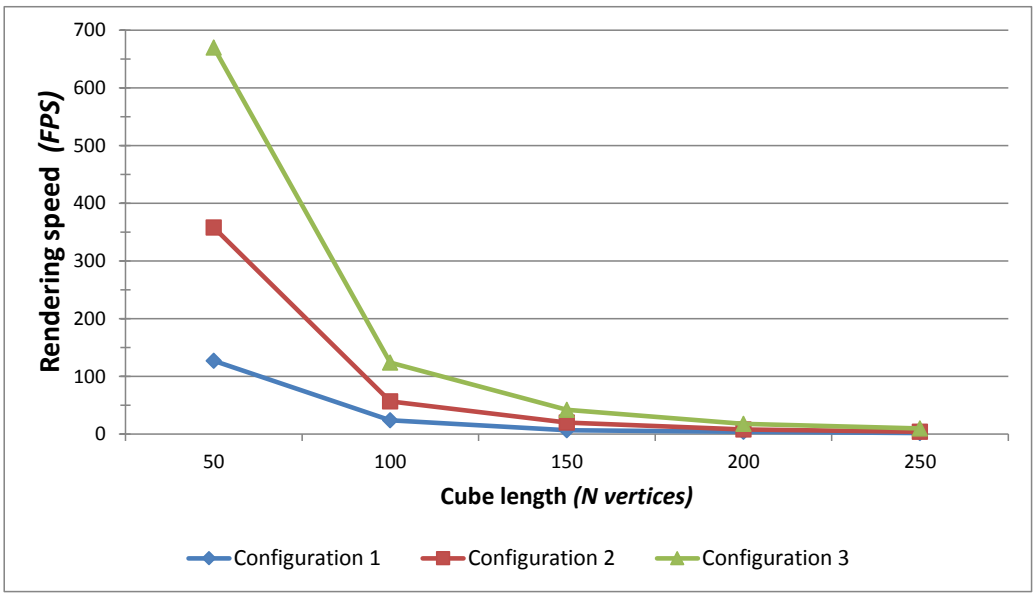


Figure 5. Graphical representation of the benchmark results

affect the performance of the program when used on large models. An implementation would probably be similar to the culling technique where a model could be divided into smaller regions and when comparing depths those that are further from the camera would not be processed.

VI. CONCLUSION

Natural user interface technologies can be more than just a platform for gaming and entertainment because they possess a great potential in revolutionizing human-machine interactions. In these types of interface designs 3D motion sensing input devices like Microsoft Kinect have proven to be a suitable and valuable hardware choice. However, the effectiveness of the hardware heavily depends on the accompanying software which still does not follow any generally accepted interaction model. This article proposes a solution to this problem in the form of a functional prototype that tries to imitate a real life experience of traditional human skills such as clay modelling.

A functional prototype presented here forms the basis for the future versions with an improved interaction system, which could provide tremendous opportunities for addressing various problems, and also opens new frontiers and new challenges in user interaction with virtual objects.

REFERENCES

All web links have been tested and functional on February 8, 2015.

[1] Oliver Kreylos' Homepage, Wiimote Hacking. http://idav.ucdavis.edu/~okreylos/ResDev/Wiimote/index.html (posted October 2007)

[2] A developer's perspective on immersive 3D computer graphics. Virtual clay modeling with 3D input devices. http://doc-ok.org/?p=493 (posted March 2013)

[3] Cho, S., Heo, Y., & Bang, H. (2012, August). Turn: a virtual pottery by real spinning wheel. In ACM SIGGRAPH 2012 Posters (p. 11). ACM.

[4] Cho, S., Baek, D., Baek, S. Y., Lee, K., & Bang, H. (2014). Three-dimensional Volume Drawing on a Potter's Wheel. IEEE Computer Graphics and Applications, (1), 1.

[5] Dave, D., Chowriappa, A., & Kesavadas, T. (2013). Gesture interface for 3d cad modeling using kinect. Computer-Aided Design and Applications, 10(4), 663-669.

[6] Ryan, D. J. (2012). Finger and gesture recognition with microsoft kinect. http://brage.bibsys.no/xmlui/bitstream/handle/11250/181783/Ryan%2c%20Daniel%20James.pdf?sequence=1

[7] Lorensen, W. E., & Cline, H. E. (1987, August). Marching cubes: A high resolution 3D surface construction algorithm. In ACM siggraph computer graphics (Vol. 21, No. 4, pp. 163-169). ACM.

[8] Geiss, R. Generating Complex Procedural Terrains Using the GPU. In GPU Gems 3; Nguyuen, H., Ed.; Addison Wesley Professional, USA, 2008; pp 7-39.

Seismic data visualisation

M. Ivančić*, Ž. Mihajlović*. and I. Ivančić**
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** University of Zagreb, Faculty of Science, Department of Geophysics, Zagreb, Croatia
matejka.ivancic@gmail.com, zeljka.mihajlovic@fer.hr, iivancic@gfz.hr

Abstract - The focal mechanism of an earthquake and fault plane solutions provide insight into the geology of an area. The objective of this paper is to give a means of better understanding fault movement when an earthquake occurs. In this paper we present an interactive 3D visualisation of possible fault plane solutions, which are usually represented as 2D "beach-ball" plots, with their respective block models. Additionally a 3D alleviation display of an area, with respective beach ball earthquake representations, provides a more in depth picture of seismological maps under the earth's surface. The program solution allows a better grasp of space, which is useful to students of seismology and anyone who wants to visualize fault movement of an earthquake.

Key words: interactive visualisation, fault plane, focal mechanism,

I. INTRODUCTION

Seismology is the scientific study of earthquakes. An earthquake is the result of a sudden release of energy in the Earth's crust in the form of seismic waves. The most common earthquakes are tectonic earthquakes, caused by rupture of geological faults. Faulting is relative movement of two geological blocks (walls), by creating a new fault or shifting along an existing fault plane [1].

The full process of an earthquakes hypocentre is very complicated and difficult to describe exactly. However, some of the properties observed on seismograms can be described by mathematical models of a point source with the conditions to observe phenomena far from the source and in the short time after the earthquake begins [2].

When the earth suddenly moves during an earthquake, energy is released and seismic waves radiate from the Earthquakes hypocentre in all directions [1]. The waves travel through the Earth's interior towards the surface. There are two types of waves: the faster longitudinal waves (P-waves) and slower transverse waves (S-waves). The waves are recorded by seismographs, and correspond to up and down movements on the seismogram [3].

II. FAULT MECHANISM

In the early 20th century, by observing the pattern of first motions (first arriving P waves), seismologists noticed that the first movement of a seismographs pen (recorded direction of the first shift of a geological block)

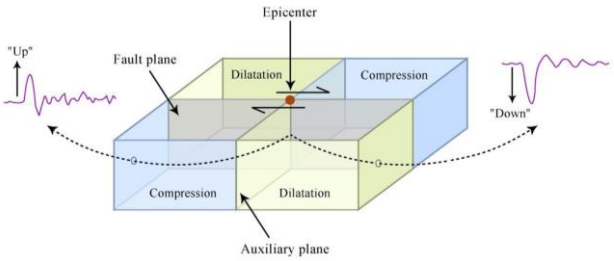


Figure 1. First motions observed from different directions[2].

depends on where the station is located relative to the epicentre [2].

The fault walls move in different directions which creates a different P wave polarity (seen in Fig. 1). If a block moves towards a seismological station, the seismograph will record the first shift as a compression with an up movement, but if a block moves away from the station, it is called dilatation and it will be recorded with a down movement. On a map, these areas can be divided by lines into four quadrants, each of which has a different polarity, but without further information there is no way to decide what type of a fault it is [3].

Seismologists examine the first motion of the P waves from local and remote stations to reconstruct the initial forces that triggered the event [1].

A. *Fault types*

There are three main fault types [2]:

- strike-slip, where the offset is predominately horizontal, parallel to the fault trace.
- dip-slip, offset is predominately vertical and/or perpendicular to the fault trace.
- oblique-slip, combining significant strike and dip slip.

An earthquake is described by the faults orientation and shift direction (seen in Fig. 2) [1]:

- Strike (φ) is the azimuth of a line created by the intersection of a fault plane and a horizontal surface, 0º to 360º, relative to North

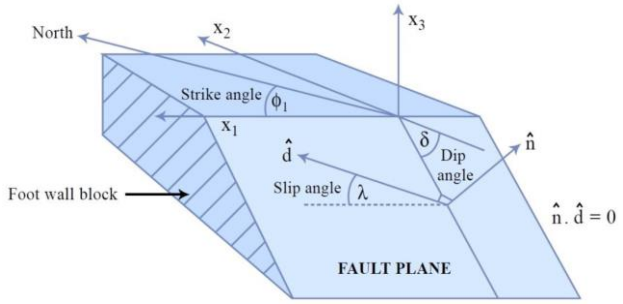


Figure 2. Fault geometry: fault orientation determined by the direction of the fault, dip angle and rake angle [3].

- Dip (δ) is the angle between the fault and a horizontal plane, 0º to 90º
- Rake (or slip, λ) is the direction a hanging wall block moves during rupture, as measured on the plane of the fault relative to fault strike, -180º to 180º (CCW)

The orientation is defined such that a fault dips to the right side. The hanging wall block is therefore always on the right, and the foot wall block on the left [1].

If the fault breaks the surface, the angles can be determined by geological methods, but in most cases, earthquakes do not break the surface and seismological data must be used to determine fault plane solutions [3].

B. Fault mechanism in practice

After an earthquake occurs, seismologists create graphics of focal mechanisms by mapping the P wave information on a focal sphere, informally referred to as a *beach ball*, to show the faulting motions that produced the earthquake. Focal mechanisms, also known as fault-plane solutions, are based on the direction of the first arriving P wave [1-4].

The orientations of the first P waves are gathered from a number of stations and plotted onto the focal sphere. If there are sufficient observations, two planes are drawn to divide the compressive from the tensional observations and these are the nodal planes (seen in Fig. 3). By convention the compressional quadrants are color-filled and the tensional (dilatation) left white [3,5].

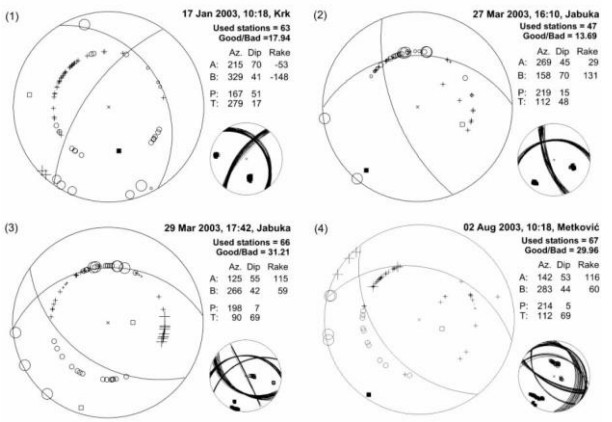


Figure 3. Focal mechanism examples for Croatian earthquakes that took place in Middle Dalmatia, the island of Krk and Jabuka [6]. The crosses represent compression and circles dilatation.

Figure 4. Example of ambiguity

The fault plane responsible for the earthquake will be parallel to one of the nodal planes, the other being called the auxiliary plane. Unfortunately it is not possible to determine solely from a focal mechanism which of the nodal planes is in fact the fault plane [3, 5].

A typical example can be seen in Fig. 4. A strike-slip fault with shift direction towards north (rake $\lambda = 0º$, strike $\varphi = 0º$) is indistinguishable from the strike-slip fault with shift direction towards west (rake $\lambda = 180º$, strike $\varphi = 90º$) [7].

To remove the ambiguity other geological or geophysical evidence is needed. There are a number of methods such as the terrain method (knowing the geology of the area will help identify the most likely fault geometry) or by comparing the two fault-plane orientations to the alignment of small earthquakes and aftershocks [7].

III. IMPLEMENTATION

Interpreting fault plane solutions can be difficult, especially because they are generally given as two-dimensional stereographic projections of the lower hemisphere of a focal mechanism sphere. In order to make it easier to understand and visualize, we present an interactive three-dimensional visualisation of faults and their matching beach balls.

The program solution offers the possibility of entering fault parameters (strike, dip and rake) with the option of seeing a matching beach ball as a three-dimensional sphere and the stereographic projection that is usually used on seismic maps. Furthermore a 3D relief display of Croatia, with respective beach ball earthquake representations of selected earthquakes provides a more in depth picture of seismological maps under the earth's surface.

To make our solution easily available for learning purposes, the technologies used were WebGL with the Three.js JavaScript library.

A. Fault visualisation

To simplify an earthquake process, geological blocks are displayed as cuboid bodies (seen in Fig. 5). A transparent plane represents the fault plane and a helper

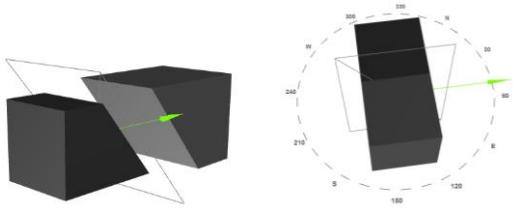


Figure 5. Fault visualisation

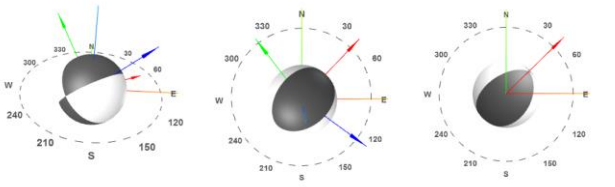


Figure 6. Beach balls with local and global axis

vector shows the direction of the hanging wall. Additionally, a compass points around the fault blocks helps to show the azimuth relative to north. The hanging wall shift is emphasized for the sake of demonstration.

B. *Focal sphere visualisation*

The focal sphere, informally referred to as a beach ball, is a sphere divided into four parts. The color fields describe a shift from the focus (compression), while white fields describe towards the focus (dilatation).

For better understanding, the beach ball is shown both as a 3D interactive sphere and as a generally used two-dimensional stereographic projection of the lower hemisphere.

C. *Fault visualisation with the focal sphere*

To better illustrate the focal mechanism solutions, we take as an example the earthquake that took place near the island Jabuka on 27th March 2003 (which focal mechanism solution can be seen in Fig.3) and show both of the possible solutions as a block model with the respective beach balls. The ambiguity between the two possible physical solutions obtained from seismic data and seen in Fig. 6 can now easily be seen from the presented block model and the corresponding focal spheres.

D. *Relief map*

One of the methods to remove the ambiguity is the terrain method, where knowing the geology of the area will help identify the most likely fault geometry. For this

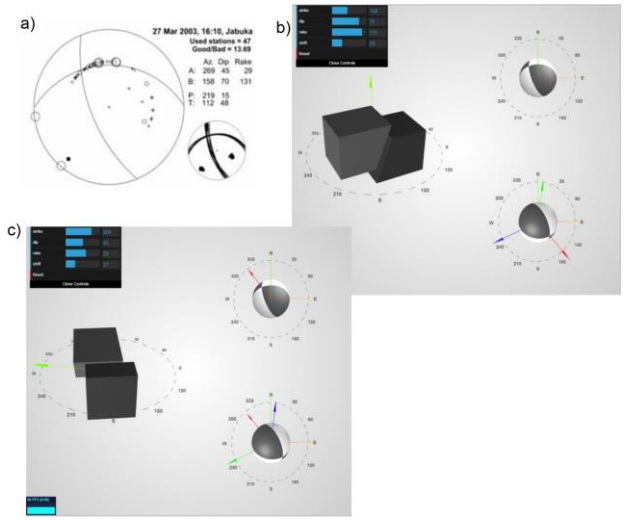


Figure 7. Visualized data from the earthquake that occurred near the island of Jabuka on 27th March 2003

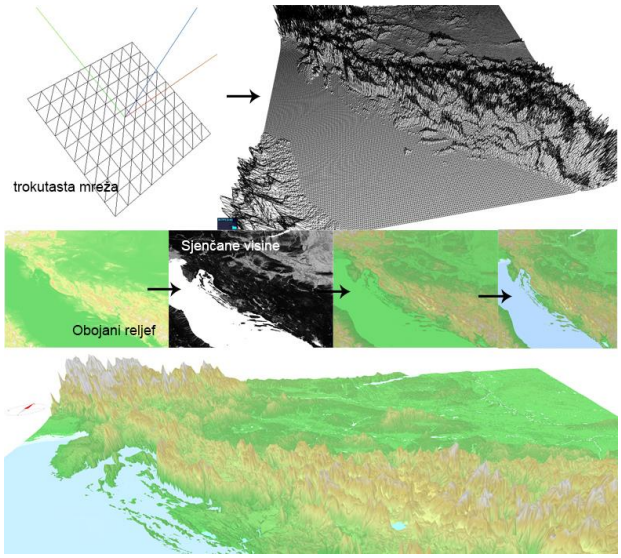


Figure 8. The process of creating the terrain map

purpose we made a map of Croatia on which we placed focal mechanism solutions of representative earthquakes from the Archives of the Department of Geophysics [8].

The height map and terrain texture were made based on data from OpenTopography [9-13] and Corine Land Cover dataset from European Environment Agency (EEA) [14]. The area covered is from 42 ° to 47 ° N and from 13 ° to 20 ° E. The process and final map can be seen in Fig. 8 [15-18]. The relief was accentuated for demonstration purposes. Selecting a beach ball shows information about the earthquake (date, magnitude, coordinates and depth), the block model and corresponding beach ball (seen in Fig. 9).

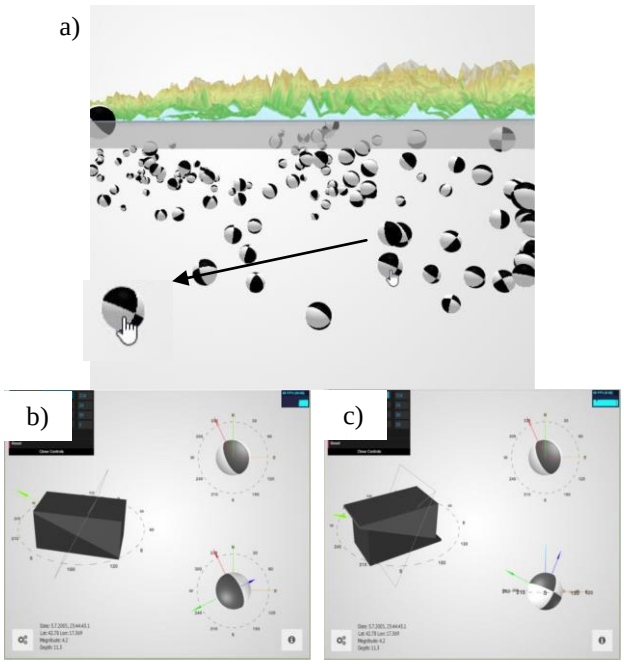


Figure 9. a) Selecting a particular earthquake shows b) the focal mechanism of the earthquke, where one can change the view angle and visualise the shift of the fault plane in space. This example shows an eathquake that took place on 23rd March 2005 at 23h 44m, magnitude 4.2 ($\varphi = 334°$, $\sigma = 24°$ and $\lambda = 90°$).

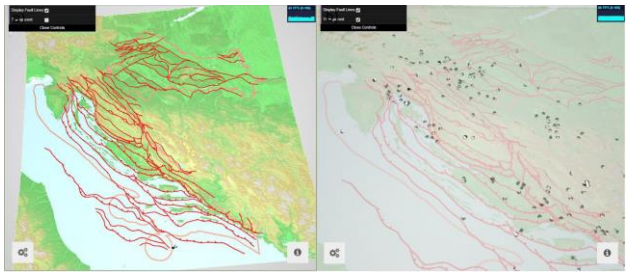


Figure 10. The map with the Geological map of Croatian faults (shown with full opacity and partly transparent so the beach balls can be seen from all angles)

Furthermore, we add a second texture showing the Geological map of Croatian faults (seen in Fig. 10) [19].

Fig. 11 shows how this allows for a better visualisation of fault plane solution spatial distribution.

E. User interface

The program was implemented using WebGL and Three.js JavaScript library as a website [20]. The menu is in the left bottom corner, where the user can change between the focal mechanism visualisation and catalogue of Croatian earthquakes in map view. It takes a little time for beach balls representing the earthquakes to load and apply needed transformations, as there are quite a number of them.

IV. CONLUSION

Focal mechanisms are determined based on earthquake analysis from a number of seismic stations. Its importance lies in the geological and seismological information that is gained about an earthquake's physical process. Fault plane solutions allow insight into the geological structure of an area.

The main objective of this paper was to give a means of better understanding fault movement by giving an interactive 3D visualisation of possible fault plane solutions (usually represented as 2D plots) with their respective block models. The program solution allows a better grasp of space, which is useful to students of seismology and anyone who wants to visualize fault movement of an earthquake.

Additionally a 3D relief display of Croatia with respective beach ball representations provides a more in depth picture of seismological maps under the earth's surface.

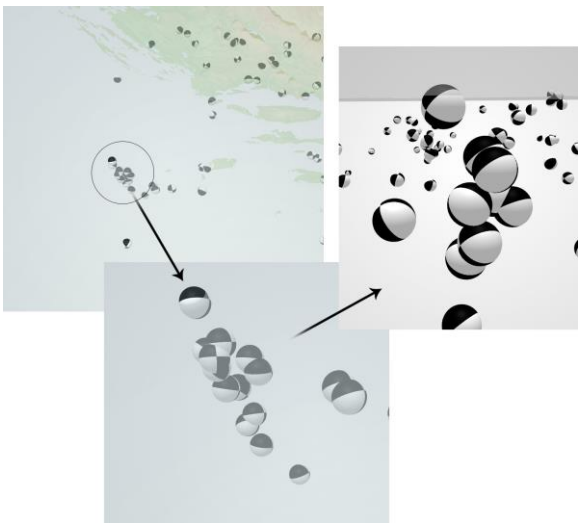


Figure 11. Different viewpoints of Jabuka epicenter area

REFERENCES

[1] M. Herak and I. Dasović, "Selected chapters of seismology" in Selected topics Geophysics, Department of Geophysics, Faculty of Science, University of Zagreb ,2014.

[2] S. Stein and M. Wysession, "An introduction to seismology, earthquakes, and earth structure", Oxford, Blackwell Publishing, 2003.

[3] J. Havskov and L. Ottemöller, "Routine Data Processing in Earthquake Seismology", Springer Science+Business Media B.V., 2010.

[4] Focal Mechanism, Incorporated Research Institutions for Seismology, [online], http://www.iris.edu/hq/programs/education_and_outreach/animations/ (Accessed: 1 Februry 2015).

[5] Focal mechanism, Wikipedia, [online] 22 September 2014, http://en.wikipedia.org/wiki/Focal_mechanism (Accessed: 2 Februry 2015).

[6] I. Ivančić, D. Herak, S. Markušić, I. Sović, I. and M. Herak, Seismicity of Croatia in the Period 2002 – 2005, Geofizika, Vol. 23, 87– 103., 2006

[7] E. A. Okal, "Earthquake focal mechanisms" in: Encyclopedia of Solid Earth Geophysics, ed. by H. Gupta, Earth Sciences Series, Springer, 2011, pp. 194-199.

[8] Focal mechanism of earthquakes in Croatia and neighboring areas (1909-2013), Archives of the Department of Geophysics, Faculty of Science, University of Zagreb

[9] Farr, T. G., i M. Kobrick, Shuttle Radar Topography Mission produces a wealth of data. Eos Trans. 2000, AGU, 81:583-583.

[10] Farr, T. G. et al. The Shuttle Radar Topography Mission, 2007, Rev. Geophys., 45, RG2004, doi:10.1029/2005RG000183.

[11] Kobrick, M., On the toes of giants--How SRTM was born, 2006, Photogramm. Eng. Remote Sens., 72:206-210.

[12] Rosen, P. A. et al., Synthetic aperture radar interferometry, 2000, Proc. IEEE, 88:333-382. http://dx.doi.org/10.5069/G9445JDF.

[13] Topography material provided OpenTopography Facility with support from the National Science Foundation under NSF Award Numbers 1226353 & 1225810.

[14] Corine Land Cover dataset, European Environment Agency, 22.8.2012., http://www.eea.europa.eu/data-and-maps/data/corineland-cover-2000-clc2000-seamless-vector-database-4 (Accessed 28.4.2014.)

[15] Sandvik, B., Thematic mapping blog, Terrain building with three.js, 9.10.2013., http://blog.thematicmapping.org/2013/10/terrain-buildingwith-threejs.html, (Accessed 28.4.2014.)

[16] Sandvik, B., Thematic mapping blog, Textural terrains with three.js, 14.10.2013., http://blog.thematicmapping.org/2013/10/textural-terrains-withthreejs.html, (Accessed 28.4.2014.)

[17] Sandvik, B., Creating color relief and slope shading with gdaldem, 30.6.2012., http://blog.thematicmapping.org/2012/06/creating-colorrelief-and-slope-shading.html, (Accessed 28.4.2014.)

[18] Sandvik, B., Land cover mapping with Mapnik, 2.7.2012., http://blog.thematicmapping.org/2012/06/creating-color-relief-andslope-shading.html, (Accessed 28.4.2014.)

[19] E. Prelogović, B. Tomljenović, M. Herak and D. Herak, Geologycal map of Croatian faults, according to the Map of active faults, 2006, COST – 625, unpublished

[20] M. Ivančić, Seismic data visualisation, M.S. thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, 2014, http://www.zemris.fer.hr/predmeti/ra/Magisterij/14_Ivancic/

Situation and location awareness in harsh environment

Miklos Kozlovszky[1,2], Daniela Zavec Pavlinić[3], Andreja Oder[4] ,Gábor Fehér[5], Pál Bogdanov[2]
[1] Óbuda University/John von Neumann Faculty of Informatics, Budapest, Hungary
[2] MTA SZTAKI/Laboratory of Parallel and Distributed Computing, Budapest, Hungary
[3] Titera Technical Innovative Technologies Ltd., Slovenia
[4] Prevent&Deloza Ltd., Celje, Slovenia
[5]OMT-LAB, Budapest University, Hungary

{kozlovszky.miklos@nik, bogdanov.pal@biotech}.uni-obuda.hu, info@titerad.com, andreja.oder@prevent-deloza.si, gabor.feher@omt-lab.hu

Abstract - Enhanced environment awareness is a key element of life protection in dangerous situations. Thus the Personal Protection Equipment (PPE) clothing industry is forced in the direction to develop highly innovative "intelligent" products, and equip their products with sensors and actuators combined with smart technologies. Besides basic functions of protection, thermal and ergonomic comfort, these new generation protective garments have additional functions such as monitoring and/or alarming for prevention in case of emergency situation. Such systems are costly because even if the clothing and engineering parts of the development are feasible, their combination is far from trivial. The different system parts have different requirements, and these requirements are sometimes not compatible with each other. Parameters such as cost, bulkiness, accuracy, independency, scalability, robustness are key parameters, which can cause high acceptance or usage prevention rapidly by the end-users. Our research is focusing on the seamless integration of the various subsystems to develop effective and user friendly, intelligent situation aware personal protection systems. Our solution helps to receive accurate information about user performance and activity conditions. The main goal of our work is to adapt the already existing monitoring solution and provide it to people working in harsh environment. The paper describes the results of our development with the description of the main functionalities.

Keywords

Sensor data acquisition, location and situation awareness, health monitoring

I. INTRODUCTION

The surrounding micro end macro environment changes all the time. Measurements convert the environment parameters into information, what we can process and evaluate to better understand the environment, and predict its parameter changes. Enhanced environment awareness (situation, location, etc.) is a key element of life protection in dangerous situations. Thus the Personal Protection Equipment (PPE) clothing industry is forced in the direction to develop highly innovative "intelligent" products, and equip their products with sensors and actuators combined with smart technologies. Design requirements such as cost, bulkiness, accuracy, independency, scalability, robustness are key parameters, which can cause high acceptance or usage prevention of location sensor devices rapidly [1]. We are designing and developing an intelligent Personal Protective System (PPS) integrated with a data acquisition (DAQ) system and we are using enabling technologies to collect location information and a wide range of environment information. In this paper we show how our first prototype of environment and location monitoring is designed and developed. In this paper we provide a detailed overview how information acquisition is realized within the PPS to monitor and provide situation and location awareness to the PPS user. Besides basic PPE functions of protection, thermal and ergonomic comfort, the new generation protective garments have additional functions such as monitoring and/or alarming for prevention in case of emergency situation. This paper shows how we have built up our biosignal, micro and macro environment monitoring solution to support human operations in harsh environments using intelligent PPS. In the second section we provide an overview how our system fits into the generic data acquisition concept. In the third section an overview is given about sensor data integration and sensor data fusion and some related problems described. In section four we give sensor fusion.

II. STATE-OF-THE-ART

A. DAQ

1) Single sensor DAQ

A single sensor measurement is the simplest method to receive information about the environment. In such case

the accuracy of the sensor determines the maximum accuracy of the measurement.

2) Multi-sensor DAQ

In case of multi-sensor measurements the inclusion of additional sensor provides different effect on the measurement and on the accuracy as well.

Multi-sensor data acquisition requires by definition complex solutions. Sensors can be operated inside as a homogenous set, or they can vary in types (multi-modal), resolutions, they can differ in monitoring target, temporal

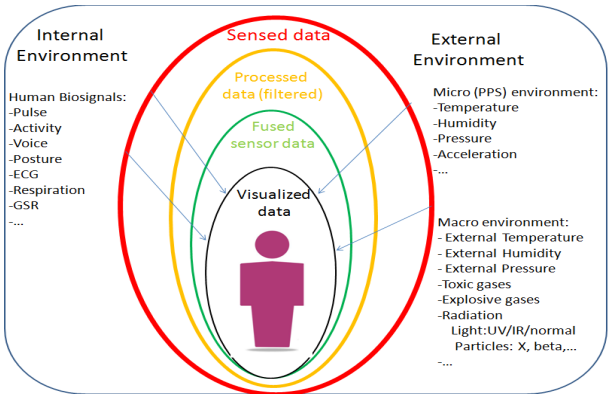


Figure 1. Multi modal DAQ of the PPS

or spatial parameters. Formal frameworks are available for multi-sensor data fusion. [2]

a) Multi-modal DAQ

Biosignal, micro and macro environment DAQ

Biosignal, micro and macro environment data acquisition is achieved using different type of sensors. The possible sensor configurations can be differentiated according to the Durrant-Whyte [3] classification, sensors can be: complementary, competitive or cooperative. Internal environment parameters of the PPS user are acquired by biosignal sensors, such as pulse sensors, activity sensors, ECG, GSR sensors, respiration sensors and posture sensors. The external environment is divided to two main areas. The PPS environment information is acquired by sensors attached to the internal layer of the PPS such as: temperature, humidity, pressure, acceleration sensors. The environment which is located outside the PPS is monitored by external sensors, such as: toxic/explosive gas sensors, radiation sensors (IR, gamma, beta, etc.) (in Figure 1.) Audio and video streams are also captured about the environment in real-time.

Location DAQ

Location information acquisition is a challenging task [4]. The location sensors are ranging from low accuracy sensors (with almost global coverage) to high accuracy (short range) solutions. The proposed localization solution based on a combination of different technologies. Most of the technologies are using radio wave communication. We are tracking the position of the PPSs, combining absolute and relative positioning techniques. Relative positioning is measuring signal strength difference; thus works with relative distance and direction of arrival measurements. Absolute measurements are done when signal from known position fixed elements can be detected. Fixed elements are the satellites emitting GPS signals and fire-engines at known locations providing ranging functions in the UHF band. We acquire periodically from the different sensor devices location information. For outdoor environment (if available) the system relies on GPS satellite signals and GSM signals, both techniques have large coverage, but their accuracy is poor. In indoor environments the distance measurements are more suitable, and positions are estimated from triangulations among the different signal senders. The distance measurements are done parallel with UHF and UWB radio technology. UHF distance measurements provides below 5 m accuracy in a long range (200 m-1500 m depending on environment) while UWB distance measurements give an excellent 1 m accuracy in a shorter range (20 m-100 m depending on environmental conditions). For offline movement tracking and for the support of position estimation in extreme conditions we acquire also Inertial Measurement Unit (IMU) information, which provides multi-axis acceleration information.

3) Remote DAQ in harsh environment

Remote data acquisition systems can effectively collect sensor data and transfer to remote locations. Autonomous data acquisition requires significant power supply and data storage, its transparency (location independency) and user friendliness is crucial from usability point of view.

There are already existing solutions to acquire sensor data in harsh environment such as simple sensing capability from RECCO, or pilot solutions for Firefighters and Civil Emergency Workers by ProeTEX (IP Firefighter), or product level solutions from Viking Life-Saving Equipment A/S and Globe Manufacturing Co. [11, 12, 13, 14].

III. DATA EVALUATION AND IDENTIFIED PROBLEMS

A. Definitions

Multi-sensor integration: The data integration is a usage of sensor data set to realize a particular task.

Multi-sensor data fusion: The data fusion is a process of combining input data into a common representation format. Basically multi/sensor data fusion can improve sensing in the following four different ways: Representation, Certainty, Accuracy, Completeness [5] Multi-sensor data fusion system classification can be done in various ways, one good example is Dasarathy's [6] I/O Data fusion model.

B. Performance, accuracy and scalability issues

The large number of data sources, with high sampling rates can produce huge data quantities, what could be difficult to transmit, to process, to visualize or even to store. Additionally multi-sensor data processing requires complex solutions for data representation, sensor value normalization and sensor data fusion. More sensors can provide more sensing information however one should be careful here, because this is not directly lead to any gain.

Due to the multiplicative effect caused by measurement errors/problems a single sensor can outperform a multi-sensor data fusion system (e.g.: catastrophic fusion).

C. Environment awareness

To gain environment awareness we not only just need to collect information about the surrounding environment, but we need to process it (fuse, filter, etc.) and merge it with our already existing knowledge. During sensor data processing we need to provide a common representation format of the sensor data.
Symbolically we are representing a sensor observation by adding more parameters to the generic five-tuple [5] equation. Our modified equation is the following:

$$O =< E, S, x, Tm, Tp\{\}, Y, Yu, \Delta Y > \quad (1)$$

Where E is the entity name (Sensor ID), S is the sensor type, x is the spatial location (position in space),T_m is the time instant when the measurement was occurred, T_a is the list of time instances when the measurement was processed, Y is the measurement (the physical property as measured by the sensor element),Y_u is the measurement unit, Δy is the error set which influences y caused by the measurement setup (including measurement errors, calibration errors, etc.).
The used biosignal, micro and macro environment sensors are complementary and cooperative type sensors.
The primary fusion algorithm converts the received sensor information and construct a common view in the time domain. Beside the accuracy parameter, all sensor data has a decay parameter as well.

D. Location awareness

Location awareness is a crucial to effectively work or even to survive in harsh environment.
Main requirements here are the specification of the followings:

- absolution location of the individual sensor owner
- location of other objects and entities
- relative directions and distances of other objects and entities

We are using a primary fusion algorithm to convert the received location information. We are using competitive and cooperative sensors and constructing a common coordinate system. The different technologies provides different operational parameters (such as availability, accuracy, etc.), which are related directly to the surrounding PPS environment. We are also evaluating the correctness of the received location data, because extreme ever-changing conditions can bring weird sensing situations.

IV. SYSTEM OVERVIEW

As the large scale data acquisition systems are producing vast amount of data, we need a multilayered approach for data processing. Due to the different - in most cases proprietary and incompatible- sensor technologies and solutions, it is challenging to realize generic, multi-modal data acquisition system [9-11]. The first level of data evaluation has multiple layers: the basic layer handles single source, primitive data types (integers, float, double, etc.) and provides range focused evaluation; the second layer of data evaluation is targeting the complex data types (e.g.: ECG, EEG curves), where complex pattern searching is required. The second level of data evaluation is used by trend analysis using calculation of derivatives in higher dimensions.
Our proposed solution combines existing; self developed hardware units, with client and server side applications and frameworks. Security is a pivotal point of the system. Sensor data is transported end-to-end in an encrypted way.

A. PPS side hardware

PPS side DAQ hardware is modular, and based on the embedded sensor infrastructure. The PPS internal communication system is capable to use a vast amount of technologies:

- wired: USB, I^2C, 1w, CAN
- wireless: WiFi, GSM/3G, XBee, Bluetooth 2/4

B. PPS side software

On the PPS (client) side we are using our own DAQit modular, multi-platform, generic data acquisition software framework, with in-built data processing, visualization and management capabilities. The framework:

- collects “personalized” biosignal (digital fingerprint of the patient’s/person’s psychophysical status and performance), and micro/macro environment information
- archive the data locally and at remote locations
- process, analyze and visualize the acquired data
- sends the acquired data towards to the datacenter for further processing.
- uses the following communication technologies at software level: ssh, https

Other basic software functionalities and features:

- Multi-language support (English, German, Italian, and Hungarian).
- Silent sensor DAQ mode (automatic data collection).
- Fused sensor data visualization.

- Data compression and encryption during data transmission.
- Automatic sensor data pre-evaluation.
- Store and forward mode to enable offline data collection

C. *Server side*

On the server side we have designed a modular, generic data processing, visualization and management system. The GUI is a Liferay [7] based open source, online, service-rich front-end with the following features:

- Data management and data sharing

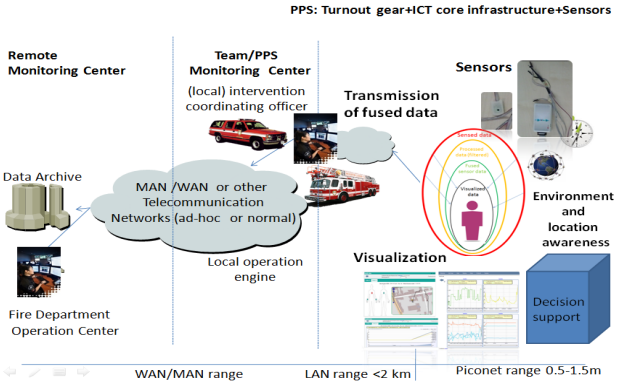


Figure 2. System overview of the PPS and remote monitoring centers

- Data visualization (multi-sensor)
- Real-time monitoring (environment and location)
- Strong authentication

Beside near real-time monitoring, the portal users can access the acquired, processed, fused and archived sensor data. The system has the capability to share information among other similar server instances. A hierarchical server cascade can built up from such servers in a scalable way. The sensor data processing and data fusion services are located separately. High-throughput real-time sensor data processing requires large computational resources [8]. Appropriate HPC infrastructure can be added to the system in a loosely coupled way to cope with the high processing demand.

V. SUMMARY AND FUTURE WORK

We have designed and realized an intelligent PPS solution using combined multi-modal sensor infrastructure for emergency service personnel. Intelligent PPS systems which are capable to support environment and location awareness are costly because even if the clothing and engineering parts of the development are feasible, their combination is far from trivial. Our research is focusing on the seamless integration of the various textile materials, DAQ/ICT subsystems, and sensors to develop effective and user friendly PPS. Our solution helps to receive accurate information about user performance, activity conditions, environment and location parameters autonomously.

During sensor fusion we are converting the data to a common representation format. The acquired data is processed on different levels, and we evaluate/filter sensor data in near real-time. The combination of the sensors provides better environment and location awareness even in special locations/ extreme environments for the PPS users. The acquired location data can be visualized by the PPS user and/or can be transmitted towards a remote location to let visualize the emergency service team activities. Sensor data acquisition and fusion enables both PPS end-users and emergency team officials to receive near real-time, highly accurate monitoring data. The server side system enables seamless data archiving to do comparative performance analysis between the PPS users or retrospective situation analysis if needed.

VI. ACKNOWLEDGMENTS

The projects have been supported by the European Union. The authors would like to thank GOP-1.1.1-11-2012-0055- Dialogic „DIALOGIC – Mathematical model based decision support system for diabetes monitoring” for its financial support. This work makes use some of the results produced by the AALAMSRK project (Research & development of assisted living patterns and their integration into decision support system to promote the quality of life at home), Hungarian National Technology Programme, A1, Life sciences project.

REFERENCES

[1] D.A. Trovi, G.V. Hadjisophocleou. Research in Protective Clothing for Firefighters: State of the Art and Future Directions, FireTechnology35(2), 1999, 111-130.

[2] Shahbazian,E:Introduction to DF:Models and processes, architectures,, techniques, and applications. In:Hyder, A.K.,Shahbazian E.(eds.) Multisensor Data Fusion, Kluwer Academic Publication (2002)

[3] Durrant-Whyte, H.F.:Sensor models and multisensor integration. Int. Journal Robotics Res. 7., 97-113(1988)

[4] Anthony LaMarca, Eyal de Lara; Location Systems: An Introduction to the Technology Behind Location Awareness; Morgan & Claypool Publishers, 2008, ISBN 9781598295825

[5] H.B.Mitchell; Data Fusion:Concepts and Ideas, Springer Heidelberg second edition,ISBN 978-3-642-43730-4

[6] Dasarathy, B.V.; Decision Fusion, IEEE Computer Society Press , 1994

[7] http://www.liferay.com/documentation/liferay-portal/6.2/user-guide [acc. 25.02.2015]

[8] M. Bräger, M.Brightwell,E.Koufakis, R.Martini,A. Suwalska;High-availability monitoring and big data: using Java clustering and caching technologies to meet complex monitoring scenarios,ICALEPCS13 International Conference on Accelerator & Large Experimental Physics Control Systems (2013), pp.:439-443,ISBN 978-3-95450-139-7

[9] Kozlovszky, M.; Pavlinic, D.Z.; Environment and situation monitoring for firefighter teams, 2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI), 19-21 Nov. 2014,pp. 439 – 442

[10] A. Lymberis,D. de Rossi ;Wearable EHealth Systems for Personalised Health Management: State of the Art and Future Challenges

[11] PROETEX IP Firefighter - http://www.proetex.org/objectives.htm [acc. 25.02.2015]

[12] RECCO® Rescue System - http://www.recco.com/the-recco-system [acc. 25.02.2015]

[13] VIKING Fire Suit - http://www.viking-life.com/viking.nsf/public/newspress-built-inintelligenceoptionforallvikingfiresuits.html [acc. 25.02.2015]

[14] WASP -Wearable Advanced Sensor Platform - http://www.globeturnoutgear.com/innovations/wasp

Error Correcting Codes and Their Usage in Steganography and Watermarking

H. Kostadinov
Institute of Mathematics and Informatics, BAS
Acad. G. Bonchev St., Bl.8, 1113 Sofia, Bulgaria
Email: hristo@math.bas.bg

N. L. Manev
USEA "Lyuben Karavelov" and
Institute of Mathematics and Informatics, BAS
Acad. G. Bonchev St., Bl.8, 1113 Sofia, Bulgaria
Email: nlmanev@math.bas.bg

***Abstract*—Steganography and Digital Watermarking are concerned with embedding information in digital media such as images, audio signals and video. Both scientific disciplines develop methods for conceal message (a sequence of bits) by modifying the host (cover) digital object but their goals are slightly different.**

In this work we discuss two methods of information embedding in spatial domain: a matrix embedding by q-ary codes and a method based on pseudo-noise patterns and which explodes the erasure capability of error correcting codes.

We have made numerous experiments with picture from several galleries (with grey-scale and color images) and many error control codes for both type of algorithms. Our observations show that q-ary codes are good choice in the case of syndrome embedding.

I. Introduction

Data hiding technique provide a way to embed extra information within the original contents, without serious degradation of the quality. With the development of proper and robust data hiding systems, more technologies have found new and promising applications.

Steganography and Digital Watermarking are concerned with embedding information in digital media such as images, audio signals and video. Both scientific disciplines develop methods for conceal message (a sequence of bits) by modifying the host (cover) digital object but their goals are slightly different. Digital watermarking is a data hiding technique with a trade-off between the embedding capacity, robustness and quality. Its goal is to prevent piracy or to prove the ownerships by imperceptibly altering the digital media object. The purpose of steganographic techniques is to alter the cover object in undetectable manner, that is, no one but the intended recipient to be able to detect the altering of the cover work. Both digital watermarking and steganography are subject of a strong interest and research activity especially in the last decade. A lot of techniques as well as commercial realization of some of them have been proposed. A comprehensive overview of the mathematical methods and the core techniques can be found, e.g., in [3], [7].

The application of error correcting codes to steganography and watermarking started with so called matrix embedding introduced by Grandall [5]. Since that time it has been popularized, developed, and carefully analyzed by many authors. We shall mention just few of them here [9], [10], [11].

In this paper we discuss two methods of information embedding in spatial domain: the aforesaid matrix embedding by q-ary codes and a method based on pseudo-noise patterns and which explodes the erasure capability of error correcting codes.

II. Preliminaries

The process of hiding information can be described in general as a mapping (***embedding process***)

$$\mathcal{E}: \left| \begin{array}{lcl} (\mathfrak{P},\mathfrak{T},\mathfrak{K}) & \rightarrow & \mathfrak{P} \\ (\mathbf{c}_0,\mathbf{m},\mathbf{k}) & \rightarrow & \mathbf{c}_m=\mathcal{E}(\mathbf{c}_0,\mathbf{m},\mathbf{k}) \end{array} \right. ,$$

where $\mathbf{c}_0$, the cover object, is an element of the set $\mathfrak{P}$ of possible digital objects (images), $\mathbf{m} \in \mathfrak{T}$ is the message intended for embedding, and $\mathbf{k}$ is a key randomly taken from the space of keys $\mathfrak{K}$.

Respectively, the process of revealing the hidden text consists of applying a ***decision function*** (which is based on the chosen ***detection metric***) to the received, eventually distorted copy $\mathbf{c}_{mn}$ of the sent work $\mathbf{c}_m$:

$\mathbf{m}'=\mathcal{D}(\mathbf{c}_{mn})$, or $\mathbf{m}'=\mathcal{D}(\mathbf{c}_{mn},\mathbf{c}_0)$, if $\mathcal{D}$ needs the original. In many cases the distortion can be considered as an additive noise $\mathbf{n}$, i.e, $\mathbf{c}_{mn}=\mathbf{c}_m+\mathbf{n}$. In our study $\mathbf{c}_{mn}=\mathbf{c}_m$.

In the described considerations and experiments the embedding process is a composition of two mappings. The first one, $\mathcal{M}$, transforms the cover image into a matrix over Z_q, the ring of integers modulo q, or into the finite field $GF(q)$ of q elements. We shall denote the set of such matrices by $\mathfrak{M}$, that is,

$$\mathcal{M}: \ (\mathfrak{P},\mathfrak{T},\mathfrak{K}) \ \rightarrow \ \mathfrak{M}.$$

The transformation $\mathcal{M}$ depends on the cover object and the chosen key. After careful analysis of the image some pixels are market as "wet" pixels which are not used. The order of the embedding into the rest ("dry") pixels is also determined according to the chosen key. The second transformation realizes the embedding algorithm based on the use of error control codes. Usually it is referred to as ***embedding function*** and we shall keep the notation $\mathcal{E}()$ for it.

Herein we discuss two methods of concealing data into images and both are realized by embedding into the spatial

domain: A) the aforesaid syndrome (matrix) embedding using q-ary codes, and B) a method based on pseudo-noise patterns.

A. Syndrome Embedding

Syndrome embedding (also known as matrix embedding) is a method in which the parties agree on a parity-check matrix $\mathbf{H}$ of a linear code and the secrete message is extracted as a sequence of syndromes (with respect to $\mathbf{H}$) from the cover digital object. With a few exceptions ([10], [8]) the codes used in the proposed algorithms are mainly binary codes, e.g., BCH, Reed-Solomon, random codes, etc.. In these case $\mathcal{M}$ transform the digital object that is in fact a matrix of 8, 16 (images), or 12 (CD audio) bits long unsigned integers (i.e., elements of $\mathbb{Z}_{2^b}, b = 8, 12, 16$) into a binary matrix $\mathbf{D}$. All these algorithms are based on looking among the elements of a given coset for a vector that minimize distortion in the image. The covering radius of the code is also very important characteristic.

More detailed description of this method in the form used by us is given in Section 3.

B. Noise patterns embedding

In most general form this embedding algorithm can be described as

$$\mathbf{c}_m = \mathcal{E}(\mathbf{c}_0, \mathbf{m}) = \mathbf{c}_0 + \alpha \mathbf{w}_m, \tag{1}$$

where $\mathbf{w}_m$ called ***message pattern*** is a function of the message and the set of predefined reference patterns (and eventually of $\mathbf{c}_0$). The scale constant α controls the tradeoff between visibility and robustness of the embedded data. The set of predefined reference patterns $\mathbf{W} = \{\mathbf{w}_{r1}, \mathbf{w}_{r2}, \ldots, \mathbf{w}_{rk}\}$ consists of matrices whose entries have a given probability distribution, most often normal or uniform distribution. These patterns are pair-wise orthonormal according to the chosen detection metric $\delta(\ ,\ ) : \mathfrak{M} \times \mathfrak{M} \to \mathbb{R}$. In fact it is enough $\delta(\mathbf{w}_{ri}, \mathbf{w}_{rj})$ to be relatively small, not necessarily exact zero.

Assume that the message is a binary vector $\mathbf{m} = (m_1, m_2, \ldots, m_k)$. Then

$$\mathbf{w}_m = \gamma\left(\epsilon_1 \mathbf{w}_{r1} + \epsilon_2 \mathbf{w}_{r2} + \cdots + \epsilon_k \mathbf{w}_{rk}\right),$$

where

$$\epsilon_i = \begin{cases} 1, & m_i = 1 \\ -1, & m_i = 0 \end{cases}.$$

The coefficient γ, when it differs from 1 is used for a kind of normalization of $\mathbf{w}_m$, for example, $\gamma = \frac{1}{\sqrt{k}}$ normalizes the variance to 1.

The process of revealing message is also based on the detection metric. For $i = 1, 2, \ldots, k$ the receiver calculates $\delta(\mathbf{c}_{mn}, \mathbf{w}_{ri})$ and according to the sign of its value decides what is the i-th bit m_i. Here is the most used detection metrics:

- ***Linear correlation***:

$$lc(\mathbf{x}, \mathbf{y}) = \frac{1}{N}\mathbf{x} \cdot \mathbf{y} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i,$$

where N is the length of the vector;

- ***Normalized correlation***:

$$nc(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}.$$

- ***correlation coefficient***:

$$cc(\mathbf{x}, \mathbf{y}) = \frac{\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}}{|\hat{\mathbf{x}}| \cdot |\hat{\mathbf{y}}|},\ \hat{\mathbf{x}} = \mathbf{x} - \mathrm{E}[\mathbf{x}],\ \hat{\mathbf{y}} = \mathbf{y} - \mathrm{E}[\mathbf{y}]$$

III. Syndrome embedding and q-ary codes

In this section we describe in short our approach to syndrome embedding underling the differences with the usual one. Assume $\mathbf{c}_0$ is a matrix of elements of $\mathbb{Z}_{256}$. We will skip descriptions of $\mathcal{M}$ in concern of the use of the key and determination of the "dry" pixels, and will consider only the part that concerns embedding function. The aforesaid feature of $\mathcal{M}$ are implemented in the developed software, but in most of experiments they are skip and realize only what follows.

Let $q' \geq q$ be an integer less than 255. Consider $\mathbf{D}' = \mathbf{c}_0 \pmod{q'}$. This is the target matrix of the embedding. If q' is a power of 2 the action is equivalent to embedding in the several less significant bits. In this case binary codes can also be used. In order to prevent the embedding to be visually detected we must have $q' < 2^5$.

Algorithm:

Let $\mathbf{H}$ be a $r \times n$ parity-check matrix of a linear code over $\mathbb{Z}_q$, or a finite field $\mathbb{F}_q$ of q elements. Let $\mathbf{D}$ be a $n \times N$ matrix over $\mathbb{F}_q$ or $\mathbb{Z}_q$ obtained from the target image by the transformation $\mathcal{M}$. Let the message (randomized and eventually encrypted) be also transformed into a $r \times N$ matrix $\mathbf{m}$ over $\mathbb{F}_q$.

1) Compute $\mathbf{S} = \mathbf{m} - \mathbf{HD}$, i.e., $\mathbf{m} = \mathbf{S} + \mathbf{HD}$
2) Find $n \times N$ matrix $\mathbf{E}$ such that $\mathbf{S} = \mathbf{HE}$
3) Compute $\mathbf{V} = \mathbf{D} + \mathbf{E}$
4) Construct an image $\mathbf{c}_m$ such that $\mathcal{M}(\mathbf{c}_m, \mathbf{k}) = \mathbf{V}$ and send $\mathbf{c}_m$.

The receiver

1) Determines $\mathbf{V} = \mathcal{M}(\mathbf{c}_w, \mathbf{k})$
2) computes $\mathbf{HV} = \mathbf{HD} + \mathbf{HE} = \mathbf{HD} + \mathbf{S} = \mathbf{m}$

Embedding efficiency in the case of syndrome encoding using a code with $r \times n$ parity check matrix is

$$E_f = \frac{r \log_2 q'}{n} \quad \text{bits/pixel}$$

Note that the embedding efficiency does not depend on which field (ring) the code is defined over. The number of its elements is usually $q = q'$, but it can be in partial $q = 2$ when q' is a power of 2.

Our approach simplifies the process of determining the matrix $\mathbf{E}$. The representative of the coset that corresponds

to a given syndrome is chosen to be with the minimum Lee weight and its entries to be with minimum absolute value. For example, between the $\mathbb{Z}_4$-vectors $(0,3,3)$ and $(2,0,0)$ the first vector is preferable ($3=-1$ in $\mathbb{Z}_4$).
Here is a simple example

Example 1. Let $q=4$ and $\mathbf{H}=(1\,2)$ be the parity check matrix of $[2,1]$ code over $\mathbb{Z}_4$. The set of syndromes $Synd=\mathbb{Z}_4=\{0,1,2,3\}$. The matrix of error patterns corresponding to the syndromes is

$$\mathbf{Eh}=\begin{pmatrix}0&1&0&3\\0&0&1&0\end{pmatrix},\quad \text{i.e,}\quad \mathbf{H}.\mathbf{Eh}=(0\ 1\ 2\ 3)$$

The target image $\mathbf{I}$ ($a\times b$ matrix over $\mathbb{Z}_{256}$) is transform to $\mathbf{Der}\equiv\mathbf{I}\pmod 4$ and then by rearranging to $2\times ab/2$ matrix over $\mathbb{Z}_4$. The message is transform to $1\times ab/2$ matrix $\mathbf{m}$ over $\mathbb{Z}_4$ and then we follow the described above algorithm. The j column of the matrix $\mathbf{E}$ is the $\mathbf{S}(j)+1$ column of $\mathbf{Eh}$, where $\mathbf{S}=\mathbf{m}-\mathbf{HD}$.

IV. Pseudo-noise patterns and erasure codes

A. *Algorithm*

Embedding

A-1. Let C be a binary $[n,n-r]$ linear code. Encode the source message into a binary sequence $\mathbf{m}$.

A-2. Starting with a given state (used as password) of the random number generator generate t reference patterns of size $a\times b$: $\mathbf{W}=\{\mathbf{w}_1,\mathbf{w}_2,\dots,\mathbf{w}_t\}$.

A-3. Divide (in some way) the cover work into N blocks, $\mathbf{c}_1,\dots,\mathbf{c}_N$, each of size $a\times b$.

A-4. (optional) Replace the set $\mathbf{W}$ by the set of patterns $\{\mathbf{h}_i\}$ which are orthogonal to all blocks $\mathbf{c}_1,\dots,\mathbf{c}_N$.

A-5. In each block $\mathbf{c}_j,\ j=1,2,\dots,N$, embed t bits of the message sequence $\mathbf{m}$ by

$$\mathbf{c}_{jw}=\mathbf{c}_j+\frac{\alpha_j}{\sqrt{t}}\left(\epsilon_1\mathbf{w}_1+\epsilon_2\mathbf{w}_2+\cdots+\epsilon_t\mathbf{w}_t\right),$$

where $\epsilon_i=\begin{cases}1, & m_{ji}=1\\ -1, & m_{ji}=0\end{cases}$.

The scale constant α_j controls the tradeoff between visibility and robustness of the hiding data.
Embedding efficiency in this case is

$$E_f=\frac{(n-r)tN}{nab}\ \text{bits/pixel.}$$

Detection and Decoding

A-6. The recipient divides the received image into N blocks $\{\tilde{\mathbf{c}}_j\}$ and knowing the reference patterns (or the key to generate them) calculates

$$\delta(\tilde{\mathbf{c}}_j,\mathbf{w}_i)=\delta(\mathbf{c}_j,\mathbf{w}_i)+\frac{\alpha_j}{\sqrt{t}}\delta(\mathbf{w}_i,\mathbf{w}_i)\,\epsilon_i,$$

where $j=1,2,\dots,N,\ i=1,\dots,t$, and $\delta(,)$ is the chosen detection measure. (Indeed $\tilde{\mathbf{c}}_j$ are noise versions of $\mathbf{c}_{jw}$) Then recover the message:

$$\tilde{\mathbf{m}}_{ji}=\begin{cases}1, & \text{if}\quad \delta(\tilde{\mathbf{c}}_j,\mathbf{w}_i)>\tau\\ 0, & \text{if}\quad \delta(\tilde{\mathbf{c}}_j,\mathbf{w}_i)<-\tau\\ \text{an erasure} & \text{if}\quad -\tau\le\delta(\tilde{\mathbf{c}}_j,\mathbf{w}_i)\le\tau\end{cases},$$

A-7. The error control code decoder corrects errors and erasures. Its output can be

- "there is no watermarking or hidden message", when the number of erasures is $>N/2$;
- a decoded message (a sequence of bits);
- a decoded message with warning "errors are possible".

B. *Error analysis*

The expected value of $\delta(,)$ is $\mu_1=\mu=\frac{\alpha}{\sqrt{t}}$, when $m_i=1$, and $\mu_0=-\mu=-\frac{\alpha}{\sqrt{t}}$, when $m_i=0$ is embedded, respectively (see A-5 and A-6). Let us assume that $\delta(,)$ is normal distributed. Then the variance is $\sigma^2=\sigma^2_{\boldsymbol{w}_i}(\sigma^2_{\boldsymbol{c}}+\sigma^2_{\boldsymbol{n}})$, where $\sigma^2_{\boldsymbol{w}_i}=1$, and $\sigma^2_{\boldsymbol{c}}$ and $\sigma^2_{\boldsymbol{n}}$ are the variance of the cover work and the channel noise, respectively. Usually $\sigma^2_{\boldsymbol{c}}\approx(60/255)^2$. Let $p_c,\ p_{er}$ and p_{es} be the probability of correct detection, of error, and of an erasure, respectively. Then in both cases, when $m_i=1$ and $m_i=0$ is embedded:

$$p_c=\frac{1}{2}\operatorname{erfc}\left(\frac{\tau-\mu}{\sigma\sqrt{2}}\right);\qquad p_{er}=\frac{1}{2}\operatorname{erfc}\left(\frac{\tau+\mu}{\sigma\sqrt{2}}\right);$$

$$p_{es}=1-p_c-p_{er}=\frac{1}{2}\left[\operatorname{erf}\left(\frac{\tau-\mu}{\sigma\sqrt{2}}\right)+\operatorname{erf}\left(\frac{\tau+\mu}{\sigma\sqrt{2}}\right)\right].$$

The probability of **a false positive decision** ($\mu=0$) is given by

$$P_{fp}=\sum_{i=0}^{\lfloor\frac{N}{2}\rfloor}\binom{N}{i}p^{N-i}(1-p)^i, \tag{2}$$

where $p=\operatorname{erfc}\left(\frac{\tau}{\sigma\sqrt{2}}\right)$

The embedded n bits can be correctly decoded with a probability

$$P_{corr}=P_1+P_2+P_3,$$

where

$$P_1=\sum_{s=0}^{\lfloor\frac{d-1}{2}\rfloor}\binom{n}{s}p_{es}^s\left(\sum_{t=0}^{\lfloor\frac{d-1-s}{2}\rfloor}\binom{n-s}{t}p_{er}^t p_c^{n-t-s}\right),$$

$$P_2=\sum_{s=\lfloor\frac{d-1-s}{2}\rfloor+1}^{d-1}\binom{n}{s}p_c^{n-s}p_{es}^s,$$

$$P_3=\sum_{s=d}^{n}\binom{n}{s}p_c^{n-s}(q-p_c)^s,$$

d is the minimum distance of the code, and

$$q = \frac{1}{2}\operatorname{erfc}\left(\frac{-\mu}{\sigma\sqrt{2}}\right)$$

(this is p_c with $\tau = 0$) is the probability of positive (resp. negative) value of $\delta(\tilde{\mathbf{c}}_{wn}, \mathbf{w}_i)$.

(**b**) The probability of correct decoding in the case when the codes are used only in error correcting mode is given by

$$Q_{corr} = \sum_{t=0}^{\lfloor \frac{d-1}{2} \rfloor} \binom{n}{t} q^{n-t}(1-q)^t.$$

(**c**) If more than $N/2$ erasures are marked for a given watermark pattern $\mathbf{w}_i$ then the detector outputs "there is no watermark", that is, it makes **false negative decision**. The probability, P_{fn}, for such an output is given by

$$P_{fn} = \sum_{i=0}^{\lfloor \frac{N}{2} \rfloor} \binom{N}{i} p_{es}^{N-i}(1-p_{es})^i. \qquad (3)$$

Since pixels can be interpreted (are presented) as elements of $\mathbb{Z}_q$ (usually $q = 256$) a function (called *distortion function*) ρ

$$\rho : \left| \begin{array}{ccc} \mathbb{Z}_q & \rightarrow & \mathbb{R} \\ x & \rightarrow & \rho(x) \end{array} \right.$$

can be used as a measure of the embedding impact

One natural choice of ρ for algorithms that change the pixel value by up to $\pm x$ is

$$\rho(x) = [1 - \Pr(-x \le t \le x)]^{-1} = \left[\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right)\right]^{-1}$$

which is the reciprocal value of the probability of the random variable's deviation (dispersion) in the standard normal distribution to be greater than x.

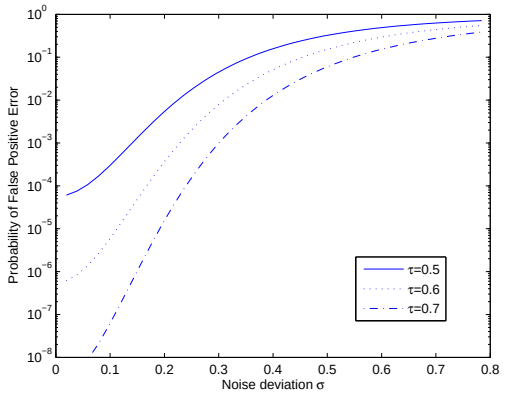
V. Experiments and Conclusions

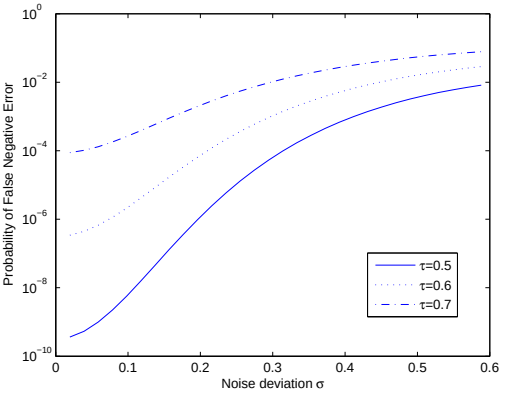
We have made numerous experiments with picture from several galleries (with grey-scale and color images) and many error control codes for both type of algorithms. Also, we have tested the cover images with the available in the internet software for stego-analysis. Nevertheless, the channel is assumed noiseless and detection blind, that is, the receiver and the opponent don't have the original, we have compered the results of stego-analysis on both original and cover objects.

Our observations show that q-ary codes are good choice in the case of syndrome embedding. For both algorithms it is better to use

VI. Acknowledgements

This work was partially supported by the National Science Fund of Bulgaria under Grant DFNI-I02/8 and the EU project OUTLAND, funded under the territorial cooperation programm Bulgaria-Greece 2007-2013.





a) b)

Fig. 1. a) False Positive Error for $n = 8$ b) False Negative Error for $n = 13$

References

[1] W.C. Huffman, Codes and Groups, *Handbook of Coding Theory,* V. S. Pless and W. C. Huffman, Eds., Elsevier, Amsterdam, 1998, Ch. 17, 1345–1440.

[2] S. V. Avgustinovich and F. I. Solov'eva, Perfect Binary Codes with Trivial Automorphism Group, in *Proc. Int. Workshop on Inform. Theory, Killarney, Ireland, 1998,* IEEE, Piscataway, 1998, 114–115.

[3] I.J. Cox et al., *Digital Watermarking and Steganography*, Morgan Kaufmann Publ., 2008.

[4] F. Galand and G. Kabatiansky, Information hiding by coverings, in *Proc. IEEE Inf. Theory Workshop, 2003*, 151–154.

[5] R. Grandal, Some notes on stegsnography, avilable from http://os.inf.tu-dresden.de/ westfeld/crandall.pdf

[6] F.J.MacWilliams and N.J.A.Sloane, *The Theory of Error-Correcting Codes*, North Holland, Amsterdam, 1977.

[7] P. Moulin and R. Koetter, Data-hiding codes, *Proceedings of IEEE,* vol. **93**, 2083–2126, 2005.

[8] H. Rifa, J. Rifa, and L. Ronquillo, Perfect $\mathbb{Z}_2\mathbb{Z}_4$-linear codes in steganography, *Adv. Math. Commun.*, **5**, (3), 425–433, 2011.

[9] A. Westfeld, High capacity despite better steganalysis (F5 - A steganographic algorithm), vol.2137 of Lecture Notes in Computer Science, pp. 289-302, Springer-Verlag, 2001.

[10] F.M.J. Willems and M. van Dijk, Capacity and codes for embedding information in grayscale signals, *IEEE Trans. on Inf. Theory*, **51**, (3), 1209–1214, March 2005.

[11] W. Zhang, X. Zhang, Sh.Wang, Near-optimal codes for information embedding in gray-scale signals, *IEEE Trans. on Inf. Theory*, vol. 56, (3), 1262–1270, March 2010.

Model for calculation of post service efficiency by using GIS

Mirjana Kranjac*, Uroš Sikimić**, Marija Vujaković***
* Faculty of technical sciences/Department for transport, Novi Sad, Serbia
** Politechnico di Milan/Department for engineering, Milan, Italy
*** Center for standardization and certification, Novi Sad, Serbia
mirjana.kranjac@ftn.uns.ac.rs, uros_sikimic@hotmail.com, marija.vujakovic@bsccentar.rs

Abstract - The goal of the paper is to find model of using GIS in monitoring and evaluation of postal service management. The methodology which is used is two visualize some data about work efficiency by creating separate layers and analyzing them. The analysis of results shows big potential of GIS as a tool for economic analysis. In the case of postal system of Serbia it proves that cross cutting analysis of various factors and their visualization gives better model for economical improvements of system management. The scientific contribution of this paper is to introduce GIS as valuable tool for economical research.

I. INTRODUCTION

The geographic information systems are family of software for visualization of geo referenced data. They could be used for different analysis which offers visualized solutions. Geographic referenced data are more acceptable for human use.

Geographic location is the element that distinguishes geographic information from all other types of information. Without location, data are termed to be non-spatial and would have little value within a GIS. Location is, thus, the basis for many benefits of GIS: the ability to map, to lap layers, to measure distances and to tie different kinds of information together because they refer to the same place as in [1].

II. LITERATURE OVERVIEW

Some authors discuss the use of GIS in econometrics. At the begining, GIS was widely used in business, government and in a range of academic disciplines and its application in economics has been more limited. The most frequent application in economics is the use of GIS for vizualisation or mapping of economic data which have spatial components. GIS use in econometrics begins with "plot the data" at an early stage of analysis to help identify trends, outliers etc. GIS provides a simple and efficient way to identify some economic trends. Less common, but more interesting, is the use of GIS to generate additional data as inputs in to further statistical analysis by using storage and management functions. This enables use of the GIS to cross cut spatial data from many sources. Some of them, like sampling and census data, are familiar to economists [2]. Rosenthal and Strange [3] provide an example of the use of GIS to implement such an instrumental variables strategy. They are interested in whether density of employment helps determine wages. The problem is that higher wages should attract more workers and lead to higher employment densities. That is, density may be caused by wages and not vice-versa.

GIS is used to construct area zones or to define neighbourhoods ("buffers") around objects. For example in [4] authors use such overlapping zones in their study of urban spreading

There is ability of GIS to combine spatial data from various sources that allows the creation of new data sets. It is combining socio-economic data from different spatial units like, population data from US census with employment data. Many economists use mappings between sets of spatial units and merging of such data. GIS offers flexibility in allowing users to create their own cross maps of different geographical units with data from different sources [5] .

For telecomm industry GIS is very important tool. Even the smallest companies use GIS for automatisation of localized tasks such as network design and planning and gain strategic advantage by standardising information throughout their organisations. Marketing staff is using the same data for mapping of customer distribution and engineers are designing fibre networks in GIS. All data are stored in structured database with many interfaces [6].

Postal operators use GIS for many services like address code, routing, marketing. For example, in the USA for the fiftieth anniversary of the ZIP Code, the Office of the Inspector General at the USPS has released a paper exploring ways to improve the ZIP code, "both to save postal costs and to enhance the opportunity for third party innovators to discover new uses and applications." In the report, entitled "*The Untold Story of the ZIP Code*", the executive summary recommends linking the USPS address database with GIS to create a more efficient ZIP Code system:

One such enhancement would be to combine the ZIP Code with the precision of geocodes (latitude and

longitude coordinates). This could have a direct impact on the U.S. Postal Service's bottom line by facilitating delivery route reconfiguration. Historically, the Postal Service has delivered mail via set delivery routes where each carrier visits the same homes every day. With the anticipated growth of parcel delivery, including the possibility of same-day delivery, the Postal Service will find it necessary to develop dynamic carrier routing. Combining the ZIP Code with the geocoding of addresses could ease the processing necessary to improve the efficiency of parcel delivery [7] .

Some papers deal with complex postal routing problems and with servicing of limited number of known location from one or more depots, in presence of certain limitations regard to demands of service and available recources by using GIS [8] .

Post of Serbia has developed the following services based on the Geographic Information System (GIS):

- Sale of coordinate position of a house number on the map
- Sale of coordinate position of streets on the map
- Sale of coordinate position of PAK on the map
- Generation of alphanumerical data based on coordinate position of house number
- Generation of alphanumerical data based on coordinate position of PAK
- Sale of data on number of households on the level of house number*
- Delivery of data on number of households on the level of PAK.

The Post of Serbia has been developing the Geographic Information System since 2002 with goal to analyse the territorial availability of postal service to the citizens and to rationalise transport and delivery of postal items.

The Post has geo-positioned the complete street network in Serbia, that means 103,000 kilometres of streets and roads. The company delivered coordinates for about 900,000 houses.

As regards the rural areas, where the houses are not marked with numbers, the Post created system wich coordinates positions of PAKs. In this way, the Post of Serbia geo-referenced over 113,000 PAKs on the Serbian territory.

Post of Serbia has developed a GIS portal to enable citizens and post service users to have spetial information of the area in which the services are performed.

The Post of Serbia offers to its users the address standardisation service which provides that any detail of our user which includes address detail shall be geo-positioned.

Based on the data of Geographic Information System, various analyses are possible like this presented in Figure 1:

- analyses of territorial accessibility to facilities and users,
- planning of advertising material distribution,
- planning of routes,
- generation of address lists by zones,
- analysis of population density aiming to plan infrastructure capacities, or selection of the most favourable sites for opening of new facilities, etc. [9].

.

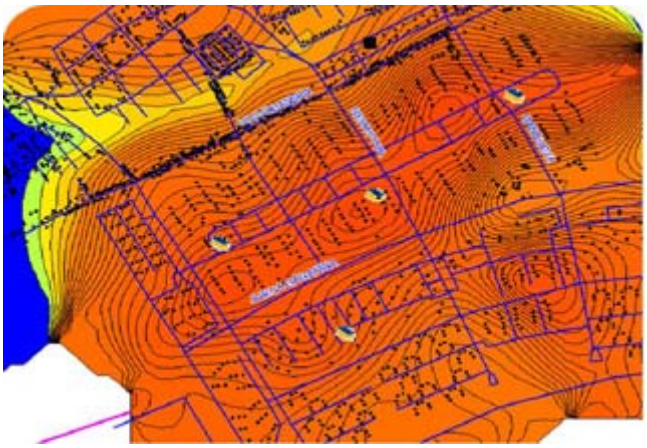


Figure 1. Analysis of population density and position of post offices

Reference [10] gave a contextualized professional development approach for geographic information technologies as a project-based science.

Innovative approach for analysis by using GIS is described in articles of reference [11].

Geographic information systems (GIS) are presented and explained as decision support in the paper referred in [12, 13].

III. GEOGRAPHIC INFORMATION SYSTEM

Due to the increasing development of science and technology and development of human society, the importance of information about the area becomes larger and extends from the idea to implementation of a project to change the space that surrounds us. Therefore, there are developed different survey methods to collect data on the area, and the different ways of storing and sorting data, collected them and a better analysis according to various criteria.

The spatial data describe the location, shape and orientation of objects in space. They are known as geospatial data. Geospatial data can describe the characteristics of many different types of objects on the surface. These objects can be tangible, physical things such as an office building, landscape or abstract forms such as the imaginary line that marks the political boundary between the states.

From the moment when first started collecting spatial data and display them on maps, there is a tendency for

them to be systematized and made available. Over a long historical period, the most effective way to display spatial information was analogous to the map. Map was the forerunner of the original spatial database, the original spatial information systems, and in some ways a forerunner of spatial data infrastructure. Map is important to display the spatial distribution, connectivity and interaction of objects and phenomena, as well as view qualitative and quantitative alteration of the condition over time.

Technologies of GIS are the "main culprit" for the change. Geographic information integrated into other products and software applications have become mass-market product. Forerunner of spatial data infrastructure in today's terms it is probably the concept of integrated mapping of different thematic layers of data from the sixties. Thanks to information and communication technologies, conventional way of presenting information about the area is the past. Today, spatial data is usually collected, stored, processed, analyzed and presented in digital form through a number of applications. Spatial data types provide fundamental abstraction mechanisms for modeling the geometric structures of spatial data, its attributes, operations on them, and the relations between them.

"A GIS is a collection of software that allows you to create, query and analyze geospatial data. Geospatial data refers to information about the geographic location of an entity. This often involves the use of a geographic coordinate, like a latitude or longitude value. Spatial data is another commonly used term, as are: geographic data, GIS data, map data, location data, coordinate data and spatial geometry data." [9] .

GIS is a system for managing spatial data and their associated properties. In the strictest sense, it is a computer system capable of integrating, storing, editing, analyzing and displaying geographic information. In a broader sense, GIS is a "smart card" that gives users the ability to set interactive queries (user created surveys), analyze spatial information, and edit data. The technology of geographic information systems can be used for scientific investigations, resource management, property management, development planning, regional planning, mapping and infrastructure planning. GIS is often used for the purposes of market research, then in geology, civil engineering, but also in all areas that use data related to the map.

A. Work in GIS

Information is data on artificial and natural objects in an area. These are, for example, infrastructure facilities, housing, sports and other buildings, construction of reservoirs, rivers, land use, elevation terrain data, geology, terrain, etc.. They are called features in GIS. Each feature has its attributes. If feature is mobile telecomm operator in Serbia, than the features could be: number of users per towns, number of different services...

Information on the geometry of the object can be in the form of:

- Grid and
- Vector.

Grid consisting of rows and columns of cells, called pixels, where each of these cells has a single, particular, digital value. In the case of images, the numerical value represents the number of colors (colors are coded with numbers). In the case of display other information, the pixel value is not only the color but also represents the spatial data. Examples, in some areas there have been excesses in the form of emission of toxic gases. Raster that shows the amount of pollution, consisting of pixels whose numerical values carry spatial information on the concentration of toxic substances in the air in each fraction of the space.

Displaying information in vector form refers to the geometry of the shape (length, height, and shape), either in terms of line or polygon entities and their spatial position (the position of the coordinate system).

In addition to this information in GIS is no additional information, so-called. spatial data, which can bind to a particular raster or vector data. Example: noise can be spatially displayed in the form of irregular polygons (polygons entity) and non-spatial data can be information about the types of trees, their numbers, percentages, different types (Figure 2), etc.

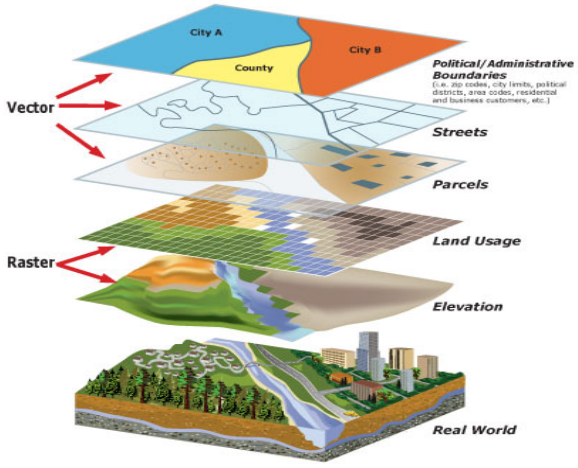


Figure 2: Example of vector and raster layers in GIS

B. Application of geographic information systems

GIS is an essential tool in all areas of design, planning, management and analysis. Around the world, used by electrical, mechanical and civil engineers, architects, bankers and economists, journalists, teachers, geodetic engineers, environmentalists, politicians, criminologists, planners, health professionals and others.

Business people see the world as a collection of information about sales, customers, warehouses, demographic profiles and much more. The basis for all of this information is the address, sales region, or transport routes of delivery which can all be displayed and interactively operated on the map.

The planning and engineering tasks that can be easily solved with the help of GIS are:

- location analysis,
- supervision and monitoring
- automatic mapping,
- analysis of water quality,
- modeling field
- geo-bibliography,
- soil analyzes (soil analysis)
- inventory and stock control.

Competitive pressure and new legislation causes management to an efficient and responsible manner. This requires access to information based on a geographically distributed elements and operations. In today's world, concurrent, successful management requires the maximum of all the resources, people, equipment and information. Using GIS to integrate geography with other relevant data, gives the system fully equipped for this task.

IV. ANALYSIS OF POSTAL SERVICES EFFICIENCY: CASE OF THE POST OF SERBIA

This paper refers to the use of GIS for analysis of the postal services in the Public enterprise PTT Srbija. The intention of the authors was to find different possibilities and to focus on the application of the GIS in monitoring and evaluation of postal service management. The essential problem of managing postal system is distribution of employees as function of distribution of residents. Hereof visualization is recognized as a good tool which is used to present some data of work efficiency by creating separate layers and analyzing them. The analysis of results shows big potential of GIS as a tool for economic analysis. Although, the postal system of Serbia like many others, uses GIS only to locate post offices, customers, fleet management systems, this paper indicates big potential for economical research. It proves that cross cutting analysis of various factors and their visualization gives better model for economical improvements of postal system management. The scientific contribution of this paper is to introduce GIS as a valuable tool for economical research and decision makings.

A. GIS maps

Some data which are available are analyzed. The map in Figure 3 shows density of inhabitants of Serbia. The most density is in Belgrade region and then in Južna Bačka region in Vojvodina area. The graduated presentation is used.

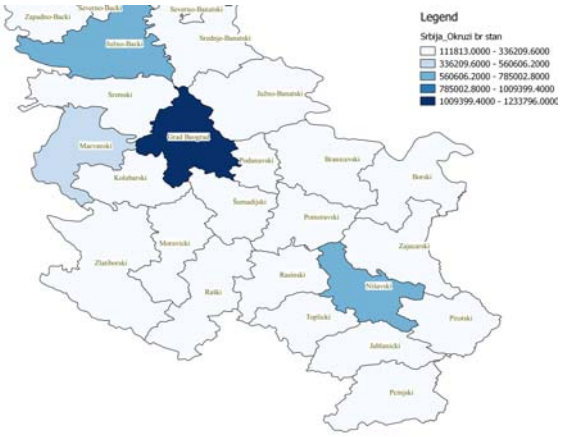


Figure 3. Inhabitants density according to counties

Next Figure is showing number of post offices in Serbia. From the graduated map can be seen that the most offices are in:

1. Belgrade county
2. Južna Bačka county
3. Srem and Južni Banat counties.

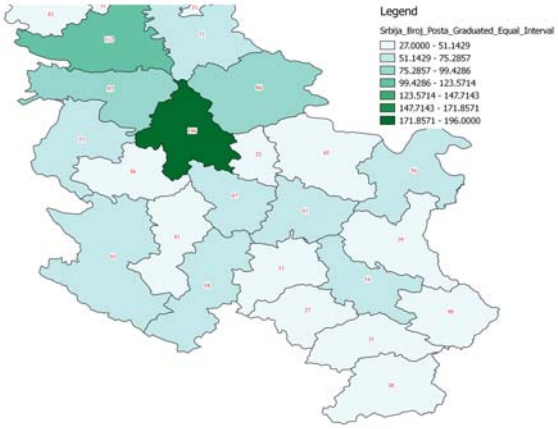


Figure 4. Post offices density according to counties

Figure 5 presents number of employees in post offices in Serbian counties:

1. Belgrade region and Južna Bačka ounty
2. Južni Banat i Srem counties.

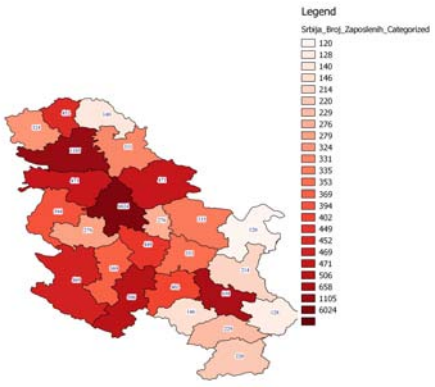


Figure 5. Number of employees in post offices in counties

The next map in Figure 6 presents average number of inhabitants per employed in postal system at certain county.

Results are:

1. Bor and Severni Banat
2. Jablanički and Pčinjski
3. Podunavski and Pirotski counties.

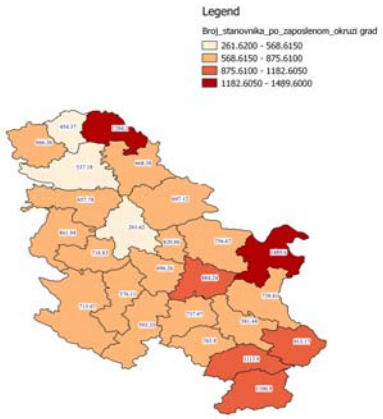


Figure 6. Average number of inhabitants per employed in counties

Figure 7 is giving the picture about average number of employees in a post office.

Belgrade is with the most workers and than a group of regions is with similar average number (the same group in a graduated presentation).

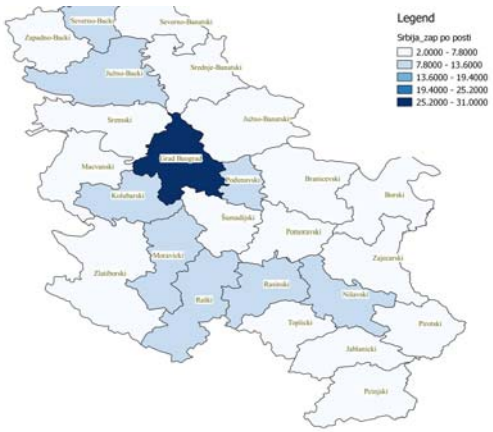


Figure 7. Average number of employees in a post office according to counties

V. CONCLUSION

The presented figures offer set of different analysis which are possible only with geo refernced spatial data. Layers created as vector layers in GIS enable comparisson of data. The most significant conclusions can be drown when verlapping differnt layers and making cross cutting analysis.

Conclusions made from presented analysis of efficiency of postal system in Serbia are:

The most density of inhabitants per employees are in Bor and Severni Banat. Bor is a county which belongs to the most developed counties, above the Republic average. The Severni Banat county is among the less developed. This needs further analysing. Severni Banat is probably with less economic turn over and this could be reson that it has less employed.

Reason that Bor has not many employed could be efficiency of workers, good management of working chain, well educated employers. Post office reflects the general picture of the county.

The biggest number of employees are in post offices on the territory of Belgrade and Južna Bačka counties. Explanation is that these counties have the biggest population and big settlements which require big post offices with more workers.

The most post offices are in the north part of the country. Niš county, where is the third large town in Serbia, has not enough number of post offices.

These are some conclusions:

Financial data are not available for external use and thez would be additional support for the concusions. Adding them wouldmake picture more clear. The proposal is to create some additional financial maps with data like:

- Postal income per county, per office, per employee, client
- Profit per county, per office, per employee

and some similar. These would offer cross cutting analysis which could give more interesting results. Such results could signal some changes in work management and improve work results. Corrections will bring new picture which must be presente in GIS, again and this will lead to continous monitoring of effects that changes caused. Precondition are databases with real data and carefully selected data. Investment into GIS and educated people who could use it in the right way brings improvement of organisation functioning

VI. REFERENCES

[1] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, "Geographic information systems and science", John Wiley, pp. 290-300, Chichester, 2001.

[2] Geographical Information Systems (GIS) and Economics1 Henry G. Overman (London School of Economics) 5th January 2006.

[3]Rosenthal, S. and W. Strange, "The Attenuation of Agglomeration Economies: A Manhattan Skyline Approach" ,Processed University of Toronto, 2005.

[4]Overman, H. G. and Y. Ioannides, "The Spatial Evolution of the US Urban System", Journal of Economic Geography 4 (2), 131-156, 2004.

[5] Bateman, I., A. Jones, A. Lovett, I. Lake and B. Day , "Applying Geographical Information Systems (GIS) to Environmental and Resource Economics. Environmental and Resource Economics ", pp. 219–269, 2002.

[6] http://www.geos.ed.ac.uk/ (accessed 03.12.2015)

[7]http://www.gislounge.com/usps-making-the-case-for-gis/ (accessed 03.12.2015)

[8] Z. Marković, Lj. Ostojić, "Ptt route – softver za projektovanje ruta vozila i izradu reda prevoza poštanskih pošiljaka jp ptt saobraćaja JP "Srbija", XXIV Simpozijum o novim tehnologijama u poštanskom i telekomunikacionom saobraćaju – PosTel 2006, Beograd, 2006.

[9] http://www.posta.rs/struktura/eng/it/usluge-GIS-a.asp (accessed 12.12.2015)

[10] RS Bednarz, SW Bednarz, "The importance of spatial thinking in an uncertain world Geospatial Technologies and Homeland", Springer, 2008.

[11] T. Randall, Cameron, Ch.,.B., J Churchill, Brian Baetz, "Geographic information system (GIS) based decision support for neighbourhood traffic calming" Canadian Journal of Civil Engineering, 32(1): 86-98, 10.1139/l04-085, 2005.

[12] Y. Kuo, C. Yu, "3G Telecommunication operators' challenges and roles: A perspective of mobile commerce value chain" Technovation 26 (12), pp. 1347-1356, 2006.

[13] Rochet J. C., J. Tirole, "Platform competition in two sided markets", Journal of the European Economic Association 1(4), pp. 990–1029, 2003.

Generalization of hypersphere function

D. Letic* and I. Berkovic **
University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
* dletic@open.telekom.rs, ** berki@sbb.rs

Abstract - In this paper are presented the results of theoretical researches of hypersphere function on the basic of generalization of two known functions referring to hypersurface $S(k,r)$ and hypervolume $V(k,r)$ of sphere and supposed recurrent relation between them. On the basis of the two introduced freedom degrees is performed the generalization of these functions. So, we have got a special continual function, i.e. generalized hypersphere function. Symbolic evaluation and numerical experiment are realized by the program packages MathCAD Professional and Mathematica.

I. INTRODUCTION

Hypersphere function is a hypothetical function can necked with multidimensional space. It belongs to the group of special functions, so its testing is being performed on the basic of the known functions as the types *gama, psi, beta* and the like. The most significant value is in its generalization from desertion to continuum. In addition to it we can from the scope (fields?) of natural integer of values, pass to the set of real and noninteger values. So, there gather conditions as well for its graphic interpretation and more concise analysis. For the development of the hypersphere functions theory are especially responsible: *Neville* [12], *Manning* [9], *Rucker* [13], *Conway* [2], *Sommerville* [14], *Wells* [15], *Maunder* [10], *Hinton* [5], *Freden* [3], *Hocking* and *Young* [6].

A. Hypersphere Surface and Volume

As analyses present, the surface of the hypersphere of the diameter r can be defined as a form function [19]:

$$S(k,r)=\frac{k\sqrt{\pi^k}\,r^{k-1}}{\Gamma\left(\frac{k}{2}+1\right)}=\frac{2\sqrt{\pi^k}\,r^{k-1}}{\Gamma\left(\frac{k}{2}\right)}, \qquad (1)$$

where $\Gamma(x)$ ($x\neq 0$) is *gama* function, while k is specially defined dimension degree ($k\in N$). In the intergration result, we get the term for hypersphere volume (V), of r radius [16]:

$$V(k,r)=\int_0^r S(k,r)dr=\frac{\sqrt{\pi^k}\,r^k}{\Gamma\left(\frac{k}{2}+1\right)}. \qquad (2)$$

The obtained terms are key ones for defining more general function that would enclose both previous ones, and so would be performed a unique interpolation between these vectors by which are being described surfaces, i.e. volumes of hypersphere.

B. Hypersphere function with two degrees freedom

Previous results (1) and (2), as it is known, present, respectively two-dimensional (surface-surf), that is three-dimensional (volume-solid) entities. In addition to certain generalizations, these terms may create a notion of a hypersphere functions family, whose development would now flow along "horizontal" dimensions or freedom degrees n , where is $n\in N$. Within every "horizontal" freedom degrees n , exists as well "vertical" one of freedom, k (that is orthogonal in relation to n) (Figure 1 and 2).

C. General Form of Hypersphere Function

The function (2) will be in continuation marked with $HS(k,3,r)$ (*Hypersphere*), where k is the dimension within the third degree of freedom ($n=3$), and r hypersphere radius. General mark would be now presented by the functions of three arguments, in the form of $HS(k,n,r)$. With the hypersphere of the third degree of freedom, the point is, as it is emphasized, about volumes or solids, while with the hypersphere of the second, to which is adequate the function $HS(k,n,r)$. Consequently, on the basic of previous, there are:

$$HS(k,2,r)=S(k,r) \quad \text{and}$$

$$HS(k,3,r)=V(k,r). \qquad (3)$$

The transition anto the higher function than the appropriate one (e.g. from surf to solid), is reached by integrating, so that the previous functional equation $HS(k,3,r)$ can be expressed by the next form:

$$HS(k,3,r)=\int_0^r HS(k,2,r)dr=\frac{\sqrt{\pi^k}\,r^k}{\Gamma\left(\frac{k}{2}+1\right)}=\frac{2\sqrt{\pi^k}\,r^k}{k\Gamma\left(\frac{k}{2}\right)}. \qquad (4)$$

The *freedom degree* is here conditionally defined and presents dimensions level of hypersphere entity. So, in the zero ($n=0$) and the first degree of freedom ($n=1$) is included the entity out of man's experience perception. It is related as well to higher freedom degrees that are included from the fourth, the fifth to the n -th degree. These concepts (notions) are precisely defined in topology and geometry, and in this paper are specially approached

(treated). For the fourth horizontal freedom degree, follows the hypersphere form function:

$$HS(k,4,r)=\int_0^r HS(k,3,r)dr=\int_0^r\int_0^r HS(k,2,r)drdr$$

$$=\frac{2\sqrt{\pi^{\,k}}\,r^{\,k+1}}{\Gamma\left(\frac{k}{2}\right)k\,(k\,+\,1)}\,, \qquad (5)$$

where: $HS(k,2,r)$ is hypersphere function of the second freedom degree. The integration procedure to higher freedom degrees would lead to the n -th hypersphere function, and general value of the placed integral. This is also a way how, and the basic of recurrent relation based a differentiation, respectively integrating (integralling), to perform hypersphere function generalization. So, there are:

$$HS(k,n,r)=\int_0^r HS(k,n-1,r)\,dr\,. \qquad (6)$$

This result can be achieved in many equivalent ways.

On the basis of general integral value follows final or *generalized form of hypersphere function*, as:

$$HS(k,n,r)=\frac{2\sqrt{\pi^{k}}\,r^{k+n-3}\Gamma(k)}{\Gamma(k+n-2)\Gamma\left(\frac{k}{2}\right)}\quad (k,n\in N\,). \qquad (7)$$

Concluding about general value of this integral, is supported by the fact that:

$$\frac{1}{\prod_{i=0}^{n-3}(k\,+\,i)}=\frac{1}{k\,(k\,+\,1)(k\,+\,2)\,\cdots\,(k\,+\,n\,-\,3)}=$$

$$\frac{1}{\prod_{i=0}^{n-3}(k+i)}=\frac{1}{k(k+1)(k+2)\cdots(k+n-3)}=\frac{\Gamma(k)}{\Gamma(k+n-2)}=\frac{1}{(k)_{n-2}}\,. \qquad (8)$$

The previous functional equation can be presented as well by so-called *Pochhammer's* symbols [20]. At the same time, with the deductions according to the radius r, we can, as well, perform "motions" to lower *freedom degrees*, starting from the n -th one, on the basic of recurrent relations of the type

$$\frac{\partial}{\partial r}HS(k,n,r)=HS(k,n-1,r)\qquad \text{and}$$

$$\frac{\partial^n}{\partial r^n}HS(k,n,r)=HS(k,0,r)\,. \qquad (9)$$

The example of derivation of known sphere functions shows that the second ($n=2$) freedom degree is achieved, and it is level lower than level of volumes (columns).

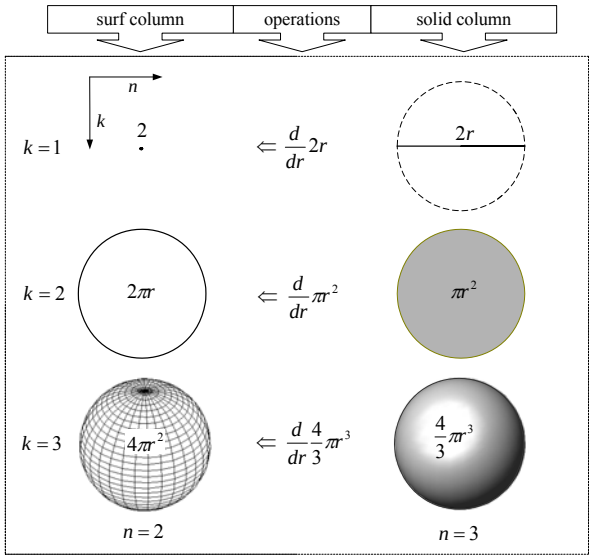


Figure 1. Moving through the vector of real surfaces: Deducting one freedom degree k of surface sphere we get circumference, and for two (degrees) we get the point 2 (left). Moving through the vector of real solids, i.e. by deduction of one freedom degree k from the sphere, we get a circle (disk), and for two (freedom degrees), we get a line segment or diameter (right column)

The right column in Figure 1 relates to sphere solids. For example *HS* - solid with the freedom degree $k=n=3$ is a sphere solid or a sphere. Within this column, for $k<3$ every primitive is called a hypersphere solid. For $k>3$ ($k\in Z$) these objects are hypersphere primitives. Consequently, the first hyporsphere solid is a circle (or a disc), while the second is diameter. Formally, they are solids with deducted one, respectively two freedom degrees k. When k is $\in R$, then, also, are stake hypersphere objects. Similar division is related as well to the freedom degree $n=2$, that relates to the surface sphere objects, or surf (the left column).

II. THE MATRIX FORM OF HYPERSPHERE FUNCTION

A. *The Summary of Hypersphere functions*

Taking into consideration that the hypersphere function is two-dimensional (leaving reserve toward radius r, as the third dimension), and that its values change depending on number of freedom degrees n and k ($i=0,1,2,3,\ldots n$ and $j=0,1,2,3,\ldots k$), cache of them can be presented, in discrete form, by the next matrix

$$\left[M_{ij}\right]=M_{kxn}\,. \qquad (10)$$

For example for two adjacent columns of this matrix $[M]^{<i-1>}$ and $[M]^{<i>}$ follows:

$$[M]^{<i-1>} = \frac{\partial}{\partial r}\begin{bmatrix} \frac{r^{i-3}}{\Gamma(i-2)} \\ \frac{2r^{i-2}}{\Gamma(i-1)} \\ \frac{2\pi r^{i-1}}{\Gamma(i)} \\ \frac{8\pi r^{i}}{\Gamma(i+1)} \\ \vdots \\ \frac{2\sqrt{\pi^k}\, r^{k+i-3}\Gamma(k)}{\Gamma(k+i-2)\Gamma(k/2)} \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{r^{i-4}}{\Gamma(i-3)} \\ \frac{2r^{i-3}}{\Gamma(i-2)} \\ \frac{2\pi r^{i-2}}{\Gamma(i-1)} \\ \frac{8\pi r^{i-1}}{\Gamma(i)} \\ \vdots \\ \frac{2\sqrt{\pi^k}\, r^{k+i-4}\Gamma(k)}{\Gamma(k+i-1)\Gamma(k/2} \end{bmatrix}, \text{ for } \quad i = \overline{0,n}\,. \quad (11)$$

Interesting results can be obtained on the basic of *horizontal* (n) or *vertical* (k) *freedom degrees*. Thus, for example, it follows that

$$\left.\frac{2\sqrt{\pi^k} r^k}{k\Gamma(k/2)}\right|_{k=3} \vee \left.\frac{8\pi r^n}{\Gamma(n+1)}\right|_{n=3} \vee \left.\frac{2\sqrt{\pi^k} r^{k+n-3}\Gamma(k)}{\Gamma(k+n-2)\Gamma(k/2)}\right|_{k=3\wedge n=3}$$

$$\Rightarrow \qquad \frac{4}{3}\pi\, r^{\,3}\,. \qquad (12)$$

Consequently, this property is very essential, because it can bring the same result on the basic of two special formulas, or using only one general (formula).

On the basic of general form (10) of the matrix M_{kxn} $(k, n \in N)$ follow concrete values (Figure 2) for the selected submatrix 11x12 ($j = -3, -2, \ldots 7;\ \ i = -5, -4, \ldots 6$).

B. Hypersphere Function Gradients

Gradient can be applied to hypersphere function, taking into consideration its multi-dimensionality. As this function contains three independently changeables: k, n and r, follows the solution of the gradient functions $\nabla_{k,n,r}$ in vector form:

$$\nabla_{k,n,r}\{HS(k,n,r)\}$$

$$= HS(k,n,r)\begin{bmatrix} \ln 2r\sqrt{\pi} + \frac{1}{2}\psi_0\left(\frac{k+1}{2}\right) - \psi_0(k+n-2) \\ \ln r - \psi_0(k+n-2) \\ \frac{1}{r}(k+n-3) \end{bmatrix}. \quad (13)$$

This function has special meaning with the establishment of the extreme of contour HS -functions.

C. The graphics of hypersphere functions

The graphics of the function $HS(k,n,r)$ can be presented for the field of real numbers, including both freedom degrees, and accepting radius as parameter.

Graphic generating (Figure 3) has been performed by the program package *Mathematica*. In $k, n \in N$ domain and with the unit radius, the function is convergent. In the other three quadrants of the coordinate system, the oscilations are characteristical, the appearance of zeros and the singularities of the function $HS(k,n,1)$, so it is demanding for analytical testings.

The same characteristics are noted on the presented matrix, too (Figure 2).

$$
M_{k,n}=\begin{array}{c|ccccccccccc|c}
 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & n/k \\
\hline
 & \frac{113400}{\pi^3 r^{11}} & -\frac{11340}{\pi^3 r^{10}} & \frac{1260}{\pi^3 r^9} & -\frac{315}{12\pi^3 r^8} & \frac{45}{2\pi^3 r^7} & -\frac{15}{4\pi^3 r^6} & \frac{3}{4\pi^3 r^5} & -\frac{3}{16\pi^3 r^4} & \frac{1}{16\pi^3 r^3} & -\frac{1}{32\pi^3 r^3} & \frac{1}{32\pi^3 r} & -5 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{12\pi^2} & -4 \\
 & -\frac{10080}{\pi^2 r^9} & \frac{1260}{\pi^2 r^8} & -\frac{180}{\pi^2 r^7} & \frac{30}{\pi^2 r^6} & -\frac{6}{\pi^2 r^5} & \frac{3}{2\pi^2 r^4} & -\frac{1}{2\pi^2 r^3} & \frac{1}{4\pi^2 r^2} & -\frac{1}{4\pi^2 r} & \infty & \infty & -3 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2\pi} & -\frac{r}{2\pi} & -\frac{r^2}{4\pi} & -2 \\
 & \frac{720}{\pi r^7} & -\frac{120}{\pi r^6} & \frac{24}{\pi r^5} & -\frac{6}{\pi r^4} & \frac{2}{\pi r^3} & -\frac{1}{\pi r^2} & \frac{1}{\pi r} & \infty & \infty & \infty & \infty & -1 \\
 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & r & \frac{r^2}{2} & \frac{r^3}{6} & \frac{r^4}{24} & 0 \\
 & 0 & 0 & 0 & 0 & 0 & 2 & 2r & r^2 & \frac{r^3}{3} & \frac{r^4}{12} & \frac{r^5}{60} & 1 \\
 & 0 & 0 & 0 & 0 & 2\pi & 2\pi r & \pi r^2 & \frac{\pi r^3}{3} & \frac{\pi r^4}{12} & \frac{\pi r^5}{60} & \frac{\pi r^6}{360} & 2 \\
 & 0 & 0 & 0 & 8\pi & 8\pi r & 4\pi r^2 & \frac{4\pi r^3}{3} & \frac{\pi r^4}{3} & \frac{\pi r^5}{15} & \frac{\pi r^6}{90} & \frac{\pi r^7}{630} & 3 \\
 & 0 & 0 & 12\pi^2 & 12\pi^2 r & 6\pi^2 r^2 & 2\pi^2 r^3 & \frac{\pi^2 r^4}{2} & \frac{\pi^2 r^5}{10} & \frac{\pi^2 r^6}{60} & \frac{\pi^2 r^7}{420} & \frac{\pi^2 r^8}{3360} & 4 \\
 & 0 & 64\pi^2 & 64\pi^2 r & 32\pi^2 r^2 & \frac{32\pi^2 r^3}{3} & \frac{8\pi^2 r^4}{3} & \frac{8\pi^2 r^5}{15} & \frac{4\pi^2 r^6}{45} & \frac{4\pi^2 r^7}{315} & \frac{\pi^2 r^8}{630} & \frac{\pi^2 r^9}{5670} & 5 \\
 & 120\pi^3 & 120\pi^2 r & 60\pi^2 r^2 & 20\pi^3 r^3 & 5\pi^3 r^4 & \pi^3 r^5 & \frac{\pi^3 r^6}{6} & \frac{\pi^3 r^7}{42} & \frac{\pi^3 r^8}{336} & \frac{\pi^3 r^9}{3024} & \frac{\pi^3 r^{10}}{30240} & 6
\end{array}
$$

(The $n \in N$ axis runs horizontally across the columns; the $k \in N$ axis runs vertically down the rows.)

Figure 2. *Submatrix* $HS(k,n,r)$ *of the function that covers one area of real freedom degrees (* $k,n \in R$ *). There are also noticeable the coordinates of six characteristically sphere functions (*∞ *are singular values, and* 0 *are zeros of this function)*

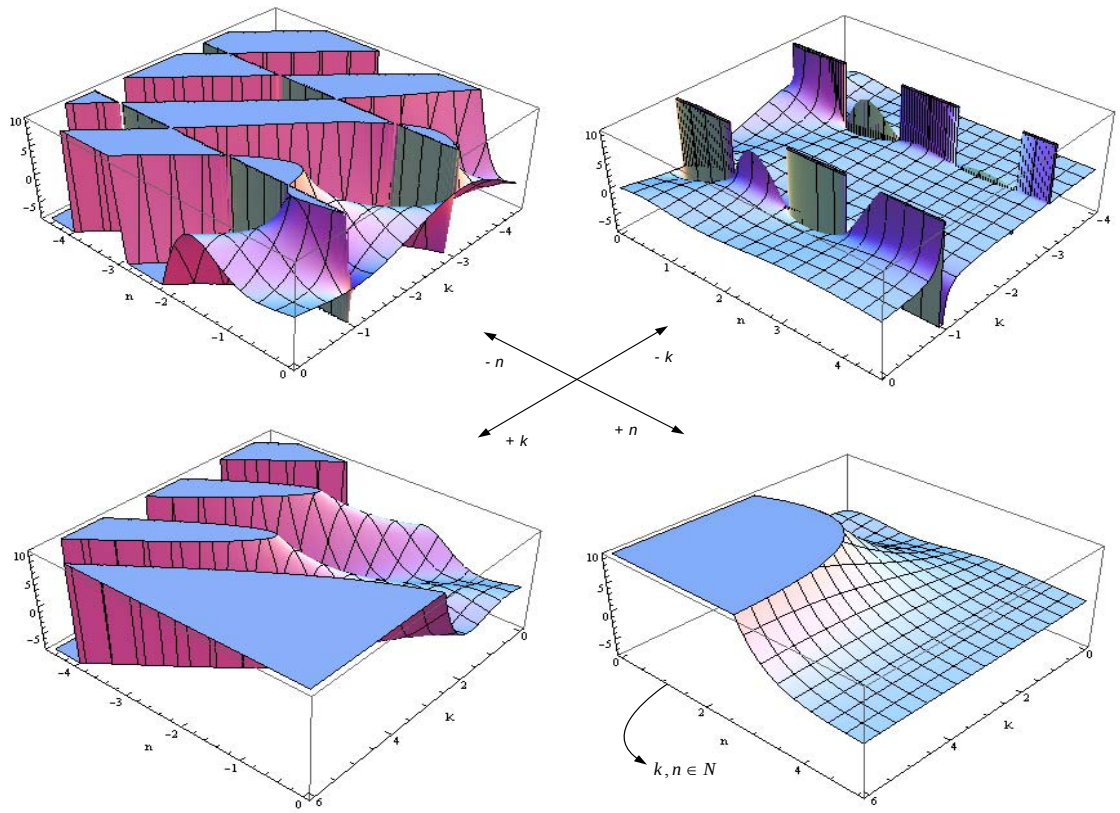


Figure 3. Visualization of the hypersphere function in the field of real numbers, for all four quadrants

III. CONCLUSION

In researching the hypersphere function there are left at stake the next expanding: dimensions k and n higher than three. Continual dimensions that include noninteger values, zero dimension, the domain of lower than zero dimensions, consequently, the area of complex freedom degrees and the like theoretically they exist, and the generalization achieved by this paper, directs to generality in phenomene that are met in multi-dimensional geometry and topology. In any case, the matrix of negative dimensions demands special structuring and deeper mathematical analysis. Hyperspheres present one part of multi-geometrical objects. In the introduction is stated that sphere is the most symmetrical geometric body. Some other bodies have such property, but not completely. Namely, cube is symmetrical along orthogonal coordinates, if its center is in the concurrence of big diagonals, and the axes of coordinate system are parallel to cube edges. This property have some other polyhedrons. Cube, because of its orthogonallity, has retained the property, that its functions derivations of solid entity (create) generate surf-entity functions. With the objects that have not the property that recurrency is being made on the basic of derivation function, have to be introduced new freedom degrees depending on object complexity. The next object that has some properties similar to sphere, is cilynder (this similarity is caused by the equality of cilynder height and diameter). Its symmetry is twofold and is not any more unambigous along all coordinate axes as with sphere. These properties are essential suppositions for defining recurrent relations that would give the connestion among functions of some entities and along the columns and along progressions (lines) of cube, respectively cylinder. The examined analysis of multi-dimensional space and formula of this geometry, leads to the conclusion about its complexity and connections with spacial functions and the other areas of mathematics.

ACKNOWLEDGEMENT

Paper presents results of project TR32044 "The development of software tools for business process analysis and improvement" financed by Ministry of science and technology development, Republic of Serbia.

REFERENCES

[1] N. Cakić, "A note Stirling numbers", The Fibonacci Quarterly, 40, 4-5, 1998.

[2] J. H. Conway, "Sphere packing, lattices and groups", 2nd ed. New York: Springer-Verlag, p. 9, 1993.

[3] E. Freden, Summing a series of volumes (Problem 10207), Amer. Math. Monthly 100, 882, 1993.

[4] G. H. Hardy, "Ramanujan, Twelve lectures on subjects suggested by his life and work", Cambridge, 1940.

[5] C. H. Hinton, "The fourth dimenzion", Pomeroy, WA: Health Research, 1993.

[6] J. G. Hocking and G. S. Young, "Topology", New York: Dover, 1988.

[7] D. Letić, "Generalization of hypersphere function", (scientific monograph), Technical Faculty "M. Pupin", Zrenjanin, 2009.

[8] D. Letić, and D. Cakić, "Algorithms of mathematical constants", Technical Faculty "M. Pupin", Zrenjanin, 2009.

[9] H. Manning, "Geometry of fourth dimension", New York: Dover, 1956.

[10] C. R. F. Maunder, "Algebraic Topology", New York: Dover, 1997.

[11] S. D. Mitrinović and D. J. Kečkić, "Introduction into special functions", Scientific Book, Belgrade, 1991.

[12] E. H. Neville, "The Fourth Dimension", Cambridge, England: Cambridge University Press, 1921.

[13] R. Von B. Rucker, "The Fourth Dimension: A Guided Tour of the Higher Universes", Boston, MA: Houghtson Miffin, 1984.

[14] D. M. Y. Sommerville, "An Introduction to the Geometry of n Dimensions", New York: Dover, P. 136, 1958.

[15] D. Wells, "The Penguin Dictionary of Curious and Interesting Numbers", Middlesex, England: Penguin Books, 1986.

[16] http://mathworld.wolfram.com/Ball.html.

[17] http://mathworld.wolfram.com/Four-DimensionalGeometry.html.

[18] http://mathworld.wolfram.com/Hypercube.html.

[19] http://mathworld.wolfram.com/Hypersphere.html.

[20] http://mathworld.wolfram.com/Pochhammer Symbol.html.

Visualization and multimedia presentation of cultural heritage

S. Rizvic, V.Okanovic and A. Sadzak*
* Faculty of Electrical Engineering Sarajevo, Sarajevo, Bosnia and Herzegovina
srizvic@etf.unsa.ba, vokanovic@etf.unsa.ba, asadzak@etf.unsa.ba

Abstract - Understanding the past is one of the key factors of human culture. Cultural heritage contributes to the preservation of collective memory. It is difficult to imagine the original appearance of monuments while observing their archeological remains. Digital technologies are an efficient tool for visualization and multimedia presentation of cultural heritage. This paper describes the use of these technologies through the workflow of the White bastion 4D visualization project, from collecting information, through 3D modeling and texturing to interactive web implementation using the most advanced web 3D technologies. We discuss the advantages and drawbacks of each technology for this particular purpose.

I. INTRODUCTION

Cultural heritage (CH) is a silent witness of our past. We are surrounded by remains that the previous generations left behind, as remainders of the glorious times, events and cultures. Most of those remains do not reflect the magnificence of the objects in their time. Computer graphics and digital technologies provide tools to recreate the original appearance of cultural heritage objects and take the visitors to the journey through the past.

There are many techniques and methods used for visualization of cultural heritage. This field is still a very hot research topic. However, most of the CH visualization professionals use similar workflows to obtain the desired goal. Visualization can be performed for various purposes, such as archaeological scientific research, archiving and presentation. In this paper we will focus on cultural heritage visualization for presentation purposes, particularly the interactive visualization.

The goal of interactive CH visualization is to enable the user to interact with the digital model of CH object. In case of small objects, for example museum artifacts, the user can turn them around and click on their various parts, change the illumination and open the additional digital content for learning about their history and purpose. In case of buildings and architectural complexes, interactive visualization enables the user to virtually walk through the objects and explores their various sections, learning about them through the additional digital content such as stories, avatars and 3D models of their smaller elements.

We will present the workflow of a 3D interactive visualization using the case study of the White Bastion 4D digitization project. We will identify the common problems and issues arising in the process of conversion of the 3D geometry from non-interactive into interactive format. We will analyze different web 3D technologies for presenting 3D models online and discuss their advantages and drawbacks. At the end we will offer our conclusions as a contribution to the future CH visualization projects.

II. RELATED WORK

Interactive visualization of cultural heritage has been implemented in the last decade through many projects and case studies. They range from 3D scanning and virtual restoration of damaged artworks and artifacts to interactive virtual reconstruction of whole buildings and monuments.

The Digital Michelangelo project [1] presents a hardware and software system for digitizing the shape and color of large fragile objects under non-laboratory conditions using laser triangulation rangefinders, laser time-of-flight rangefinders, digital still cameras, and a suite of software for acquiring, aligning, merging, and viewing scanned data. As a demonstration of this system, they digitized 10 statues made by Michelangelo, including the well-known figure of David, two building interiors, and all 1163 extant fragments of the Forma Urbis Romae, a giant marble map of ancient Rome. It was one of the first achievements in cultural heritage digitization, introducing the issues of large data sets and their management in the research. Some of those issues were addressed and solved through MeshLab, a system for the processing and editing of unstructured 3D triangular meshes [2].

Ancient history comes to life through virtual walks around Ancient Olympia [3] and valley of pyramids in Giza [4]. At the Virtual Reality programme in the Hellenic Cosmos Theater in Athens, there is a tour at the peninsula and the city of Miletus, as it was 2000 years ago. Visitors can explore the virtual city and learn about the most important public buildings. Starting from the gate at the Port of the Lions, they enter the sanctuary of Apollo Delphinios. Then, they visit the Ionian Stoa, which housed the stores of the city, the Hellenistic Gymnasium, the North Agora and the Bouleuterion, being able to see even the details of the buildings' architecture. They can “fly” over the city for one last panoramic view of the perfect city planning design, or even “dive” into its port.

For the famous Giza Pyramids, the Sphinx, and surrounding tombs and temples, just west of modern Cairo (3rd millennium BCE), the Giza Project at Harvard University is putting together older traditional archives

(digital photos, archaeological drawings, object metadata) with realistic 3D visualization of the site. This project provides revolutionary access to Giza, its statues, hieroglyphic inscriptions, architecture, and wall decoration. Real-time immersive models allow the visitors to pose new research questions, provide interactive classroom instructions, and investigate diachronic approaches to Giza's evolution over several millennia.

The potential of 3D technologies is proven also in protection of cultural heritage sites from destruction caused by many visitors. The Pure Land project presents "a full-scale augmented digital facsimile of Cave 220 from the UNESCO World Heritage site, the Mogao Grottoes, Gansu Province, northwestern China. The peerless treasuries of paintings and sculptures at Dunhuang are extremely vulnerable. Comprehensive digitization has become a primary method of preservation at the site." [5]

Virtual Museum Transnational Network V-MusT.net has integrated a number of interactive visualizations of cultural heritage into the international multimedia exhibition Keys To Rome [6], held at the same time in Rome, Amsterdam, Alexandria and Sarajevo. The exhibition presents these 4 locations within the Roman Empire through an interactive storytelling application enabling the visitors to walk through Roman villas and thermal baths, basilicas and temples, collecting objects and learning about their history.

The mentioned projects represent just a small part of a rich collection of virtual cultural heritage applications. Through this work some open issues aroused and have been identified by Scopigno in [7]. The problems of obtaining interactivity without sacrifice on quality (simplification and multi-resolution), web and mobile presentations of the interactive virtual environments, finding the best interaction and navigation methods and integration with other media are still subjects of intensive scientific research. We hope to offer some solutions through the case study presented in this paper.

III. CH VISUALISATION WORKFLOW

The fortification known as „White Bastion" is one of the most impressive and important historical sites in Sarajevo. It is located on the southeast outskirts of the City, with an overview on the city valley (Fig. 1). The remains found on the site date from 14th till 18th century. The goal of the White Bastion 4D digitization project is to present the development of this valuable cultural heritage object through history using interactive 3D visualization combined with digital storytelling.

Although the cultural heritage visualization applications differ in technologies they use and purpose for which they have been created, we can define a common workflow they all follow during their development. This workflow consists of steps presented in Fig. 2.

Collection of materials for the White Bastion project included the drawings and blueprints of the archaeological remains, consultations with historians and archaeologists working on excavations and review of available literature from historical sources. This step resulted with identification of 6 different building shapes to be modeled and presented in time within the interactive application.

Figure 1. White bastion fortress, Sarajevo, Bosnia and Herzegovina

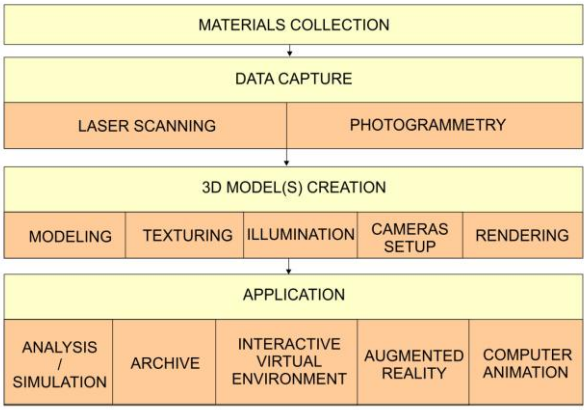


Figure 2. Cultural heritage visualization workflow

For the fortress geometry we skipped the data capture step, as we had no equipment available for laser scanning the remains, while capturing the geometry using photogrammetry was not possible due to the large dimensions of the object and difficulties of approach from some sides.

However, the selected archaeological findings, such as remains of weapons, helmets and similar objects will be digitized using photogrammetry and added to the 3D models' geometry.

IV. 3D MODELLING

In order to present the development of the White bastion through history, we consulted the archaeologist who was leading the excavations on the site. He identified six 3D models of the object to be created, some of them different from each other by the addition of interior objects, and exterior extensions.

The modeling process was started by building the terrain. In order to create precise terrain model the Digital Elevation Model over Balkans area from the GMES RDA project [8] was used. This geotiff map was then imported in Global Mapper - GIS application for reading and

conversion of different kinds of spatial datasets. Sarajevo area terrain was extracted from the map and exported as 3D VRML file which was then used in Cinema 4D as the precise Sarajevo terrain model.

3D models of White bastion are created in Cinema 4D, while Adobe Photoshop is used for creating and adjusting the textures. Six final 3D models correspond to six different construction phases, the first one, from the medieval period, looking completely different from the last one from Austro-Hungarian period.

In order to achieve photo-realism in renders of Bastion, the atmosphere with vegetation and Miljacka river flow was added and rendered at high quality (Figures 3 and 4).

Figure 3. Exterior and interior 3D models of fortress from the medieval period

Figure 4. 3D model of fortress from the first phase of the Ottoman period

V. CONVERTING TO HYPERMEDIA

For the purpose of web presentation of 3D models a few different web technologies we have used. Firstly, we have used MeshLab for conversion the 3D models of White bastion to PLY format which was then used in 3D Heritage Online Presenter (3DHOP) [9]. 3DHOP constructs the model web presentation through the WebGL API technology and SpiderGL JavaScript library. Although we have used 3DHOP for the model web presentation in the other projects before, this time it was unusable. We have previously used the much simpler 3D models, but this time the 3D models (of White bastion) were quite complex and 3DHOP was unable to load and present them.

After an unsuccessful attempt with 3DHOP, we have decided to use XML based X3D files which are well supported in HTML5. MashLab is again used for conversion the 3D models of White bastion to X3D format. When we export X3D file, we can convert X3D format to HTML5 webpage, for example using InstantReality X3D converter [10]. However, similar to the previous case, and this time, due to the complexity of the models and size of the XML file, conversion to HTML webpage was not successful.

After all, we have decided to use Unity Web Player [11] with whom we have had the most success in the other projects. 3D models of White bastion, created in Cinema 4D, were exported as FBX file which was then used in Unity Web Player. From Unity Web Player we have build web presentation of the models without any problems.

VI. INTERACTIVE DIGITAL STORYTELLING APPLICATION

The final output planned in this project is an interactive digital storytelling application which will combine the geometry of the 6 models with stories about the object, its history and related events and characters. This part of the project is still in progress.

Our idea is that the user of the application will watch interactive digital stories interlinked in a common structure with interactive 3D models. The stories consist of the main story presenting the insight in the chronology of White bastion's development and sub-stories presenting details about different time periods in object's history,

legends related to the object and characters who left their mark in development of the fortress. Each digitized archaeological artifact will be adjoined with the story about its purpose and history. From each sub-story the user will be able to open the link to the corresponding 3D model of the fortress and explore it virtually.

VII. CONCLUSION

In this paper are discussed digital technologies that are an efficient tool for visualization and multimedia presentation of cultural heritage. The paper described the use of these technologies through the workflow of the White bastion 4D visualization project, from collecting information, through 3D modeling and texturing to interactive web implementation using the most advanced web 3D technologies. Benefits and limitations of using these technologies for visualization and multimedia presentation of cultural heritage are indicated.

ACKNOWLEDGMENT

The project "Visualization of White Bastion fortress based on 3D Spatial Documentation of Material Cultural Heritage And Interactive Digital Storytelling" is a case study of the MPNS COST Action TD1201 "Colour and Space in Cultural Heritage" (COSCH).

REFERENCES

[1] M. Levoy, The Digital Michelangelo Project: 3D Scanning of Large Statues, Proceedings of Siggraph 2000

[2] P Cignoni, M Corsini, G Ranzuglia, Meshlab: an open-source 3d mesh processing system, Ercim news 73, 45-46

[3] A. Gaitatzes, D. Christopoulos, G. Papaioannou, Virtual Reality Systems and Applications: The Ancient Olympic Games, Advances in Informatics, Lecture Notes in Computer Science Volume 3746, 2005, pp 155-165

[4] P. Der Manuelian, Giza 3D: Digital archaeology and scholarly access to the Giza Pyramids: The Giza Project at Harvard University, Digital Heritage International Congress (DigitalHeritage), 2013 (Volume:2), pp 727 – 734, IEEE, ISBN 978-1-4799-3168-2

[5] S. Kenderdine, "Pure Land": Inhabiting the Mogao Caves at Dunhuang, Curator: The Museum Journal, Volume 56, Issue 2, pages 199–218, April 2013

[6] Keys To Rome exhibition, V-MusT.net, www.keystorome.eu

[7] R. Scopigno, Interactive Visualization for Cultural Heritage - Current Capabilities and Open Issues, VISIGRAPP 2013, Barcelona

[8] GMES RDA project (EU-DEM), http://www.eea.europa.eu/data-and-maps/data/eu-dem

[9] 3DHOP (http://vcg.isti.cnr.it/3dhop/)

[10] Instant Reality X3D converter (http://doc.instantreality.org/tools/x3d_encoding_converter/)

[11] Unity Web Player (http://unity3d.com/webplayer)

Using Entropy Information Measures for Edge Detection in Digital Images

Diego Sušanj[*], Vjeran Tuhtan[*], Luka Lenac[*], Gordan Gulan[#], Ivica Kožar[†] and Željko Jeričević[‡]
[*]Engineering Faculty Students, Rijeka, Croatia
[#]Medical School Rijeka, Croatia
[†] Department of Computer Modeling/Civil Engineering Faculty, Rijeka, Croatia
[‡]Departmentof Computer Engineering/Engineering Faculty, Rijeka, Croatia
zeljko.jericevic@riteh.hr

Abstract - Shannon information entropy measures were used as filters of different kernel sizes to detect edges in digital images. The concept is based on communication theory with splitting of edge detection kernel into source and destination parts. The arbitrary shape of the kernel parts and the fact that information filter output is a real number with reduced problem of edge's continuity represents the major advantage of this approach. The results are compared with traditional edge detection algorithms like Sobel to illustrate performance and sensitivity of the information entropy filters. Besides the well known test image Lena, the real life examples are taken from medical X-Ray imaging of knee joints in order to illustrate the algorithm performance on real data.

I. INTRODUCTION

Shannon developed his information theory [1] for the analysis of communication channels. Central to his theory is the concept of information entropy (1) which is a measure of unpredictabilty of information. However, the concept of information entropy and entropy measures are used in other fields besides communications. Some examples are statistical inference, natural language processing, cryptography, evolutionary biology and various forms of data analysis [2],[3].

The edge detection is one of the most important procedures in image processing. It is usually a first step in image processing sequence to acieve segementation of images, extraction of details, search for a pattern, data reduction and many other complex tasks.

The idea to use information entropy has been proposed more twenty five 25 years ago in digital microscopy [4], and in various forms has been in sporadic use since then [5],[6]. It is still active area of research [7],[8] wiith open possibility for improvement and wider application.

The main goal in our work presented here was to achieve continuity of the edge in low contrast and noisy X-ray images, which is a first necessary step for shape analysis of femur (thigh bone) in a knee joint. The success opens the possibility for automatic shape analysis of femur.

II. THEORY

The information entropy H, is defined in (1) for a system containing n discrete states, each state i with its own probability p_i.

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

Some information theory measures which are used for simultaneously analyzing two distributions are presented bellow. Traditionally, those two distributions (X & Y) are source and destination in communication theory. The measures used in this paper are joint entropy (2), conditional entropy (3) and mutual information (4).

Joint entropy

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i, y_j\right) \log_2 p\left(x_i, y_j\right) \tag{2}$$

Joint entropy $H(X,Y)$ is the measure of how much information is contained in a system of two random variables. It ranges from $H(X)$ for the same variables to $H(X)+H(Y)$ for independent variables.

Conditional entropy

$$H(X \mid Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i, y_j\right) \log_2 p\left(y_j \mid x_i\right) \tag{3}$$

Conditional entropy $H(X|Y)$ is a measure of information still contained in X after Y is known. It ranges from zero for the same variables to $H(X)$ for independent variables.

Mutual information

$$I\left(X,Y\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} p\left(x_i, y_j\right) \log_2 \frac{p\left(x_i, y_j\right)}{p\left(x_i\right) p\left(y_j\right)} \tag{4}$$

Mutual information $I(X,Y)$ is a measure of mutual dependence of variables X and Y. It ranges from zero (for independent variables) to $H(X)$ for the same variables.

Sobel operators

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \qquad (5)$$

Sobel operator is a simple filter for finding edges in digital images. It has 3 by 3 kernels (5) which are convolved with a testing image [9].

III. METHOD

One of the reasons for using information entropy kernel is that their size and/or shape can be very arbitrary. The only restriction is that two variable kernels *H(X,Y)* (2), *H(X|Y)* (3) and *I(X,Y)* (4) have to be split into two equal parts with pixels in one to one correspondence (source and destination parts). Information entropy kernel *H(X)* has no such restriction (1).

To illustrate improvements of our entropy based methodology over Sobel filtering the series of pictures is presented here. Main test image is the well known test picture Lena found in most digital image processing books [9].

The Fig. 1 shows application of Sobel operator on Lena picture. The main problem is discontinuity and noisy appearance of the edges.

Fig. 2 shows the result for application of information entropy operator *H(X)* in the direction of columns, while Fig. 3 shows the result for application of information entropy operator *H(X)* in the direction of rows.

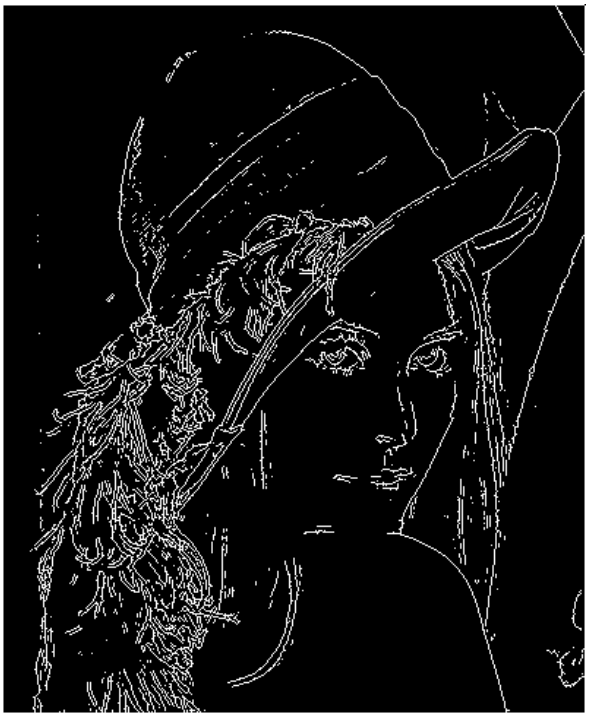


Figure 1. Sobel operator (5) applied to Lena picture

Figure 2. Entropy operator (1), columns direction

Figure 3. Entropy operator (1), rows direction

Figure 4. Conditional entropy operator *H(X|Y)* (3), columns direction

The application of information entropy operator *H(X)* (1) in both directions Fig. 2 and Fig. 3. shows better continuity of the edges and lower level of high frequency noise.

Fig. 4 shows the result for conditional entropy operator H(X|Y) while Fig. 5 shows the result for joint entropy operator H(X,Y)

Figure 5. Joint entropy operator *H(X,Y)* (2), columns direction

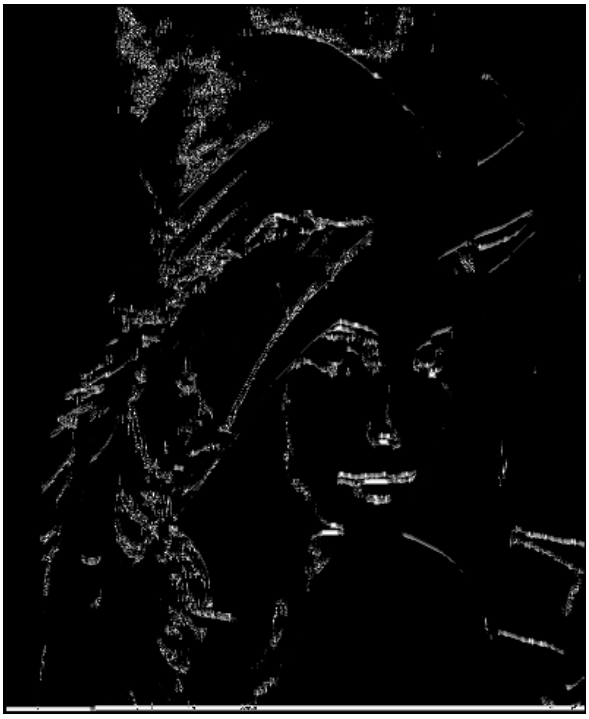


Figure 6. Mutual information *I(X,Y)* (4), columns direction

Fig. 6 shows the result for mutual information operator *I(X,Y)*. All information entropy kernels are of size 5 by 5.

The results on test images illustrate that the main purpose of designing entropy based edge detection has been achieved. With the exception of mutual information operator (4) whose results are shown in Fig. 6, all entropy operators have a better continuity of edges then Sobel operator.

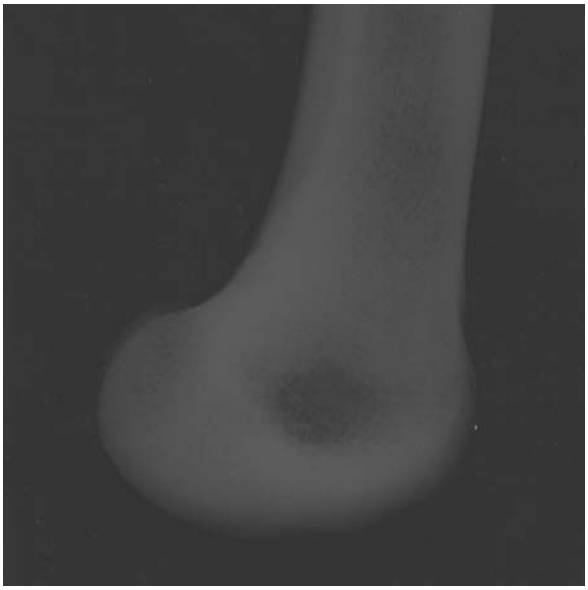


Figure 7. X-Ray image of knee bone

The entropy operators are also applied to the X-ray pictures of bone from knee joint. This type of images are often low contrast and noisy, but entropy algorithms can be adjusted to work on them. The example of such an

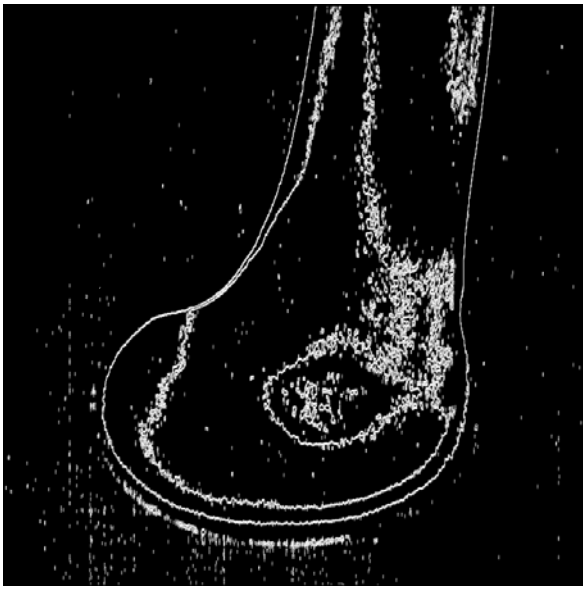


Figure 8. Entropy *H(X)* (1), applied on image in Fig 7, column direction

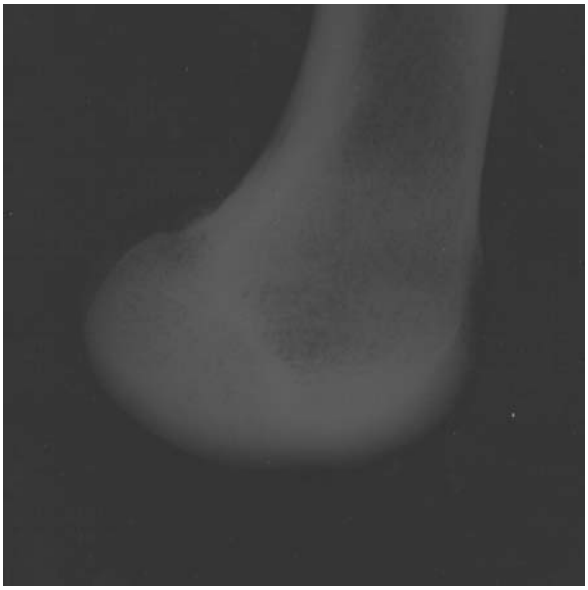


Figure 9. X-Ray image of knee bone

image is shown in Fig. 7 and application of entropy operator in Fig. 8.

Another example of femur image is shown in Fig. 9 and application of entropy operator in Fig. 10.

IV. CONLUSION

The information entropy based approach using four entropy based measures for detecting edges in X-ray digital images has been outlined and tested. First testing was done on a well known image Lena, then on the X-ray images of femur bones. Our results strongly indicate that the better edge continuity is achievable with entropy operators then with Sobel operator. The only exception was mutual information operator (4). The better performance is due to considerable flexibility which

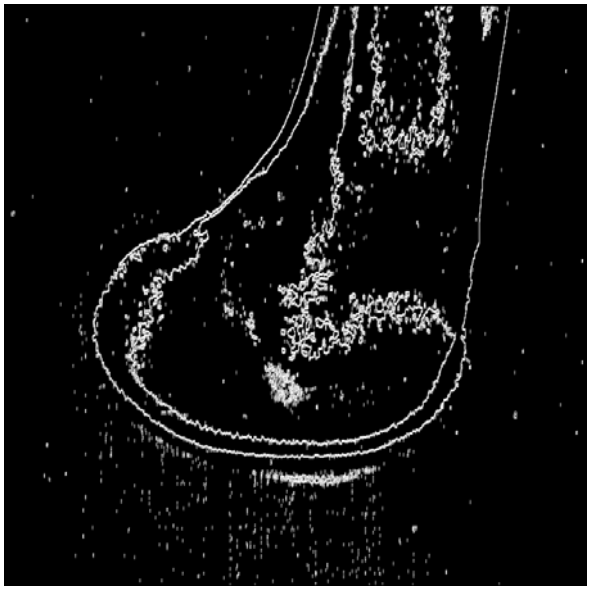


Figure 10. Entropy *H(X)* (1), applied on image in Fig 9, column direction

entropy based operators have in sense of size, shape and orientation, as well as the fact that their result is a real number from predefined range. The edge continuity opens a possibility for automatic edge detection and bone shape extraction and analysis which is the main goal of our future work. From the medical standpoint, the shape analysis of multiple knee joints will provide orthopedic surgeon with detailed statistics about variability in the shape of a knee joints.

Future work would also include construction of user friendly software (on the basis of our present research version), targeted for the use by medical personnel.

REFERENCES

[1] Shannon, C.E. "A Mathematical Theory of Communication". *Bell System Technical Journal*, **27** (1948): 379–423.

[2] Gray, R.M. "Entropy and Information Theory", 2nd Ed, 2011, Springer-Verlag, New York, pp 409.

[3] Luenberger, D.G., "Information Science", 2006, Princeton University Press, Princeton, pp 423.

[4] Barba, J., et all, "The use of local entropy measures in edge detection for cytological image analysis", Journal Microscopy (Oxford), **156** (1989) 125-134.

[5] Lenac, L., "Detekcija rubova slike upotrebom informacijske entropije", Završni rad, 2010, Tehnički fakultet, Sveučilište u Rijeci

[6] El-Sayed, M.A., "A New Algorithm based Entropic Treshold for Edge Detection in Images", International Journal of Computer Science **8** (2011) 71-78.

[7] Elaraby, A.E.A., et all, "New Algorithm For Edge Detection in Medical Images Based on Minimum Cross Entropy Thresholding", International Journal of Computer Science Issues, **11** (2014) 196-200

[8] Ashour, A.S., El-Sayed, M.A., "New Method Based on Multi-Threshold of Edges Detection in Digital Images", International Journal of Advanced Computer Science and Applications, **5** (2014), 90-99.

[9] Gonzalez, R.C, Woods, R.E., "Digital Image Processing", 3rd Ed. 2008, Pearson Prentice Hall, Upper Saddle River, pp 497.

BIOMEDICAL ENGINEERING

Wireless Electrocardiographic Monitoring in Veterinary Medicine

M. Brložnik* and V. Avbelj **

* PRVA-K, Small Animal Clinic, Gorkičeva 6, Ljubljana, Slovenia

** Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia

majabrloznik@gmail.com, viktor.avbelj@ijs.si

Abstract - **A comfortable option for long term monitoring of heart activity will hopefully soon be available in veterinary medicine also. This is the first report of the ECG data obtained in animals with a wireless body electrode attached to the skin and able to convey measured data to a smart device via low power Bluetooth technology. The ECG data were obtained from two dogs, one with sinus rhythm and the other with atrial fibrillation and ventricular extrasystoles. Beside determination of average heart and respiratory rates, identification of sinus rhythm, atrial fibrillation and ventricular premature complexes, the device was able to identify ventricular pre-excitation and asynchronous depolarization of ventricles. With appropriate software this device offers an immense potential for the veterinary cardiology. Due to its simplicity, long autonomy and high reliability it can substitute the conventional Holter monitoring in veterinary medicine.**

I. INTRODUCTION

In a patient with clinical signs caused by a cardiac arrhythmia that occurs only intermittently, a few minutes ECG performed in the veterinary hospital may be normal if the arrhythmia is not occurring at the time of measurement. A long-term monitoring is therefore required in many cases. In veterinary medicine conventional 24-hour Holter monitoring is often not easily available. The measured ECG interpretation can be obtained with human cardiology software only, which is not suitable. As a result, long-term monitoring of cardiac rhythm is expensive, time-consuming and not supported by reliable interpretation rules. Additionally, it is uncomfortable for the patient. All this set aside, the diagnostic yield can still be low due to the monitoring of a relatively short period of time. Therefore, cost-effective wireless body electrodes that enable long-term monitoring of cardiac rhythm in real-world environment are being developed. They are suitably small and attached directly to the skin with a patch carrier, therefore dubbed ECG patch monitors. They are waterproof and include wireless communication technology. Beside the wearing comfort, which is achieved by reducing the number and length of the wires, these devices have very high reliability [1, 2, 3, 4, 5, 6, 7].

In human medicine supporting clinical validation data on wireless body electrodes grow [1, 5, 6, 7, 8, 9, 10, 11, 12], but in veterinary medicine this is to our knowledge the first report of the ECG data obtained in animals with a wireless body electrode attached to the skin and connected to a smart phone (tablet) via low power Bluetooth technology. The aim of these pilot measurements was to find out whether the limitations of the conventional Holter monitoring could be surmounted.

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095.

II. PRESENTATION

A. Materials and Methods

The wireless body ECG sensor incorporating an electronic module with battery and a pair of self-adhesive skin electrodes, on a distance of 9 cm, was attached to the skin of the dog. The wireless connection of ECG sensor with the smart device (tablet, Nexus 7) was established via low power Bluetooth technology. The electrodes recorded one bipolar lead (Fig.1). Alternatively, a wire extension was used to record the standard lead I (Fig. 2). The wireless communication with a smart device allowed an immediate display and evaluation of the ECG. This device records real-time data from the ECG sensor and is described with more details in [12, 13, 14, 15]. Detailed analysis of ECG recordings was made off-line using NevroEKG custom research software for ECG interpretation [16].

Fig. 1: Wireless body ECG sensor is placed on the left side of the thorax and real-time ECG data is transmitted to the tablet on the floor in front of the patient.

Fig. 2: Wireless body ECG sensor with a wire extension is used to record the standard ECG lead I.

The ECG data were obtained from two dogs, one with sinus rhythm and the other with atrial fibrillation (AF), ventricular premature complexes (VPCs) and dilated cardiomyopathy (DCM). Different positions of electrodes have been tested and the recordings have been made while dogs were lying, sitting and walking. From recorded measurements, average heart rate and respiratory frequency have been determined. In addition, a comparison of the recordings with standard 6-lead ECG has been made.

B. Results

In the dog with sinus rhythm respiratory sinus arrhythmia (RSA) is prominent, which is evident from Fig. 3. Beat to beat frequency varies for more than two times (Fig. 4). The RSA is a vagally mediated modulation of heart rate that is only partially understood. It is synchronous with respiration. However, it remains unclear whether this is due to a central or baroreflex mechanism [17].

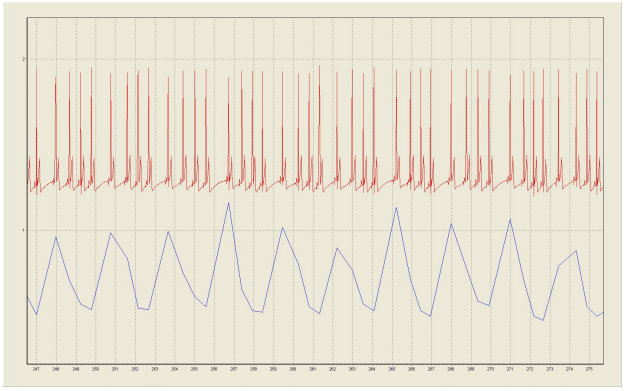


Fig. 3: Display of the recording with NevroEKG software in the dog with sinus rhythm. Upper trace is ECG signal (red line), lower trace shows the time between successive RR intervals (blue line).

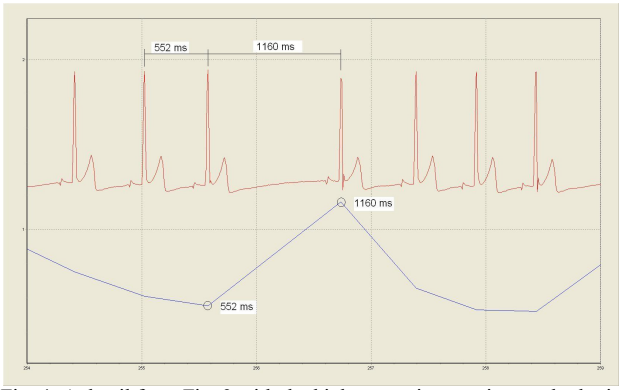


Fig. 4: A detail from Fig. 3 with the highest respiratory sinus arrhythmia (RSA). Durations of two consecutive intervals between beats are indicated. Beat to beat frequency varies for more than two times.

The sinus arrhythmia could be interpreted as a physiological irregular rhythm. To such an extent as in dogs (Fig. 3), RSA is never present in human ECG recordings.

In the measurement of the dog with AF and VPCs the AF can be recognized due to the absence of P waves, supraventricular QRST morphology and the non physiological heart rate variability, which is visible in Fig. 5. Atrial fibrillation is caused by multiple reentry circles. Ventricular rate is dependent on the refractoriness of the atrioventricular node. The VPCs can be differentiated from atrial premature complexes (APCs) due to a wide and bizarre appearance of QRST waves and the compensatory pause (Fig. 6). The frequency, configuration, consecutiveness and the rate of the VPCs can be evaluated, and as such, graded. In this case there were 3.6 VPCs/min (18 VPCs in 5 min recording), there were no couplets or triplets, no ventricular tachycardia and no fast rate.

Beside the correct identification of atrial fibrillation and ventricular premature complexes, as shown in Fig. 5, the device was able to identify also an asynchronous depolarization of ventricles (R notch), in the dog with VPCs, which is evident from Fig. 6. In standard 6-lead ECG R notch was seen only in lead III. A moderate slope at the start of R wave (before the encircled notch in Fig. 6) has as well been observed, which might indicate pre-excitation of the ventricles.

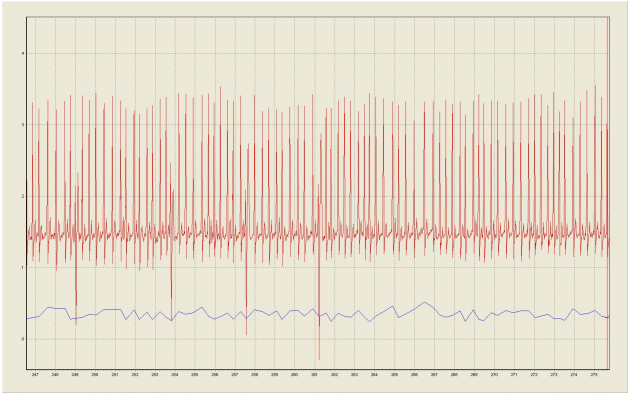


Fig. 5: Atrial fibrillation and 4 ventricular premature complexes in a dog with DCM. The electrodes were placed on the left side of the thorax, negative on atrium and positive on ventricle. ECG (red line), beat to beat heart rate (blue line).

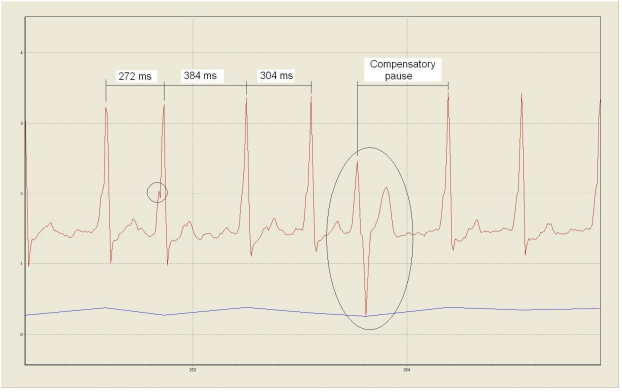


Fig. 6: A detail from Fig. 5. Duration of three consecutive intervals among heart beats. A notch in R wave and a VPC are circled.

This slurred upstroke of the R wave has not been seen in standard 6-lead ECG measurement. Since there are no P waves due to AF, a further analysis is needed to determine if this could be a delta wave as in the Wolff-Parkinson-White syndrome, where an accessory path is present between atrium and ventricle. In such a case the refractoriness of the atrioventricular node is decreased and a higher heart rate can thus be explained.

C. Discussion and Further Work

In human medicine the wireless ECG monitoring has proven to be highly reliable. It can be used during various activities and for prolonged cardiac monitoring having therefore a high diagnostic yield and a high patient compliance. It can exceed and could thereby replace conventional Holter monitoring in human patients [5, 7, 12]. A more definite comparison of the new ECG monitor with a reference device in a larger population of animals is pending. However, it is reasonable to assume that wireless ECG monitoring could replace conventional Holter in veterinary medicine, as well.

In animals, hand-held wireless monitors are already used for reliable differentiation of various arrhythmias as a few minute screening of the patients [18]. There is also a report of a non-invasive wearable sensor used to monitor vital signs, including ECG, remotely and continuously, but even in this telemetric method wires were used on the body of the dogs [19].

The prototype of the described wireless ECG sensor has been successfully applied in the synthesis of a 12-lead ECG with an insignificant loss of diagnostic information. Smaller number of measurement sites (electrodes) and fewer wires is a big advantage of these systems that can also synthesize the multi-lead ECG [20, 21, 22, 23]. Further development and validation are expected in this interesting field.

For the time being, in veterinary medicine only human cardiology software is available for Holter monitoring, which is mostly unsuitable. Due to big differences in normal heart rates and prominence of respiratory sinus arrhythmia, overinterpretations or underinterpretations are common. Therefore, it remains veterinarian's obligation to assess the ECG segments [3]. For this reason, the diagnostic method is very time-consuming and expensive. The development of appropriate software, tailored to veterinary medicine, could significantly lower the cost of the long-term cardiac monitoring in veterinary medicine and would make such diagnostic methods much more available and popular.

Besides the differences in heart rate and rhythm discussed above, other significant variations are also present in dogs when compared to humans. The myocardial infarctions are extremely rare. The continuous ECG monitoring is indicated in animals with intermittent weakness, collapse, and syncope. It is used in animals with arrhythmia before and during the treatment with antiarrhythmic drugs. In some breeds (Boxers, Dobermans) long-term ECG monitoring is utilized as screening for arrhythmogenic forms of cardiomyopathies and for detection of arrhythmia that could be the cause of sudden death in dogs with increased risk due to a previously diagnosed or occult disease [2, 3].

III. CONLUSIONS

The improvement in information and communication technology offers immense opportunities for the development of more accurate, cheaper and wearable devices that are urgently needed in veterinary medicine also. The wireless electrocardiographic monitoring presented in this work has a huge potential. It enables a long-term monitoring of cardiac rhythm in real-world environment with appropriate accuracy and reliability. Since there are crucial differences in human and veterinary electrocardiography, the future development of appropriate software for reading and interpretation of the veterinary measurements would even further increase the potential of the presented device.

REFERENCES

[1] S. S. Lobodzinski, "ECG patch monitors for assessment of cardiac rhythm abnormalities," Prog. Cardiovasc. Dis., vol. 56, 2013, pp. 224–229.

[2] G. Wess, A. Schulze, N. Geraghty, K Hartmann, "Ability of a 5-minute electrocardiography (ECG) for predicting arrhythmias in Doberman Pinschers with cardiomyopathy in comparison with a 24-hour ambulatory ECG," J. Vet. Intern. Med., vol. 24, 2010, pp. 367–371.

[3] E. Côté, "Electrocardiography and cardiac arrhythmias," in Textbook of veterinay internal medicine, 7th ed, vol. 3, S. J. Ettinger, E. C. Feldman, eds. St.Louis: Elsevier Saunders, 2010: pp. 1159–1186.

[4] I. Tomašić, R. Trobec, "Electrocardiographic systems with reduced numbers of leads-synthesis of the 12-lead ECG," IEEE Rev. Biomed. Eng., vol. 7, 2014, pp. 126–142.

[5] P. M. Barret, R. Komatireddy, S. Haaser, S. Topol, J. Sheard, J. Encinas, A. J. Fought, E. J. Topol, "Comparison of 24-hour Holter monitoring with 14-day novel adhesive patch electrocardiographic monitoring," Am. J. Med., vol. 127, 2014, pp. 95 e11–7.

[6] M. P. Turakhia, D. D. Hoang, P. Zimetbaum, J. D. Miller, V. F. Froelicher, U. N. Kumar, X. Xu, F. Yang, P. A. Heidenreich. "Diagnostic utility of a novel leadless arrhythmia monitoring device," Am. J. Cardiol, vol. 112, 2013, pp. 520–524.

[7] D. Schreiber, A. Sattar, D. Drigalla, S. Higgins, "Ambulatory cardiac monitoring for discharged emergency department patients with possible cardiac arrhythmias," West. J. Emerg. Med., vol. 15, 2014, pp. 194–198.

[8] S. L. Higgins, “A novel patch for heart rhythm monitoring: is the Holter monitor obsolete?,” Future Cardiol., vol. 9, 2013, pp. 325–333.

[9] M. Rosenberg, M. Samuel, A. Thosani, P. Zimetbaum, “Use of a noninvasive continuous monitoring device in the management of atrial fibrillation: a pilot study,” Pacing. Clin. Electrophysiol., vol. 36, 2013, pp. 328–333.

[10] G. Hindricks, E. Pokushalov, L. Urban, M. Taborsky, K. H. Kuck, D. Lebedev, G. Rieger, H. Pürerfellner, XPECT trial investigators, “Performance of a new leadless implantable cardiac monitor in detecting and quantifying atrial fibrillation: Results of the XPECT trial,” Circ. Arrhythm. Electrophysiol., vol. 3, 2010, pp. 141–147.

[11] M. Schrivastav, S. Padte, V. Arora, M. Biffi, “Pilot evaluation of an integrated monitor-adhesive patch for long-term cardiac arrhythmia detection in India,” Expert. Rev. Cardiovasc. Ther., vol. 12, 2014, pp. 25–35.

[12] R. Trobec, A. Rashkovska, V. Avbelj, “Two proximal skin electrodes – a respiration rate body sensor,” Sensors (Basel), vol. 12, 2012, pp. 13813–13828.

[13] A. Rashkovska, I. Tomašić, R. Trobec, “A telemedicine application: ECG data from wireless body sensors on a smartphone,” Proceedings of MEET & GVS on the 34th International Convention MIPRO 2011, Opatija, Croatia, 2011, pp. 293–296

[14] R. Trobec, M. Depolli, V. Avbelj, “Wireless network of bipolar body electrodes,” Proceedings of the 7th International Conference on Wireless On-demand Network Systems and Services, WONS 2010, pp. 145–149.

[15] R. Trobec, V. Avbelj, A. Rashkovska, "Multi-functionality of wireless body sensors", The IPSI BgD transactions on internet research, 2014;10, pp. 23–27.

[16] R. Trobec, V. Avbelj, B. Meglič, V. Švigelj, "Analysis of baroreflex sensitivity", Proceedings of IASTED Conf. on Biomedical Engineering, Innsbruck, 2006, pp. 276–281.

[17] D. L. Eckberg. Pont:Counterpoint: "Respiratory sinus arrhythmia is due to a central mechanism vs. respiratory sinus arrhythmia is due to the baroreflex mechanism", J Appl Physiol, vol 106, 2009, pp. 1740–1742.

[18] M. S. Kraus, F. C. Brewer, M. Rishniw, A. R. Gelzer, “Comparison of the AliveCor® ECG device for the iPhone with a reference standard electrocardiogram,” ACVIM Seattle 2013: abstract.

[19] R. Brugaloras, J. Dieffenderfer, K. Walker, A. Wagner, B. Sherman, D. Roberts, A. Bozkurt, “Wearable Wireless Biophotonic and Biopotential Sensors for Canine Health Monitoring,” IEEE 2014.

[20] I. Tomašic, R. Trobec, V. Avbelj, “Multivariate linear regression based synthesis of 12-lead ECG from three bipolar leads,” Proceeding of the 3th Int. Conf. Health Informatics, HealthInf 2010, pp. 216–221.

[21] R. Trobec, I. Tomašić, “Synthesis of the 12-lead electrocardiogram from differential leads,” IEEE Trans. Inf. Technol. Biomed. vol. 15, 2011, pp. 615–621.

[22] I. Tomašić, S. Frljak, R. Trobec, “Estimating the Universal Positions of Wireless Body Electrodes for Measuring Cardiac Electrical Activity,” IEEE Trans Biomed Eng 2013, vol X, pp. XX

[23] A. Rashkovska, I. Tomasić, K. Bregar, R. Trobec, “Remote monitoring of vital functions - Proof-of- concept system,” Mipro Proceedings 2012, pp. 463–467.

Software for acid-base status disorders management

Mensur Mandžuka[1], Edin Begić[2], Zijo Begić[3], Dušanka Bošković[1]

[1] Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
[2] Faculty of Medicine, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
[3] Paediatric Clinic, Clinical Center University of Sarajevo, Sarajevo, Bosnia and Herzegovina
mensur.mandzuka@gmail.com

Abstract - Normal acid-base status represents the value of hydrogen ions in extracellular fluid is within pH range of 7.36 (44 nmol/L) and 7.44 (36 nmol/L). Acid-base balance disorders may be a result of enhancement or reduction in amount of hydrogen ions, and include four basic disorders: metabolic and respiratory acidosis, metabolic and respiratory alkalosis. First step implies diagnosing the type of acid-base status disorder. The application input consists of blood analysis data and body weight info. Next step is determining the significance of the disorder and finally recommending future treatment. Software offers the optional functionality of entering mineral level values to improve treatment recommendations. Following the latest trends, along with the aim of increasing availability and mobility of the system, the software is developed for the Android platform using the Java programming language.

Keywords - acid-base balance, disorder, treatment, mobile application

I. Introduction

Normal organism acid-base implies fluctuation in hydrogen ions in extracellular fluid within narrow limits of pH, from 7,36 (44nmol/L) to 7,44 (36nmol/L) [1].

Acid-base status of trans-cellular spaces is regulated utilizing active transfer of acids and bases outside of cellular space. Pathological conditions with an increased concentration of hydrogen ions are called acidosis. Conversely, states with reduced concentrations of hydrogen ions are called alkalosis. Acidosis and alkalosis are terms that are associated with systemic disorders in concentration of hydrogen ions. Disturbances in the concentration of hydrogen ions, occurred in blood, are called acidemia and alkalemia.

In order to keep acid-base status within physiological limits, organism engages powerful regulatory mechanisms: buffer systems, lungs, kidneys and skeletal system. Buffer systems act immediately, but have relatively small capacity. Respiratory system is activated after several hours. Most powerful buffer capacity lies in the kidneys and skeletal system, but they also require most time to activate. Buffer systems act within extracellular and intracellular fluid. Most important extracellular system is the bicarbonate buffer system. On the other hand, the most important intracellular system is the protein and phosphate buffer system [2]. Considering their ability to remove carbon dioxide from the organism, lungs can have a significant impact on acid-base status. If lung ventilation doubles, pH blood value will increase (alkalosis). Conversely, the reduction in lung ventilation will decrease pH value (acidosis). The total buffering power of respiratory system is almost twice the power of all chemical buffer systems combined.

Renal system buffering effect is in conserving filtered bicarbonates, regulating titratable acidity and forming ammoniac ion.

Skeletal system has a role in correcting chronic disorders of acid-base status. In conditions of chronic acidosis, demineralization of bone tissue occurs (release of calcium and bicarbonate), which can compensate the disturbance of acid-base status.

II. Acid-base status disorders

Causes of acid-base disorders may be metabolic disorders or dysfunction of the respiratory system. Thus disturbances of acid-base status include four basic disorders: metabolic and respiratory acidosis, and metabolic and respiratory alkalosis. According to the change of pH levels (if compensatory mechanisms are present), the mere compensation can be: full - full normalization of pH with the preservation of compensatory mechanisms, sub compensation - insufficient correction of pH value with preservation of compensatory mechanisms and decompensation- altered pH and lack of compensatory mechanisms.

Metabolic acidosis (pH <7,35, reduced level of bicarbonate <24 mmol/L, compensatory hyperventilation-reduced pCO_2) occurs as a result of increase in the amount of hydrogen ions and due to loss of bicarbonate ions (increased intake of acid, increased synthesis of volatile acids, the accumulation of metabolic acids, chronic renal failure, distal renal tubular acidosis, diarrhea, malignant processes, proximal renal tubular acidosis, Fanconi syndrome, Wilson's disease).

Base deficit is the basis of metabolic acidosis, and is due to the above mentioned etiological factors. At the expense of reduction of plasma bicarbonate concentration (used to buffer the hydrogen ions), there is an increase in the concentration of other anions (in order to achieve and maintain electro neutrality). The anion may be chloride (Cl^-) or other anion of a metabolic product acid. Based on this, we distinguish: hyperchloremic metabolic acidosis (kidney disarrangements-interstitial nephritis with uremia, hydronephrosis, proximal and distal renal tubular acidosis, gastrointestinal causes-diarrhea, uretero-enterostomy) metabolic acidosis with anion deficit (renal failure, diabetic ketoacidosis, L-lactacidemia, ingestion of salicylates) and their combination (L-lactacidemia) [3]. Anion gap higher than 15 mmol ⁄ L indicates metabolic acidosis, and greater then 25 mmol ⁄ L confirmes metabolic acidosis. Furthermore, based on the value of the anion gap it can be determined whether acidosis is with elevated or normal anion gap.

$$\text{Anion gap} = Na\ (\text{mmol} \diagup L) - Cl^-(\text{mmol} \diagup L) + HCO_3(\text{mmol} \diagup L) \quad (1)$$

The primary cause of respiratory acidosis (pH <7,35, elevated pCO_2, compensatory decrease levels of bicarbonate <mmol/L 24) is an acute or chronic respiratory system insufficiency (acute airways obstruction, acute restriction lung disease, acute circulatory insufficiency, depression of the respiratory center , neurological diseases-poliomyelitis, polyneuropathy, myasthenia gravis, chronic obstructive pulmonary disease, chronic inhibition of respiratory center of the chest wall diseases, diseases of the lung parenchyma). Metabolic alkalosis (pH> 7,45, serum bicarbonate levels> 24 mmol/L, compensatory hypoventilation - increased pCO_2) occurs due to loss of acid outside of cellular liquids (loss of HCl by vomiting, hypokalemia, hypercalcemia, use of diuretics who do not save potassium - furosemide, ethacrynic acid), and because of the influence of factors that increase the concentration of bicarbonate in the extracellular fluid (milky-alkali syndrome, Conn's disease, Liddle's syndrome, the therapeutic use of the mineralocorticoid, the application of organic acids, post hypercapnic condition). Metabolic alkalosis can be divided into those that are salt dependent - Cl responsible (contraction alkalosis, vomiting, use of diuretics, antacids entry - chlorides correct them) and to those not dependant of salt - Cl resistant (sodium chloride is not used in treatment, potassium chloride is used - presence of chloride in the urine). Respiratory alkalosis (pH> 7,45, hyperventilation - decreased pCO2, compensatory reduction in the level of bicarbonate <24 mmol/L) is such a disturbance of acid-base balance in which increased alveolar ventilation causes a decrease in pCO2, and consequently reduce the level of bicarbonate, and increasing pH blood values. It is divided into acute (hyperventilation due to anxiety, high temperatures, salicylate poisoning, sepsis) and chronic (brain tumors, encephalitis, liver cirrhosis, coma, pregnancy).

TABLE I. DISORDERS IN ION CONCENTRACION AND IN PARTIAL PRESSURE OF CO_2 DUE TO ACID-BASE DISORDERS (R - REDUCED, I - INCREASED, * - SIGN FOR PRIMARY DISORDER)

	pH	**H^+**	**pCO_2**	**HCO_3^-**
Norman	7,4	40 nmol⁄L	5,3 kPa	24 nmol⁄L
Respiratory acidosis	R	I	I*	I
Respiratory alkalosis	I	R	R*	R
Metabolic acidosis	R	I	R	R*
Metabolic alkalosis	I	R	I	I*

In everyday practice, complex forms of acid-base status disorders are not rare (a combination of these conditions). Complex forms may be additive, as a combination of respiratory and metabolic acidosis (in diabetes that lasts for a long time - a combination of diabetic ketoacidosis and in uremic acidosis due to diabetic nephropathy - severe academia). Respiratory and metabolic alkalosis are usually seen in patients who are on mechanical ventilation and nasogastric suction - heavy alkalemia. Combinations of respiratory alkalosis and metabolic acidosis occurs in septic shock and is characterized by a reduced concentration of bicarbonate in plasma and reduced pCO_2 . Respiratory acidosis and metabolic alkalosis occurs in chronic obstructive pulmonary disease, and during the use of diuretics - elevated bicarbonate concentration in plasma and increased pCO_2 .

Condition for proper treatment of acid-base disorders is an accurate diagnosis. Based on patient history, disease symptoms and laboratory tests, we can determine if there is a disorder of body fluids and acid-base status. Arterial blood is taken for laboratory tests, and the analysis of pH, bicarbonate concentration and pCO_2 is necessary [4]. Treatment of all disorders is reflected in the etiological treatment as well as in the treatment of associated disorders that can lead to their aggravation, with a special overview of trace elements' concentration in human body. Treatment of metabolic acidosis is based on eliminating the metabolic disorder that led to this condition. Acute metabolic acidosis is frequently encountered in critically ill patients [5]. Therapy with $NaHCO_3$ is indicated when the pH is lower than 7,20. Compensation is determined according to the following formula:

$$NaHCO_3\ (\text{mmol}) = BE\ (\text{mmol}) \times 0.3 \times BW\ (\text{kg}) \quad (2)$$

BE – base excess
BW- body weight

The treatment implies giving 8,4 % $NaHCO_3$, where 1 ml of the solution corresponds to 1 mmol in the formula. Half of the solution is given immediately, and the rest is given according to the disorders' improvement. Hyperkalemia is an accompanying symptom of metabolic acidosis , and for its treatment calcium gluconate in combination with insulin is used. Glucose administration helps the entry of potassium into the cells and reduces the

concentration of potassium in plasma. If the acidosis is of lactate origin and if it is associated with cardiogenic shock , treatment with bicarbonates can only aggravate the existing condition.

Treatment of metabolis alkalosis is suggested depending on its type. Chloride responsible ones are treated with NaCl (serum Cl is reduced), and alkalosis with Cl^- the resistant ones are treated with KCl with obligatory monitoring of electrolytes levels, and continuous ECG monitoring. Amount of Cl ions necessary is calculated using the following formula [6]:

$$\text{Corrected Chloride} = \text{Cl} \times (\text{Na normal} / \text{Na observed}) \quad (3)$$

Severe metabolic alkalosis (pH > 7,7) is treated by the administration of isotonic HCl. (iv . 150mmol - l). A pH greater than 7,65 may carry a high risk of mortality (up to 80%) [6]. Bicarbonate excess is calculated using the following formula:

$$\text{HCO3 (excess)} = 0.5 \times \text{BW (kg)} \times (\text{HCO3 found} - \text{HCO3 desired}) \quad (4)$$

The goal in treatment of respiratory acidosis is primarily removing the cause. By improving the ventilation, the increased CO_2 is removed. In case of a very low pH (< 7,2), the patient must be connected to mechanical ventilation. In case of pH lower than 7.1 and HCO_3 below 15 mmol/ L , treatment with $NaHCO_3$ is indicated. Excessive correction of chronic hypercapnia leads to alkalosis. Although the symptoms, signs, and physiologic consequences of respiratory acidosis are numerous, the principal effects are on the central nervous and cardiovascular systems [7].

Respiratory alkalosis is an extremely common and complicated problem affecting virtually every organ system in the body. This article reviews the various facets of this interesting problem. Respiratory alkalosis produces multiple metabolic abnormalities, from changes in potassium, phosphate, and calcium, to the development of a mild lactic acidosis [8]. Respiratory alkalosis is treated by removing the cause. Heavy alkalemia pH > 7,6 requires therapy with arginine or ammonium chloride. In psychogenic disorders, use of sedatives is giving good results.

III. Software Diagnostics of Acid-Base Status Disorders

With the aim of maximizing the availability, the application was developed for the Android platform using the Java programming language. In 2012, the market share of Android devices accounted 75% of the world market, with 136 million devices shipped worldwide [9]. The application requires entry of key parameters necessary for diagnosis of acid-base disorders:

- blood pH (normal value 7,35 to 7,45),
- partial pressure of carbon dioxide in the blood (expressed in kilopascals) (normal 4,3-6,4 kPa),
- partial pressure of oxygen in the blood (expressed in kilopascals) (normal 7,3-10,6 kPa),
- bicarbonate value (expressed in mmol/L) (normal 24 mmol/L),
- base excess and
- body weight of the patient.

Diagnosis of acid-base disorders is made based on parameters obtained from the arterial blood sample (Figure 1.).

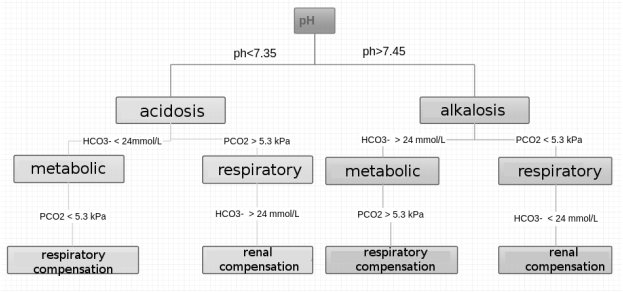


Figure 1. Diagnosis of acid-base disorders

Home screen of application (Figure 2. and Figure 3.) contains fields for entering the above mentioned data.

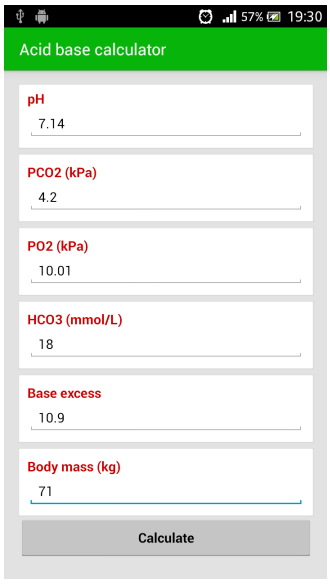


Figure 2. Application home screen, metabolic acidosis input

The data are typical for particular disorder or combination of disorders, and based on them the type of disorder can be determined, and also compensatory responses, with defense mechanisms of the organism. (Figures 2, 3, 4, 5)

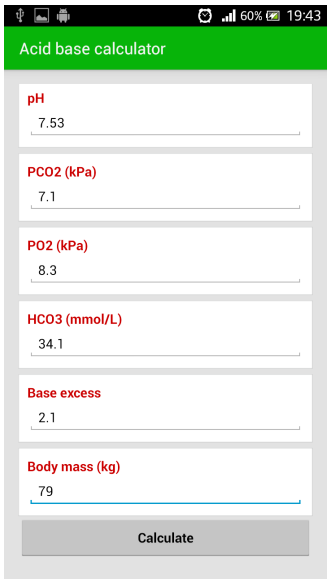


Figure 3. Application home screen, metabolic alkalosis input

In the next step application determines the type, significance and compensation of disorders. Figure 2. and Figure 4. represent one patient, and Figure 3. and 5. second patient. The next step is to enter pathological values of certain elements in the body (Figure 4. elevated level of potassium - reference value is from 3,5 to 5,2 mmol/L).

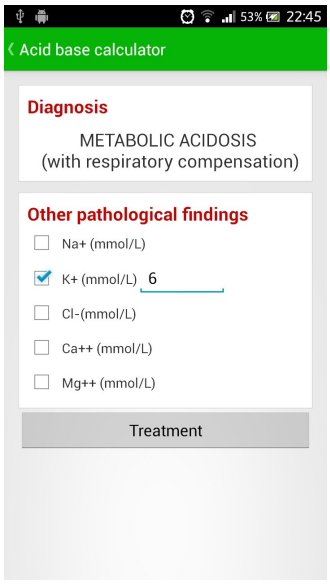


Figure 4. Metabolic acidosis diagnosis

The user can select and enter disturbed values of one element or combination of disturbed values of more elements. Reference values for sodium are 137-146 mmol/L, for chloride 101-111 mmol/L, for calcium 2,14-2,53 mmol / L, for magnesium 0,65-1,05 mmol/L.

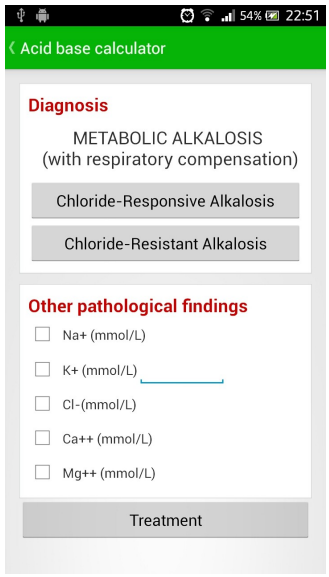


Figure 5. Metabolic alkalosis diagnosis

The user can optionally enter diagnosis of the patient, if it is already known, and the application will take that into account when determining the disorder and displaying the appropriate therapy (Figure 6.) .

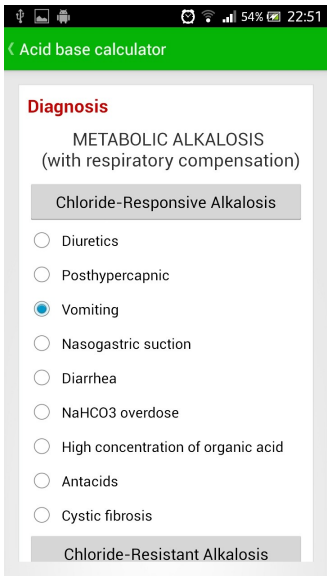


Figure 6. Additional input

Based on the data entered in Figure 2. and Figure 3. application will recommend the treatment for the patient, with the exact values, based on the set of mathematical formulas and recent guidelines for the treatment of acide-base disorders (Figure 7.).

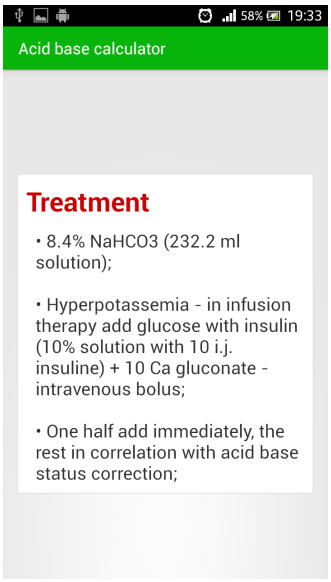


Figure 7. Treatment recommendations

Based on the characteristic values, the application recognizes and takes into account combination of acid-base disorders.

By browsing online solutions and literature in the field of telemedicine, it was realized that there are many Android and web applications which in the form of a calculator give the option of differentiation of disorders, and some of them, determining its compensates. This application should be a pioneer in advising therapy, taking into account the current state of the patient, their medical history and any pre-existing co-morbidities. With the aim of automating the process, future versions of the application will have the functionality of photographing findings of arterial blood. Within the application optical character recognition would be performed, with the aim of recognizing the necessary parameters. Library for optical character recognition would enable autofill fields, necessary to diagnose a disorder. The functionality would be provided through an open-source Tessaract library, originally developed by Hewlett-Packard, and since 2006. maintained by Google.

IV. Conclusion

The application would provide assistance to doctors in clinical practice and is focused on treatment in the intensive care unit. It is easy accessible, practical, and is extremely useful in the treatment of metabolic acidosis (reducing the time required to calculate the necessary therapy-when calculating the required values of sodium bicarbonate) and in the treatment of metabolic alkalosis (calculating the excess bicarbonate). The solution contains guidelines for the therapeutic effect of each disorders, with reference to fluctuations in trace elements, and therapeutic effect in the event of their disorders, which in modern medicine is a challenge in therapy. Combinations of acid-base disorders can greatly complicate the work of the doctor, and such combinations represent a challenge when making software solutions of this type.

References

[1] S. Gamulin, M. Marušić, Z. Kovač and associates, Pathophysiology, 7th ed., vol. 1, Medicinska naklada, Zagreb, 2011.
[2] S. Simonović-Živančević, General pathological physiology, University of Kragujevac, Kragujevac, 2006.
[3] M. Jukić, V. Gašparović, I. Husedžinović and associates, Intensive Care Medicine, Medicinska naklada, Zagreb, 2008.
[4] C.A. Guyton C, E.J. Hall, Medical Physiology, Medicinska naklada, Zagreb, 2012.
[5] S. Lim, "Metabolic acidosis", Acta Med Indones. 2007 Jul-Sep;39(3):145-50.
[6] S.K. Epstein, N. Singh, "Respiratory acidosis" Respir Care. 2001 Apr;46(4):366-83..
[7] S. Tripathy, "Extreme metabolic alkalosis in intensive care", Indian J Crit Care Med 2009;13:217-20
[8] G.T. Foster, N.D. Vaziri, C.S. Sassoon, "Respiratory alkalosis" Respir Care. 2001 Apr;46(4):384-91.
[9] J.J. Drake, Z. Lanier, C. Mulliner, P.O. Fora, S.A. Ridley, G. Wicherski, Android Hacker's Handbook, Wiley, USA, 2014.

Impact of electrode misplacement on the accuracy of 12-lead ECG synthesized from differential leads.

I. Tomasic*, R. Trobec **, A. Rashkovska **and M. Lindén*
* Mälardalen University/Biomedical Engineering Research Group, Västerås, Sweden
** Jožef Stefan Institute/Department of Communication Systems, Ljubljana, Slovenia

Abstract – The sensitivity to electrode misplacement of the 12-lead ECG, synthesized from differential leads, was analyzed by simulating the electrode misplacement, with a help from interpolated body surface potential maps. The electrodes forming optimal differential leads were independently moved only on the lines defined by gradient vectors, calculated on the integral body surface potential map, obtained from juxtaposed average beats for each of the 47 multichannel ECGs used in the study. The 12-lead ECGs were synthesizes from the simulated moved differential leads and compared to the ECGs synthesized from the nominal electrodes' positions, by using correlation coefficients and RMSD. The results show that synthesized 12-lead ECGs are less accurate if electrodes with bigger associated gradients are misplaced, but also suggest that small displacements of up to 1 cm do not have a big influence on the synthesis quality. From the perspective of ECG synthesis, the presented methodology can be used to define an additional criterion for optimal triplets of differential leads selection, which will ensure that the differential lead triplet are optimal not only in terms of the quality of the synthesized leads, but also in terms of being minimally sensitive to electrode misplacement.

I. INTRODUCTION

A number of previous studies have shown that electrode misplacement of standard electrode positions, for the acquisition of the 12-lead ECG, can cause differences in ECG signals and their interpretation [1]. It has also been shown that misplacement of electrodes employed by a reduced-lead set {I, II, V2, V5}, has effects on synthesized 12-lead ECGs, in terms of increased root-mean-square deviation (RMSD), as the electrode distances from their nominal positions increase [2].

Body surface potential maps (BSPM) [3] have been used before for the purpose of simulating the electrode misplacement in standard 12-lead electrode set [4] or in a reduced-lead set [2]. They enable the measurements obtained from simulated misplaced electrodes to be synchronous in time with signals from other electrodes, but also enable position changes over distances smaller than the once possible to realize physically.

In this study, BSPM are used to investigate how the electrode misplacement effects 12-lead ECGs synthesized from differential leads (DLs). Differential leads [5] are bipolar leads that measure the potential between two closely placed body-surface electrodes. The DL measurements can be obtained by so called wireless body electrodes (WBEs) – novel devices that enable minimal obtrusion and wireless transmission of recorded signals [6]. Since there are an infinite number of possible directions of electrode misplacement, normally authors choose to restrict the simulation of electrode misplacement to only vertical and/or horizontal directions. In this study however, we consider the most significant direction for the electrode misplacement to be in the direction of the BSPM gradients on each electrode.

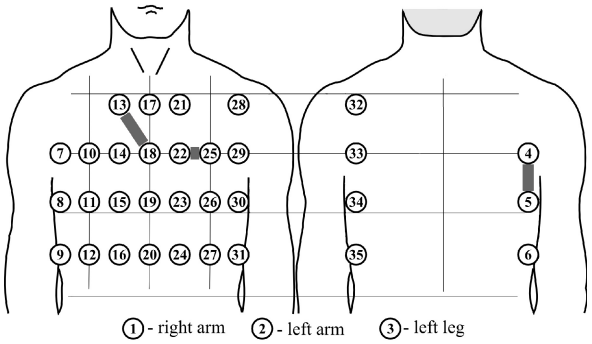


Figure 1. Schematic locations of multichannel ECG (MECG) electrodes on the chest (left) and the back (right). The gray lines represent the calculated optimal universal leads for the 12-lead ECG synthesis: {(13, 18), (22, 25), (4, 5)}.

In our previous studies, we have shown that it is possible to synthesize quality 12-lead ECG from three DLs [5]. By using our algorithm for selection of optimal DLs, we have also identified optimal universal positions of three DLs (see Fig. 1) from which 12-lead ECG can be synthesized by employing universal or personalized transformations [7]. In this study, we investigate the impact DLs' electrodes misplacement has on the 12-lead universal and personalized transformation. There is also the third approach to the synthesis in which both DLs positions and transformations are personalized. We have not investigated this fully-personalized approach for the electrode misplacement effect since for that approach a multichannel ECG (MECG) needs to be obtained from each person, which makes it possible to mark on the skin the positions of electrodes of the optimal personalized

DLs and thereby minimize the possibility of misplacement.

II. METHODS

A. Studied data

Twenty healthy volunteers (13 male, 7 female, mean age ± SD = 50.6±9) with no previous medical record related to heart disease and with a normal 12-lead ECG, and twenty-seven patients (22 male, 5 female, mean age ± SD = 58±10), scheduled for a coronary artery bypass surgery, were included in our study. Informed consent was obtained from all the subjects before the study. A single MECG measurement was obtained from each person. For the data recording device and procedure please refer to [8]. The measurements were obtained during our previous studies [9]. The positions of the MECG electrodes are specified in Fig. 1.

The length of each measurement was 360 seconds. They were processed using MatLab (MathWorks, Inc.) where the average beats were obtained for each lead of the baseline corrected MECGs, by the procedure described in [7]. Note that the high-frequency noise is satisfactorily damped by the averaging.

B. Generic differential leads

A method of juxtaposing measurements was used: the average beats from each MECG were concatenated for each lead thus forming generic MECG leads. The generic DLs are formed by taking differences of juxtaposed MECG leads.

C. Body surface potential maps

For the interpolation of potentials between MECG electrodes, the cubic spline interpolation has been used to provide smooth transition of approximated data between measurement positions. The interpolated potential maps [10] were used for the simulation of electrode misplacements in each time instant. Additionally, an integral BSPM [11] was calculated for the generic differential leads. We will refer to this map as generic integral BSPM (GIBSPM). Since the GIBSPM is calculated from juxtaposed average beats from each person, it can be thought of as estimation of average integral map (for the whole beat) across all the persons covered by the study. The estimation is of course more precise as the variance of average beats lengths across different persons is smaller.

D. Experimental procedure

Gradient vectors were calculated from GIBSPM for each MECG electrode position. We have considered the worst-case scenario of electrode misplacement to be if the electrodes are moved on the line defined by associated gradients, since the gradients define the directions of fastest changes in potential. We have defined the movement of each electrode {13, 18, 4, 5, 22, 25} that makes the three optimal differential leads (see Fig. 1) in both directions of the gradient defined lines, up to 3 cm, in 0.5 cm steps. The electrodes were moved one at the time, i.e. while one electrode was moved the positions of all other electrodes were kept nominal.

For each person, five second intervals were identified in the MECGs and associated 12-lead ECG, that were not used for the calculation of universal positions and transformations used for the synthesis. For each person and for each independent movement of the six electrodes, the 12-lead ECG was calculated using the universal transformation, and by using personalized transformation. Here the simulated moved electrode measurement is taken from integral map for each time instant. The correlation coefficients (CCs) and RMSDs were calculated between the obtained synthesized 12-lead ECGs and 12-lead ECGs synthesized from the nominal electrode positions.

III. RESULTS

Fig. 2 shows the calculated GIBSPM. The gradient vectors are oriented in the direction of increasing potential

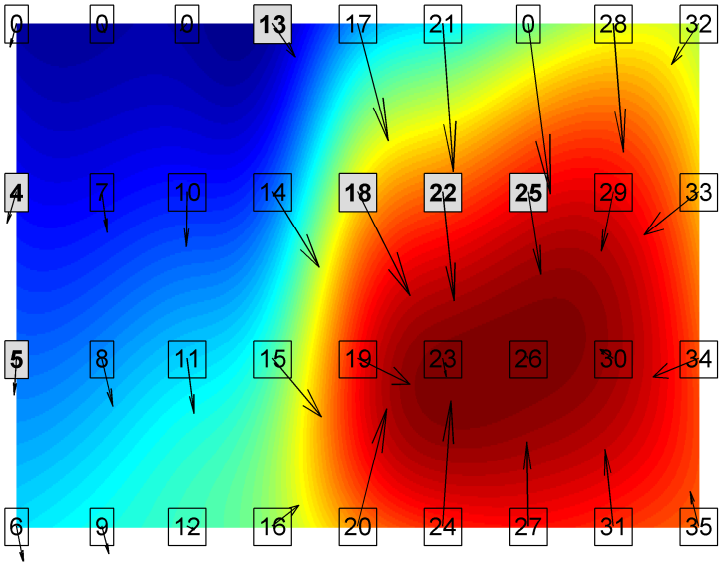


Figure 2. Generic integral BSPM (GIBSPM) for the juxtaposed whole average beats. Electrodes are organized in a matrix as defined by Fig.1. Positions without electrode are denoted with 0. Electrodes forming the optimal lead tripelt, the moved electrodes, are denoted with gray boxes.

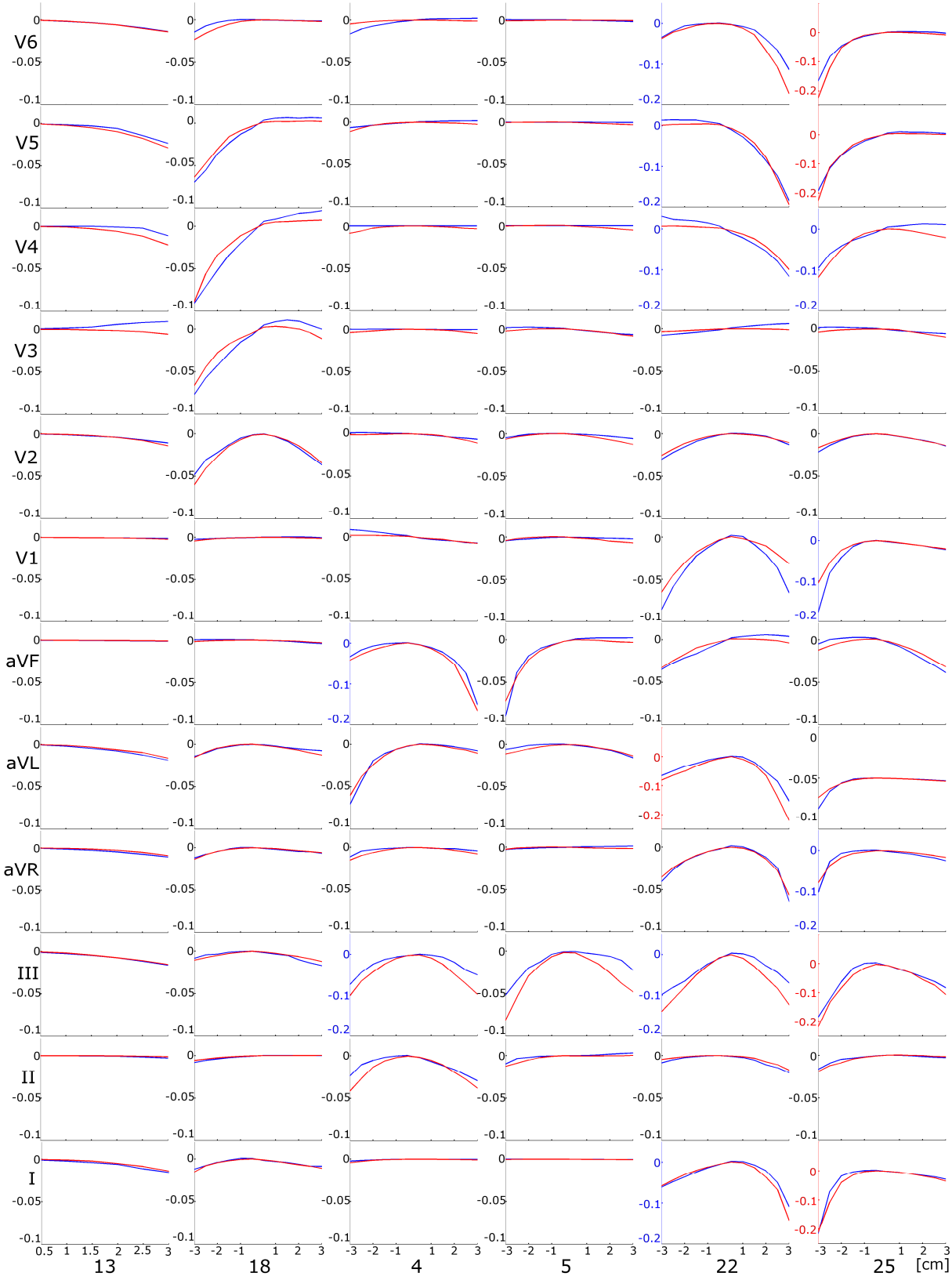


Figure 3. Median differences (dCCs) between CCs for misplaced position and the CCs obtained for nominal electrode positions (lower values imply lover synthesis quality). The CCs are calculated between target and synthesized leads. The columns represent the six electrodes. Each electrode was moved from -3 to 3 cm in steps of 0.5 cm (x axes), except for the electrode 13 which was moved from 0.5 to3 with the same step (0.5). Three scales used in the graphs: black for the dCCs above -0.1, blue for dCCs above -0.2, and red for dCCs below -0.2. Blue curves correspond to dCCs for universal transformation whereas the red line is for the personalized transformation.

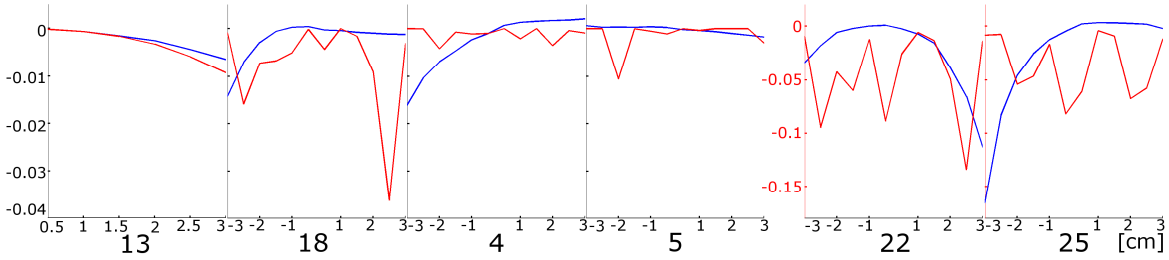


Figure 4. Median differences (dCCs) between CCs for misplaced position and the CCs obtained for nominal electrode positions across all leads of the 12-lead ECG (lower values imply lower synthesis quality). The CCs are calculated between target and synthesized leads. The columns represent the six electrodes. Each electrode was moved from -3 to 3 cm in steps of 0.5 cm (x axes), except the electrode 13 which was moved from 0.5 to 3 with the same step (0.5). Three scales used in the graphs: black for the dCCs above -0.04, red for dCCs below -0.15. Blue line corresponds to universal transformation whereas the red line is for the personalized transformation.

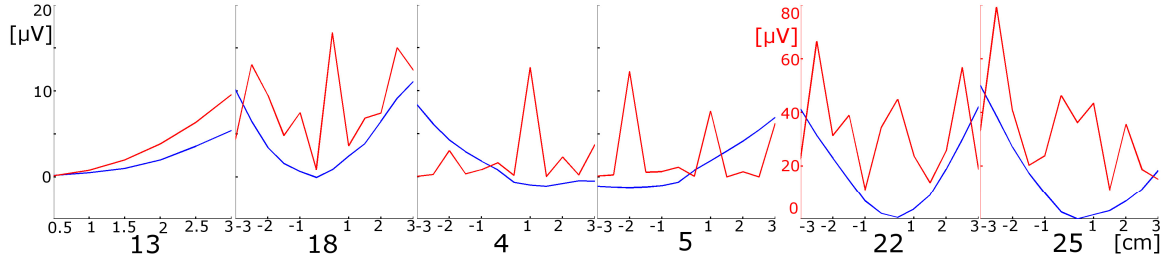


Figure 5. Median differences between RMSDs for misplaced position and the RMSDs obtained for nominal electrode positions across all leads of the 12-lead ECG (higher values imply lower synthesis quality). The RMSDs are calculated between target and synthesized leads. The columns represent the six electrodes. Each electrode was moved from -3 to 3 cm in steps of 0.5 cm (x axes) , except the electrode 13 which was moved from 0.5 to 3 with the same step (0.5). Two scales used in the graphs: black for the RMSDs below 20μV, red for RMSDs abow 60μV. Blue line corresponds to universal transformation whereas the red line is for the personalized transformation.

(positive direction). Gradient vectors for electrodes 4 and 5 point only slightly outside of the interpolated map. Therefore for these two electrodes, instead of using extrapolation, the gradients were just postulated to be pointing straight downward to keep them inside the map. For electrode 13 (also a boundary electrode) the misplacements were calculated only in the direction of its gradient (positive direction), and not in the negative direction since it falls out of the map.

Medians of calculated CCs between synthesized and target 12-lead ECGs are presented on Fig. 3 for each of the six electrodes and for each misplacement distance. The CCs were subtracted by the CCs obtained for the corresponding nominal electrodes positions and therefore negative values represent decrease in synthesis quality as each electrode is moved.

We have also calculated the median CC differences across all synthesized leads (Fig. 4). The analogous RMSD differences are presented on Fig.5.

IV. Discussion

It can be suspected by analyzing Fig. 3 that in most cases the synthesis using personalized transformation loses more in quality when electrodes are misplaced. This becomes obvious form Fig. 4 and Fig 5. Fig 4. shows that CCs distances from the CCs obtained for nominal positions are in general larger when personalized syntheses parameters are used (red lines). Similarly, Fig. 5 shows that RMSDs deviations from the RMSDs obtained for nominal electrode positions, are in general larger for the personalized transformation parameters (red lines).

These results are a bit unexpected since one would expect a more personalized synthesis algorithm to perform better in any aspect, compared to the universal one. Even though this shows that the approach with personalized transformation is more susceptible to electrode misplacement, it does not say that the approach with personalized transformation is inferior by any means. The results certainly show that the personalized approach loses more on the synthesized leads quality, but then again it has more to lose since it provides better synthesized 12-lead ECGs, from the nominal electrode positions [7].

All three figures show that as the misplacement increases, the synthesis quality decreases. Still, Fig. 3 points out that in general small misplacements in range [-1, 1] cm do not influence the synthesis significantly. It is also possible that a small misplacement even increases the synthesis quality (for example, lead V3 for electrode 18). There is however quite a big variation between the electrodes and between the synthesized leads. For example, lead V2 is much less vulnerable for any of the six electrode misplacements compared to lead III (Fig. 3).

Even more obvious is that electrodes 22 and 25 have the biggest influence on the synthesis quality (Fig, 3, 4, and 5). This can be explained by the fact that they together form one DL and are located in the GIBSPM area with fast potential changes – they both have big gradients (Fig. 2). Electrode 18 also has a much bigger gradient than electrodes 4, 5, and 13, and hence a higher impact on the synthesis quality (Fig. 4 and Fig. 5).

We have considered only displacements in the direction of gradient vectors obtained from GIBSPM. This

makes a good sense since the gradient is the direction of fastest potential change. Still the gradients were obtained for the whole beats but are in general different for different ECG segments or intervals. To apply the same approach, but concentrating only on a part of the beat, one would need to construct integral map only for the segment/interval of interest and calculate their gradients.

V. CONLUSION

The results suggest that synthesized 12-lead ECGs are more distorted if electrodes with bigger associated gradients are misplaced. Additionally the results confirm what could be expected: as the electrodes are moved more from the nominal positions, the quality of the synthesized leads decreases in terms of CC and RMSD. Small displacements however, show only small decrease in CC and RMSD.

In this study, only one electrode is moved at a time. A possible extension of this work is to simulate simultaneous movements. Furthermore, this study employs only the CC and RMSD for comparing waveforms, but they can be rather crude in expressing similarities or finding differences between detailed waveform patterns. A small RMSD or a large CC can lead to differences in the diagnostic content of the synthesized and the target 12-lead ECG. Further work is therefore needed to assess the diagnostic accuracy of the ECGs synthesized from displaced electrodes.

Finally, the approach presented here can be employed as an additional criteria for selection of optimal DLs, for 12-lead ECG synthesis. As there are a number of combinations of three DLs that can be used for the synthesis of 12-lead ECG [7] (in this work only one combination was used), by applying sensitivity to misplacement as the additional criteria, another triplet can be obtained that is not only suitable in terms of the quality of the synthesized leads, but also in terms of being minimally sensitive to electrode misplacement.

ACKNOWLEDGMENT

This work was partly supported by the KKS financed research profile Embedded sensor systems for health, and partly by the Slovenian Research Agency under Grant P2-0095.

REFERENCES

[1] B. J. Schijvenaars, G. van Herpen, and J. A. Kors, “Intraindividual variability in electrocardiograms,” *J Electrocardiol,* vol. 41, no. 3, pp. 190-6, May-Jun, 2008.

[2] D. D. Finlay, S. P. Nelwan, C. Nugent, and S. H. Meij, “Effects of electrode misplacement on the reconstruction of the 12-lead ECG,” in Computers in Cardiology, 2009, pp. 281-284.

[3] R. L. Lux, "Body Surface Potential Mapping Techniques," *Comprehensive Electrocardiology*, P. W. Macfarlane, A. v. Oosterom, O. Pahlm, P. Kligfield, M. Janse and J. Camm, eds., pp. 1362-1374: Springer-Verlag London, 2011.

[4] B. J. Schijvenaars, J. A. Kors, G. van Herpen, F. Kornreich, and J. H. van Bemmel, “Effect of electrode positioning on ECG interpretation by computer,” *J Electrocardiol,* vol. 30, no. 3, pp. 247-56, Jul, 1997.

[5] R. Trobec, and I. Tomasic, “Synthesis of the 12-Lead Electrocardiogram From Differential Leads,” *IEEE Trans Inform Technol Biomed,* vol. 15, no. 4, pp. 615-621, 2011.

[6] K. Bregar, and V. Avbelj, “Multi-functional wireless body sensor - Analysis of autonomy,” in 36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013, pp. 322-325.

[7] I. Tomasic, S. Frljak, and R. Trobec, “Estimating the Universal Positions of Wireless Body Electrodes for Measuring Cardiac Electrical Activity,” *IEEE Trans Biomed Eng,* vol. 60, no. 12, pp. 3368-3374, 2013.

[8] V. Avbelj, R. Trobec, B. Gersak, and D. Vokac, “Multichannel ECG measurement system,” in Tenth IEEE Symposium on Computer-Based Medical Systems, 1997, pp. 81-84.

[9] S. Frljak, V. Avbelj, R. Trobec, B. Meglic, T. Ujiie, and B. Gersak, “Beat-to-beat QT interval variability before and after cardiac surgery,” *Comput Biol Med,* vol. 33, no. 3, pp. 267-76, May, 2003.

[10] R. Trobec, “Computer analysis of multichannel ECG,” *Comput Biol Med,* vol. 33, no. 3, pp. 215-26, May, 2003.

[11] L. d. Ambroggi, and A. D. Corlan, "Integral maps," *Comprehensive Electrocardiology*, P. W. Macfarlane, A. v. Oosterom, O. Pahlm, P. Kligfield, M. Janse and J. Camm, eds., pp. 1377-1378: Springer-Verlag London, 2011.

Extranodal Pacemaker Activity During Sleep - a Case Study of Wireless ECG Sensor Data

V. Avbelj
Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
viktor.avbelj@ijs.si

Abstract - **Sleep apnea is a breathing disorder during sleep which is characterized by breathing cessation lasting at least 10 seconds. Sleep apnea is associated with sudden cardiac death. Every apnea event is accompanied by cardiovascular changes. In polysomnography, to perform cardiorespiratory sleep study in today's sleep laboratories, many parameters such as ECG (electrocardiogram), EEG (electroencephalogram), respiration effort, oronasal airflow, snoring (by microphone) and oxygen saturation in blood are recorded. We present a case where only single channel ECG obtained from wireless electrodes was used to study heart-rate dynamics during sleep apnea events. A sudden change of morphology of atrial P wave in ECG recording was observed during sleep. This observation supports the hypothesis that sinus node is not the only physiological pacemaking structure in the atrium.**

I. Introduction

Sleep apnea is characterized by breathing cessation during sleep, lasting at least 10 seconds. These events are accompanied by cardiovascular changes and are associated with sudden cardiac death. Many parameters are measured in polysomnography in modern sleep laboratories (ECG, EEG, respiration, airflow, snoring, oxygen saturation in blood) to identify the possible sleep apnea [1]. Whether these complex measurements could, at least partially be replaced by simpler methods, is the question of this text.

We present a case study where only one ECG channel is used during sleep. A detailed analysis of the recording shows that a variety of information can be obtained from the bare ECG signal. Low sampling rate prevented direct application of the ECG signal for obtaining surrogate respiration signals. Appropriate interpolation solved the problem partially. Amplitude and morphology of the ECG waves probably changed according to changes of the positioning of the body. If another sensor for 3D acceleration (gravity) was available, the movements and position changes could reliably be measured also. In the ECG signal, the skeletal muscle signals are also visible (normally as unwanted noise), but in this case study we use them as signs of the effort to change the position of the body during sleep. Detailed examination of the recording revealed unexpected extranodal pacemaker activity. This extranodal pacemaking is not recognized as a normal physiological activity of human atrium, but here we present an example that extranodal pacemaking could be present widely, although broadly unrecognized [2].

II. Presentation

A. Materials and Methods

In this case study an ECG measurement of a woman (58 years) is presented. The measurement was made during night by small body ECG sensor (mass 15 g) with a wireless connection to a smart device (tablet, Nexus 7) via low power Bluetooth technology [3]. The two self-adhesive Ag/AgCl electrodes (Skintact T-401, distance 9 cm) were placed horizontally approximately 10 cm below the left clavicle. The ECG data were collected overnight in the tablet for off line analysis. The ECG signal was sampled at 125 Hz with a 10 bit A/D converter on the ECG sensor. For derivation of the respiration from the ECG signal, a few preprocessing steps have been done.

The amplitude of the ECG waves is modulated by breathing due to the changes in the lung and thorax volume. The ionic current sources in the heart produce electric potential difference between the locations of the two electrodes. This potential difference is modulated due to the varying volume impedance. At 125 Hz sampling rate, the R wave is represented by only a few values (Fig. 1). The amplitudes of the R waves, taken from measured values are coarse.

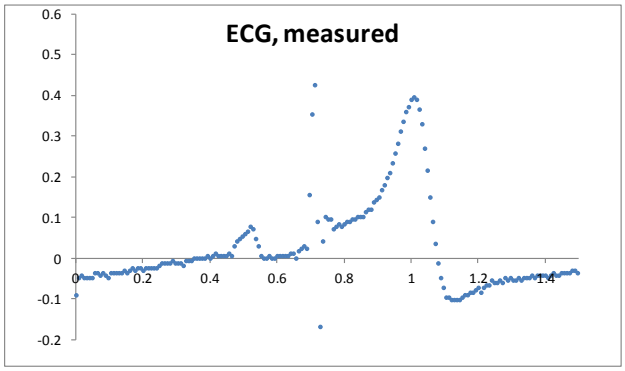


Fig. 1: The ECG data points at sampling rate of 125 Hz. The R wave is represented by only 4 data points. The peak of the R wave is at time 0.712 s where the value is 0.426 mV. The minimum value (peak of the S wave) is at time 0.728 s, where the amplitude is -0.168 mV.

By interpolation techniques better amplitude fidelity could be obtained from the sampled data. We used cubic spline interpolation with 8-fold increase in sampling frequency. The result is present in Fig. 2.

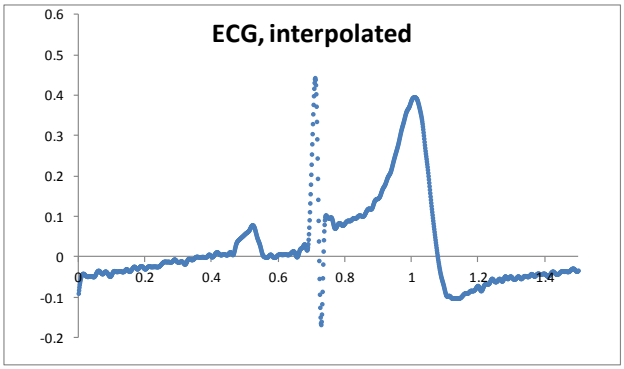


Fig. 2: The same signal as in Fig. 1 where interpolated samples are inserted between the measured ones. Synthetic sampling frequency is 1000 Hz and the peak of the R wave is at time 0.710 s where the amplitude is 0.4437 mV (4% more than in Fig. 1). The minimum value (the peak of the S wave) is the same as in Fig. 1.

If the PQ interval is clamped to zero potential (correction of baseline wandering) the peak value of the R wave could be determined precisely. The successive peaks of the R waves represent an ECG derived respiration signal. The process of interpolation is shown in Fig. 3, where original and interpolated samples are shown separately.

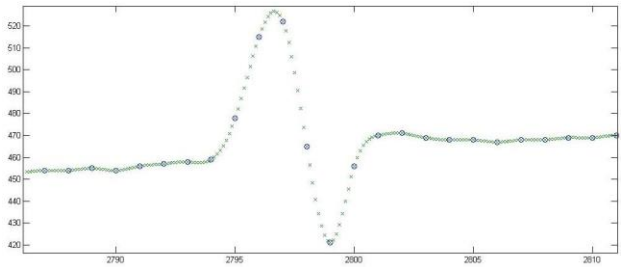


Fig. 3: The R and the S waves of the ECG where the measured values are marked with circles and interpolated with crosses. The time and amplitude resolution is enhanced. Horizontal axis counts the measured samples of the recording while the vertical axis represents direct A/D converter values.

Using NevroEKG software, we performed the following steps in the processing of the ECG recording: determination of successive RR intervals, baseline correction adjustment and the determination of the peak of the R waves that represented an ECG derived respiration signal [3]. The result is a graph with three channels: ECG, RR intervals and R peaks (respiration surrogate). The graph of this ECG recording was examined in detail regarding long-term ECG amplitude variations, respiration, and details of ECG waves around abrupt changes in the signals.

B. Results

In Fig. 4 overnight ECG is presented beginning at 1:57 AM. Initially (first 1300 s) the ECG had fast varying amplitude that was probably associated with frequent changes of the body position before falling asleep. Subsequently we observed blocks of fairly constant amplitude of the ECG signal that was changing abruptly. Large spikes were present where the ECG signal was lost due to the unfavorable position of the transmitting sensor relative to the receiving tablet. There was also a large interval with no signal in the 7/8 of the recording.

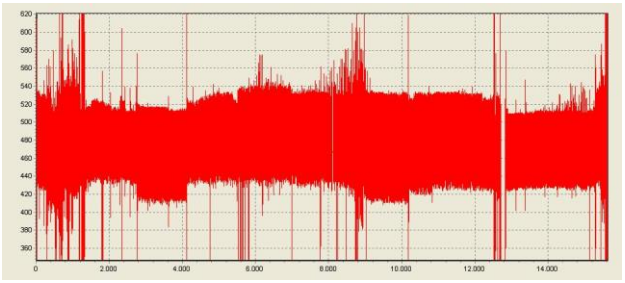


Fig. 4: Amplitude of the ECG signal of the whole overnight recording. There is an interval of around 2 minutes, where the signal from the electrode was completely lost (see the interval in 7/8 of the recording).

The next step was to find out what is the reason of abrupt changes in the amplitude of the ECG recording. An expressive event of this type is noticed some time after 4000 s of the recording and is shown in Fig. 5. Before the event the R wave is smaller than the S wave and the T wave is much higher than the R wave. After the event the opposite is true. The R wave is larger than the S wave while the T wave stays approximately the same. During the change the large noise signals (not shown) are present, connected probably with moving of the electrodes during the change of the body position in the bed.

Fig. 5: Abrupt changes of the ECG waves, especially the R and the S wave before and after the change of the position of the body.

There were similarities in other discontinuities of the ECG amplitudes in the cases shown in Fig. 4.

There were many fast changes of the heart rate during the night. The most striking event could be observed around 10800 s (at 5 AM). Before the change, the heart rate was 48 beats/min and increased to 86 beats/min in less than 15 s. Then the heart rate quickly dropped to 55-58 beats/min, and finally dropped abruptly to a value from before the episode which lasted 76 s. The relative amplitudes of the QRST waves before and after the episode remained unchanged (Fig. 6). It is not known what the reason of the heart rate speed-up was. When the heart rate started to rise, the muscle noise was visible for 5 s, the QRS waves became smaller and then returned to previous values while the heart rate still continued to rise (not shown in Fig. 6). An abrupt change of the morphology of the atrial P wave occurred in the speed-up of the heart rate.

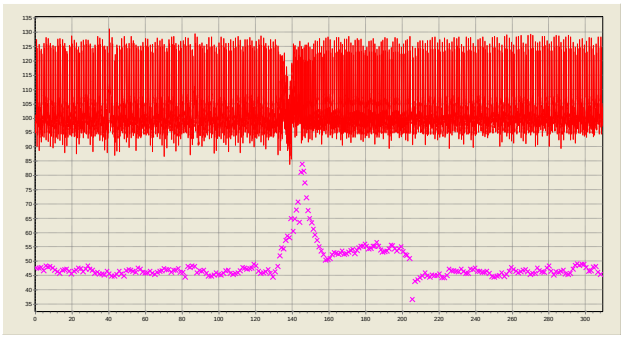


Fig. 6: The ECG (upper trace) and the heart rate (lower trace) during a fast change in the heart rate around 5:00 AM. In the beginning of the episode the heart rate changed from 48 to 86 beats/min.

The changed P wave's morphology remained the same until the end of the episode when the heart rate returned to the previous value. The change in the morphology of the P wave means that the leading pacemaker region changed the position. It is believed that the sinus node is the origin of each heart beat in normal cases, although there is growing evidence for the physiological operation of extranodal pacemaker sites of a widely distributed atrial pacemaker complex in the human heart [2]. In a previous study, we documented such a pattern in the process of deceleration of the heart rate during spontaneous cardioinhibitory syncope [4].

In establishing the reason for the episode shown in Fig. 6, we examined the surrogate respiratory signal (amplitude of the R waves). Although the signal amplitude diminished in the seconds before the acceleration of the heart rate, it cannot be reliably detected as a sleep apnea episode. Again, this signal diminished immediately after the end of the episode as it can be seen in Fig. 7.

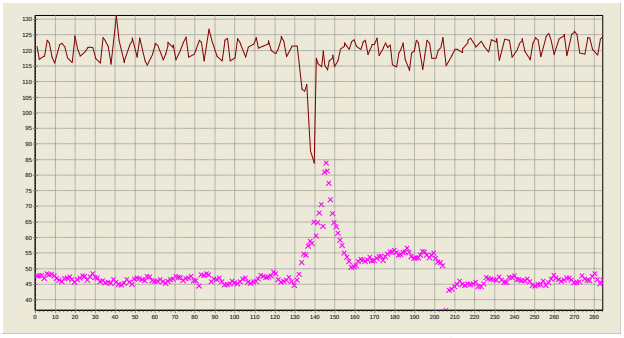


Fig. 7: The surrogate breathing signal (amplitude of the R waves in the upper channel) and the heart rate (lower channel) around 5:00 AM.

III. CONLUSIONS

The case of long-term one channel ECG recording was presented, where only the slow sampling rate (125 Hz) ECG signal was available. With interpolation the signal was enhanced to serve also for the extraction of breathing information - ECG derived breathing. Even more precise data could be obtained should the amplitude of the R waves be determined in the ECG sensor at a higher sampling frequency. It has been shown that with thorough examination of the ECG signals, the morphology of the P waves could be studied in detail. In our case, an unexpected pattern of pacemaking structures in the atrium was observed during sleep. Many ECG recordings should be thoroughly examined for minute changes in the morphology of the P waves to find out whether this was a common phenomenon during sleep. It would be a substantial improvement if the ECG sensor had an accelerometer to find out the position of the body and the movements during sleep.

REFERENCES

[1] E. Sforza, C. Jouny, V. Ibanez, "Cardiac activation during arousal in humans: further evidence for hierarchy in the arousal response," Clin Neurophysiol., vol. 111, 2000, pp. 1611-1619.

[2] J. P. Boineau, T. E. Canavan, R. B. Schuessler, M. E. Cain, P. B. Corr, J. L. Cox, "Demonstration of a widely distributed atrial pacemaker complex in the human heart," Circulation, vol. 77, 1988, pp. 1221-1237.

[3] R. Trobec, A. Rashkovska, V. Avbelj, "Two proximal skin electrodes – a respiration rate body sensor," Sensors (Basel), vol. 12, 2012, pp. 13813–13828.

[4] V. Avbelj, R. Trobec, "A closer look at electrocardiographic P waves before and during spontaneous cardioinhibitory syncope," Int J Cardiol. vol. 166, 2013, pp. e59-e61.

A New Color Assessment Methodology using Cluster-based Features for Skin Lesion Analysis

M.J.M. Vasconcelos*, L. Rosado * and M. Ferreira**
* Fraunhofer Portugal AICOS, Porto, Portugal
** Portuguese Institute of Oncology, Porto, Portugal
{maria.vasconcelos, luis.rosado}@fraunhofer.pt

Abstract - Melanoma is considered the most dangerous form of skin cancer, however if detected in early stages there are high success rates of recovery, making prevention essential. The risk assessment of skin lesions usually follows the ABCD rule (asymmetry, border, color and dermoscopic structures). This paper presents a methodology to assess the number of ABCD rule colors of skin lesion images. It starts by extracting 660 color features and several feature selection and machine learning classification methods are tested in order to minimize classification errors. The developed methodology has the advantage of being adaptable to the dataset used. Two publicly available dermoscopic image datasets and one mobile acquired dataset were used to test the methodology, achieving accuracy rates of 77.75%, 81.38% and 93.55%, respectively, for ABCD rule color feature assessment.

Keywords – Image analysis, Color Features, Melanoma, Teledermatology.

I. INTRODUCTION

Skin cancer is currently one of the most common types of cancer in Caucasian population and its increasing incidence has been leading to multiple public campaigns with the goal of raising awareness of the general population. Although Malignant Melanoma (MM) accounts for only a small percentage of skin cancer, it is far more dangerous than other skin cancers, while being responsible for most skin cancer related deaths [1]. Therefore, considering the high success rates of recovery when the malignancy is detected during the early stages of its development, the early diagnosis of MM is extremely important.

The majority of the computer-aided diagnosis (CAD) systems to analyze skin images automatically proposed hitherto in the literature are based on dermoscopic images. In [2] it is presented a systematic review of dermoscopy and CAD systems. Dermoscopy is a non-invasive method [3] that allows the in vivo evaluation of colors and microstructures of the epidermis, the dermo-epidermal junction and the papillary dermis not visible to the naked eye, opening a new dimension in the examination of pigmented skin lesions and, especially, in the identification of the early phase of MM.

The risk assessment of skin lesions in primary care centers is often conducted by looking at changes in the ABCD rule of dermatoscopy [4]. This algorithm addresses quantitatively the issue of whether a lesion is benign, suspicious or highly suspicious of Malignant Melanoma based on the analysis of four visual features: Asymmetry, Border, Color and Dermoscopic structures. The Total Dermoscopy Score (TDS) is given by the multiplication of each feature score by a weight factor (1.3, 0.1, 0.5 and 0.5, respectively), and depending on the obtained values, a grading of the lesions is possible with respect to their malignant potential.

The ABCD rule proposes the identification of six different colors in skin lesions considered highly relevant for melanoma diagnosis: white, black, blue-gray, red, light brown and dark brown. Different approaches have been suggested to assess the colors' presence in skin lesions [5]: by using a color palette; performing color histogram analysis; or using statistical parameters from different color spaces. In [6], the authors created a color palette constituted by six group colors which involved different color patches for each color group and assigned all the pixels of the skin lesion to the most relevant color patch in order to infer about the melanoma colors presence. Alternatively, a pattern recognition software toolkit was used to classify clinical skin lesions using relative color features obtained by histogram analysis [7]. Finally, in [8], the authors extracted features like mean and standard deviation of three lesion regions (lesion, inner periphery and outer periphery) over the three channels of six different color spaces to infer on the malignancy of the lesion.

In this work, a methodology to assess the number of ABCD rule colors present on a skin lesion image is presented, based on cluster characteristics analysis, having the advantages of being adaptable to different datasets. The paper is organized as follows: section II describes the datasets used; section III presents and explains the proposed methodology; section IV presents the obtained results; at last, conclusions are drawn in section V.

II. DATASET

The methodology suggested in this study was applied in two dermoscopic image sets and one mobile acquired dataset, described next. Examples of skin lesions from both datasets are depicted in Fig. 1.

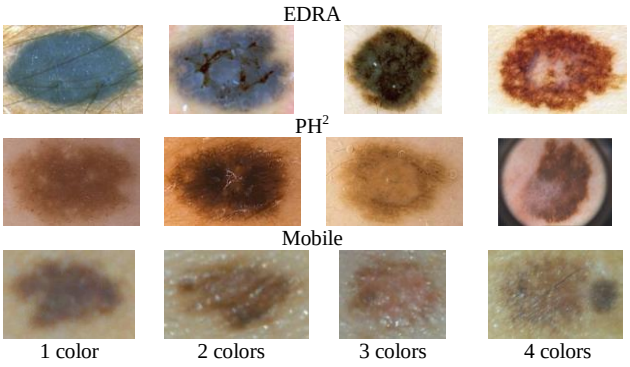


Figure 1. Examples of skin lesions from EDRA, PH^2 and Mobile datasets classified by medical experts according the number of colors

A. EDRA

The dermoscopic image set used consists on images obtained from two databases. The first is the CD-ROM Interactive Atlas of Dermoscopy [9], which is an image collection acquired in three institutions: University Federico II of Naples, Italy, University of Graz, Austria, and University of Florence, Italy.

A total of 296 images were used in this study. Moreover, all images were annotated by medical experts regarding the lesion color with the following distribution: 11 images with 1 color; 130 images were classified with 2 colors; 127 with 3 colors; and 28 images with 4 colors.

B. PH^2

The second dataset considered is the PH2 database with images collected in the Dermatology Service of Hospital Pedro Hispano (Matosinhos, Portugal) under the scope of the ADDI project [10].

A total of 160 images were used in this study. Analogously, all images were annotated by medical experts regarding the number of colors of the skin lesion: 52 images with 1 color; 88 images were classified with 2 colors; 15 with 3 colors; and 5 images with 4 colors.

C. Mobile

This dataset was collected at the Portuguese Institute of Oncology of Porto, under the scope of the project Melanoma Detection [11]. The dataset contains 80 24-bit color images with 652x652 pixels of resolution, acquired with a HTC One S mobile phone.

The dataset is composed by 15 images of 1 color, 51 images with 2 colors, 13 images with 3 colors and 1 image with 4 colors.

III. METHODOLOGY

The main goal of this methodology is to determine the number of ABCD colors present in a skin lesion image, ranging from 1 to 6 colors.

The first step of this methodology consists in performing the skin lesion segmentation. In the case of dermoscopic images, an adaptive threshold segmentation algorithm based on Otsu's threshold method was applied [12]. The threshold parameters used in this segmentation algorithm are dynamically estimated according to the image pixel resolution. In the case of lesion images acquired via mobile devices, a mobile application for the Android OS was used to help the image acquisition process, which performs real-time detection of the region of interest of the target skin lesion and segments the skin lesion [13].

Afterwards, the proposed methodology extracts 660 color features pooled in two different groups: global and relative features. Several feature selection techniques are later combined with different machine learning classifiers, in order to obtain the optimal subset of color descriptors, as well as the respective best classifier for a specific dataset (see Fig. 2).

A. Feature Extraction

The first step for the extraction color features consists in transforming the original image of the skin lesion, given in the RGB color space, to the Lab color space and also to the device independent $LC_{ab}H_{ab}$ color space [14]. In order to facilitate the clustering process, a mean shift filtering algorithm is applied to the three color spaces of the original image. Mean shift filtering removes noise by smoothing the image but preserves discontinuity, i.e. smoothing near edges is reduced [15].

The k-means clustering is then performed, with k varying from 1 to 6 clusters. Afterwards, a total of 660 features are extracted from two distinct groups: one that gathers the global characteristics of the skin lesion; and another uses relative features obtained from the Euclidean distances between the clusters centers, as summarized in Table 1.

For the relative color features, the mean, $\mu_{cluster}$, and standard deviation, $\sigma_{cluster}$, of each cluster previously found by k-means are calculated and also $measure_1$ and $measure_2$, that are defined as:

$$measure_1 = \frac{\mu_{cluster} - \mu_{skin}}{0.5(\mu_{cluster} + \mu_{skin})}, \quad (1)$$

$$measure_2 = \frac{\mu_{cluster}}{\sigma_{cluster}} - \frac{\mu_{skin}}{\sigma_{skin}}, \quad (2)$$

where μ_{skin} and σ_{skin} represent the mean color and standard deviation of the skin surrounding the lesion, respectively. The design of these cluster-based features was motivated by the need of using features with the potential of being adaptable to different skin lesion images datasets.

To better comprehend these cluster-based features, consider the case that k=4 where 4 clusters are obtained from the skin lesion. For each of the 4 clusters obtained, four metrics are calculated (mean, standard deviation, $measure_1$ and $measure_2$), the Euclidean distances of each pair of clusters (meaning 6 combinations) are obtained and the distances rearranged in descending order. In addition to the former features, the least square regression line from the resulting distances is calculated, and two extra features are extracted, namely the slope and coefficient of determination. The former procedure is repeated for the three color spaces with k varying from 1 to 6.

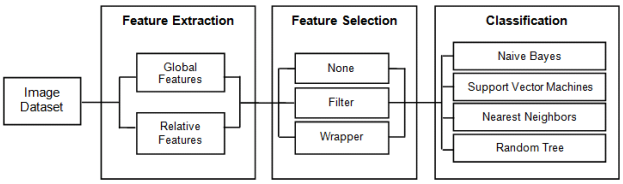


Figure 2. Block diagram of the color assessment methodology

TABLE I. Summary of the extracted color features

Group	Color Space / Channels	k	Descriptors
Global	R, G, B L,C_{ab}, H_{ab}	k=1	mean, standard deviation, energy, entropy, skewness, kurtosis, L1 norm, L2 norm
Relative	RGB Lab L$C_{ab}H_{ab}$	1≤k≤6	Euclidean distances: mean, standard deviation, measure$_1$ and measure$_2$
		3≤k≤6	slope and coefficient of determination: mean, standard deviation, measure$_1$ and measure$_2$

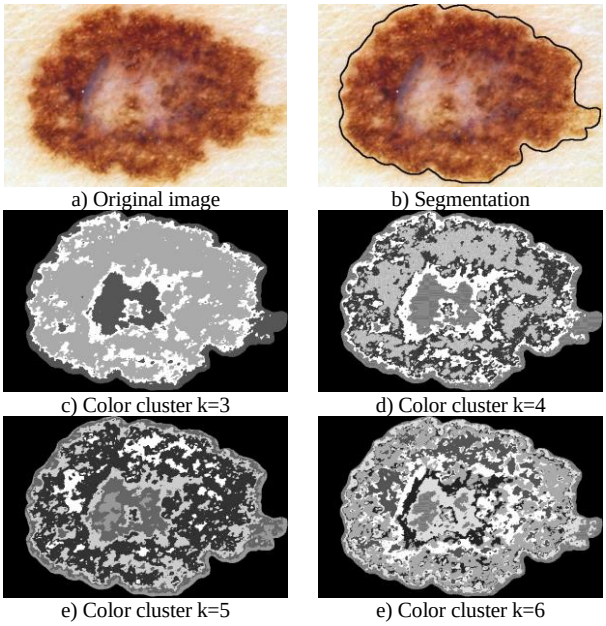


Figure 3. a) Example of skin lesion from the dataset, b) the segmentation obtained, c-e) color cluster masks obtained for 3≤ k≤6.

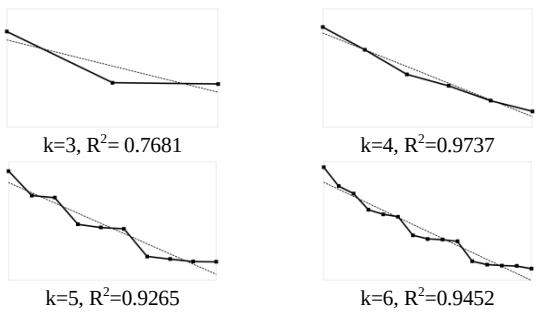


Figure 4. Example of Euclidean mean distance values obtained for 3≤ k≤6, rearranged in descending order, and respective regression lines and coefficient of determination ,R^2, of a skin lesion with 4 colors

The formulation of these cluster-based features was based on the fact that, for the ideal k value of a particular image, i.e. its actual number of ABCD colors, the clusters will be significantly distant from each other. This means that the regression line extracted from the referred cluster distances can present particular characteristics when traced for the ideal k value. As an illustrative example, in Fig. 3 and Fig. 4 the mask clusters obtained as well as the regression line are shown, for different k values, of a skin lesion image with 4 ABCD rule colors. For instance, the coefficient of determination with a value closest to one is obtained for the ideal k value.

B. Feature Selection

Feature selection algorithms can be classified into two main categories according to their evaluation criteria: filters and wrappers [16, 17]. In order to choose the best method of feature selection for each classifier, after a thorough research about the most effective and common approaches [17], three different feature selection methods were chosen: one filter-based method and two wrappers. The filter-based method used was the Correlation based Feature Selection (CFS). Regarding wrapper methods, the search methods used were Linear Forward Selection (LFS) and Greedy (G). In addition, the feature selection step using wrapper methods was performed independently for the four selected classifiers.

C. Classification

The last step of our approach consists in training a classification process with the feature attributes selected. Several techniques have been adopted to diagnose skin lesions [18] and so far no consensus has been reached on the best one to choose for. Due to the former reasons, four different classifiers were adopted: Bayes, Support Vector Machines (SVM), K-Nearest Neighbors and Random Tree. The first classifier was the Naïve Bayes that has the advantage of requiring a small amount of training data to estimate the parameters necessary for classification. In the second classifier, the Sequential Minimal Optimization (SMO) algorithm was applied for training a SVM, through a polynomial kernel. For the nearest neighbors, K=3 was used, meaning that the algorithm locates the 3 nearest instances to the query instance and determines its class by identifying the single most frequent class label. The last classifier considered was the Random Tree, also widely used in this field.

During the experiments, 10-fold cross-validation was used, meaning that 90% of the samples were used for training and the lasting 10% were used for testing purposes.

IV. RESULTS

The methodology presented was applied in two dermoscopic image sets, EDRA and PH^2, and one mobile acquired dataset. In order to deal with class imbalance, a widely used synthetic over-sampling technique, SMOTE [19], was applied. Other studies like [8, 20] also use this technique, which over-samples the minority classes introducing synthetic samples using the k minority class

nearest neighbors, in this study k=3 was used. Table 2 presents the dataset distribution before and after applying SMOTE.

The classification results are presented in terms of three metrics ordinarily used for this purpose: 1) accuracy (AC), i.e. the percentage of lesions corrected classified overall; 2) kappa statistic is a measure of agreement, where a value close to 1 mean better reliability; and 3) ROC area, that quantifies the overall ability of the test to discriminate the number of colors, being a good classifier when the value is near 1.

Tables 3 and 4 summarize the results obtained with the various feature selection methods and classifiers adopted using each dermoscopic dataset and Table 5 presents the results obtained using the mobile acquired dataset. Regarding the overall results one can confirm that feature selection besides reducing the number of features to be analyzed (from 660 features to less than 100), improves the performance of the classifier, except for the SVM classifier. Another common result is the performance of the Random Tree classifier that is significantly lower than the others.

Moreover, the best performances for both dermoscopic datasets are given by the SVM classifier using all 660 features (accuracy of 77.75% and 81.38%). For the EDRA dataset, the second best classification is given by the Naïve Bayes classifier using the wrapper evaluator with greedy search method, that selects 12 features and achieves 74.22% of accuracy, while for the PH^2 dataset the second best classifier is the Ib-3 using the wrapper evaluator with greedy search method, that selects 9 features and achieves 77.78% of accuracy. The selected features are different in the datasets, include features from both global and relative groups, choosing features from the 3 color spaces and mainly focusing on metrics like slope and coefficient of determination.

Regarding the classification results obtained with the Mobile dataset, the Naïve Bayes classifier using the wrapper search method with LFS attains the best results, with 93.55% of accuracy, using 23 features. The second best classifier was the SVM using all 660 features achieving 89.03% of accuracy. For the Mobile dataset the selected features also include features from the 3 color spaces and focus on metrics like slope and coefficient of determination, likewise as the case of dermoscopic images.

Thus, the former results confirm that the proposed methodology is adaptable not only to dermoscopic image datasets acquired with different dermoscopes, which might present considerable differences in terms of image quality and illumination conditions, but also to a mobile acquired image dataset.

TABLE II. Dataset distributions before and after SMOTE data sampling application

Number of colors	EDRA		PH^2		Mobile	
	before	after	before	after	before	after
1	11	110	52	83	15	52
2	130	130	88	88	51	51
3	127	128	15	82	13	52
4	28	113	5	80	1	-

TABLE III. Classification results obtained with the EDRA dataset for each classifier according the feature selection method adopted

Classifier	Feature Selection	AC (%)	Kappa (%)	ROC (%)area	# Features
Bayes	none	57.38	43.53	81.8	660
	CFS	71.73	62.25	89.5	74
	W + LFS	64.24	52.32	84.3	5
	W + G	74.22	65.59	89.6	12
SVM	none	77.75	70.34	87.8	660
	CFS	71.10	61.49	85.2	74
	W + LFS	68.40	57.94	83.5	17
	W + G	71.93	62.58	85.3	16
Ib-3	none	65.49	54.21	84.6	660
	CFS	65.90	54.83	85.6	74
	W + LFS	64.66	52.87	83.0	8
	W + G	66.94	55.95	83.4	10
Random Tree	none	56.13	41.45	70.4	660
	CFS	61.54	48.67	74.0	74
	W + LFS	62.58	50.04	74.8	11
	W + G	62.79	50.37	74.9	4

TABLE IV. Classification results obtained with the PH^2 dataset for each classifier according the feature selection method adopted

Classifier	Feature Selection	AC (%)	Kappa (%)	ROC (%) area	# Features
Bayes	none	62.16	49.76	85.0	660
	CFS	72.07	62.85	91.1	95
	W + LFS	69.37	59.21	87.3	16
	W + G	75.68	67.56	90.4	12
SVM	none	81.38	75.19	89.9	660
	CFS	81.08	74.80	89.6	95
	W + LFS	68.77	58.45	81.9	10
	W + G	78.38	71.20	87.9	21
Ib-3	none	75.38	67.26	89.7	660
	CFS	74.47	66.11	89.8	95
	W + LFS	72.37	63.20	85.2	9
	W + G	77.78	70.39	88.5	9
Random Tree	none	62.46	49.94	74.8	660
	CFS	65.77	54.36	77.1	95
	W + LFS	64.96	53.11	76.4	8
	W + G	74.17	65.57	82.7	5

TABLE V. Classification results obtained with the Mobile dataset for each classifier according the feature selection method adopted

Classifier	Feature Selection	AC (%)	Kappa (%)	ROC (%)area	# Features
Bayes	none	71.61	57.43	88.1	660
	CFS	80.65	70.96	94.9	49
	W + LFS	93.55	90.32	97.9	23
	W + G	88.39	82.58	91.3	12
SVM	none	89.03	83.53	91.3	660
	CFS	87.74	81.6	92.7	49
	W + LFS	83.87	75.78	88.4	22
	W + G	81.94	72.88	86.6	11
Ib-3	none	68.39	52.44	86	660
	CFS	73.55	60.22	91.0	49
	W + LFS	80.00	70.00	90.1	8
	W + G	80.00	70.00	88.8	9
Random Tree	none	63.23	44.79	72.4	660
	CFS	69.68	54.50	77.3	49
	W + LFS	72.90	59.34	79.7	5
	W + G	72.90	59.34	79.7	4

V. CONCLUSIONS AND FUTURE WORK

In this work, a methodology to assess the number of ABCD rule colors of skin lesions from images is presented. Most of the methodologies presented to date for color skin lesions analysis are highly dependent of the image quality and illumination conditions of the images

used, like the popular color palette approach, which has the disadvantage of not being adaptable to datasets acquired with different dermoscopes.

The novelty of methodology proposed in this work consists on the granular color skin lesion analysis approach, by assessing the ABCD rule color score of the skin lesions. In order to be adaptable to different datasets and different image types, another differential factor of this methodology is the usage of cluster-based features, whose design was motivated by the need of using features with the potential of being independent of the characteristics of the used image dataset. Following this new cluster-based approach, two new descriptors were proposed ($measure_1$ and $measure_2$) as well as existing descriptors were also calculated for each cluster. Moreover, and since no consensus exists on the best classifier to apply in this type of studies, four different classifiers were tested.

The results confirm that the methodology accurately classifies the number of ABCD rule colors, and is adaptable for different dermoscopic image datasets and the mobile dataset used. For the EDRA dataset, the SVM classifier obtains the best performance with 77.75% of accuracy extracting all features, while Bayes classifier using wrapper feature selection with greedy method obtains 74.22% of accuracy using 12 features. For the PH^2 dataset, the SVM classifier obtains the best performance with 81.38% of accuracy extracting all features, followed by the Ib-3 classifier using wrapper feature selection with greedy method obtains 77.78% of accuracy using 9 features. In comparison with a similar study [21] that presents 62.5% of accuracy in the evaluation of the number of colors in a skin lesion, the proposed approach achieves better results. However, most of the studies in the literature focus on classifying the skin lesions in terms of melanoma or non-melanoma, whether than concentrating on each criteria of the ABCD rule of Dermatoscopy.

For the Mobile dataset, the Bayes classifier using wrapper feature selection with LFS method obtains 93.55% of accuracy using 23 features, followed by the SVM classifier with 89.03% of accuracy extracting all features. For dermoscopic images the possible range of colors varies from 1 to 6. According to a dermatology expert, the number of colors possible to distinguish in mobile acquired images is at most 4 colors. One hand, the dermoscopic images uses epiluminescence microscopy that allows the observation of morphologic key structures, with more detail and thus more colors, located in the epidermis down to the upper dermis. One the other hand, mobile acquired images capture information only from the epidermis, less structures can be observed with fewer colors. Thus, the clusters color variability is lower in mobile acquired images, which can explain the higher accuracy rates obtained using this type of images.

As future work, the authors plan to develop a similar methodology to assess the differential structures score, with the main goal of assembling a risk triage framework for skin lesion analysis based on ABCD rule that scores each feature independently.

ACKNOWLEDGMENT

This work was done under the scope of the project "SMARTSKINS: A Novel Framework for Supervised Mobile Assessment and Risk Triage of Skin Lesions via Non-invasive Screening" with reference PTDC/BBB-BMD/3088/2012 financially supported by Fundação para a Ciência e a Tecnologia in Portugal.

REFERENCES

[1] D. Whiteman, A. Green, " Epidemiology of Malignant Melanoma. Skin Cancer – A World-Wide Perspective". Springer-Verlag Berlin Heidelberg, 2011.

[2] A. Masood, A. Ali Al-Jumaily, "Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms," Int. journal of biomedical imaging, article ID 323268, 22 pages, 2013.

[3] W. Stolz, P. Bilek, M. Landthaler, et al, "Skin surface microscopy," The Lancet 334.8667, 1984, pp. 864-865.

[4] W. Stolz, A. Riemann, A. Cognetta, L. Pillet, W. Abmayr, D. Holzel, et al., "Abcd rule of dermatoscopy-a new practical method for early recognition of malignant-melanoma," European Journal of Dermatology, vol. 4, pp. 521-527, 1994.

[5] R. Garnavi, "Computer-aided diagnosis of melanoma," PhD thesis. The University of Melbourne, 2011.

[6] S. Seidenari, G. Pellacani, C. Grana, "Computer description of colours in dermoscopic melanocytic lesion images reproducing clinical assessment," British Journal of Dermatology, vol. 149, no. 3, 2003, pp. 523–529.

[7] Y.I. Cheng, R. Swamisai, S.E. Umbaugh, et al., "Skin lesion classification using relative color features," Skin Research and Technology, vol. 14, no. 1, 2008, pp. 53–64.

[8] M.E. Celebi, H.A. Kingravi, B. Uddin, H. Iyatomi, Y.A. Aslandogan, W.V. Stoecker, R.H. Moss, "A methodological approach to the classification of dermoscopy images," Computerized Medical Imaging and Graphics vol. 31, pp. 362-373, 2007.

[9] G. Argenziano, H.P. Soyer, V. Giorgio, D. Piccolo, P. Carli, M. Delfino, et al., "Dermoscopy: a tutorial," Milan EDRA Medical Publishing & New Media, 2002.

[10] T. Mendonça, P.M. Ferreira, J.S. Marques, A.R. Marcal, J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5437-5440, 2013.

[11] P. Fraunhofer, Melanoma detection, internal project. http://www.fraunhofer.pt/en/fraunhofer_aicos/projects/internal_research/melanoma_detection.html, 2015.

[12] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man and Cybernetics, vol. 9(1), pp. 62-66, 1979.

[13] L. Rosado, M.J.M. Vasconcelos, "Automatic segmentation methodology for dermatological images acquired via mobile devices" In Proceedings of 8th International Conference on Health Informatics, Lisbon, pp. 246-250, 2015.

[14] T. Gevers, A. Gijsenij, J. Van de Weijer, J.M. Geusebroek, "Color in computer vision: fundamentals and applications," vol. 24, John Wiley & Sons, 2012.

[15] D. Comaniciu, P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 603-619, 2002.

[16] M.A. Hall, G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," IEEE Transactions on Knowledge and Data Engineering 15, pp. 1437-1447, 2003.

[17] H. Liu, L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, pp. 491-502, 2005.

[18] J. Scharcanski, M.E. Celebi, "Computer vision techniques for the diagnosis of skin cancer", 2013.

[19] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–57, 2002.

[20] G. Capdehourat, A. Corez, A. Bazzano, P. Musé, "Pigmented skin lesions classification using dermatoscopic images," Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 537-544, Springer Berlin Heidelberg, 2009.

[21] J.L.G. Arroyo, B.G Zapirain, "Automated Detection of Melanoma in Dermoscopic Images", Computer vision techniques for the diagnosis of skin cancer, pp. 139-192, 2013.

Automatic adaptation of filter sequences for cell counting

Uroš Čibej†, Jasna Lojk*, Mojca Pavlin*, and Luka Šajn†
†Faculty of computer and information science, University of Ljubljana
*Faculty of electrical engineering, University of Ljubljana
Ljubljana, Slovenia

***Abstract*—Manual cell counting in microscopic images is usually tedious, time consuming and prone to human error. Several programs for automatic cell counting have been developed so far, but most of them demand some specific knowledge of image analysis and/or manual fine tuning of various parameters. Even if a set of filters is found and fine tuned to the specific application, small changes to the image attributes might make the automatic counter very unreliable. The goal of this article is to present a new application that overcomes this problem by learning the set of parameters for each application, thus making it more robust to changes in the input images. The users must provide only a small representative subset of images and their manual count, and the program offers a set of automatic counters learned from the given input. The user can check the counters and choose the most suitable one. The resulting application (which we call Learn123) is specifically tailored to the practitioners, i.e. even though the typical workflow is more complex, the application is easy to use for non-technical experts.**

I. Introduction

Cell counting on microscopic images of various types is one of the fundamental tasks for many researchers and practitioners in all life-sciences. Being such a fundamental task, many automated solutions exist. Despite this abundance of tools for automated counting, the majority of counting tasks are still done manually by the practitioners. The main reasons for this can be boiled down to two problems. The first problem is the inaccuracy of the automated counters and the second one the difficulty to use for non-technical experts. Our work is focused on overcoming both of these problems, i.e. to develop an accurate tool, and a tool that is simple to use out-of-the-box.

Currently several programs for cell counting are already available. From many commercial solutions, such as MetaMorph, BioQuant, Image-Pro and SymenTec to free software, such as ITCN ImageJ Plugin [5], CellProfiler [4], UTHSCSA ImageTool [1], and CellC [10]. However, all of the existing solutions require some additional input or even some image preprocessing of the raw images, which is not only time-consuming but can also be too complex to some biology field experts. [3].

We already developed a cell counting tool called CellCounter [7], with which we demonstrated that for a particular application (i.e. cell viability estimation), a fixed set of filters with specifically tailored parameters can give great results, making the human counting almost obsolete and thus speeding up the analysis of the experiment results in life-sciences by an order of magnitude. This result is very important in practice, however it is still not the silver bullet for all the problems one might encounter in every such application.

The program was developed in collaboration with a research group, who used cell counting as a way to determine cell viability [2] and transfection efficiency [8] for certain experimental settings. By fluorescently staining cell nuclei they obtained fluorescent images with higher contrast compared to BF or PC images, but it still left them with large number of microscopic images to count, requiring a more efficient way to process the images than manual counting.

To get an intuitive picture of all the problematic features of the counting of cells, let us mention some of the problems, that typically arise in practice: different types of assays resulting in a large variety of images which are hard to generalize, inability to detect individual nuclei in multinucleated cells, overlapping cells, uneven illumination and other equipment related factors, such as electronic and/or optical noise [6], that lead to images having variable contrast and quality, low contrast of the image, different cell shapes, internal cell structures, extracellular debris, and varying density of cell culture.

Due to all these possibilities, there is a large set of variations that prevent a general solution for any circumstance to be developed. The optimal tool would detect all these variations without any user intervention. However, the many failed attempts to find such a tool demonstrate that a different approach is needed. The main goal of this article is to present such an application, which is able to adapt to the specific use-case, with (as little as possible) guidance from the user. The user only provides a set of examples, from which the application learns the general counting as accurately as possible.

Since one of the major objectives is also a widespread usability of this application, the following goals are sought:

- the learning should be done as intuitively as possible,
- the result of the learning should be easily verifiable, to ensure that the obtained results are sound,
- the workflow must not be too complex, avoiding any technology related terms,
- the learning has to be a repeatable and incremental process,
- the results of the learning and counting must be usable also in other de-facto standard applications of the user-base.

Automatic learning is a widespread area of research, so there are plenty of possible approaches to developing such an application. We define the learning as an optimization problem,

where the goal is to optimize the difference between the counting results defined by the user, and the counting obtained by the automatic counter. As mentioned above, the CellCounter application used a predefined sequence of filters, where the parameters were fine-tuned for a particular application. We used this sequence as a general framework, and the parameters of the image filters represent the vector which needs to be optimized.

As the optimization method we chose an evolutionary approach, because of the many positive features. Namely, the evolutionary approach is very flexible, we can easily change the goal function, stop the process at any point and get feasible solutions, start from existing solutions and improve the solution by interacting with the user.

The remainder of the article is structured as follows. Section II. presents the learning application, section III. describes all the implementation details, the framework, and the technologies used, section IV. describes the experimental evaluation of the application, and finally, section V. concludes the paper and gives directions for our future work.

II. Description of the Application

In this section we will first describe the application from a purely user perspective. Since one of our basic goals is to make the application as simple as possible for the end-user, we managed to strip down all the unnecessary concepts. The user does not have to be aware of any underlying algorithm and technology and in the workflow it can only focus on the images and the task at hand. This is an extremely important aspect for the practical usability of such application. Many similar applications overflow the user with technical terms and make them specify various parameters of algorithms, of which the end user should not be aware of.

The application is subdivided into two sub-application. The first one being the learning part (Learn123) and the second the counting part (Count123) of the application. In what follows we describe in more detail the tasks that constitute a typical workflow, starting with the learning part of the application.

A. Choosing a subset of images

The first task of the user is to select a representative subset of images. The subset needs to be relatively small (typically around 10 images) and the images spawn as many different aspects of the entire set (i.e. different sizes of object, different lighting conditions, different shapes, etc.).

B. Manual count

The second step for the user is to manually count the objects in the images. The information that it provides are only the approximate centers of the object. This is another very useful simplification, since some counting tools require the user to provide the contours, or at least the center and the radius of the object. Providing only the centers speeds up this step significantly.

C. The learning process

Once all the objects in the images have been manually annotated, the learning process can be initiated. The length of the learning process depends mostly on the size of the image subset, the typical span of the learning is from 5 minutes to about an hour. But the user can also participate in this ongoing process, since a set of candidate counters is immediately offered. This set is the set of currently best known counters and the user can assess the given counters. If the user is satisfied with the results he/she can interrupt the learning and use the counter.

D. Assessing the quality of the counters

When the program finishes its learning process, the set of 10 best counters is presented to the user. The user can see the count of the automatic counter visually and compare it to the counts made manually. If none of the counters is satisfactory, the user can add new images to the testset, remove some old images, and restart the learning process. Fig. 1 shows an example of a counter assessment. Notice that the learning does not start from scratch, it uses most of the information learned from the previous learn run.

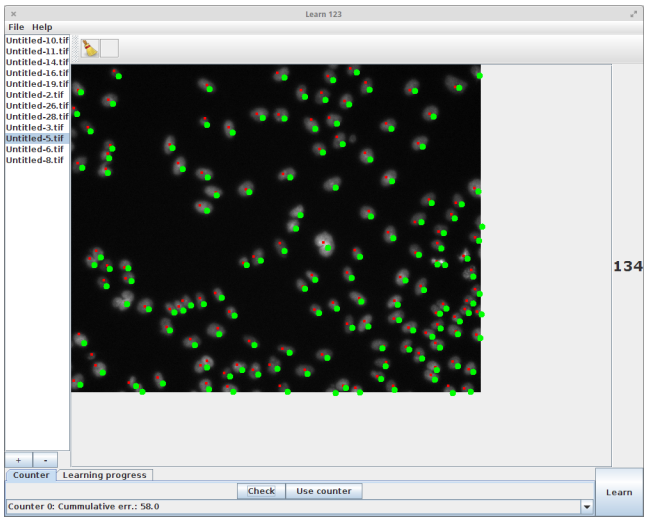


Fig. 1. The comparison of the manual and automatic count in a test image. The small red dots represent the manual count, the larger green dots represent the automatic counter.

E. Counting application

When a suitable counter is found by the above process, it can be used in the second sub-application, the Count123. The user can choose all the images in the set and the application will visually present the results. All the results can thus be checked and any point can be added or removed to correct any errors.

The assessment can also find some images on which the automatic counter behaves badly. These images (or one typical representative) can be added to the testset and the new learning process will adapt also to this type of images.

If the obtained results are satisfactory, they can be easily exported and used in other applications for further data analysis.

Fig.2 show an overview of the workflow. Every part of the workflow is technically neutral, agnostic about any algorithms used behind the scenes.

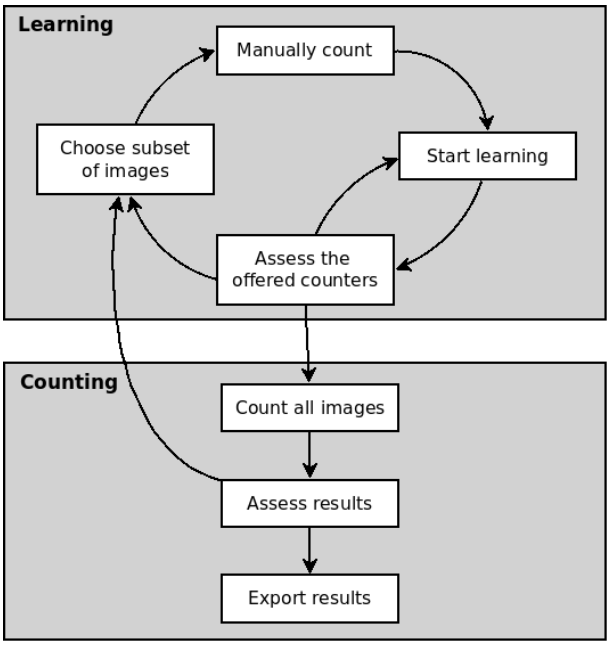


Fig. 2. The user workflow in the Learn123 application.

III. Implementation details

The implementation details can be boiled down to three aspects: (1) the modeling of the learning as an optimization problem, (2) the algorithm used for the optimization, and (3) the technologies used for the implementation.

A. The optimization problem

Intuitively, the objective of the learning process is to find a counter (specified by a (numeric) vector) with minimal counting error on the learning set. There are many possibilities how to encompass this problem into a mathematical model, the presented optimization model empirically demonstrated very good results, showing its good applicability to real problems.

As the input, a set of k images is given:

$$Img = \{img_1, img_2, \ldots, img_k\}.$$

For every image img, the manual count is given by a sequence of points in the plane, indicating the (approximate) centers of the identified objects:

$$C_{man}(img) = \{p_1, p_2, \ldots, p_n\}$$

The subject of the optimization is a vector of parameters

$$P = \{param_1, param_2, \ldots, param_l\},$$

which uniquely define the automatic counter. These parameters can be integer values, boolean values, real numbers, or certain objects from specific sets, but without any loss in generality we can assume only integer valued vectors.

The result of applying the counter defined by the vector P to the image img will be denoted by the set of points:

$$C_P(img) = \{p_1, p_2, \ldots, p_m\}.$$

Now for the most important part of the model, we have to define the goal function (criteria), which will model what we consider to be an accurate counter. Intuitively, there are two criteria that a good counter should fulfill:

- the difference of the obtained counts (the numbers) should be as small as possible, and
- the positions of the found objects should be as close as possible.

These two intuitive criteria have to be modeled appropriately and combined into a suitable goal function. We will first define this function for one image img. The first criterion is simple to model as a simple absolute difference between the number of points found with the manual counts and number of points with automatic counts on the same image:

$$\Delta(C_{man}(img), C_P(img)) = ||C_{man}(img)| - |C_P(img)||.$$

The second criterion compares all the points found with the manual count and see how far are they from the points found with the automatic counter.

The goal function for the second criterion was chosen as

$$D(C_{man}(img), C_P(img)) = \sum_{p \in C_{man}(img)} d(p, closest(p, C_P(img)))$$

,

where d is the Euclidean distance between the two points, and $closest$ is a function that finds the closest point from a given point to the given set.

This goal function is simple and fast to compute, but it can run into problems on some corner cases where it prefers degenerate counters. An example of the problem with this function is given in Fig. 3. This problem will be amended by combining the two goal functions. How these two goal functions are combined is described together with the algorithm, since it is more specific to the way the algorithm works.

The functions Δ and D model the differences in counts for a single image. Now we need to choose a suitable function that combines the results of these two function on all images. We experimented with a few variations, such as the maximum value, the median value, and the average value over all images. Empirically the simple average gives the best results, so the goal functions which combine Δ on all images is

$$G_\Delta(P, Img) = \sum_{im \in Img} \Delta(C_{man}(img), C_P(img)),$$

and an equivalent definition is used for G_D.

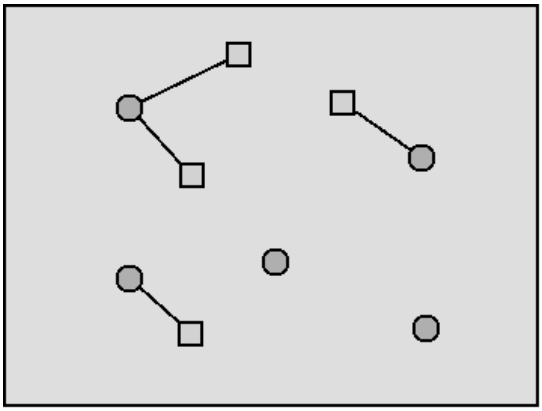


Fig. 3. An example of the D criterion matching between manual counts and automatic counts. The circles represent automatic counts, the square represent manual counts. This example also demonstrates the shortsightedness of this goal function, which has to be amended with the function Δ.

1) Evolutionary algorithm: An evolutionary algorithm is used for the minimization of the above defined optimization problem. The main reason for this choice is the general applicability of this algorithm, robustness, and the fact that at any point there are feasible solutions available, which can be immediately presented to the user for assessment. The framework of each counter is a sequence of filters for which we demonstrated in [7] to be a very adequate sequence for counting cells in electronic microscope images. This sequence of filters was manually fine-tuned for the specific application. Namely, each filter has a set of parameters, and the final count is very sensitive to the specific values of these parameters. The sequence of filters (i.e. image-processing algorithms) used for counting is as follows:

1) Enhance contrast - the images have typically a very low intensity and a very low contrast. By enhancing the contrast, many more features of the image are visible, and the results of the counting improve dramatically.
2) Threshold - in the second step, the image is converted to binary format. For this a threshold is set, by using either a specific value of the threshold or one of the 13 possible algorithms.
3) Watershed - by thresholding we typically loose the boundary between different cells, and by watershedding [9] this merged objects are divided again.
4) Holefill - after the application of the above filters, the objects are not uniformly filled. With this filter, the holes in the objects are filled, making it more suitable for the next detection step.
5) Particle analysis - the last step of the algorithm detects the remaining blobs of active pixels in the image and decides whether to add an object to the count or not. There are a few parameters upon which this decision is made - the most important being the minimum radius, the maximum radius, and the eccentricity of the object. With this parameters the counter can adapt to count only specifically shaped objects or objects with specific sizes.

The parameters for these filters represent the chromosome, as it is shown in Fig. 4, and a typical evolutionary algorithm

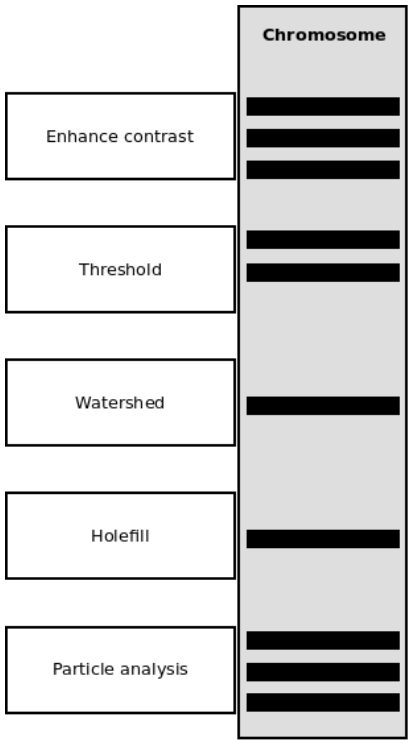


Fig. 4. An overview of the basic chromosome.

is used to evolve from completely random counters to very successful ones. The pseudocode for the algorithm is shown in Algorithm 1

Algorithm 1 The outline of the learning algorithm

```
1:  population = randomPopulation(populationSize);
2:  for gen ∈ [1, 2, ..., numGen] do
3:      if gen ≤ numGen/10 or gen%2 = 0 then
4:          goal = G_Δ
5:      else
6:          goal = G_D
7:      end if
8:      rankedPop = sortByFitness(population, goal)
9:      pop1 = select(rankedPop,populationSize)
10:     pop2 = select(rankedPop, populationSize)
11:     population = crossover(pop1, pop2)
12:     population = mutate(population)
13:     population = take(population, populationSize)
14: end for
```

Lines 3-6 of the pseudocode demonstrate the aforementioned combination of functions Δ and D. The first 10% of the generations, the fitness function is only Δ. After this starting generations, the D function is used in every other generation, eliminating possibly corrupt counters that accidentally count a similar number of cells, but on completely different positions that those obtained by the manual count.

The selection (lines 8-10) of the counters which are used for creating a new generation is done with rank selection, i.e. the probability that a counter is chosen is proportional to its rank in the current population.

The crossover (line 11) used is a one-point crossover, i.e. a random position in the chromosome is chosen and the two chromosomes exchange the parameters to the left of that position. The mutation (line 12) is not a bit-level mutation (as it is usually the case with genetic algorithms), but we used a mutation on the parameter level, where a uniformly random value is chosen from the parameter domain.

Ultimately, to obtain the next generation, only the best counters are chosen in line 13.

B. The technologies used

The program was developed in the programming language Scala. Scala compiles to the JVM, making the Learn123 application completely platform independent. Furthermore, a wide range of de-facto standard Java tools can be used as a part of the application. We integrated the entire ImageJ application into Learn123; for the users to help them further manipulate the images as well as for the algorithm itself, since most of the filters in the counting algorithm are directly used in ImageJ. Since a lot of researchers use this tool for their everyday work, they can use Learn123 just to find the appropriate counter and then use it entirely in ImageJ.

IV. Experiments

The algorithm was developed and optimized based on microscopic images obtained from several experiments the collaborating research group provided, but one experiment was selected for the evaluation of the program efficiency.

The experiments were done on Chinese hamster ovary cells (CHO) grown in Ham's tissue culture medium for mammalian cells (HAM) supplemented with 10% fetal bovine serum (FBS) at 37 °C in 5% CO_2-enriched air at saturation humidity. All experiments were performed on 24h old cell cultures in exponential growth phase.

The cell viability experiment was performed as described previously [2]. Briefly, cells were incubated with increasing concentration of polycationic polymer coated magnetic nanoparticles for 24 h and stained with two fluorescent dyes; Hoechst 33342, which stained cell nuclei, and propidium iodide (PI), which differentially stained dead cells. At least 15 visual fields at 200 × magnification were taken of each sample for each used fluorescent dye using a fluorescent microscope (Zeiss 200, Axiovert, Germany). The images were recorded by MetaMorph imaging system software (Visitron, Germany) and saved in lossless TIFF format.

For the Hoechst dye 173 images were obtained and for the PI dye 159 images were obtained. All these images were manually counted. Then (for each dye separately) a small training set was chosen (15 images for PI and 13 images for Hoechst). The manual counting was done again using Learn123 and the learning process offered a set of the best counters (according to the function G_Δ). The best offered counters were used to perform the automatic count on the entire set of images. Figures 5 and 6 show a scatterplot of the comparison between automatic count (x-axis) and manual count (y-axis).

The results show a very good correlation between the manual and automatic counts. Only a couple of images in the entire set exhibit a somewhat larger error. But even these errors can be easily detected by quickly skimming through the results in Learn123, and manually add or remove the few erroneous points.

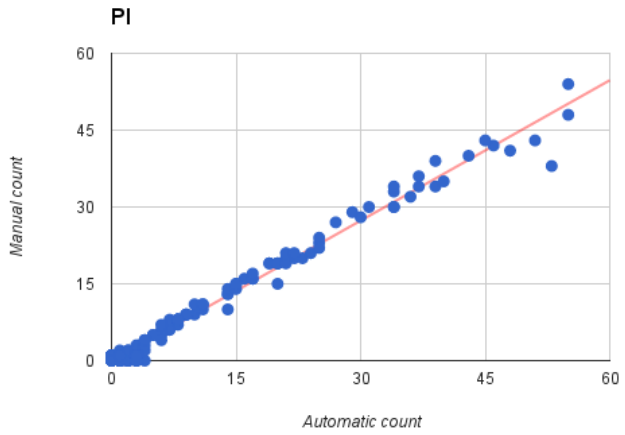


Fig. 5. Scatterplot showing a very good correlation between manual counting and automatic counting.

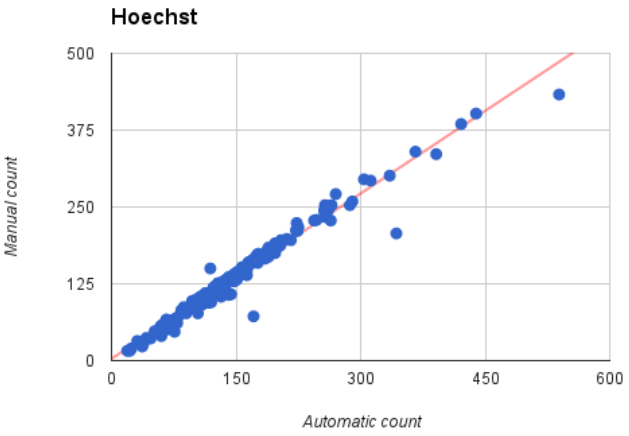


Fig. 6. Scatterplot showing a very good correlation between manual counting and automatic counting.

V. Conclusion

We presented a novel application for cell counting. The application enables regular, non-technical users to find better suited automatic counters, tailored specifically for their set of images. This is one step towards overcoming the problem of huge variations of the problems set, which can only be solved by a great adaptability of the application. Learn123 offers a simple workflow in which the user can teach the program how to count, by only giving example counts on a small subset of all the images.

The presented method was used on a fixed sequence of filters, however it can easily be extended to any sequence. Our future goal is to extend this application for any sequence

of filters that can be defined as an ImageJ macro. With this extension Learn123 will become a truly flexible tool, usable not only for cell counting, but for many other counting applications.

ACKNOWLEDGEMENT

This work was supported by Slovenian Research Agency within projects J4-4324,J2-6758,J3-6794 and young researchers program.

REFERENCES

[1] Uthscsa imagetool, home page. web, 2014. http://compdent.uthscsa.edu/dig/itdesc.html.

[2] Vladimir Boštjan Bregar, Jasna Lojk, Vid Šuštar, Peter Veranič, and Mojca Pavlin. Visualization of internalization of functionalized cobalt ferrite nanoparticles and their intracellular fate. *International journal of nanomedicine*, 8:919–931, 2013.

[3] Jiyun Byun, Mark R Verardo, Baris Sumengen, Geoffrey P Lewis, BS Manjunath, and Steven K Fisher. Automated tool for the detection of cell nuclei in digital microscopic images: application to retinal images. *Mol Vis*, 12:949–960, 2006.

[4] Anne Carpenter, Thouis Jones, Michael Lamprecht, Colin Clarke, In Kang, Ola Friman, David Guertin, Joo Chang, Robert Lindquist, Jason Moffat, Polina Golland, and David Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100, 2006.

[5] Sean R Gallagher. Digital image processing and analysis with imagej. *Current Protocols Essential Laboratory Techniques*, pages A–3C, 2010.

[6] Kang Li, Eric D Miller, Lee E Weiss, Phil G Campbell, and Takeo Kanade. Online tracking of migrating and proliferating cells imaged with phase-contrast microscopy. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 65–65. IEEE, 2006.

[7] J. Lojk, L. Šajn, U. Čibej, and M. Pavlin. Automatic cell counter for cell viability estimation. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 239–244, May 2014.

[8] Igor Marjanovič, Maša Kandušer, Damijan Miklavčič, Mateja Manček Keber, and Mojca Pavlin. Comparison of flow cytometry, fluorescence microscopy and spectrofluorometry for analysis of gene electrotransfer efficiency. *Submitted for publication*, 2014.

[9] Roerdink and Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. *FUNDINF: Fundamenta Informatica*, 41, 2000.

[10] Jyrki Selinummi, J Seppala, Olli Yli-Harja, and Jaakko A Puhakka. Software for quantification of labeled bacteria from digital microscope images by automated image analysis. *Biotechniques*, 39(6):859, 2005.

Assessing Blood Pressure Unobtrusively by Smart Chair

D. Zazula, J. Kranjec, P. Kranjec and B. Cigale
University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
damjan.zazula@um.si

Abstract – We developed a smart chair with unobtrusive sensors that measure functional-health parameters of a person sitting on the chair. Capacitive sensors are placed in the chair's backrest and seat, while the armrests support a combination of U-shaped electrodes and incorporated photoplethysmographic (PPG) sensors for synchronous electrocardiographic (ECG) and PPG measurements. In a set of experiments with 11 young males, the two types of signals were acquired. Time distances were estimated between the ECG R-wave peaks and the PPG foots as detected by our heartbeat search, yielding the so called pulse transit times (PTTs). The experiments were conducted in two phases: the first one at rest and the second one after a minute of intensive squats. Referential systolic and diastolic blood pressures were taken just before every trial by a Critikon Dinamap Pro 300 sphygmomanometric device. Our goal was to model the relationship between blood pressure and estimated PTTs in both experimental phases. The obtained linear models revealed an interesting observation. Subjects reacted in two different physiological ways; 6 out of 11 participants conformed to a different model as the other 5 did. The main difference is the rate of decrease of either systolic or diastolic pressure per 1-ms change of the PTT.

Key-words: Smart chair, Unobtrusive sensing, Pulse transit times, Electrocardiogram, Photoplethysmogram, Blood pressure

I. INTRODUCTION

Today's common practice in home care and home health monitoring still depends largely either on the skills of the observed persons or help of caregivers. As for the elderly or impaired persons, even simplest tasks, such as pressing a button, may not be feasible. Also persons without physiological or mental problems often feel discomfort if they are expected to take care about a measuring device, measuring procedure, and data forwarding or interpretation. For example, most telemetric projects currently underway include interventions of people being monitored. A person is expected that he or she regularly measures his/her blood pressure, which means taking and switching on a device, mounting a pressure cuff, activating and monitoring the measurement, reading out the results and forwarding it over the phone or in an e-mail. This is very hard for elderly, if not impossible, and very error-prone. And due to severe worldwide healthcare budget cuts, it became unreal to count on constant help of caregivers.

It is then obvious that any home measurement must be completed automatically, without a need of any personal awareness or interaction. This, of course, implies unobtrusiveness. Sensors, or sensing devices, must reside in dwelling environments in such a way that daily activity of people living there, and also all abnormal situations, activate measurements in parallel. If, for example, sensors (and algorithms) can detect heartbeat and breathing unobtrusively, a measuring device can be built into the bed and transmit signals whenever a person in bed. Similar solutions are available for chairs, shoes, etc. Not only long-term and unobtrusive monitoring of functional-health parameters is feasible, but also all deviations from normal can be detected and automatically reported.

By the end of 2013, we developed a system of unobtrusive sensors and devices for observing functional health, which was funded within the framework of the Biomedical Engineering Competence Centre. The system includes measuring of biopotentials from human body parts, in particular palms and fingers, mechanical and acoustic vibrations due to the activity of internal organs, such as the heart and lungs, mechanical movements of the body and body parts, changes of pulsative pressure in blood vessels, blood oxygenation, the temperature of body parts, and face colour and emotions. Sensors were placed into household appliances, beds, mirrors, and shoes [1].

A logical step forward was to build a smart chair (Fig. 1), which is a perfect example of unobtrusive solution with sensors the persons are not even aware of when performing their daily activities at home. When seated, the invented device and methods detect the subject's heartbeat, compute heart rate, estimate the level of oxygen in the blood and blood pressure, acquire several nonstandard electrocardiographic (ECG) leads and the temperature of fingers, follow respiration by the breathing curve, detect movement and the subject's posture and position of their body and body parts.

In this paper, we are going to concentrate on measurements made by a combined sensory device in the chair armrests. It captures photoplethysmographic (PPG) [2] and ECG signals in parallel, which enables assessment of various cardiovascular parameters. In Section II, the construction of the device and signal-processing methods are revealed. Section III describes experiments and results, while Section IV discusses the results and concludes the paper.

II. DEVICES AND METHODS

A common posture when sitting in a chair is with hands leaned on the armrests. If the armrests have rounded and rather thin endings, sitting persons usually grasp it and bend their fingers under it. We took advantage of this fact and developed a supporting U-shaped dry electrode that is fixed under the chair armrest. A PPG sensor is built in the bottom of the electrode carrier. Its shape guarantees a stable and good contact between the subject's skin and the electrode conductive surface, which improves the quality of ECG measurements, while at the same time minimizes the influence of ambient illumination on PPG sensing (Fig. 2). Both the left and right dry electrodes incorporate LEDs with two different wavelengths, three optical sensors (photodiodes) to acquire light reflected from person's fingers, and two thermistors to measure finger temperature.

Figure 1. Modified gaming chair: smart, unobtrusive sensors are built in the backrest, seat, and armrests.

Figure 2. A detail: dry U-shaped ECG electrode having incorporated components for photoplethysmographic and temperature measurements.

Experiments described in this paper analyze parallel recordings of the PPG and ECG signals only. All acquired measurements are synchronized by a microcontroller build in the chair and wirelessly transmitted to a host computer. Analysis algorithms extract relevant features and estimate parameters of functional health. These are rendered by user interface that is adapted to run on different mobile devices.

A. *Detection of heartbeats*

We have developed a robust approach to the heartbeat detection from the PPG signals and published it in [3]. The same idea was used with the chair PPG sensors.

In every trial, 6 PPG signals were acquired in each hand with 3 photodiodes at two different light wavelengths. The latter was not important in our experiments, except that the heartbeat detections were more robust due to the observation of the same phenomenon by 3 sensors in parallel and at two different light sources. To avoid high-frequency disturbances, we applied low-pass filtering to every signal and averaged the six acquired in each hand to a single recording. Denote the obtained PPG signal by $\mathbf{s} = [s(0), \ldots, s(N\text{-}1)]$. The PPG values decrease fast to the PPG foot at every systole and increase slower to the PPG peak during diastolic intervals. Fast systolic changes can be detected by differentiating the signals:

$$d(n-1) = s(n) - s(n-1);\ n = 1,\ldots,N-1 \qquad (1)$$

Differences in $d(n)$ from Eq. (1) introduce high-frequency components due to the differentiation. There is also a higher contribution of the frequencies caused by regular heartbeats. These depend on heart rate, which is physiologically limited to a range between 40 and 210 beats per minute, i.e. 0.67 and 3.5 Hz. Heartbeat detection is then looking for these frequencies, and to do it most efficiently all other frequencies should be suppressed. We completed the job by constructing an adequate band-pass filter based on wavelets. A suitable range of scales α was applied, so that a set of wavelets, $\psi_\alpha(n)$, was built. A sum of the wavelets over the selected scales produced desired filtering band-pass characteristic:

$$\psi(n) = \sum_{\alpha \in S} \psi_\alpha(n);\ n = 0,\ldots,N_\alpha \qquad (2)$$

where ψ_α designates the chosen mother wavelet at scale α, ψ the constructed band-pass filter, S a set of preselected scales, and N_α the length of the wavelet at highest scale chosen.

The filter from Eq. (2), when applied to the PPG gradient **d,** maximizes the response at different frequencies dependent on heart rates. Optimum heartbeat detections, therefore, reside on the following signal:

$$p(n) = \sum_{k=0}^{N-1} d(k)\psi(n-k);\ n = 0,\ldots,N-1 \qquad (3)$$

Local minima in $p(n)$ represent heartbeats as detected in the PPG signal.

B. *Blood pressure and pulse transit times*

Following the example from [7], we computed pulse transit times (PTT) as a time interval between an ECG R-wave peak and the nearest subsequent heartbeat location in the analyzed PPG. PTTs are inversely proportional to the instantaneous blood pressure [4, 5, 6]. However, there are several properties of the individual physiology that influence the relationship between blood pressure and PTT. Referring to [8], PTTs change with transmural blood pressure, i.e. the pressure that appears as a difference of inner distending blood pressure and outer contact pressure across the vessel walls. If the transmural pressure becomes negative, which happens when the contact forces from outside overwhelm the pressure inside blood vessels, PTT doesn't change with it anymore and is constant for all negative transmural pressure values. On the contrary, when transmural pressure is positive, PTT decreases with the increasing pressure.

The functional relationship of PTT versus transmural pressure depends mainly on three individual physiology properties. It depends on maximum blood vessels compliance, maximum arterial volume, and the arterial volume at zero transmural pressure. These characteristics are not easily determined with independent measurements. Most common estimation uses PPG measurements at known transmural pressures.

Due to the fact that we didn't calibrate individual vascular parameters prior to our smart-chair experiments, we were not in the position to estimate blood pressure from the measured PTTs. Therefore, we used obtained PTTs in comparisons with referential blood pressure measurements and observed differences in the individual characteristics of tested subjects.

III. Experiments and Results

Our experiments were organized in two-phase measurement sessions where subjects relaxed on our smart chair by grasping properly the U-shaped sensors under the both armrests by fingers. In the first phase, referential blood pressure was verified before a 3-minute smart-chair signal acquisition at rest. In the second phase, subjects were asked to keep squatting for one minute and, immediately after that, referential blood pressure was verified again. A second 3-minute interval of measurements followed with subjects relaxing on the chair.

Their ECGs were recorded between the two hands and, in parallel, also the PPG signals (Fig. 3).

All the PPG and ECG signals were synchronized by the microcontroller firmware and resampled to a common sampling frequency of 200 Hz.

The R-wave peaks were detected in the ECGs by well-known Pan-Tompkins algorithm [9]. All R-wave detections were checked and, if necessary, corrected manually. Heartbeats were also detected from the PPG signals by our own algorithm based on continuous wavelet transform [3] as described in Subsection II.A. Low-pass filtering was applied to the signals with a cut-off frequency at 10 Hz prior to the differentiation. After it, a band-pass filtering was based on a combination of Morlet wavelets at scales 60, 72, 84, 96, 108, 120, 132, 144, 156, and 168.

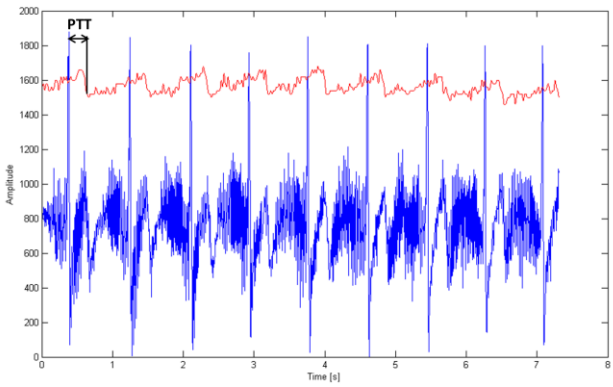


Figure 3. A combination of ECG (blue) and PPG measurements (red) that enable an estimation of blood pressure based on PTTs. The principle of PTT estimation is marked as a distance between the ECG R peak and a subsequent PPG foot

Eleven healthy male subjects participated in experiments. Their mean age is 28.9±6.6 years, their height 180.6±5.1 cm, and their weight 90.2±16.9 kg. All gave their informed consent before the experimental trials. Referential data were also taken for blood pressure before every trial. A Critikon Dinamap Pro 300 was applied for a sphygmomanometric measurement of systolic and diastolic blood pressure in every participant.

After detecting ECG R waves and PPG-based heartbeats, we were ready to compare referential blood pressure in tested subjects with computed PTTs. We measured time distances from the referential R peaks to the PPG foots as detected by our heartbeat search. PTT values were estimated from 10 consecutive heartbeats at the beginning of every trial. PPG signals from the left hand were taken into account. Inverse values of PTTs are expected to be proportional to mean blood pressure [6, 12].

We modeled the relationship between PTTs and subject's blood pressure by linear model in four situations: for systolic and diastolic referential pressures at rest and after the exercise. Obtained formulas transform PTTs to blood pressure estimates in the least-square-error sense. Table I summarizes estimated PTTs for all 11 participants at rest and after exercise.

TABLE I. Systolic and Diastolic Blood Pressures Measured Before Every Trial and Mean PTTs Estimated at the Beginning of Every Trial

Subject	At rest		After exercise	
	***Systolic/diastolic BP* [mmHg]**	***Mean PTTs* [*ms*]**	***Systolic/diastolic BP* [mmHg]**	***Mean PTTs* [*ms*]**
1	129/70	236.4	137/85	195.9
2	153/87	197.3	143/67	160.9
3	121/73	233.7	119/67	269,1
4	110/69	250.2	120/74	233.2
5	129/79	285.0	163/79	168.2
6	126/78	244.1	134/84	213.4

Sub-ject	At rest		After exercise	
	Systolic/dias-tolic BP [mmHg]	*Mean PTTs* [*ms*]	*Systolic/dias-tolic BP* [mmHg]	*Mean PTTs* [*ms*]
7	135/72	268.6	212/78	201.4
8	161/94	190.0	240/108	135.6
9	134/65	259.8	153/72	172.3
10	140/84	191.6	140/77	174.8
11	121/65	270.0	150/72	225.2

Best-fit analytical models, i.e. formulas that transform PTTs into corresponding blood pressure, are depicted in Table II along with mean-square errors. These were computed between the referential blood pressures and their estimates according to obtained analytical models. Agreement between the references and modeled estimates is not accurate enough for systolic, but is acceptable for diastolic pressures.

TABLE II. ANALYTICAL RELATIONSHIPS BETWEEN REFERENTIAL BLOOD PRESSURE MEASUREMENTS AND ESTIMATES OF PTTS QUANTIFIED BY MEAN SQUARE ERRORS (MSE)

BP type	At rest		After exercise	
	BP [*mmHg*] *vs. PTT* [*ms*] *relationship*	*MSE* [*mmHg*]	*BP* [*mmHg*] *vs. PTT* [*ms*] *relationship*	*MSE* [*mmHg*]
Sys-tolic	$BP = -0.1352 \cdot PTT + 167.87$	4.75	$BP = -0.6103 \cdot PTT + 275.63$	8.43
Dia-stolic	$BP = -0.2041 \cdot PTT + 124.54$	1.84	$BP = -0.1589 \cdot PTT + 109.50$	2.80

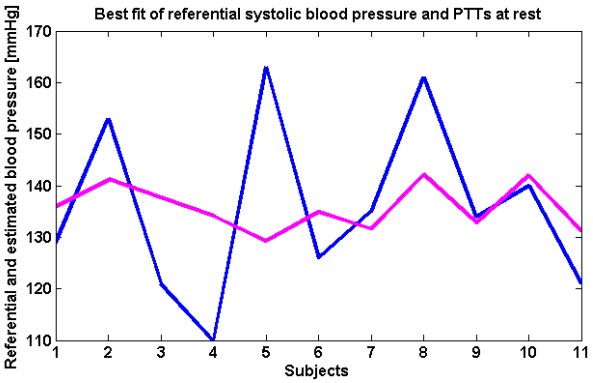


Figure 4. Referential systolic blood pressure (in blue) and its estimate according to PTTs (in magenta) measured at rest.

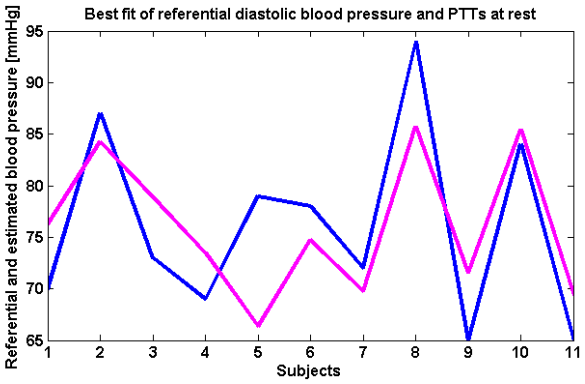


Figure 5. Referential diastolic blood pressure (in blue) and its estimate according to PTTs (in magenta) measured at rest.

Figs. 4, 5, 6, and 7 show the referential blood pressures as measured before the experimental trials by the Critikon Dinamap device in blue and, in parallel, their analytical estimates computed with the corresponding formulas from Table II.

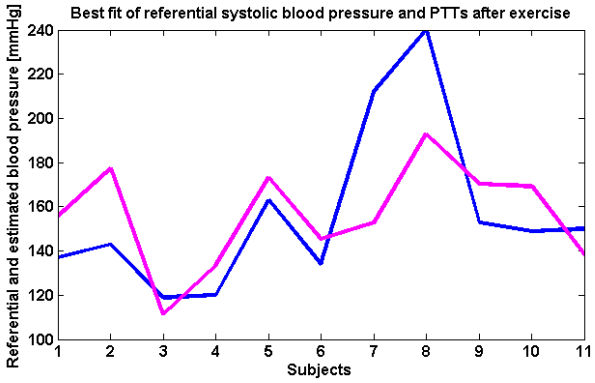


Figure 6. Referential systolic blood pressure (in blue) and its estimate according to PTTs (in magenta) measured after exercise.

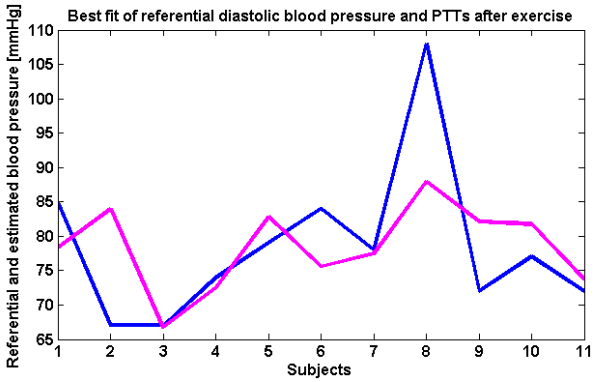


Figure 7. Referential diastolic blood pressure (in blue) and its estimate according to PTTs (in magenta) measured after exercise.

IV. DISCUSSION AND CONCLUSIONS

Contact forces on fingers change transmural pressure and also PTTs [10]. At the same time, PPGs from oppressed fingers show weaker dynamics and lower changes between the systolic and diastolic levels. This aggravates correct heartbeat detections. In spite of being instructed not to do so, participants in our tests may have pressed by their fingers against the armrest sensors on our smart chair. We couldn't collect any information on that, so we expect that transmural pressures in our experiments equal distending blood pressures in arteries. Anyway, errors neglected this way in the PTT estimation do not exceed 18 ms at 0.5 N of external contact force, which equals about 80 mmHg when 2 cm^2 of finger's surface are oppressed [8]. We are quite positive that much lower forces were unconsciously exerted in our experiments, which alludes that errors of this type remained in PTTs blow 5%.

Relationship between referential blood pressures and estimated PTTs unveiled inferior fitting of systolic blood pressures, both at rest and after exercise. We suspect the reason might be individual differences in vascular parameters that, as we explained in Subsection II.B,

influence this relationship. Therefore we tried to exclude these differences by modelling changes of blood pressures from the rest to post-exercise condition with changes of PTTs for each subject. We expected this might circumvent interpersonal variations. Figs. 8 and 9 depict differences between post-exercise and rest blood pressures and PTT-based estimates for systolic and diastolic measurements, respectively.

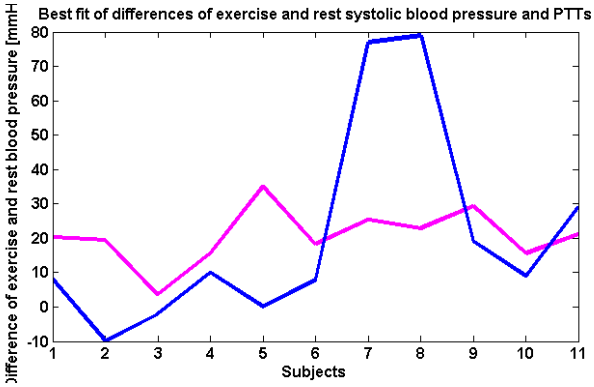


Figure 8. Changes of systolic blood pressures (in blue) from resting to post-exercise period are aligned with changes of estimated blood pressures as modelled by PTTs (in magenta.)

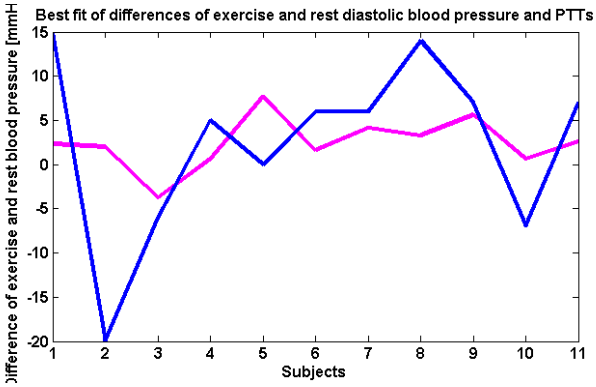


Figure 9. Changes of diastolic blood pressures (in blue) from resting to post-exercise period are aligned with changes of estimated blood pressures as modelled by PTTs (in magenta.)

Figs. 8 and 9 indicate different physiological reactions of individuals. While the predictions of blood pressures based on the PTTs in subjects 1, 3, 4, 9, 10, and 11 follow the direction of changes in referential blood pressures, the same doesn't hold for subjects 2, 5, 6, 7, and 8. Two groups with different response in their blood pressure and estimated PTTs are formed.

Therefore, we repeated modelling systolic condition for the first group of subjects 1, 3, 4, 9, 10, and 11, due to inferior results obtained by analytical estimates of blood pressures based on the PTTs both at rest and after exercise. New models decreased MSE to 3.45 mmHg at rest, and to 3.47 mmHg after exercise.

We are not able to explain different observations in the two groups of participants. They do not differ significantly either by their body indexes or referential blood pressure changes or heart rate changes when comparing the outcomes at rest and after exercise. Additional research is planned with more participants for better statistical assessment. Deeper insights will be induced into the vascular parameters and also more attention paid to possible stress and muscle tonus of participant during experimental measurements.

ACKNOWLEDGMENT

Authors acknowledge a financial support of the Slovenian Ministry of Education, Science, and Sport within the Programme Funding Scheme P2-0041.

REFERENCES

1. The Biomedical Engineering Competence Centre, "Connected Home Devices in Support of Independent and Healthy Living," http://storm.uni-mb.si/kcbme/index-en.php.
2. B. D. Zislin and A. V. Chistyakov, "The History of Oximetry," Biomedical Engineering, vol. 40, no. 1, pp. 53-56, 2006.
3. C. Pirš, B. Cigale and D. Zazula, "A Feasibility Study of Heartbeat Detections from Photoplethysmograms With Fingers Compressed," Proc. of the 5th WSEAS International Conference on Sensors and Signals, Sliema, Malta, 2012, pp. 47-52.
4. J. Spigulis, R. Erts and U. Rubins, "Micro-circulation of skin blood: optical monitoring by advanced photoplethysmography techniques," Proc. of the SPIE, vol. 5119, Bioeng. and Bioinspir-ed Syst., Maspalomas, Canary Islands, Spain, 2003, pp. 219-225.
5. Y. Choi, Q. Zhang and S. Ko, "Noninvasive cuffless blood pressure estimation using pulse transit time and Hilbert–Huang transform," Computers and Electrical Engineering, vol. 39, pp.103-111, 2003.
6. D. B. McCombine, A. T. Reisner and H. H. Asada, "Motion Based Adaptive Calibration of Pulse Transit Time Measurements to Arterial Blood Pressure for an Autonomous, Wearable Blood Pressure Monitor," Proc. of the 30th Annual International IEEE EMBS Conference, Vancouver, 2008, pp. 989-992.
7. A. Arza, J. Lazaro, E. Gil, P. Laguna, J. Aguilo and R. Bailon, "Pulse Transit Time and Pulse Width as Potential Measure for Estimating Beat-to-Beat Systolic and Diastolic Blood Pressure," Computing in Cardiology, vol. 40, pp. 887-890, 2013.
8. X.-F. Teng and T. Zhang, "Theoretical Study on the Effect of Sensor Contact Force on Pulse Transit Time," IEEE Trans. Biomed. Eng., vol. 54, no. 8, pp. 1490-1498, 2007.
9. J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Trans. on Biomed. Eng., vol. 3, pp. 230-236, 1985.
10. S. Rhee, "Design and Analysis of Artifact-Resistive Finger Photoplethysmographic Sensors for Vital Sign Monitoring," PhD Thesis, Massachusetts Institute of Technology, Boston, 2000.
11. D. Zazula, C. Pirš, K. Benkič, D. Đonlagić, B. Cigale, "Short-Term Photoplethysmography Embedded in Household Appliances," 6th European Conference of the International Federation for Medical and Biological Engineering, MBEC 2014, Dubrovnik, Croatia. IFMBE proceedings, vol. 45, pp. 922-925, Sept. 2014.
12. H. Gesche, D. Grosskurth, G. Küchler, A. Patzak, "Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method," Eur J Appl Physiol, DOI 10.1007/s00421-011-1983-3, May 2011.

Towards noninvasive bioimpedance sensor design based on wide bandwidth ring resonator

M. Pavlin* and F. Novak**
*Hyb d.o.o., Levičnikova 34, SI-8310 Šentjernej, Slovenia
marko.pavlin@inmedica.si
**Jozef Stefan Institute, Jamova 39, Ljubljana, Slovenia
**Jozef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia
franc.novak@ijs.si

Abstract - Miniature and noninvasive sensors play a major role in healthcare, by monitoring vital health parameters in real time. Furthermore, energy efficiency is an important issue in human-centric embedded systems. In this paper we describe a new approach for a contact-less bioimpedance measurement. The basic idea stems from the permittivity measurement of electronic circuit substrate. A common procedure in this case is to design and fabricate a substrate with a given resonator structure where the dielectric constant can be calculated from the measured resonant frequency and known resonator geometry and substrate thickness. Similar principle can be employed for bioimpedance measurement. The resonator can be designed such that the resonant frequency lies within the responsive range of the target tissue where the resonant frequency changes are highest following physiological processes. The changed value of the frequency reflects the dielectric properties of the target object. The measured frequency thus implicitly reflects biological and physiological processes within the target object. Some preliminary experiments are described in this paper.

I. INTRODUCTION

Human body or part of the body can be represented as distributed volume conductor. Spatial impedance variability can be measured to detect different physiological phenomena. Measuring chest impedance is one of the most popular techniques to detect respiration activity. The respiration and ECG are usually measured simultaneously. Impedance measurements can be implemented with several non-invasive measurement principles. Bioimpedance sensors can be integrated within other sensors (e.g. ECG electrodes can be used to measure impedance of the thorax) or as separate sensors for more specific measurements.

Impedance pneumography is commonly-used technique to measure respiration rate. The operation principle is based on changing impedance of the thorax. The thorax impedance analysis can be described using proper impedance model. The thorax represents volume conductor with constant and varying component of volume impedance [1].

Therapeutic decisions in intensive care unit are based on measurement results from different sensors indicating the status of a patient's vital signs. One of the important parameters is cardiac output. The cardiac output is calculated from stroke volume measurements. The stroke volume is difference between end diastolic volume and end systolic volume. Stroke volume can be measured with invasive and non-invasive methods. One of the non-invasive methods to determine cardiac output is thoracic electrical bioimpedance measurement [2].

Proper impedance model is required to investigate the impedance variations within thorax due to different physiological processes. There are many different models for whole human body or just thoracic volume with accurate tissue properties [3-6].

Standalone ultra-low power wireless bioimpedance sensors are rare or non-existent. However, the combination of biopotential+bioimpedance sensors is more common. They can be seen as wireless ECG with integrated respiratory functions. Integrated circuits supporting such combination were developed and are commercially available from different vendors.

Another option for investigation in this area can be oriented on sensor, which is more intrinsic and can operate with lower power. Microwaves are used to measure dielectric properties of the materials. The permeability perturbations are known in biological tissue caused by (patho-) physiological processes [7]. Ring resonators are simple and efficient components for measuring permeability. They can be realized in micro-strip technology. A tuned micro-strip ring-resonator oscillator was reported to have a wide tuning bandwidth of 30%[8]. Both building blocks can be used for low power wireless bioimpedance sensor. A GHz range oscillator can reach power consumption in sub-0,100mA range [9]. What is even more motivating, the concept of microwave impedance measurement of biological cells have been reported reaching very high sensitivity capabilities [10].

Another system can detect and display the detected heartbeat and respiration in real time was implemented as Doppler radar system measuring periodic movement [11]. The authors reported using short wavelengths to improve sensitivity and reduce the power consumption.

More than one vital sign can be measured with micorwave non-invasive sensor. A multi-purpose, low-cost and non-invasive vital signs sensor that was capable of measuring heart-rate, respiration rate, changes in the lung water content, and the heartbeat waveform was presented based on the measurement of reflection coefficient [12].

II. BIOIMPEDANCE MEASURING CIRCUIT

Low power contact-less bioimpedance circuit for measuring changes in dielectric properties of a observed tissue was constructed. The circuit measures resonant frequency deviations in microwave resonator. Similar method was introduced by several authors for the measurement of dielectric constant of some materials [13]. The resonant frequency depends on permeability of material above microstrip planar resonator. The block diagram of the circuit is shown in Fig. 1.

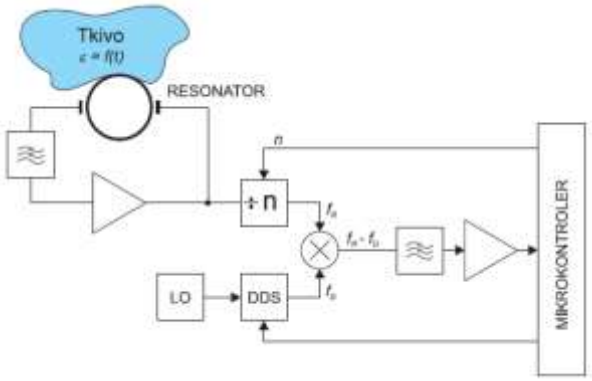


Fig. 1 - Large tuning range resonant sensor

As an example, electrical impedance measurements of the lungs at lower frequencies were reported. Significant dependence of impedance on air content was measured in the frequency range 9.6 kHz–1.2 MHz [7]. This phenomenon indicates the relationship between impedance and air content of lung tissue. The tissue impedance in microwave region depends mostly on salt content. The conductivity is proportional to molar concentration of NaCl [15].

The above idea has been developed and preliminary evaluated by numerical simulations in a commercial finite-element code (Comsol Multiphysics). The model of the resonator structure was studied under different boundary conditions and loads (i.e., different human body tissues) according to the expected application in practice. Obtained results were within the range of the human bio-impedance frequency response. Simulation results were compared with measurements on physical prototype of the resonator. Resonator measurements were performed with a vector analyzer. The key parameter for oscillator design (scatter parameter S21) was measured between 2MHz and 6GHz. The simulation and measurements results showed good match at resonant frequencies. This confirmed the feasibility of the proposed approach.

A. Resonator

The ring resonators are excellent components for many sensing applications [16]. Some typical measurement applications are optimisation of microstrip substrate thickness [17], discontinuity and equivalent circuit parameters [18], dielectric and dispersion characterization [19, 20], or broadband materials characterisation [21]. The ring resonators are also standard structures to measure dielectric properties of printed wiring board materials [22]. Resonator operation in free air was simulated with agilent ADS.

The basic ring resonator circuit is shown in Fig. 2. It consists of two feed lines with coupling gaps and the resonator ring. The ring is powered via feed lines. Separation between lines and ring is provided by coupling gaps. The gaps represent loose coupling and should be optimal to provide proper oscillation within the ring structure. The structure supports waves that are integral multiple of the guided wavelength of the circumference

$$n\lambda_g = \pi D, \quad n = 1,2,3, \ldots$$

where n is mode number and D is ring mean diameter. When the total length of the resonator, resonant frequency f and the resonance order n are known, the effective permitivity (i.e., the permittivity of the tissue, which changes with physiological state) can be calculated.

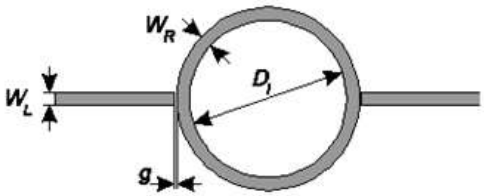


Fig. 2 - Microstrip ring resonator geometry

B. Oscillator

The ring circuit is a resonator and it can be utilized to stabilize an oscillator. In our case, the active part of the oscillator is NPN Silicon RF Transistor BFP520. Amplifier input is connected to the ring resonator via low pass filter. This filter rejects all higher harmonics above maximum operating frequency of the resonator.

The output of the amplifier feeds the resonator and provides output signal. The output of the oscillator will be connected to the next stage, the frequency divider in the first prototype.

The designed sensor will be tested first component-by-component and finally as whole module. Long term stability of the measured frequency should be tested without testing subject.

C. Test results

Preliminary measurements show detectable deviation in resonator frequency during changes in permittivity of loading media around resonator.

III. CONCLUSION

The presented sensor is based on proven microwave measuring principle. The operating frequency is based on the tissue dielectric properties, which are stable within the frequency range and mainly dependent only on the tissue composition and current physiological state. Since the tissue composition is not changing significantly, the principal sensitivity of the sensor is on the tissue physiological changes. The sensor operation was tested measuring the breathing, which was easy detectable and

phenomenon. However, the sensor utilization is not limited to simple breathing detection. Many sources describe correlation between dielectric properties and different health related states [1-7]. The presented measurement methods elsewhere were mostly with direct electrical contact between skin and sensor electrodes. Our contactless, non-invasive method can be used in many different bio-impedance measurement applications.

IV. ACKNOWLEDGMENT

This work was partially supported by the Slovenian Research Agency under contract P2-0098.

V. REFERENCES

[1] Frerichs, I. (2000), 'Electrical impedance tomography (EIT) in applications related to lung and ventilation: a review of experimental and clinical activities', *Physiological Measurement* **21**(2), R1.

[2] Patterson, R. (1989), 'Fundamentals of impedance cardiography', *Engineering in Medicine and Biology Magazine, IEEE* **8**(1), 35-38.

[3] Akhand, M.; Trakic, A.; Terril, P.; Liu, F.; Wilson, S. & Crozier, S. (2009), Optimal tissue types in the thoracic electrical impedance model for thoracic electrical bioimpedance (TEB) studies, *in* 'Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE', pp. 3913-3916.

[4] Patterson, R. P. & Zhang, J. (2003), 'Evaluation of an EIT reconstruction algorithm using finite difference human thorax models as phantoms', *Physiological Measurement* **24**(2), 467.

[5] Hartinger, A. E.; Gagnon, H. & Guardo, R. (2006), 'A method for modelling and optimizing an electrical impedance tomography system', *Physiological Measurement* **27**(5), S51.

[6] Coulombe, N.; Gagnon, H.; Marquis, F.; Skrobik, Y. & Guardo, R. (2005), 'A parametric model of the relationship between EIT and total lung volume', *Physiological Measurement* **26**(4), 401.

[7] Nopp, P.; Harris, N.; Zhao, T.-X. & Brown, B. (1997), 'Model for the dielectric properties of human lung tissue against frequency and air content', *Medical and Biological Engineering and Computing* **35**(6), 695-702.

[8] Gardner, P.; Paul, D. & Tan, K. (1994), 'Microwave voltage tuned microstrip ring resonator oscillator', *Electronics Letters* **30**(21), 1770-1771.

[9] Chee, Y. H.; Niknejad, A. & Rabaey, J. (2005), A sub-100 mu;W 1.9-GHz CMOS oscillator using FBAR resonator, *in* 'Radio Frequency integrated Circuits (RFIC) Symposium, 2005. Digest of Papers. 2005 IEEE', pp. 123-126.

[10] Gabriel, S., R. W. Lau, and Camelia Gabriel. "The dielectric properties of biological tissues: II. Measurements in the frequency range 10 Hz to 20 GHz."*Physics in medicine and biology* 41.11 (1996): 2251.

[11] Lin, J. & Li, C. (2007), Wireless Non-Contact Detection of Heartbeat and Respiration Using Low-Power Microwave Radar Sensor, *in* 'Microwave Conference, 2007. APMC 2007. Asia-Pacific', pp. 1-4.

[12] Celik, N.; Gagarin, R.; Huang, G.; Iskander, M. & Berg, B. (2013), 'Microwave Stethoscope: Development and Benchmarking of a Vital Signs Sensor Using Computer Controlled Phantoms and Human Studies', *Biomedical Engineering, IEEE Transactions on* **PP**(99), 1-1.

[13] Nafaa, M. S.; Bashir, M. K. & kassem, A. S. A. (2009), Measurement of Dielectric Constant of Some Materials Using Planar Technology, *in* 'Second International Conference on Computer and Electrical Engineering'.

[14] Stuchly, M. & Stuchly, S. (1980), 'Dielectric Properties of Biological Substances- Tabulated', *J. Microwave Power* **15**(1), 19--26.

[15] Munrot, K.; Reeves, J. W.; Birch, M. J. & Collier, R. (2001), Standing wave measurements to determine the dielectric properties of liquids and gels for use in microwave phantoms, *in* 'Antennas and Propagation, 2001. Eleventh International Conference on (IEE Conf. Publ. No. 480)', pp. 111-114 vol.1.

[16] Troughton, P. (1969), 'Measurement technique in microstrip', *Electron. Lett.* **5**, 25–26.

[17] K. Chang, F. Hsu, J. B. & Nakano, K. (1984), 'Find optimum substrate thickness for millimeter-wave GaAs MMICs', *Microwaves & RF* **27**, 123-128.

[18] Hoefer, W. & Chattopadhyay, A. (1975), Measurement of the Equivalent Circuit Parameters of Discontinuities in a Resonant Microstrip Ring, *in* 'Microwave Symposium, 1975 IEEE-MTT-S International', pp. 103-105.

[19] Wolff, I. & Knoppik, N. (1971), 'Microstrip ring resonator and dispersion measurement on microstrip lines', *Electronics Letters* **7**(26), 779-781.

[20] Hopkins, R. & Free, C. (2008), 'Ultra-wideband slotline dispersion measurements using ring resonator', *Electronics Letters* **44**(21), 1262-1264.

[21] Hopkins, R. & Free, C. (2008), 'Equivalent circuit for the microstrip ring resonator suitable for broadband materials characterisation', *Microwaves, Antennas Propagation, IET* **2**(1), 66-73.

[22] Heinola, J.-M. & Tolsa, K. (2006), 'Dielectric characterization of printed wiring board materials using ring resonator techniques: a comparison of calculation models', *Dielectrics and Electrical Insulation, IEEE Transactions on* **13**(4), 717-726.

Exploring bacterial biofilms with NAMD, a highly scalable parallel code for complex molecular simulations

Drasko Tomic

Hewlett-Packard, Zagreb, Croatia, drasko.tomic@hp.com

Abstract –Bacterial biofilms are highly complex structures. We are more and more recognizing that bacterial biofilms are predominant forms of the bacterial existence against the planktonic one. Under certain circumstances, bacteria starts to build biofilm and forms 3D structure, so called extracellular matrix. Hulled within this matrix, bacteria becomes more prone to host defense mechanisms and most antibiotics, thus expressing considerably higher virulence and antibiotic resistance than its planktonic form. Because of their importance, numerous researchers investigated bacterial biofilms in the last decades, using numerous methods, like electron microscopy, mass spectroscopy and nuclear magnetic resonance. However, neither of these methods is able to reveal an exact structure of extracellular matrix. Exploring dynamics of extracellular matrix is even more complex, and out of the reach for known analysis methods. For these reasons, there is a need for more effective method, and this could be computer driven simulation. In order to check if it could be a method of choice, we estimated the computational resources needed to simulate the bacterial biofilm. We found that possibility of performing this simulation in the reasonable time on fastest supercomputers today does not exists, and will not be available until at least 2028. For this reason, we explored possibilities of running NAMD based bacterial biofilms simulations on Cloud, and landed with the same conclusion. Besides, we found that for both approaches NAMD has to extend its scalability from about current 500.000 cores to many millions of cores in the future.

Key words: bacterial biofilm, extracellular matrix, NAMD, high performance computing, cloud computing

I. INTRODUCTION

In the last decades, our belief regarding the organization of the bacterial swarms in aqueous environments change dramatically. In opposite to the former view of bacteria as predominantly planktonic, we now believe that biofilms are predominant forms in aqueous environments, and that quantity of bacteria in biofilms is for several orders of magnitude higher than planktonic. Because of its very high complexity, we still do not know the exact structure of biofilm. Besides, bacterial biofilms are highly diverse in nature, which means that the biofilm for two different bacterial species differs in its architecture and ingredients. The complexity of bacterial biofilm is even higher, as its structure is dependent of environmental parameters, like the concentration of oxygen, temperature and ph [1]. In addition, bacterial biofilms are not static, but dynamical organization of many bacteria, embedded in the so-called extracellular matrix. Dynamical means that it changes over the time, started from the initial attachment of the bacteria to the surface, their growth and production of adhesins, which enable bacteria for the initial attach to the surface and interaction with other bacteria in its proximity [2]. After the initial phase of attachment and growth, this structure is going to mature, from time to time opening its surface to dissipate bacteria in its proximity. Under certain circumstances, biofilm may disperse as well [3]. What we well understand today is that it is 3D matrix like structure, and its main ingredients are polysaccharides, proteins, DNA and RNA [4]. Usually, we call this structure extracellular matrix. In Fig. 1, there is a typical seeing of the biofilm surface, under electronic microscope:

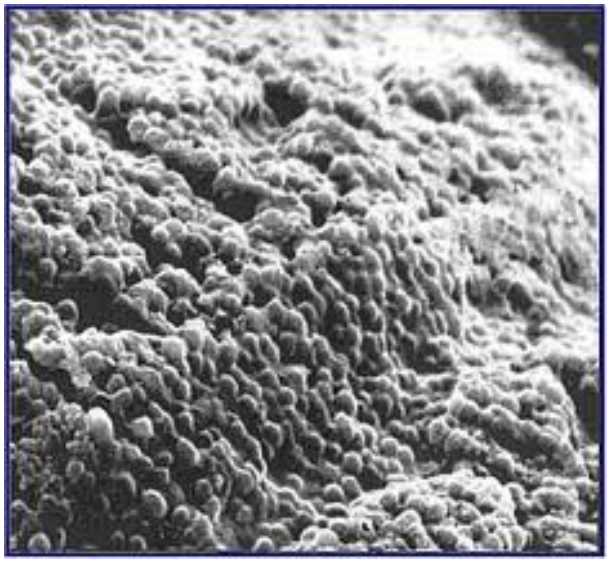


Fig.1: Bacterial biofilm image under electronic microscope

However, many bacterial species developed additional surviving strategies, like ability to penetrate deep in the tissues, autonomous movement, and production of toxins able to damage not only host, but also other competitive bacteria. Today, we know several dozens of elements that contribute to the initial bacterial adhesion to the surface, biofilm growth and its maturation [5]. To make things even more complex, different bacterial species can form common biofilm. Fig.2: schematically describes life cycle of bacterial biofilm:

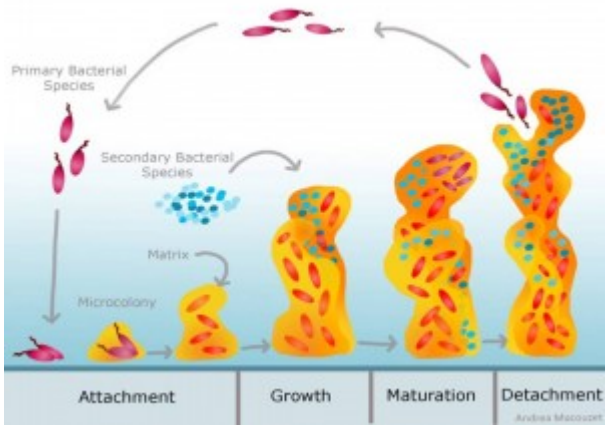


Fig.2: The life cycle of the bacterial biofilm

After initial attachment of bacteria to the surface, growth phase follows. In this phase, other species eventually join the primary one and contribute to the biofilm growth and formation. Maturation phase is probably the most complex phase, as it involves synchronization and message exchange between bacteria and involves many protein types and other molecules. Biofilm detachment is the last stage, where some of bacteria are spreading out of biofilm, and seeking other patches for the attachment, and eventually biofilm growth. Nevertheless, it is important to stress out that biofilm growth needs not to start even if the attachment was successful. It is not so clear what triggers biofilm growth, but it is supposed that the main factors are environmental, like temperature, ph and osmotic properties of the surface. Besides, nowadays is more than ever clear that bacterial biofilm enables bacteria to survive in very harsh environments, hide it from the host defense system, and even enable it to escape from antibiotics. For all mentioned before, it is evident that bacterial biofilm analysis is a very hard problem, and can be even harder in case of the multi-species bacterial infection. Number of methods for analyzing bacterial biofilms were proposed and developed in the past, like solid-state NMR spectroscopy, electron microscopy, biochemistry and their combination [6]. However, neither of these methods is capable to approach bacterial biofilm in all aspects; some of them can give us some insight in the architecture and content of extracellular matrix, but are not capable to clarify or visualize biofilm development stages. As it also became more and more evident that bacterial biofilm contribute to bacterial virulence, it is crucial to develop agents that are able to destroy bacterial biofilms, or at least help antimicrobial agents through the biofilm surface and kill bacteria hidden in it [7]. Experimental methods for finding such agents have led to the development of several drugs that can do this, however these experiments are expensive and last for days. Besides, there is rarely, if ever, a possibility to observe in-vivo effect of agents to the surfaces colonized with bacterial biofilm, thus danger of side effects is extremely high. In addition, as already pointed out by Max Planck, each experiment induced some uncertainty into the observed system. For smaller system is the greater probability that observer will induce uncertainty in it. Indeed, bacterial biofilms are not so great structures; they can envelop whole organs, e.g. bladder and lungs, but they are at most several hundred micrometers high.

Therefore, the only method known today which would be able to reveal such complex structure like bacterial biofilm and its interaction with antimicrobial agents and surfaces is the computer-based simulation. Popping up as a first choice is the computer simulation of large biomolecules. Among many microbiological simulation programs, NAMD is probably one the best known [8]. Its popularity is due to several factors: excellent scaling which enable to use NAM over thousands of processor cores, plus large database of proteins, polysaccharides and genomes, which is open for anyone to use it. That all together makes NAMD as one of the most preferred choice for the potential use of the bacterial biofilm analysis. However, the complexity of the biofilms and number of atoms to be simulated can eventually prevent using NAMD for this even on the available supercomputer resources. For this reason, we will try in the next section to estimate computational resources NAMD needs to perform bacterial biofilm analysis. In addition, we will try to estimate how long it could take to perform the simulation. Lastly, we will propose some ways to improve it.

II. CALCULATING MICROBES

The mass of an average sized virus is about 1.5 femtogram (1.5*10^-15 gram) and viruses are made up primarily of proteins. Some coat themselves in sugars or other substances but these tend to be the bigger ones. Protein is roughly 70% carbon, 12% nitrogen, 10% oxygen, and 7% hydrogen by mass. We can then break down the mass of the virus into parts and get as following for carbon, nitrogen, oxygen and hydrogen parts of it::

C:1.05*10^-14g
N:1.95*10^-16g
O:1.50*10^-16g
H:1.05*10^-16g

Consulting a periodic table we find that carbon consists of 12g/mol=12g/6*10^23 atoms. It follows:

1.05*10^-14g / (12 g / 6*10^23 atoms) = 1.05*10^-14g * 6*10^23 atoms / 12 g.

Units cancel and we can multiply the terms in the numerator:

6.3*10^9 atoms / 12 = 5.25*10^8 atoms = 525,000,000 atoms of carbon.

We can now perform the same calculation for other elements:

N: 8,360,000 atoms
O: 5,630,000 atoms
H: 63,000,000 atoms

gives us approximately 600,000,000 (six hundred million) atoms. For the simulation of smaller problems, e.g. the simulation of the capsid of HIV virus, we need supercomputer at Illinois University that delivers about 1PLOFS computational power. Next figure shows simulated capsid as the brawn nucleus in the middle of HIV virus:

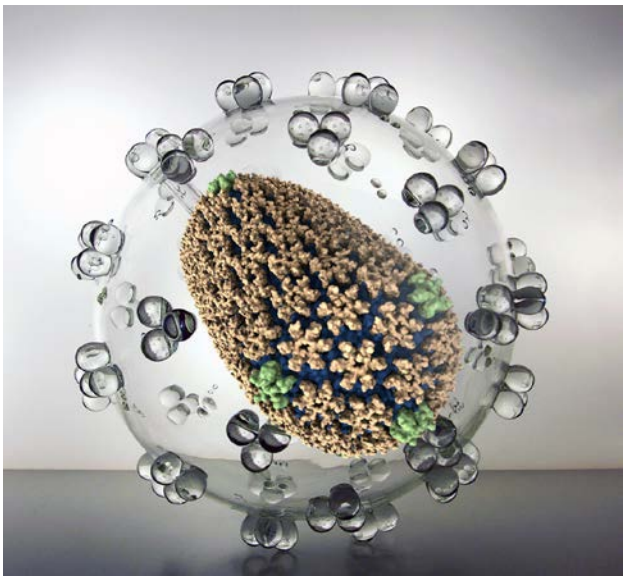


Fig.1: HIV and its capsid simulated at University of Illinois with NAMD

Now we will try to estimate number of atoms in bacteria. Our first assumption is that atoms in a bacterium are packed as closely as physically possible, which is idealization since a bacterium is mostly liquid. Besides, there are many different types of atoms in a bacterium, but the main components are hydrogen and carbon. Hydrogen has a radius of .025nm and carbon has .07nm. We know the volume of the bacterium, so one has to find out the volume of the atoms. We will estimate them as spheres with V = 4/3pi*r^3. We get an atomic volume of 4.18 x 10^-30 cubic meters. Dividing the volume of the bacterium by this we get a value of about 2.4x10^11 atoms. We know that a bacterium weights about 1

picogram (1 x 10^-12 grams). We further know its chemical composition is mostly water, and we'll estimate that the average atom has an atomic weight of 6 (which means 6 times the atomic mass unit, which is 1.6 x 10^-27). We arrive at the result that there are about 10^14 atoms in a bacterium. This means about 10^6 more atoms then in virus.

III. SIMULATING BIOFILM ON ZETTAFLOP COMPUTER

In order to find out how the need for computational resources increases with the number of atoms, we performed several standard NAMD benchmarks on Openstack based Cloud infrastructure with 4 VCPUs and 16GB RAM. Standard benchmarks were ApoA1, ATPase and STMV, with 92.224, 327.506 and 1.066.628 atoms, respectively. Elapsed times for these benchmarks on 4 VCPUs and 16GB RAM were 246, 719 and 2483 seconds, respectively. This experiment showed us that computational resources needed scale approximately linearly with number of atoms. Under this assumption, we estimated about 10 PFLOPS computational power would be needed to simulate the whole virus, and about 10^6 Pflop/s needed to perform bacterial simulation. If we take the look at top500.org, the list of most powerful computers, the fastest supercomputer today is able to deliver about 34 Pflop/s [9], enough to perform virus simulation, but far behind resources we would need to perform bacterial and biofilm simulations. Besides, by taking view at the performance development of the fastest supercomputers over time, we can see that performance increases exponentially, and we can expect that Exaflop supercomputer delivering 10^3 Pflop/s supercomputer will be available in 2018, while Zettaflop supercomputer delivering 10^6 Pflop/s will be available somewhere in 2028 year. Fig.4 shown how the performance of 500 fastest supercomputers developed in the last two decades:

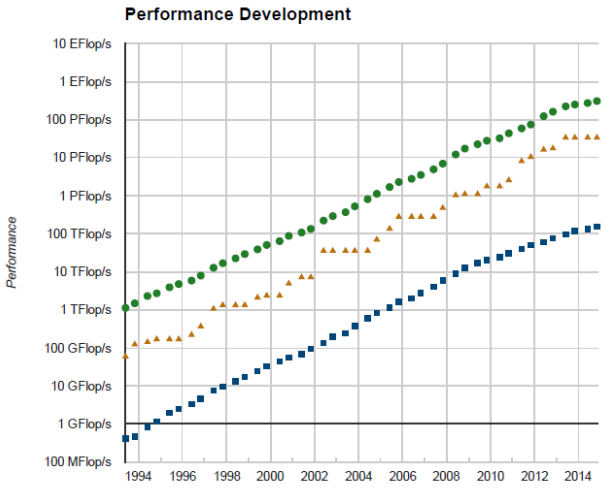


Fig.4: Performance development over the time

From other experiments, it is apparent that NAMD is not able to scale on more than several dozens of vCPUs on Cloud [10]. On the other side, NAMD proved to scale linearly over hundred thousands of CPU cores on standard supercomputers. For this reason, Cloud needs to undergone important changes if we want NAMD and other HPC applications to scale on it as on monolithic supercomputers. Besides providing support for ultra-fast optical interconnect, some other technologies like containers, SR-IOV and cluster instance placement groups must become a standard for HPC in the Cloud. We already have seen movements in these directions from Amazon, Google and other Cloud providers. Regarding installed computational power, none of these companies provides the number. Nevertheless, according to some estimations it could be up to 40 Pflop/s for some of them, and this back in 2012. It is evident that there is at least the same computing power available on Cloud as on top500.org, where the total computing power in 2012 was 100 Pflop/s. Besides, today's trend in HPC is that HPC Cloud is growing with about the same factor as the standard HPC. For this reason, we do not believe HPC Cloud growth will overtake standard HPC in the near future. Anyway, the cost of running HPC applications on Cloud is considerably lower than on Supercomputers, so it could be an additional boost to further investigate and develop Cloud based bacterial biofilm simulations. The next important factor to consider is NAMD scalability. Envisioned Zettaflop computer, either monolithic or Cloud based, will have many millions of cores. Currently, NAMD is scaling up to 500000 cores.

For it, we need improved load-balancing strategy, able to increase NAMD scalability by far beyond the current limit.

III. CONCLUSION AND THE FUTURE WORK

Until today and as of our knowledge, there is a little or no work at all in the area of computational simulation of bacterial biofilms. Microbiologists explore bacterial biofilms experimentally, and there are several reasons for this. Firstly, bacterial biofilms are highly complex structures. Moreover, although in last several decades our knowledge and understanding of bacterial biofilm grew enormously, we still do not have a clear picture of its architecture, their interaction with various environmental habits, and influence of the environmental parameters to its formation, maturation, and eventually dispersion. Thus, setting up the computational model of a certain bacterial biofilm is very challenging and complex task. Another important reason for not performing computer simulations of bacterial biofilms is the lack of computational resource available, even for the relatively simple virus simulations. Simulation of bacterial biofilm would impose the usage of costly computational resources, thus limited to the very small number of scientist. However, with the advent of cloud computing, the availability of high-performance computing resources with affordable cost has enormously increased, providing us a chance to simulate bacterial biofilms on the Cloud. Besides, before being able to simulate biofilms on Cloud, we need to extend the current scalability of NAMD from several dozens to several thousand Cloud instances by implementing new Cloud technologies like ultra-fast optical interconnect, SR-IOV, containers and cluster instance placement groups. In addition, for running bacterial biofilm simulations on either Zettaflop monolithic or Zettaflop Cloud based supercomputer, NAMD load-balancing scheme needs improvement: to enable scaling from todays 500.000 to many millions of cores in the future. Anyway, for both Zettaflop monolithic and Cloud supercomputer based bacterial biofilm simulations, we have to wait at least until the end of the next decade.

References

[1] C.Beloin, A.Roux, J.M.Ghigo, “Escherichia Coli Biofilms”, Curr Top Microbiol Immunol. 2008, 322: 249-289.

[2] T.R.Garrett, M.Bhakoo, Z. Zhang, “Bacterial adhesion and biofilms on surfaces”, Progress in Natural Science 18 (2008) 1049-1056

[3] O.Rendueles, J.B.Kaplan, J.M.Ghigo, “Antibiofilm polysaccharides”, J. Environmental Microbiology, 1462-2920, (2012)

[4] B.Vu, M.Chen, R.J.Crawford, E.P.Ivanova, “Bacterial Extracellular Polysaccharides Involved in Biofilm Formation, Molecules 2009, 14, 2535-2554

[5] I.W. Sutherland, “Biofilm exopolysaccharides: a strong and sticky framework”, Microbiology (2001), 147, 3-9

[6] O.A.McCrathe, X.Zhou, C.Reichhardt, L.Cegelski, “Sum of the Parts: Composition and Architecture of the Bacterial Extracellular Matrix, J.Mol.Biol. 2013 November 15, 425(22)

[7] W. Costerton, R.Veeh, M.Shirtliff, M.Pasmore, C.Post, G.Ehrlich, “The application of biofilm science to the study and control of chronic bacterial infections ”, J. Clin. Invest. 112:1466-1477 (2003)

[8] http://www.ks.uiuc.edu/Research/namd/

[9] www.top500.org

[10] D. Milojicic, Keynote, IEEE 8th World Congress on Services, Honolulu, Hawaii, June 26, 2012.

Information System of the Institute for Blood Transfusion of Montenegro

Doc.dr Adis Balota

Faculty of Information Technology, Podgorica, Montenegro; adis.balota@unimediteran.net

Abstract - **This paper presents a developed Information System of the Institute for Blood Transfusion of Montenegro, which functions within the Integrated Information System of Healthcare System of Montenegro. Introduction of IT support of transfusion activity provides better availability and accuracy of data on blood donors in the healthcare system, as well as tracking the blood supply in a centralized manner on the level of the state of Montenegro.**

By introducing the bar-codes on drawn blood units, Information System of the Institute for Blood Transfusion has ensured compliance with the basic transfusion principle of tracking blood from 'the donor's vein' 'to' 'the recipient's vein'. Information system will increase the possibility for control and analysis of work, rational use of resources, and it will also modernize the health statistical database from the field of blood transfusion. Administrative work of healthcare workers, which refers to multiple input of data, will be reduced, as well as the possibility of errors.

Interoperability of business processes has been fully achieved by using shared IT infrastructure of the healthcare system, and the final goal of this approach is to improve the quality, safety, efficiency, financial accountability and clinical use of blood components in Montenegro.

I. INTRODUCTION

For the purposes of the Institute for Blood Transfusion of Montenegro, information system (ISIBT) has been developed, which presents a part of Integral Health Information System in Montenegro (IHISM) [1].

ISIBT contains **business information system** (staff operations, financial operations, payroll, etc.), as well as service records and **medical information system**. Medical Information System of the Institute for Blood Transfusion includes IT support for the processes of collecting, testing, processing, storage and issuing of blood and blood components, while complying with the principle of traceability as the basic one in blood transfusion system. This allows tracking and recording blood units and blood components from donor to recipient, as well as records of side effects. By development of medical part of information system, Information System of the Institute for Blood Transfusion has been completed, which is fully integrated with the Integral Health Information System (IHIS) as a whole.

Integration involves two directions: integration on the level of Institute for Blood Transfusion and integration of ISIBT with ISs of hospitals which use blood in the treatment and with the Institute of Public Health in the part of legally defined reporting.

A. Business goals expected from ISIBT are:

1. Ensure monitoring and control of costs and quality of work as well as rationalization of activities of healthcare providers,
2. Provide a platform for exchanging information electronically,
3. Increase the possibility of systemic quality control of services provided,
4. Increase the number of available useful information required to make timely decisions,
5. Rationalization of the system of collecting health information at all levels and creation of the strategy of adaptation and integration of all health information systems,
6. Revising and modernizing the methodology and scope of collecting health statistical data.

B. Special goals of ISIBT are:

1. Greater availability and accuracy of medical information about the patient,
2. Ensure traceability - tracking of blood units from blood donor to blood recipient,
3. Better management of blood transfusion system at the national level,
4. Increase the quality of transfusion service work,
5. Automation of the reporting system to the relevant institutions of the system and international organizations,
6. Eliminate the need of multiple input of data (name, surname, ID number, ...)
7. Reduce errors in work through introduction of automation during the input,
8. Integrated collection of information necessary for efficient management,
9. Reduction of administrative work,
10. Reduce the consumption of resources and time by reducing unnecessary printing of paper.

II. SYSTEM DESIGN

Integrated Health Information System of Montenegro integrates in its composition nine different subjects which use the same shared IT resources [1]. The shared resources involve great number of codebooks, databases, server and network communication equipment. Table 1. presents current number of locations, number of computers and number of application users per individual developed information system.

TABLE 1. DEVELOPED INFORMATION SYSTEMS

Information system	Number of connected locations	Number of computers in the system	Number of application users
Health Insurance Fund	21	212	194
Pharmaceutical institution Montefarm	45	125	314
Health centres	60	1092	2145
Institute for Emergency Medical Service	19	30	291
General hospitals	7	572	1386
Dentistry	170	190	341
Institute of Public Health	1	150	154
Agency for Medicines and Medical Devices of Montenegro	1	30	36
Institute for Blood Transfusion of Montenegro	9	50	62
	333	2451	4923

Table 1. shows that the Institute for Blood Transfusion is organized through 9 dislocated units.
The implemented information system of the Institute for Blood Transfusion, among other things, provides:

1. Unbreakable chain of business processes, activities and information from the moment of drawing blood from donor to blood recipient,
2. Full integration with the existing IS in the Institute (Business Information System),
3. Full integration with institutions in the healthcare system of Montenegro,
4. Integration with institutions outside the public healthcare system (private hospitals and clinics),
5. Integration with the Health Insurance Fund,
6. Integration with the Institute of Public Health,
7. Integration with the Ministry of Health,
8. All the records will be based on valid and structured medical documents or a version thereof, where necessary,
9. Management of blood supplies and blood components on the basis of information on supplies in all organizational units of the Institute, as well as of the current needs (requisitions) by all parameters.

Figure 1. shows a schematic representation of the concept of 'unbreakable chain' of business processes, activities and information from the moment of drawing blood from donor through the stages of processing, storage, and distribution to the final stage, i.e. the blood recipient. This chain, i.e. all the stages that are listed represent the so-called 'Traceability'.

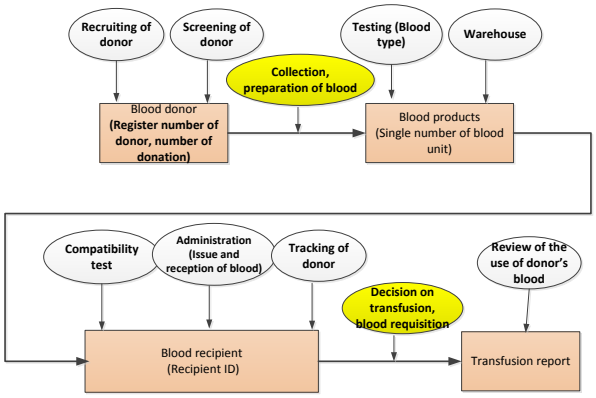


Figure 1. Unbreakable chain - traceability

Figure 2. shows a schematic representation of unbreakable chain in terms of the tracking process.

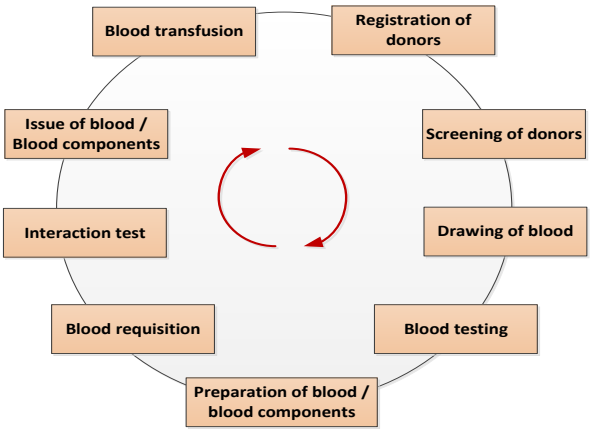


Figure 2. Unbreakable chain – process of tracking blood units from donor to recipient

Figure 3. presents logical connectivity scheme of IHISM

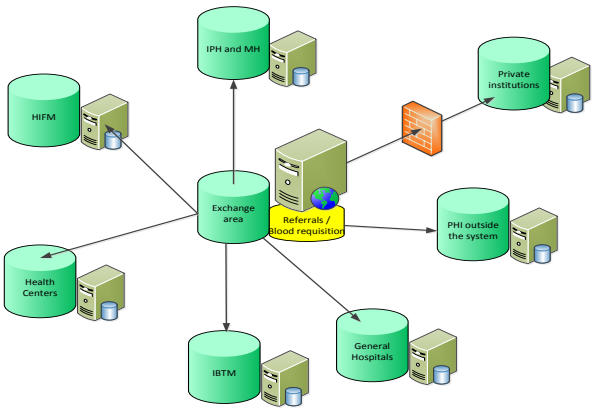


Figure 3. Logical connectivity scheme of IHISM

IT support of business processes in the Institute provides the establishment of basic registries and continuous functioning of main business functions:

1. Forming of Registries of blood,
2. Preparation of blood, processing and testing,
3. Storage and distribution of blood and blood products,
4. Preparation of blood for testing, interaction and issuing of blood,
5. Record of laboratory services,
6. Registry of blood donors.

Information System in the Institute has a function to track all the processes related to blood transfusion through all its stages, i.e. tracking of blood by the principle from the donor's vein to the recipient's vein. Special advantage of this concept represents a considerable increased level of achieved safety of blood, as well as the best way of achieving the principle of traceability of blood. Critical events may occur in the process, for example, side effects on blood transfusion, incompatibility of donor and recipient, disease transmission, etc. That is why it is necessary to provide accurate and reliable information on tracking of blood and blood products.

Implementation of information system of the Institute has allowed creation of a unique database of blood donors, i.e. Registry of blood donors.

Electronic record of blood donors (Registry of blood donors), represents a basic view of the organization of database and complete CORE application of IBT IS.

Figure 4. presents the design solution where the donor is traced from the moment of their occurrence in the system, all their activities related to blood donation, testing, processing, to final transfusion, or disposal due to date expiration [4].

According to the design solution, the data of donor are accessed with use of ID (identification number) or a unique registration number, and the administrative part of data and any other specific data become available if the user is authorized to receive such information.

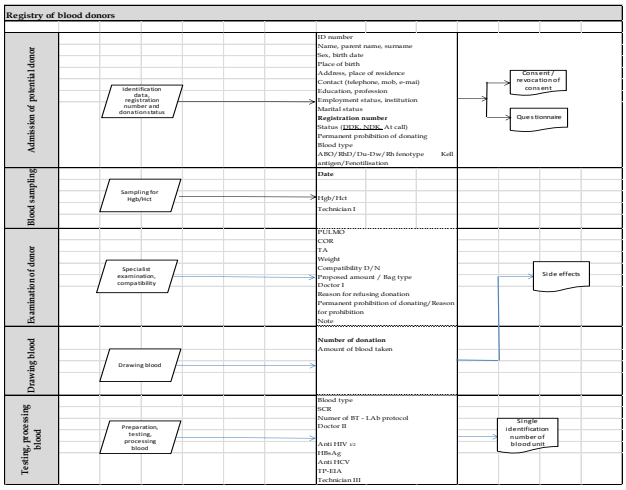


Figure 4. Registry of blood donors

Thus organized unique and centralized blood donors registry provides multiple functionalities, as follows:

- management by recruiting new donors, either from mobile collection points, in hospitals or centrally in the Institute for Blood Transfusion,
- research of the suitability of existing blood donors, whether it be the voluntary blood donors or blood donors at call.

Figure 5. is a block diagram of Registry of blood donors.

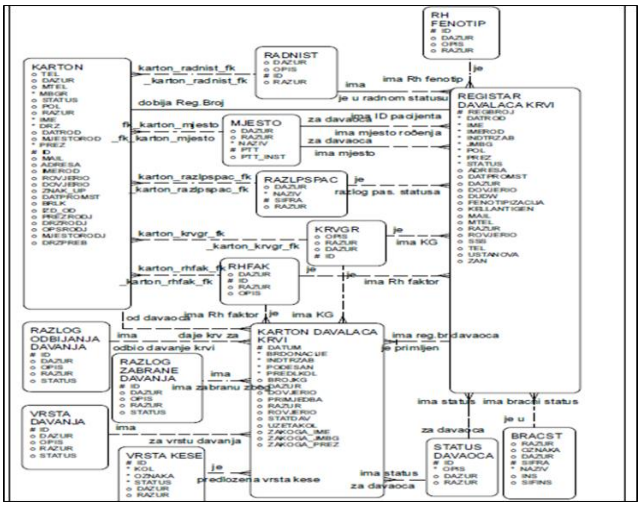


Figure 5.ERD – Registry of blood donors

I. NECESSARY PRECONDITIONS FOR THE FUNCTIONING OF ISIBT

For the implementation of software solution it was necessary to provide a direct exchange of data with the registries and codebook systems that are already formed within the Integrated Health Information System of Montenegro, and which are maintained by the Health Insurance Fund of Montenegro. Integration is achieved with the following registries: Registry of the insured, Registry of healthcare workers, Registry of healthcare institutions, Registry of drugs by ATC classification, Registry of diagnoses [2]. All locations of the Institute for Blood Transfusion are connected to server infrastructure Central office through MPLS transport links. Technical solution has also been developed for the mobile teams work, which collect blood units on the field in the form of organized actions, and which exchange information using 3G signal. Also, all departments of Hospital and Clinical Centre of Montenegro are associated with the system of the Institute, and the user application is provided, through which the data are exchanged for hospital patients with the Institute for Blood Transfusion.

This operating mode, i.e. online mode, guarantees high flexibility and availability of IS, as well as security of data and information. Also, information are taken at the place of their origin and they can be seen only by the participants in the system that have been granted the right. Hence, once entered data is no more re-entered, which significantly shortens the time of administration, as well as the possibility of errors in the administration.

II. FUNCTIONAL REQUIREMENTS

It is essential that ISIBT covers all the fields that the Institute involves, according to legal and other acts. On the basis of that, functional units of the Institute for Blood Transfusion IS are clearly identified. In order that the information system achieves its projected function, possibility of the following basic records is provided:

1. Record of blood collection,
2. Record of preparation of blood components,
3. Record of blood processing and testing (immunohaematological blood tests and tests for infection diseases) – for hospitalized patients and outpatients,
4. Record of blood storage and blood components,
5. Record of blood requisition and blood components by health institutions,
6. Record of blood requisition for the patient, tests, interactions and issues of blood by organizational units,
7. Record of issue of blood and blood components per patient (recipient).

Functions under the numbers 1-4 are in the organization of the Institute and responsibility of the Service for collecting blood, blood component preparation, testing, processing of blood units and distribution, except from number 3 for outpatients, which is the responsibility of the Service for laboratory and other diagnostics.
Functions under the numbers 5-7 are the responsibility of the Service for clinical transfusion and therapeutic services per units of the Institute which are at 9 locations.

The main function in the future functioning of the Institute is to unify blood banks in Podgorica, as the central warehouse, with the existence of auxiliary warehouses by locations (in Podgorica and within seven General hospitals and specialized hospitals in Risan).
Figure 6. shows the procedure of blood distribution and blood components in the Institute.

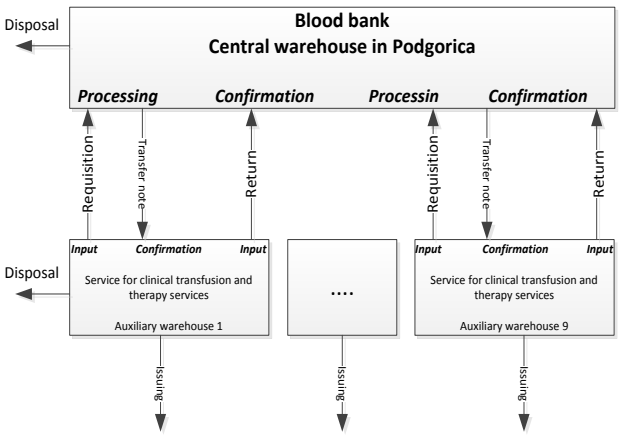


Figure 6. Procedure of blood distribution and supporting documents

In order to achieve the proposed model of establishing a centralized blood bank, it was necessary to provide a single identification of each blood unit (or component) to the level of the whole Institute, which was done.
All the necessary procedures of requisition and distribution of blood have been defined and adopted, as well as the internal documents that will accompany the requisition and distribution of blood, as well as their record within the Institute IS.

Figure 7 shows a schematic representation of the procedure for collecting blood from a potential donor.

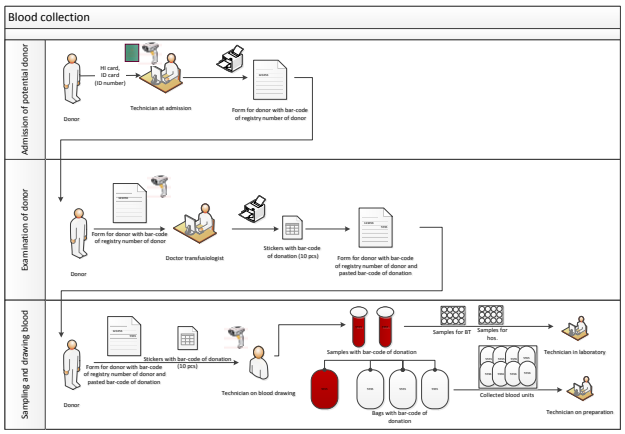


Figure 7. Procedure for collecting blood

III. CONCLUSION

By introducing bar-codes on drawn blood units, the implemented Information System of the Institute for Blood Transfusion provides compliance with the basic transfusion principle of tracking blood from donor's vein to recipient's vein. The information system covers all business functions of the Institute, and it also provides full integration with other parts of the health system. The system provides higher availability and accuracy of medical data about the patient. Also, registry of blood donors is developed, and a single electronic record for blood donors is implemented. There is also provided easy management of the storage of blood supplies and blood components, which is now centralized. Integrated collection of information for the management is achieved, which is necessary for efficient management.
The administrative work of healthcare workers, which refers to multiple input of data, will be reduced, and the possibility of errors. Modern transfusion system requires efficient and adequate information system in order to further improve the quality and efficiency of transfusion service in Montenegro. Developed information system of the Institute is one of the main parts of the Integrated Health Information System of Montenegro. The main task of the system is to improve the quality, security, efficiency, financial accountability and clinical use of blood components in Montenegro.

BIBLIOGRAPHY

[1] Prof.dr Ramo Šendelj, Adis Balota MA, Ivana Ognjanović MA: "Analyzes of the Information System of Primary Health Care of Montenegro"' *2nd International Conference on Software Engineering , Management & Applications *(* ICSEMA 2011*) Switzerland on 15th ~ 16th August, 2011 Zurich

[2] Balota Adis ma, Reković Denis, Glomazić Zoran, Šendelj Ramo: "Electronic prescriptions in primary health care in Montenegro" Yu Info 2009. Kopaonik, 08/03- 11/03, 2009, Serbia

[3] Adis Balota MA, Zoran Glomazić, Denis Reković: "Electronic prescription in primary health care in Montenegro" Congress of clinical pharmacology and pharmacotherapy of Southeast Europe in health care, Novi Sad, Serbia, June 01-03, 2008.

[4]http://www.who.int/healthinfo/statistics/toolkit_hss/EN_PDF_Toolkit_HSS_InformationSystems.pdf

[5] IDEF1 Information Modelling, Mayer R., http://www.staratel.com/iso/IDEF/IDEF/IDEF1/IDEF1-1.pdf

[6] http://agimo.gov.au/government/australian_government_arhitecture

CTI

International Conference on

TELECOMMUNICATIONS & INFORMATION

Steering Committee

Chairs: Branko Mikac, University of Zagreb, Croatia
Tihana Galinac Grbac, University of Rijeka, Croatia
Stjepan Golubić, Zagreb, Croatia
Dragan Jevtić, University of Zagreb, Croatia
Mladen Sokele, Croatian Telecom, Zagreb, Croatia
Renato Filjar, Ericsson Nikola Tesla, Zagreb, Croatia

Members: Slaviša Aleksić, Vienna University of Technology, Austria
Željka Car, University of Zagreb, Croatia
Saša Dešić, Ericsson Nikola Tesla, Zagreb, Croatia
Vlasta Hudek, Rezultanta, Zagreb, Croatia
Darko Huljenić, Ericsson Nikola Tesla, Zagreb, Croatia
Robert Inkret, Ericsson Nikola Tesla, Zagreb, Croatia
Admela Jukan, Institute of Computer and Network Engineering, Germany
Zdravko Jukić, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia
Marko Lacković, IBM Croatia, Croatia
Erich Leitgeb, Graz University of Technology, Austria
Dražen Lučić, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia
Davor Šarić, Erste Group Card Processor, Zagreb, Croatia
Dina Šimunić, University of Zagreb, Croatia
Velimir Švedek, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia
Antonio Teixeira, University of Aveiro, Portugal

PAPERS

TELECOM MARKET AND REGULATIONS

Effective Strategies for Implementation of Customer Experience Management in MSPs in Bosnia and Herzegovina

N. B. Ćurguz* and D. Ilišević *
* Telekomunikacije RS A.D., Banja Luka, Bosnia and Herzegovina
natasa.banovic-curguz@mtel.ba, dijana.ilisevic@mtel.ba

Abstract – One of global trends and key areas of interest for mobile service providers (MSPs) today is introduction of Customer Experience Management (CEM). MSPs that build high-performing networks, recognize and predict user's needs, provide real-time information and offer more innovative services will secure their market position.

In this paper, we have presented short overview of three different approaches for CEM implementation in telecom area. We have analyzed current state in performance/customer management segment for mobile service providers in Bosnia and Herzegovina in terms of giving the most important CEM recommendations.

I. INTRODUCTION

Many different network and system modernization and transformation programs have been implemented by the region's mobile service providers (MSPs) in order to migrate towards full service operation whilst attempting to balance improvement of service levels and at the same time increase revenues and customer loyalty. While the traditional mobile and broadband services still account for most of MSP's revenue, there is a significant growth in new services such as IPTV, online gaming, Internet search, social networking and digital music. Services are consumed through increasingly various devices and they are also more interactive, in real-time, with online gaming, chat, VoIP, social networks. As quickly as customers can consume data, more data is produced. For the last few years, mobile operators have been talking about developing new sources of revenue since the costs of voice and data services are being constantly pressed through competition. This driver, combined with technological advances, has led to new products, services and markets, as well as innovative approaches for operational efficiency. These can be summarized as machine-to-machine and cloud based services, analytics for marketing and monetization of consumer data, shared networks, operational efficiency, and growth into broadband services to complement existing services and generate new revenue.

Service assurance investments are currently focused on improving customer experience. Mobile suppliers are putting greater importance on solutions supporting customer experience than ever before. The transition to service operations centers is creating a demand for unified service assurance solutions that combine performance, fault and service management functions. Heavy investments into network maintenance and upgrade are not the only way to improve network performance and services. Another option is forcing more effective internal processes and their management in terms of monitoring, controlling and predicting service levels along all touch points in the customers' lifecycle.

The new approach requires a greater focus on the customer point of view. The performance of a service is not only measured based on a set of underlying systems and network elements working correctly, but also based on customers' perception. Customer Experience is defined as the sum of all experiences a customer has with a company, over the duration of their relationship [1]. From customer's point of view the main reasons for telecom services churn are pricing, poor coverage, voice quality, data connection speed and reliability, billing accuracy, ease of recharging/reloading of prepaid credit and service bundling offer. In order to reduce the churn rate, MSPs have to scale down their efforts to the individual customer level. New processes have to monitor Quality of Experience (QoE) based on individual customers. With customer oriented Operation and Maintenance (OAM), greater importance is attached to customer QoE. All OAM resources are centered on improving customer experience. The difference between traditional network oriented and customer oriented OAM is given in Table I, [2].

TABLE I. THE DIFFERENCE BETWEEN CUSTOMER ORIENTED AND TRADITIONAL NETWORK ORIENTED OAM.

	Network Oriented OAM	**Customer Oriented OAM**
Orientation	Equipment or network oriented	End-to-end service or customer oriented
Focus	Network performance and fault	Quality of Experience (QoE)
Assessment	Key Performance Indicator (KPI) dominated	Customer Experience Index (CEI)→Key Quality Indicator (KQI)→ Key Performance Indicator (KPI)
Analysis	Bottom-up approach	Top-down approach

II. CURRENT STATE IN B&H

The global changes in telecom sector are reflected on the market in B&H. The environment is changed around the telecom operator, including the customer needs, the competitors and the regulators. Even in conditions of complex political structure of B&H with weak economic growth and very specific market, official data of Communication Regulator Agency shows that there is a potential for growth. The level of penetration of mobile telephony in B&H at the end of 2013 is 90.79%, [3]. That is much less than in neighbouring countries and EU. The ratio Contract/Prepaid shows that there is a low level of confidence between customers and operators. That is a huge potential for growth of contract users.

But, mobile operators in B&H cannot compete only on the basis of price and network quality. They try to adapt their services to the customers in order to increase loyalty and profit. If mobile operator wants to be the customer relationship leader in the field, also must take into consideration many of the specific contexts and cultural changes on B&H market to achieve this ambitious objective, [4]. Customers increasingly adopt smart phones, tablets and a variety of new mobile broadband services. Also, customer expectations can change according to their position in the relationship cycle.

In MSPs in B&H conventional vertical management structure is used for network operation and maintenance. The focus of service performance has been on the measurement of network operations performance. Mobile operators use traditional network oriented operation and maintenance (OAM) concept. Key performance indicators (KPIs) are based on alarm, performance, and network configuration data; analysis data from active/passive probing and packet capturing, and billing data. A dedicated network, such as a radio access network (RAN), core network (CN) or bearer network, has its own proprietary element management system to ensure network availability and QoS. KPIs for CN, RAN, and IP bearer networks can be very high. This means the network quality is excellent. However, it may be that calls are dropping in an abnormal cell. Customers in this cell may be experiencing very bad service, but this low-quality cell is covered by other cells with high KPIs. Therefore, network KPIs do not necessarily reflect end-to-end QoS and should not be the sole basis for enhancing QoS. High network KPIs do not necessarily mean good QoE.

Mobile operators in B&H mostly implement Customer Relationship Management (CRM) that is in charge of managing the company's interactions with customers. The methodology employed by CRM mainly focuses on sales activities, marketing, and business procedures, not taking into account the customer needs on the service quality. As a result, there is a large gap between the approach of the company and customer's expectations. CEM has the potential to overcome the limitations of CRM to deal with common customers' technical problems and performance issues of services. The differences between CRM and CEM are given in Table II, [5].

TABLE II. THE DIFFERENCES BETWEEN CRM AND CEM.

	CRM	**CEM**
Who	Target user –Sales	Target user -Marketing
What	Manage and automate the sales pipeline, store interactions with customers and prospects.	Understands the customer life cycle and provides personalized experience in all touch points.
Why	Increases satisfaction by providing better services and manages sales pipelines.	Increases loyalty, ROI and reduces churn.
How	Mostly a singular viewpoint.	Gives a general overview of the customer life time-value

CEM approach is designed to focus on procedures and a methodology to satisfy the service quality needs of each end-user. Every time a company and a customer interact, the customer learns something about the company. In CRM approach it is vice versa. In large organizations, CRM and CEM should work side-by-side and exchange customer data while each focuses on a different area of the business, [6].

III. THE TRANSITION FROM NETWORK PERFORMANCE MEASUREMENT TO CEM

A. Proactive approach

Real-time monitoring and preventive assurance on service quality and customer experience helps MSPs to have an improved experience in a short time. As it is given in [7] MSPs should shift the focus from traditional KPIs to end-user experience and from reactive response to proactive and predictable maintenance. This CEM solution focuses on proactive customer care for high-value customers, experience improvement for unsatisfied users and market intelligence analysis for the operator.

High-ARPU individuals, roamers, and enterprise groups can all be classified as high-value customers. The proactive approach in identification of problems prior to the receipt of a customer complaint, helps to time user-operator interactions so that satisfaction and loyalty are maintained. This CEM solution provides various means of identifying potentially unsatisfied users. The methodology is to identify a KQI threshold (from subjective customer survey) and an objective KQI association, and then identify the root causes/factors. Market analysis must integrate subjective inputs, such as customer behaviour and user satisfaction surveys which are the most commonly accepted method for assessing customer loyalty. To avoid extreme answers, [7] suggests operators using customer experience index (CEI) that

involves the tracking of customer behaviour, and can be used for tracing and checking individual survey results.

This CEM solution delivers the following key benefits: transformation from network-centric to service/customer-centric operations, proactive and efficient care for high-value users to maintain or improve their satisfaction and loyalty, creating a lifetime value and revenue, reduced churn through proactive identification and problem resolution for potential unsatisfied users, precise and measurable marketing campaign, reduced revenue loss through proactive fraud management.

B. Holistic approach

Customer Experience Management is defined as a holistic approach taken to monitor, measure and improve all aspects of customer interactions between the operator and the customer, [8]. Five distinct stages in the customer lifecycle are identified: initial experience (relating to product selection and signup), billing and payment process (including the related terms and conditions), account management (the ease of which information can be accessed), service and support (the ability to rapidly resolve any performance or technical issues), loyalty building (ensuring regular personalized rewards for ongoing loyalty). Identifying what is important in customer relationship requires careful consideration of each step of the lifecycle in order to be able to address any planned improvements.

The approach, recommended in [8], includes three fundamental steps to realize customer experience management concept. As it is shown in Table III, it includes identifying what is important in the relationship with the customer, monitoring those interactions and defines a clear transformation lifecycle to improve customer experience.

TABLE III. KEY QUESTIONS OPERATORS NEED TO ANSWER, IN ORDER TO FOCUS ON CE.

	Key question
Identify	What are the most important elements affecting my company's customer experience?
Monitor	How can I measure those interactions through the customer lifecycle?
Transform	How can I implement improvement programs that deliver financial benefits?

MSPs need to identify and determine the relevance of each interaction with customer and correlate satisfaction with importance. Also they need to develop quality indicators that can bridge from customer interactions to specific network and indicators, allowing operators to drill down from the customer plain to the infrastructure level for troubleshooting. The largest barrier is the gap between network element performance indicators (KPI) and actual subscribers' experience. As subscribers rapidly adopt smart phones and new applications, they become more demanding of their network experience. Network operation centers must be enhanced with customer experience assurance centers, enabling the accurate correlation of customer experience with network performance.

C. Extended CEM-oriented approach

Extended CEM-oriented approach that includes the technical and customer care operations departments is proposed in [9]. By involving the entire organization in the CEM strategy and processes, it is possible to manage all interactions along the customer lifecycle. In order to enhance customers' QoE, the technical and customer care departments have to redesign their internal processes according to this new model. In CEM approach all the company's units operate as a single entity in terms of the customer's perception instead that the various departments interact with the customer according to its own needs. Management should concentrate on better performance of communication between the technical department and customer care. Sharing information between departments about the status of the network and the performance of services in real-time is the only way for the MSP to guarantee a coherent and consistent QoE. The flow of information between customer care and the technical department should be bidirectional. Customer care, being directly in touch with customers, can provide the technical department with vital information about the customers' perception of service performance.

In order to effectively introduce CEM in their operations technical department should shift the focus to monitoring end to end service performance and customer experience and put effort into predicting customer experience. Also, it is necessary to introduce new processes and tools to monitor a new set of values that are known as Key Quality Indicators (KQIs). This approach provides a more reliable view of faults and service issues and their impact on service. Put effort into predicting customer experience to move from a traditional reactive model to a predictive one using advanced proactive model. Technical department listen not only to complaints, but to the whole customer base, correlating customer satisfaction and expectations, monitored by the KQI and/or periodical surveys (including structured surveys and unstructured social media sources), to the KPI values of the operating performances.

Accessibility and ease of use, response effectiveness and operations efficiency are fundamental drivers in designing a customer care strategy that maximizes the QoE offered to the customers [9]. Enhancing the accessibility of MSP's services in terms of the CEM means: offering multiple channel access to customer care services (social networks, live web, chat), providing a uniform view of their customer care and simplifying the process to receive support and information. Communicating information about scheduled interventions automatically may improve the performance of service, reducing customer complaints. Effectiveness in customer care services means enhancing of the quality of the information MSPs provide throughout the customer care services. Customer care operators should be able to

access a complete view of the customer during the support process and provide real time information. This requires all the internal information systems to be deeply integrated and interconnected, in order to establish an efficient internal communication flow between customer care and other departments. Monitoring and predicting QoE enables the anticipation of future customer needs.

All three above-mentioned solutions suggest proactive customer-centric model that includes analyzing of customer behaviour and user satisfaction surveys. Traditional indicators should be integrated with additional information. Make an effort into predicting customer experience to move from a traditional reactive model, in which technical operations only track customers' complaints and the company's subsequent response, to a predictive one is essential. In the predictive model, technical operations merge data from business, technical and even social media, in order to predict customer expectations and habits and effectively focus investments in CEM. Customer Care becomes more relevant with advanced devices and services. CEM makes easier than ever for operators to hear "the voice" of their customers and manage the customer experience. Defining and executing a business transformation program across the organization to address key areas of improvement is a critical step. Customer experience improvement program must be driven across the whole organization and involves multiple different processes, systems and domains. In according to [9] this can be achieved by making the technical department a key enabler in the process.

D. CEM Metrics

Customer experience metrics are aggregate calculations derived from survey results. They allow consistently performance monitoring over short and long term periods as well as performance comparing across all channels, [11]. There are different strategies for selecting proper types of CEM metrics to use for a customer experience program. They are evaluated in [12]. With numerous metrics to choose, MSPs can tailor what they would like to measure about their customers' interactions and develop solutions for improvement. It is essential to choose metrics that make the most sense for the company. Very effective metrics are:

- Customer Satisfaction (CSAT) directly measures how satisfied customers are. This classic metric makes it possible to measure and improve how happy customers are.
- Customer Effort Score (CES) is used to measure how much work is required of a customer to solve a problem. CES has its main strength in measuring the smoothness of experience customers have specifically in an interaction touch point. It is easy to link with internal metrics, such as number of calls, number of transfers on a call.
- Net Promoter Score (NPS) identify promoters, those who speak positively about the company, the detractors, those who speak negatively about it, and those who are passive. NPS is able to measure the percentage of customers who would put their own reputation on the line and recommend the company to others, which is the highest form of loyalty.
- General Question Index & Customer Experience Index (GQI&CEI) are used to measure the strength of a customer's experience with a company. GQI&CEI contain multiple questions and can be customized through weighting, allowing some questions to carry greater influence depending on the decisions of the company.
- Voice of Customer (VoC) programs are about continuous consideration of customer expectations, needs and wishes in company's business. In addition to market research surveys, VoC programs include customer data such as feedback, reclaims and also employee input.

In most cases, the best solution is to use a combination of metrics. In a multichannel environment, the challenge is determining customer expectations for each channel and understanding how they may evolve. It is important to listen the voice of every customer, everywhere and all the time. Real-time feedback from a variety of touch-points, including structured surveys and unstructured social media sources, is scaled and integrated with network's data, pushing relevant insights to all levels of business.

IV. B&H PROPOSED CONCEPT

Over the years, MSPs in B&H invested in CRM systems, but in too many cases the expected returns were not clearly visible. CRM alone can not improve loyalty and ARPU, because there is no single CRM application that can deliver all requested functionalities. CRM should be a part of an overall CEM strategy, because CRM and CEM complement one another. Based on analysis of three presented CEM concepts we recommend that mobile operators in B&H apply the combined model of previously mentioned approaches. CEM needs to be established in practice across all functions of the company with the engagement of all staff.

Introduction of CEM is a complex organizational transformation for all MSPs in B&H. The complexity of MSP's organizations is in fragmented systems, multi-vendor environments, multiple service/network layers and regulatory obligations. All of them present numerous challenges to CEM introduction. These challenges can be broadly grouped under the categories of defining, measuring and improving customer experience.

Customer experience management means handling not only the technology and the people involved, but the processes as well. CEM process consists of reactive and proactive part, as it is shown in Figure 1, [10]. MSPs need to identify which customers, products and channels

are the most profitable and determine what changes would increase that success. When compared to the traditional approach of collecting vast amounts of system logs and data points, and then attempting to correlate this vast amount of data, carefully defined measurement and monitoring enables operators to be proactive and take corrective action before customers even notice. This approach makes it possible to identify potential revenue threatening situations as well as network abnormality. MSPs need to control all interaction channels along the customer lifecycle, from service delivery to post-selling and complaint management. The priority must be improving the communication flow between the technical department and the various business functions.

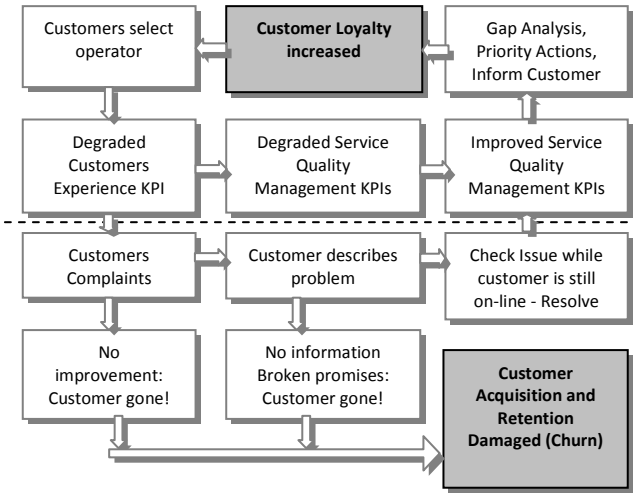


Figure 1. The Customer Experience Management Process, [10].

MSPs in B&H need to start looking at the network from the viewpoint of the quality of services delivered to the end customer. The traditional network/resource perspective, typical for fault and performance management systems, needs to be transformed. This transformation aims to translate network faults and metrics into customer service issues and metrics KQIs. Unlike KPIs, that "speak a technical language", KQIs are designed to "speak the customers' language". This means that MSPs need to measure service quality from the same perspective as the customer. The augmentation of the performance management systems with service quality management should translate technical network metrics KPIs into customer service metrics KQIs. Using the existing performance management systems entails two main benefits. Firstly, it is avoiding unnecessary investments, and secondly, service quality management is built over performance management systems, enabling customer experience issues to be tracked down to the cause of network performance problems, which is essential for end-to-end customer experience management.

A. Challenges and recommendations

In order to ensure high level of service performances MSPs in B&H need to follow this new approach: compose a complete register of services prioritized by economic relevance, compose a register of platforms and network elements underlying any service, define a register of Service Level Agreements (SLA)/KQIs for each service/customer segment, describe the end-to-end process of every service, define a technical fault prioritization matrix according to service impact. The objective of this kind of monitoring is to have real and continuous knowledge of service performance, enabling the immediate resolution of any technical problems that may have a negative impact in real time.

Improving the customer experience is a complex project that requires cooperation among different departments within an organization. MSPs in B&H have to redesign their organization according to these new trends:

- Operating an open, transparent and reputable business.
- Investing in more staff training (continues training of customer-facing employees, aligning each employee with the organization's customer-centric culture).
- Integrating all the workflows in one solution.
- Using innovative knowledge management tools with real-time update.
- Applying customer-centric behaviours and coaching across multiple touch points of interaction (incl. internet, mobile applications, interactive voice response IVR).
- Providing customers with a personalized experience. The customers should always receive the most updated coherent and consistent information from customer care across all touch points.
- Providing prompt responses to customers (inform him/her about taken actions to resolve problem).
- It is important to simplify the process of receiving support and information. Reduce the number of steps that customers have to perform to get the information they need.
- Measurement and evaluation of customer's emotions using most suitable metrics to gain feedback (comments, reactions). Predicting future customers' needs. It is essential to choose metrics that make the most sense for the concrete company.
- Automatization of processes. Reducing the number of escalations to back office through redesigning the front office role. Front office should be able to handle problem troubleshooting, provisioning verification and handling tickets.

In the case of customer experience, a transformation program takes an even more important role. A significant aspect of the transformation is enabling integration with BSS systems. It is important to automatically find, what network problem is causing the issue and start the process of problem resolution. This way, when a customer issues a complaint, he can at least be provided with information that the problem resolution is already in progress. Ideally the process should be proactive. Also KQIs can be used

to define SLAs with customers. When any extra fees for service quality apply, the customer is aware of what he is being offered for the extra money.

V. CONCLUSION

As mobile operators migrate towards full service operation, market competition is becoming fiercer. Today, operations need to be more focused on the business and customer perspective, shifting from network performance measurement to measurement and control the whole end-to-end experience that a customer has when using a service. Implementing CEM in MSPs in B&H is imperative in order to provide superior telecom services and to manage churn. This can be achieved by integrating CEM efforts throughout the organization. It requires making a larger investment in people and processes as well as technology.

REFERENCES

[1] A. Morais and other, "Managing customer experience through service quality monitoring". Future Network and Mobile Summit, 2012, Conference Proceedings.

[2] H. Hongwei and D. Xianjun, "QoE Management: Telecom services and the transition to an experience economy", ZTE Technology, vol.14 No.3 ISSUE 140, pp. 22-25., June 2012.

[3] "Annual report of Communication Regulator Agancy for 2013", CRA B&H, available http://www.rak.ba, December 2014.

[4] D. Ilisevic and N. B. Curguz, "Changing Face of Mobile Service Providers in Bosnia and Herzegovina", Telfor Belgrade, pp. 51-54, November 2014.

[5] http://www.clickz.com/column/2330421/cem-vs-crm-which-platform-is-better, January 2015.

[6] Special report, "Has CEM replaced CRM", available www.cincom.com.au, January 2015.

[7] T. Cheung, "CEM: There's no one "right" solution", Mobile World Congress, Barcelona, February 2013.

[8] S. Barros and J. Beguiristain, "Capitalizing on customer experience", Ericsson, September 2012.

[9] V.Basile, J. Portal, A. Schlautmann and L. Dadhich, " Maximizing quality of experience", Arthur D. Little, 2013., www.adl.com/improving_cex, December 2014.

[10] T. Sutter, " Big data: Is it ready for prime time in CEM?", Black Swan Telecom Journal, August 2013.

[11] Mika Uustalo, "Customer Experience Management in Telecom Operator Business", Master's thesis, Helsinki Metropolia University of Applied Sciences, May 2012.

[12] " Measuring Success: 6 Strategies for Choosing the Right Blend of CEM Metrics", ResponseTek Networks, www.responsetek.com, February 2015.

Regulatory Aspects of Child Online Protection

Suada Hadžović, Dragana Šerval and Saša Kovačević
Communications Regulatory Agency of BiH, Sarajevo, Bosnia and Herzegovina
shadzovic@rak.ba

Abstract - Every day, children around the world are exposed to many online dangers. In this paper we try to give an overview and comparison of current EU and BiH regulatory measures and self-regulatory activities regarding the protection of children against child pornography and other forms of sexual abuse and exploitation through information and communication technologies.

Communications Regulatory Agency of BiH like regulators in EU does not have jurisdiction to regulate Internet content, but it has responsibility for protecting the interests of end users of telecommunications i.e. children as the most vulnerable groups of end users too. Additionally, Communications Regulatory Agency of BiH has been involved in drafting state level action plans in order to improve the system for combating child pornography and other forms of sexual exploitation and abuse through Information and Communication Technologies in BiH.

In the view of recently adopted regulations and the latest action plan, it is expected significantly improving the level of protection of children on the Internet in BiH in the future.

I. INTRODUCTION

According to the Communications Regulatory Agency of BiH (CRA) Annual Report – Survey, licenses for providing Internet services in BiH in 2013, there were 2,188,429 Internet users in Bosnia and Herzegovina in 2013. There was a certain increase in the number of Internet users in 2013 compared to the year 2012 with an index equal to 100.65. In addition, the Communications Regulatory Agency estimated that the rate of utilization numbers of Internet in Bosnia and Herzegovina for 2013 is 56.96% [1]. Having in mind that research has shown that children in front of computer are in three times more danger than those on the street [2], adding the constant increase in the number of Internet users, the number of children who are exposed to risks on the Internet is dramaticaly increasing.

Riscs children face when accessing the Internet are numerous. One of the classification is the following [3]:

1. Content (Child as recipient of information):
   - Advertising, spam, gaming;
   - Violent/hateful content;
   - Pornographic/harmful sexual content.
2. Contact (Child as participant in communications):
   - Child being bullied, harassed or stalked;
   - Tracking/ harvesting personal information of children and misuse of this data, blackmailing;
   - Self-harm, unwelcome persuasion;
   - Child groomed online for sexual abuse offline;
   - Child sold online for sexual abuse both online and offline;
   - Child made the subjects of child abuse images;
   - Child abused through prostitution using the Internet and mobile phones;
   - Adults or young people who engage in 'cyber-sex' with children.
3. Conduct (Child as actor initiating abusive or risky behavior):
   - Creating/ uploading pornographic material;
   - Bullying or harassing another person;
   - Young people sharing personal information online;
   - Young people who place (indecent) images of themselves or other young people (unwanted) online;
   - Children who download sexually abusive images of children;
   - Children meeting 'friends' online.

This article is structured as follows. Following this introduction, Section II provides a brief overview of the recently adopted CRA rules as well as action plans concerning the protection of children on the Internet in which drafting and implementation CRA was involved. Part III performs linking EU Kids Online classification of countries in regard to the protection of children on the Internet and its mapping to BiH. Part IV is devoted to aspects of self-regulation mentioning some of the major initiatives in the EU and gives an overview of the situation in BiH. Part V represents a conclusion.

II. REGULATORY AGENCY INVOLVEMENT IN CHILD ONLINE PROTECTION IN BIH

Communications Regulatory Agency of BiH does not have jurisdiction to regulate the content that is available on the Internet. However, under the protection of end users, the CRA rules are dedicated to protect the most vulnerable groups of users, i.e. children, such as the following [1]:

- Rule 60/2012 on providing service of Internet access („BiH Official Gazette", No 36/12). According to the article 8.2 of this Rule, Internet service provider (ISP) is obliged to disable access to the Internet addresses for which is, by such special regulations or decisions of the competent institutions in BiH, determined to be harmful and unlawful. This specially refers to dissemination of child pornography and similar harmful contents, illegal

online gambling, dissemination of computer viruses or harmful programs, unlawful obtaining of personal data, dangers to public safety, unlawful use of computer programs and applications, as well as other dangers to safe use of the Internet. Therefore, if online content is declared to be harmful by the court or other competent authority, CRA is obliged to order ISPs to withdraw such content from webpages that are stored on its servers.

- Regulation 69/2013 on conditions of providing public communication services and end-users relations („BiH Official Gazette“, No 28/13). According to the article 9.2 of this rule, CRA obliges ISP to include minor protection provisions in their subscriber's contracts and also terms and conditions. Additionally, through Article 36 of this rule, CRA requires ISPs to take all necessary measures to protect minors. In order to protect end users, compliances with the Rule 69/2013 have been made, so, CRA has in past carried out an inspection of general business conditions, subscription contracts and tariffs for more than seventy operators. ISPs suggested that in the forthcoming period, their websites make and put guidelines in order to protect users on the Internet.

Additionaly, Communications Regulatory Agency of BiH has taken an active part in drafting action plans in order to improve the system for combating child pornography and other forms of sexual exploitation and abuse through IT technology in BiH. These action plans are:

- Action Plan for the Improvement of Protection System in the area of Child Pornography and other forms of Sexual Exploitation and Abuse of Children through Information and Communication Technologies in Bosnia and Herzegovina 2010 – 2012. This action plan was adopted by Council of Ministers in December 2009, and covered initiatives on prevention, harmonization of domestic laws, strengthening of institutional capacities, cooperation with ISPs, and providing support to the victims.
- Action Plan for Protection of Children and Prevention of Abuse through Information and Communication Technologies in BIH for the period of 2014-2015, which is an extension of the first 2010-2012 Action Plan, has been adopted by the BIH Council of Ministers on 4 April 2014. This Action Plan includes many activities that relevant authorities in BiH will implement by the end of 2015 in order to ensure efficient mechanisms of protection of children against pornography and sexual exploitation through technology.

The strategic objectives of new Action Plan that include Communications Regulatory Agency are following [4]:

- The strategic objective A2. Amendment to the Criminal law and alignment of subordinate legislation in Bosnia and Herzegovina. Activity A-2.3. Creation of feasibility study and the possibility of implementation of the program called "black list" which would contain information on domains where are facilities of violence against children through information and communication technologies located.
- The strategic Objective B2. Capacity building of ISPs and cooperation with law enforcement agencies and justice institutions. Activity B-4.1. Establishing cooperation between ministries of interior, the juridicial authorities, ISPs and CRA in order to treat the phenomenon of violence against children in ICT. Activity B-4.3. Creating uniform guidelines on the treatment of ISPs in cases of detection pages that show violence against children.

Communications Regulatory Agency of BiH actively participates as a member of the Safer Internet Day (SID) Committee in the preparations for the celebration on a Safer Internet Day in BiH. This body is formed in 2011 with the aim of implementing preventative and promotional activities for the protection of children and youth in their use of information and communication technologies. The committee is headed by the Coordinator for Combating Trafficking in Human Beings and Illegal Migration, and its members include representatives of the CRA, IFS-EMMAUS, Save the Children, OneWorldSEE, and Microsoft BiH. Safer Internet Day celebration is just one in a series of activities that these organizations, individually and jointly, have implemented with the aim of educating children and parents about the risks that the Internet brings, and how to adequately protect themselves from them.

Safer Internet Day 2015 in BiH was celebrated on 10 February, under the slogan „Let's create a better Internet together" through collaboration with partner CSO „Center for Missing and Exploited Children" (CNZD) from Osijek, Croatia, and through the development of special web site (www.boljiinternet2015.com) including applications for online competition for children of primary schools on Internet safety. The competition consisted of an online quiz with questions related to safe use of the internet, which was active on the February 10th. Winners were chosen randomly from participating children who had answered at least 80% of questions correctly, and these winners of prizes were announced on the aforementioned website. During the competition, through interactive maps, data was available regarding the number of children who participated in the quiz and from which cities of Croatia and BiH they came. From BiH, more then 3300 children entered the competition [5].

III. COUNTRY CLASSIFICATION

Four cluster classifications of 25 European countries that were included in the survey based on the activities undertaken, the exposure to risk and harm, and type of internet safety mediation experienced by young people across Europe, was developed by EU Kids Online [6]. A

grouping of countries in 4 distinct clusters was developed as follows:

1. Supported risky explorers (Denmark, Finland, Norway, Sweden and the Netherlands).
2. Semi-supported risky gamers (Bulgaria, Cyprus, Czech Republic, Estonia, Poland and Romania).
3. Protected by restrictions (Belgium, France, Germany, Greece, Ireland, Italy, Portugal, Spain, Turkey and the UK).
4. Unprotected networkers (Austria, Hungary, Lithuania and Slovenia).

There are cross-national differences between countries and it is concluded that some parts of Europe have better support and others receive somewhat less public support for Internet safety. Figure 1 presents ten indicators of policy implementation mentioned in EU Kids Online report [6].

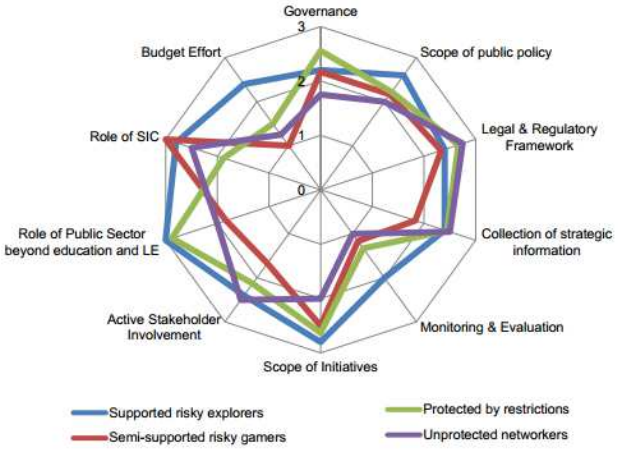


Fig. 1. Clusters comparison

In the 'Supported Risky Explorers' and 'Protected by Restrictions', countries clusters engage more initiatives in promoting safer internet practices than those in 'Semi-supported Risky Gamers' or 'Unprotected networkers'. Countries in the 'Semi-supported risky networkers' and 'Unprotected networkers' clusters shows rather uneven sort of commitments. They invest fewer than other countries, have a lower level of public sector involvement, and with less proof of coordination. By contrast, the countries in the 'Protected by restrictions' cluster have done more to create structures, and to endorse legal and regulatory frameworks around online safety even to the loss of promoting online opportunities. Countries in the 'Supported risky explorers' cluster have higher levels of public sector involvement, opposite to the Safer Internet Centers. There is more evidence of budget investment and evaluation of policy results. There are higher levels of Internet flow and digital skills among both parents and children, and a more proactive approach to arbitration.

It is also the case that countries have different starting points when it comes to internet safety: in some cases, in those countries in 'Protected by Restrictions' there is quite long history of policy involvement on the part of both the public and private sector. In other regions in Europe, developing policies and implementing strategies for Internet safety is more recent. For that reason, those parts of Europe that have had the longest experience of engaging with Internet safety policy and harmonizing the competing demands of promoting young people's opportunities while protecting against the most pressing risks provide a crucial guide for future solutions.

In terms of EU Kids Online classification, BiH belong to Unprotected networkers group of countries. But in the view of recently adopted previously mentioned rules of the Agency and the latest action plan, it is expected to significantly improve the level of protection of children on the Internet in BiH in the future.

IV. SELF – REGULATION

In line with Article 15 of EU Electronic Commerce Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 E-Commerce Directive, ISPs are not under general obligation of monitoring information they transmit or store, nor any requirement to seek proof indicating illegal activity [7].

Despite their limited responsibility under the E-Commerce Directive, Protecting Children in the DigitalWorld report of 13 September 2011 presented a survey of EU member states, which showed that ISPs are increasingly involved in the protection of minors [8].
Then, several self-regulation initiatives have been taken by the industry at European level, for example:

Self-regulatory initiative of the European mobile industry titled The European Framework for Safer Mobile Use by Younger Teenagers and Children puts forward recommendations to ensure that younger teenagers and children can safely access content on their mobile phones. Party European mobile providers, with support from signatory content providers, have proposed an EU-wide common framework to reflect developments and to inspire all relevant stakeholders to support safer mobile use. This framework will be a subject to national implementation by party providers. As of March 2014, there are 91 party companies implementing the Safer Mobile Framework through the roll-out of national voluntary agreements, known as national codes of conduct. These codes of conduct are now in place in 27 EU Member States [9]. These recommendations relate to classification of commercial content; access control mechanisms; education and awareness-raising and fighting illegal content on mobile community products or the Internet.

A number of GSMA Europe members joined the ICT Coalition for the Safer Use of Connected Devices and Online Services by Children and Young People (''the ICT Coalition''). The ICT Coalition for Children Online is made up of 22 companies from across the (ICT) sector [10].

In 2009, leading web firms signed a European agreement i.e. The Safer Social Networking Principles to improve the safety of minors who use social networking sites [11]. But, survey of 25,000 young people in 25 European countries, published by the EUKidsOnline network, shows that some of the social networking sites

that are popular among children in Europe are not signatories to the Safer Social Networking Principles [12]

European Internet Services Providers' Association is the pan-European association of ISPs for the EU and is the largest association of ISPs in the world [13].

If we consider self-regulatory role in BIH, ISPs in BIH like ISPs in EU generally have no specific mandate regarding the protection of minors.

The situation in BiH on this is diversified, i.e. on a case by case basis. Some operators have access to this problem with extreme care and offer a pretty good solution; others are only superficially touched upon the problem and offered some solutions.

According to information collected from BiH Mobile operators and ISPs web sites, ISPs do not organize themselves and do not belong to EU or international associations focused on online children protection. Consequently, they do not have such kind of codes of conduct, but they have satisfactory terms and conditions for subscribers and subscriber's contracts that include minor protection provisions according to Article 36 of CRA Rule 69/2013 [14].

V. CONCLUSION

Online child safety is listed as one of the main priorities in Digital Agenda for Europe. Recommendations state that combinations of policies are required to deliver a better Internet for children. Legal regulation remains an option if necessary, but it should be avoided in favour of more adaptable self-regulatory tools and proper education of children, parents, teachers and others on how to to use the Internet and new technologies safely and responsibly.

Communications Regulatory Agency of BiH, like EU national regulators have no authority over content regulation of the Internet. However, regulations brought aiming to protect consumers apply also to the protection on children online. Beside, Communications Regulatory Agency of BiH was involved in many activities from which most important are those relating to the creation and implementation of action plans which aim to protect childrean online.

In EU, ISPs are increasingly involved in the protection of minors, despite their limited liability and responsibility under the E-Commerce-Directive. Relating to that, better protection of minors on the Internet requires the active involvement of ISPs in BiH too.

REFERENCES

[1] Communications Regulatory Agency BiH web site, www.rak.ba

[2] Save the Children in North West Balkans, https://nwb.savethechildren.net/sites/nwb.savethechildren.net/files/library/infografika-SID14-web.pdf

[3] Dr. Michael Busch European Commission : Regional workshop "Protecting children against sexual exploitation and sexual abuse through Internet, December 2010, www.coe.int

[4] http://msb.gov.ba/PDF/140605_Nasilje_engleski_SG_ver2.pdf

[5] http://www.sigurnodijete.ba

[6] O'Neill, B. (2014). Policy Influences and Country Clusters: A Comparative Analysis of Internet Safety Implementation. LSE, London: EU Kids Online, http://eprints.lse.ac.uk/57247/

[7] E-Commerce-Directive – Directive 2000/31/EC, 2000, http://eur-lex.europa.eu/

[8] Protecting Children in the Digital World, (COM/2011/0556 final), 2011, http://eur-lex.europa.eu

[9] Fourth Implementation Review of the European Framework for Safer Mobile Use by Younger Teenagers And Children, 2014 http://www.gsma.com/gsmaeurope/safer-mobile-use/national-measures/

[10] http://www.ictcoalition.eu/members

[11] https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/sn_principles.pdf

[12] http://europa.eu/rapid/press-release_IP-11-479_en.htm

[13] http://www.euroispa.org/

[14] Suada Hadžović: '' A Safer Online Environment for Children in BiH'', Publisher: INTERNEWS in Bosnia and Herzegovina, Sarajevo 2014, www.internews.ba

Architecture and Functionality in M2M Standards

M. Ptiček*, V. Čačković**, M. Pavelić*, M. Kušek* and G. Ježić*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
** Ericsson Nikola Tesla, Krapinska 45, 10000 Zagreb, Croatia
{marina.pticek, marko.pavelic, mario.kusek, gordan.jezic}@fer.hr, vanesa.cackovic@ericsson.com

Abstract - Over the last few years OneM2M global initiative has been working on creation of global specification and standards for M2M. In order to facilitate the development of future M2M applications, which will be consolidated with this M2M standard, this paper compares OneM2M standard with ETSI M2M standard, which was defined by European Telecommunications Standards Institute, and presents ETSI M2M and OneM2M latest specification versions in the scope of architecture model and the functionality of the M2M platforms.

I. INTRODUCTION

The term Machine-to-Machine (M2M) denotes technologies that enable communication and networking of different devices and is a part of the Internet-of-Things (IoT) vision, whose main idea is networking of devices via the Internet.

The continuous growth of M2M technologies forced many standardization bodies, including European Telecommunications Standards Institute (ETSI), to define M2M standards and that lead to appearance of different standards. Because of the resulting situation, OneM2M initiative was formed with a goal to create a global M2M specification and standard. In the process of their creation, existing standards are also being considered including all of the existing specifications and they are being partially included in the global specification. A decision has been brought that big part of ETSI M2M specification will be included in the global OneM2M specification [1].

Since it is expected that future M2M applications will convey the global standard, it is necessary to compare architecture models proposed in ETSI M2M and OneM2M standards. This will simplify their creation and will also provide information necessary for the possible transition of ETSI standard compliant applications to OneM2M compliant applications.

Second chapter introduces overview of M2M standards. Third chapter presents architecture models of ETSI and OneM2M standards, while the fourth chapter explains in more details the functional requirements on M2M platforms. Fifth chapter brings overview of the recent most popular M2M platforms and the sixth chapter concludes this paper.

II. M2M STANDARDS

According to the Gartner's Hype Cycle the term Machine-to-Machine (M2M) has passed the peak and it is near hitting the lowest expectations trough disillusionment [2]. On the other hand the term Internet-of-Things (IoT) is at the peak inflated expectations. Lots of people do not find differences in those two terms. There is open debate on the Internet regarding that. We believe that the IoT and M2M have lots of parts that intersect but there are parts of M2M that is not IoT and there are parts of IoT that are not part of M2M [3].

The goal of IoT is to connect things to Internet by using standard IP protocols. The major focus of IoT is exchanging data, organizing data semantically and processing it. The underlying network is not in focus of IoT. In some cases things do not have to be connected to the network (e.g. thing tagged with NFC or QR code).

On the other hand, M2M is mostly focused on the system that is connected over public telecommunication network, usually mobile network. M2M focuses on how to connect machines (devices) and how it will impact mobile network because currently the mobile network is designed for exchanging large data over high speed. The machine that senses (e.g. temperature periodically) and sends measurement to the network occasionally connects to the network and sends few bytes of data. The mobile network and current protocols are not efficient in such environment because there is lots of data overhead. M2M also focuses on security of connection, identification and management of devices. Because of that all major standardization bodies such as ITU, ETSI, BBF, OMA, etc. started standardization processes.

ETSI started M2M standardization in 2008 by analyzing requirements and missing standards. The ETSI started defining M2M platform that will serve different verticals (smart metering, e-health, transportation, manufacturing, facility management, etc.). The ETSI did not try to standardize parts that other standardization bodies already standardized. The ETSI members involved in other standardization bodies were actually representing needs of ETSI M2M and were proposing solutions in BBF, OMA, 3GPP, etc. ETSI published first release of standard in 2011, and second release in 2013. The ETSI is European standardization body and the need for standards in the whole world was growing. In 2012 ETSI organized consortium of 7 standardization bodies from different continents and started project OneM2M in the similar organization as 3GPP. The goal of OneM2M was the same as for ETSI M2M but on the global scale. The ESTI M2M releases were the most important inputs for defining global standard. OneM2M released first global standard in the February 2015.

III. M2M STANDARDS' ACHITECTURES MODELS

This chapter describes the architecture models proposed by ETSI M2M and OneM2M standards.

A. *ETSI M2M Functional Architecture*

The high-level model of ETSI functional architecture proposed in specification "TS 102 690" [4] consists of a *Network Domain* and *Device and Gateway Domain* (Figure 1.).

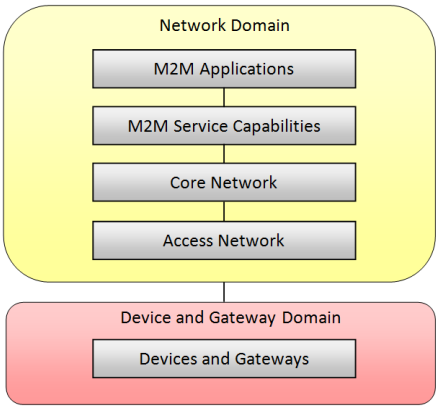


Figure 1. ETSI M2M high-level architecture

Device and Gateway Domain includes an *M2M Device*, *M2M Gateway* and *M2M Area Network*.

M2M Device runs one or more M2M Applications using M2M Service Capabilities. It can connect to Network Domain in three ways. First method of connection is "Direct Connectivity": M2M Device connects to the Network Domain via the Access network. It performs the procedures such as registration, authentication, authorization, management and provisioning with the Network Domain. It may also provide service to other devices (e.g. legacy) connected to it, but they are hidden from the Network Domain. Second method of connection is "Gateway as a Network Proxy": M2M Device connects to the Network Domain via an M2M Gateway, and connection to the M2M Gateway is accomplished using the M2M Area Network. The M2M Gateway acts as a Proxy for the Network Domain towards the M2M Devices connected to it. In that case the proxy performs the procedures of authentication, authorization, management and provisioning. The third method of connection is an extension of the second and includes multiple Gateways: M2M Devices may be connected to the Network Domain via multiple M2M Gateways.

M2M Area Network provides connectivity between M2M Devices and M2M Gateways. Examples of such network are Personal Area Network technologies such as IEEE 802.15.1, ZigBee, Bluetooth, IETF ROLL, etc.

M2M Gateway runs one or more M2M Applications using M2M Service Capabilities. It acts as a proxy between M2M Devices and the Network Domain. It may also provide service to other devices connected to it, but they remain hidden from the Network Domain. For instance, an M2M Gateway may run an application that collects and processes various information (e.g. from sensors).

The Network Domain consists of the *Access Network*, *Core Network*, *M2M Service Capabilities*, *M2M applications*, *Network Management Functions* and *M2M Management Functions*.

Access Network allows the M2M Device and Gateway Domain to communicate with the Core Network. It includes (but is not limited to): xDSL, HFC, satellite, GERAN, UTRAN, eUTRAN, W-LAN and WiMAX.

Core Network provides IP connectivity at a minimum and connectivity means or others, service and network control functions, interconnection with other networks and roaming. Each Core Network offers different features sets. Core Networks include (but are not limited to) 3GPP Core Networks, ETSI TISPAN Core Networks and 3GPP2 Core Networks.

M2M Service Capabilities provide M2M functions that will be shared by different applications and exposed through a set of open interfaces. They use Core Network functionalities and simplify and optimize application development and deployment by hiding network specificities.

M2M Applications are applications that run the service logic and use M2M Service Capabilities accessible via an open interface.

Network Management Functions consist of all the functions required to manage the Access and Core Networks. These include Provisioning, Supervision, Fault Management, etc.

M2M Management Functions consist of all the functions required to manage M2M Service Capabilities in the Network Domain. The management of the M2M Devices and Gateways uses a specific M2M Service Capability. They also include a function for M2M Service Bootstrap called MSBF (M2M Service Bootstrap Function) which is realized within a server. Its role is to facilitate the bootstrapping of permanent M2M service layer security credentials in the M2M Device or Gateway and M2M Service Capabilities in the Network Domain.

B. *OneM2M Functional Architecture*

OneM2M functional architecture proposed in published specification "TS-0001-V1.6.1" [5] comprises of several functions: *Application Entity*, *Common Services Entity* and *Network Services Entity* (Figure 2.).

Application Entity (AE) represents an instantiation of application logic for end-to-end M2M solutions. Each AE is identified with a unique AE-ID. For example, an AE can be an instance of application for fleet tracking, remote blood sugar monitoring, power metering, etc.

Common Services Entity (CSE) represents an instantiation of a set of *Common Services Functions* of the M2M environments, where each CSE is identified with a unique CSE-ID. Such service functions are exposed to other entities through reference points, where one of them is used for accessing Underlying Network

Service Entities. Service functions offered by CSE include Data Management, Device Management, M2M Subscription Management and Location Services. Such "sub-functions" may be logically and informatively conceptualized as *Common Services Functions* (CSFs). The normative Resources which implement the service functions in a CSE can be mandatory or optional.

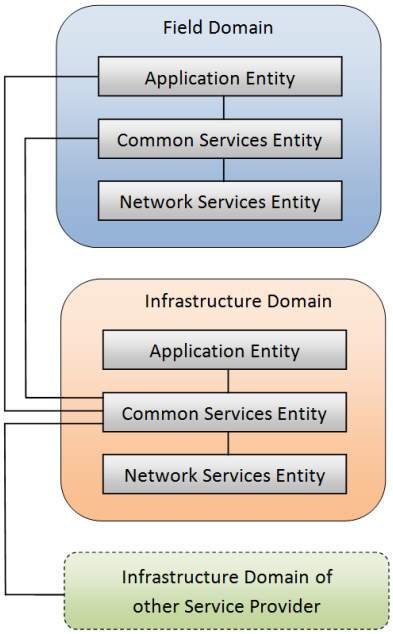


Figure 2. OneM2M functional architecture

Network Services Entity (NSE) is an underlying entity which provides services from the underlying network to CSEs. However, even though underlying networks provide data transport services between entities in the OneM2M System, such data transport services are not included in the NSE. NSE provides services of device management, location services and device triggering.

Two general kinds of OneM2M Nodes are supported, a *Common Services Entity-Capable Node* and a *Non-Common Services Entity-Capable Node*. A Common Services Entity-Capable Node is a functional entity that contains one CSE and zero or more AEs. The Application Service Node and Middle Node are examples of a Common Services Entity-Capable Nodes. CSEs resident in different Nodes could not be identical and are dependent on the services supported by the CSE in that Node. A Non-Common Services Entity-Capable Node is a functional entity that contains one or more AEs and no CSE. The Application Dedicated Node is an example of a Non-Common Services Entity-Capable Node.

OneM2M supports four basic types of Nodes: *Application Service Node*, *Application Dedicated Node*, *Middle Node* and *Infrastructure Node*. As functional objects, such Nodes may or may not be mapped to physical objects.

Application Service Node (ASN) contains one CSE and contains at least one AE. There may be zero or more ASNs in the Field Domain of an M2M System. An ASN may communicate with either exactly one Middle Node or exactly one Infrastructure Node. An ASN also communicates with NSE. Physically, an ASN could reside in an M2M Device.

Application Dedicated Node (ADN) contains at least one AE and does not contain a CSE. There may be zero or more ADNs in the Field Domain of an M2M System. An ADN communicates with a Middle Node or an Infrastructure Node. Physically, an ADN could reside in a constrained M2M Device.

Middle Node (MN) contains one CSE and zero or more AEs. There may be zero or more MNs in the Field Domain of an M2M System. A MN communicates with ASNs, ADNs, NSEs, other MNs and Infrastructure Nodes. Physically, a MN could reside in an M2M Gateway.

Infrastructure Node (IN) contains one CSE and contains zero or more AEs. There is exactly one logical IN in the Infrastructure Domain per OneM2M Service Provider of an M2M System. It may contain CSE functions not applicable to other node types. An IN communicates with one or more MNs, one or more ASNs, one or more ADNs, NSEs and INs of other M2M Systems. Physically, an Infrastructure Node could reside in an M2M Service Infrastructure.

Another type of Nodes is possible, a Non-oneM2M Node (NoDN). It does not contain any of the OneM2M Entities. Physically, NoDNs are devices attached to the OneM2M system for interworking purposes, including management. These nodes reside in Field Domain.

C. *Comparison of functional architectures*

OneM2M and ETSI have similar architectures and certain parallels can be drawn; between ETSI's M2M Applications and OneM2M's Application Entity, ETSI's M2M Service Capabilities and OneM2M's Common Services Entity, while ETSI's Core Network and Access Network could be considered correspondent to OneM2M's Network Services Entity. However, OneM2M architecture is not an extension of ETSI M2M. Both architectures consider two domains, but their internal organization differs. ETSI considers devices and gateways in the *Device and Gateway domain*. OneM2M has brought in the term of various types of "Nodes", which can reside in both of OneM2M domains, but the devices (considered as Application Service Nodes or Application Dedicated Nodes) and gateways (considered as Middle Nodes) reside in Field domain.

IV. FUNCTIONAL REQUIREMENTS ON M2M SOFTWARE PLATFORMS

M2M ecosystem consists of several components performing different functional roles. As defined by OneM2M [6], those functional roles are: *The User* (individual or company), *The Application Service Provider*, *The M2M Service Provider* and *The Network Operator*. The User uses the M2M Solution; The Application Service Provider provides an M2M Application Service and operates M2M Application; The M2M Service Provider provides M2M Services to Application Service Providers operates M2M Common Services; and the Network Operator provides Connectivity

and related services for M2M Service Providers and operates an Underlying Network (Figure 3.). Any of the above functional roles may coincide with any of the other roles.

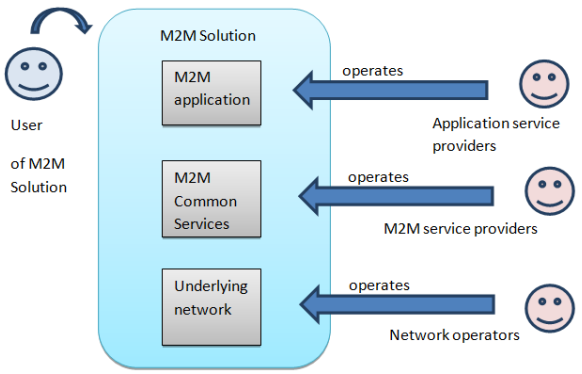


Figure 3. Functional Roles in the M2M Ecosystem

The functionalities required to achieve one M2M service span across three layers, connectivity layer, common service layer as well as the application layer. One M2M solution provider may be responsible for only one part of the solution (for example giving only the connectivity), some parts of the solution (for example giving connectivity and common service) or for the whole end-to-end solution. Depending on that one solution provider may have one or several software platforms that are handling different parts on the M2M solution.

M2M software platforms are key tools in the delivery of M2M services. Over the past several years M2M software platforms have developed rapidly and by the functional requirements can be divided as follows [7],[8]:

Connectivity Support Platforms are designed to ensure the most fundamental tasks to configure and support a machine-to-machine connection in a (typically) cellular environment. Examples of those would be connection provisioning, usage monitoring and some level of support for fault resolution. Platforms of his type are typically owned by telecom operators and are characterized by low costs and Telco grade operational excellence. The emphasis is on robustness and reliability, rather than functionality. Example of this kind of platform would be Ericsson Device Connection Platform or Jasper.

Service Enablement Platforms are flexible and have extensive horizontal capabilities in terms of solution support, reporting and the provision of a software environment and APIs to facilitate solution development for different vertical applications (industries). A key difference between the Service Enablement platform level and the Connectivity Support platform level is that the Service Enablement level will offer abstraction services: effectively a single CSP's Service Enablement Platform can potentially support access to, and offer a window onto, the Connectivity Support Platforms of multiple CSPs. This arrangement allows a single CSP to offer a multi-country solution to a client, including devices that are connected to third party networks. Together, Connectivity Support and Service Enablement functions represent the "horizontal" elements of the M2M platforms industry. Example of this kind of platform would be Ericsson Service Enablement Platform or Jasper.

Device Management Platforms are designed to support devices connected through multiple networks. They are typically designed for specific devices type and usually for a single vendor or group of devices compliant to some specific protocol suit (for example for smart meters), where those devices typically require some form of non-standard systems support (reporting, management, etc.). Device Management platforms are at this moment a niche proposition since in the future it is predicted that Service Enablement platforms will make a point of developing the capabilities required to support more and more devices. Until such time, Device Management platforms often bypass Service Enablement platforms and directly "plug in" at the Connectivity Support level. Example of this kind of platform would be Ericsson Remote Device Management Platform or Sierra.

Application Support Platforms are typically targeted at niche customer sectors. They are characterized by the provision of tailored and often specialized solutions, encompassing connected devices potentially of multiple types, potentially connected with multiple technologies, and potentially connected to the networks of multiple CSPs. Application Support platforms provide a higher level of value-added service than Device Management platforms, but still to some extent are subject to the same effect of "standardized" solution components migrating to the Service Enablement platform level. The key difference between Device Management and Application Support platforms in this regard is that this standardization and migration effect will allow an Application Support platform to focus more on their core value add proposition, whereas standardization and migration of capabilities can render corresponding Device Management platform capabilities effectively redundant. Example of this kind of platform would be Axeda or Sensor Logic.

Solution Provider Platforms used by systems integrators to support turnkey and client-specific solutions. In most cases, a solution provider M2M platform should be regarded as an enabler for a wider systems development initiative, rather than as a standalone offering. The role of Solution provider platforms is unlikely to be significantly impacted by increasing standardization within the M2M market.

Market demands software platforms that power M2M ecosystem by eliminating silos and support the creation and delivery of innovative M2M solutions, platforms that are industry-agnostic, horizontal and scalable for a wide range of industries as well as ones based on proven technologies that are deployed globally.

As already stated one M2M solution provider can operate one part of the solution or whole end-to-end solution. Based on that, one M2M solution provider will operate one or more M2M platforms. When trying to compare platforms regarding functionalities we can make a parallel with comparing different players on the M2M market. Today, most telecom operators offering M2M

services are acting as connectivity providers and operating Connectivity support platforms (TABLE I.). This includes cost efficient connectivity & more intelligent smart pipe services e.g. self-care, portals, mass provisioning, M2M charging, statistics etc. Service Enablement platforms are usually also operated by telecom operators as they are equally well suited to supporting applications that run over public or private networks and are not industry specific. Their focus on relatively complex applications and are complex to develop, however once they are developed and all the common services functionalities implemented, they can be used for different industries and use cases which makes them cost efficient and horizontal. Device Management platforms are typically operated by a M2M service company since they greatly depend on type of end user device. It can be seen that in the future Device Management platforms will support more type of devices and will move from vertical to more horizontal solutions. Application support platforms today are usually owned by M2M service companies and seldom exist as stand-alone offering since network and commons functionalities are required for end-to-end M2M service.

TABLE I. PLATFOM COMPARIOSON BY FUNCTIONAL REQUIREMENTS

	Functional layer	Typically operated by	Industry specific
Connectivity support	Underlying network	Telecom operators	No
Service Enablement	Common services	Telecom operators	No
Device Management	Underlying network	Telecom operators or Service providers	Yes
Application support	Application	Service providers	Yes

V. M2M PLATFORMS

With the development and spread of M2M technologies, many M2M platforms were developed. Regarding the support of standards, many of them support some standard partially or they do not support any of the standards at all. Since this report focuses on the standards ETSI M2M and OneM2M, this chapter will bring more details about standardized M2M platforms, most of which are open-source.

A. InterDigital

InterDigital's Standardized M2M Software Development Platform (M2MSDP) provides a standard M2M middleware solution with common APIs for scalable and horizontal M2M services. It is a scalable service platform, which features SOA and RESTful architecture, configurable charging, internetworking via Interworking Proxy Unit, lightweight device management, cognitive semantic services, automated service discovery, etc. This platform conforms to ETSI TC M2M Release 1 and Release 2 standards. It basically supports peer M2M Service Capabilities at M2M Devices, M2M Gateways and M2M Servers. It also supports four M2M reference points [9], while constrained M2M devices are supported as M2M Device Applications.

B. Eclipse OpenM2M

OpenM2M (OM2M) project is an open-source implementation of the ETSI M2M standard. It provides a horizontal M2M service platform for developing services independently of the underlying network. Its aim is to facilitate the deployment of vertical applications and heterogeneous devices. Key features of this platform are a RESTful API, modularity and extensibility [10][11]. RESTful API provides primitive procedures for machine authentication, resource discovery, application registration, etc. OM2M is built in Java as an Eclipse product using Maven and Tycho. It runs on top of OSGi Equinox runtime, thanks to which it is highly extensible via plugins. It is remotely configurable without requiring a reboot. This platform implements the ETSI M2M standard. It provides a horizontal Service Capability Layer (SCL) that can be deployed in an M2M network, device or a gateway.

C. OpenMTC

The Open Machine Type Communications (OpenMTC) Platform provides solution for conducting applied research activities and developing innovative M2M and IoT applications. It is a prototype implementation of an M2M middleware, which enables interconnection of various sensors and actuators from different vertical domains with a platform that can aggregate the collected data, forward it to the application and mediate instructions for the control of devices. OpenM2M platform aims to be aligned with multiple international standards such as ETSI M2M, OneM2M and 3GPP [12]. However, only ETSI is implemented completely. OpenMTC can be integrated with 3GPP core elements. Platform's front-end Field Domain layer and back-end cloud-based platform layer follow the OneM2M standards, proposed in oneM2M-TS-0001 (in drafting stage), and ETSI M2M standards, proposed in TS 102.921 and TS 102.690.

D. LG U+'s and SK Telecom's platform

At the end of December 2014, LG U+ and Korean SK Telecom have announced that they will create the industry's first open-source IoT platform based on OneM2M specifications [13]. Until 2012, SK Telecom was developing an ETSI TC M2M based platform, which was then commercialized and is now known as OpenM2M [14]. From 2012 until 2014, they were adding OneM2M standard to their existing IoT service infrastructure, and so did LG U+ with their platform. They united on this project of creating a first OneM2M IoT platform, which is planned to be finalized during the first quarter of 2015.

E. Comparisson of platforms

In TABLE II. is given a simple comparison of ETSI-complying platforms. Comparison of OneM2M-complying platforms is not given because LG U+ and SK Telecom platform was not announced yet, and rest of platforms are not supporting all specifications required by standards.

TABLE II. PLATFORM COMPARISON

	InterDigital	OpenM2M	OpenMTC
REST API	+	+	+
Scalability	+		+
Connectivity	HTTP CoAP	HTTP CoAP	HTTP CoAP Web sockets Diameter MQTT

Open MTC announced that compatibility with OneM2M standard will be enabled, so comparison of OneM2M platforms can be done in future work.

VI. CONCLUSION

This paper investigates current standards in M2M. The architecture of ETSI M2M and OneM2M are compared. Nodes and other parts of architecture have different name but denotes similar entities. The reason can be found in the fact that OneM2M is based on ETSI M2M. The functional comparison shows similar results.

In the last section the paper analyzed M2M platform implementations. Although OneM2M have just published their first standard, currently there are no implementations and the comparison will be in future work. That is the reason that the paper is focused on implementations of ETSI M2M. The Open MTC platform is compliant with ETSI standard and the OneM2M compliance is announced. Beside that only the Open MTC has supported connectivity by Web socket, Diameter and MQTT. In the future work we will concentrate on analysis of OneM2M platforms.

ACKNOWLEDGMENT

The authors acknowledge support of the research projects: "Managing Trust and Coordinating Interactions in the Smart Networks of People, Machines and Organizations", funded by Croatian Science Foundation and "Machine-to-Machine Communication Challenges", funded by Ericsson Nikola Tesla, Croatia.

REFERENCES

[1] FI-WARE Wiki, "Mapping Internet of Things", https://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/Mapping_Internet_of_Things, accessed March 2015.

[2] Gartner, Inc. "Hype Cycle for Emerging Techonologies", 2014, http://www.gartner.com/newsroom/id/2819918.

[3] IoT vs M2M: A subset, intersection or a value addition, http://litmusautomation.com/2014/07/iot-vs-m2m/.

[4] ETSI Technical Specification TS 102 690 V2.1.1, "Machine-to-Machine communications (M2M); Functional architecture", http://www.etsi.org/deliver/etsi_ts/102600_102699/102690/02.01.01_60/ts_102690v020101p.pdf, October 2013.

[5] OneM2M Technical Specification TS-0001-V1.6.1, "Functional architecture", http://www.onem2m.org/images/files/deliverables/TS-0001-Functional_Architecture-V1_6_1.pdf, January 2015.

[6] OneM2M Technical Specification TS-0002-V1.0.1, "Requirements", http://www.onem2m.org/images/files/deliverables/TS-0002-Requirements-V1_0_1.pdf, January 2015.

[7] Machina Research, Strategy Report: "Competitive Dynamics in the M2M Platform Space", January 2014.

[8] Internet of Things (IoT) & Machine-to-Machine (M2M) Communication Market – Worldwide Market Forecasts (2014-2019), http://www.marketsandmarkets.com, accessed February 2015.

[9] InterDigital, White paper "Standardized Machine-to-Machine (M2M) Software Development Platform", http://www.interdigital.com/wp-content/uploads/2012/10/InterDigital-M2M-white-paper_Oct2012.pdf, October 2012.

[10] Eclipse Wiki, OM2M, https://wiki.eclipse.org/OM2M, accessed February 2015.

[11] M. Ben Alaya, Y. Banouar, T. Monteil, C. Chassot, K. Drira, "OM2M: Extensible ETSI-compliant M2M service platform with self-regulation capability", Procedia Computer Science 32, 2014, pages 1079-1086

[12] OpenMTC, "Whan's OpenMTC?", http://www.open-mtc.org/index.html#openmtc-plat, accessed February 2015.

[13] OneM2M, "Operators Announce Commercial Availability of oneM2M Platforms", http://www.onem2m.org/news-events/news/49-operators-announce-commercial-availability-of-onem2m-platforms, accessed February 2015.

[14] OneM2M Showcase Presentations/ETSI M2M Workshop, "The oneM2M Platform as an enabler for Telecom Operators", http://www.onem2m.org/component/rsfiles/download-file/files?path=oneM2M_Showcase%255ConeM2M_Platform_enabler_Telecom_Operators.pdf&Itemid=198, December 2014.

Metodologija procjene kapaciteta RF spektra za potrebe implementacije mobilnih servisa u periodu do 2020. godine sa primjenom u BiH

Almedin Džebo, Amina Krivošić

Regulatorna agencija za komunikacije, Sarajevo, BiH

dzebo.almedin@gmail.com, amina.krivosic@gmail.com

Sažetak

Zahtjevi za brzi prenos velike količine podataka podrazumijevaju i povećanu potrebu za dodatnim resursima RF spektra uzimajući u obzir trendove korištenja različitih mreža, upotrebu superiornijih modulacionih postupaka, nove tehnologije, digitalnu agendu kao inicijativu u sklopu strategije razvoja Europa 2020 te niz drugih faktora. Za procjenu zahtjeva za frekvencijama, novi modeli moraju biti razvijeni i procijenjeni na način da omoguće prostornu i vremensku koralaciju između telekomunikacijskih usluga imajući u vidu uslove na tržištu i scenarije mrežne implementacije. U radu je prezentirana modifikovana verzija metodologije procjene kapaciteta spektra definirana ITU preporukama, prilagođena uslovima u Bosni i Hercegovini. Cilj je pokazati realne potrebe za RF resursima upotrebom parametara i pokazatelja rasta i razvoja mobilnih mreža koji korespondiraju sa realnim i konkretnim parametrima koji egzistiraju na tržištu telekomunikacija. Proračun potrebnog spektra za mobilne komunikacije bazira se na principima definiranja servisa, tržišnog kretanja i očekivanja te tehničkog i operativnog okvira za proračun.

Ključne riječi: procjena kapaciteta spektra, mobilni servisi, planiranje RF spektra, proširenje spektra za IMT, širina opsega

I. UVOD

Za realizaciju mobilnog servisa potrebni su odgovarajući radiofrekvencijski resursi. Obzirom da novi servisi zahtijevaju brzi prenos velike količine podataka, potreba za dodatnim resursima je u stalnom porastu. Stoga se ovaj ograničeni resurs mora efikasno planirati. Da bi se identificirao spektar na međunarodnom i evropskom nivou, a zatim namijenio i dodijelio na nacionalnom nivou, potrebno je nekoliko godina. Zbog toga je potrebno već sad započeti planiranje spektra za IMT za 2020. godinu.

Planiranje i alociranje RF opsega je pitanje o kojem se, zbog svoje važnosti, osim na tehničkom nivou raspravlja i na nivou političkih i drugih interesnih grupacija. Stoga je u strategiju razvoja pored realnih tehničko-eksploatacionih, involvirano i niz faktora koji su nastali kao produkt lobiranja i interesa kako proizvođača opreme tako i drugih grupacija zainteresovanih za porast korištenja RF spektra za ovu vrstu servisa.

Realna procjena potrebnih kapaciteta za razvoj mobilnih komunikacija vrši se na ekspertnom nivou uzimajući u obzir trendove korištenja različitih mreža (npr. nelicencnih RLAN sistema), novih tehnologija koje upotrebom superiornijih modulacionih postupaka na efikasniji način koriste spektar, potrebu za razvojem drugih servisa, odgovarajuće strategije prezentirane u EU dokumentu „Digital Agenda 2020“ i još mnogo faktora koji utiču na potrebu korisnika za mobilnošću.

U prošlosti, procjena zahtjeva za RF spektrom za razvoj bežičnih aplikacija odnosno servisa razmatrala se u okviru jednog sistema i jednog scenarija na tržištu. Pojavom konvergencije između mobilnih i fiksnih telekomunikacijskih i multi-mrežnih okruženja, kao i međupovezivanjem različitih komplementarnih pristupnih sistema nametnut je novi kompleksniji pristup problematici razvoja metodologije za procjenu potrebne količine spektra za mobilne komunikacije.

U okviru Međunarodne unije za telekomunikacije, ITU radna grupa JTG 5 koja je zadužena za pitanja koja su vezana za IMT (International Mobile Telecommunications), priprema prijedlog koji treba da dimenzioniše potrebe za novim frekvencijskim opsezima za implementaciju mobilnih komunikacijskih servisa. Prijedlog bi trebao da sadrži konkretne podatke o potrebnim resursima za period do 2020. godine sa specificiranim RF opsezima koji bi mogli da zadovolje buduće potrebe. Prijedlog će se razmatrati na Svjetskoj radio konferenciji WRC-15 (World Radio Conference) koja će se održati u Ženevi krajem 2015. godine i predstavlja jednu od najvažnijih kandidovanih tema konferencije. Ključni element za procjenu predstavlja izbor modela proračuna odnosno korištenje odgovarajuće metodologije. Metodologija procjene potrebnih kapaciteta propisana je ITU preporukama serije M u kojima su razmatrani različiti scenariji budućeg razvoja mobilnih mreža i servisa. Evolucija se odvija na način da se vrši stalno praćenje i poređenje rezultata dobijenih proračunom sa stvarnim saobraćajnim parametrima te se u skladu s tim vrši korekcija algoritma procjene.

Procjena potreba za spektrom razmatra se kroz ITU WP 5D za vrijeme pripreme za WRC-15. Agenda Item 1.1. nalaže razmatranje dodatne namjene za mobilne servise na primarnoj osnovi i identifikaciju dodatnih frekvencijskih opsega za Međunarodne mobilne telekomunikaje (IMT) i odgovarajuće regulatorne odredbe u cilju omogućavanja razvoja zemaljskih mobilnih

širokopojasnih aplikacija u skladu sa Rezolucijom 233 sa WRC-12.

Rezolucija 233 sa svjetske radio konferencije 2012. je pozvala na istraživanje i donošenje studija i prijedloga za buduće zahtjeve za spektrom i potencijalne opsege koji mogu biti kandidovani za IMT i druge zemaljske mobilne širokopojasne aplikacije [1].

II. PROŠIRENJE OPSEGA ZA MOBILNE MREŽE

Prema asocijaciji GSMA mogući opsezi koji će biti kandidovani na WRC-15 u okviru teme 1.1. su navedenim u slijedećoj tabeli:

Opis	Spektar	Korištenje	WRC-15
Opseg (<1GHz)	Dijelovi opsega 500-600MHz [470-694MHz]	TV PMSE	WRC-15 regionalna identifikacija za IMT korištenje Potrebna kooperacija sa radiodifuznom industrijom
	700MHz [694-790MHz]	TV PMSE	WRC-15 regionalna IMT identifikacija: Region 1 (AI 1.2)
Opseg (1GHz-3GHz)	Dijelovi opsega 1.4 GHz [1350-1525MHz]	D-Radio Fixed Link Scientific	WRC-15 globalna identifikacija za IMT korištenje Korištenje u naučne svrhe, samo u dijelovima opsegai u nekim dijelovima regiona
	2700-2900 MHz	Radar	WRC-15 globalna identifikacija za IMT korištenje
Opseg (3GHz-6GHz)	3.4-3.6 GHz	IMT (u nekim zemljama) Sat.	WRC-15 globalna identifikacija za IMT korištenje
	3.6-3.8 GHz	IMT Sat.	WRC-15 globalna identifikacija za IMT korištenje
	Dijelovi opsega 3.8-4.2 GHz	Sat.	WRC-15 globalna identifikacija za IMT korištenje
	Dijelovi opsega 4.4-4.99 GHz	Sat.	WRC-15 globalna identifikacija za IMT korištenje

Tabela 1. Potencijalni opsezi za mobilne komunikacije kandidovani za WRC-15

(izvor CPM-15: A Guide to WRC-15 Agenda Item 1.1)

Za harmonizaciju IMT na globalnom nivou, najveću mogućnost imaju slijedeći opsezi:

- Sub-700MHz UHF (470-694/698 MHz)
- L-Band (1350-1400 & 1427-1518 MHz)
- 2.7-2.9 GHz
- C-Band (3.4-3.8 GHz & 3.8-4.2 GHz)

Zahtjevi za spektrom za IMT-2000

Generalno problem procjene zahtijevane širine radio spektra sastoji se od slijedećih podproblema:

- Razvoj scenarija i određivanje vrijednosti ulaznih parametara
- Proračun zahtijevanog kapacitita sistema u skladu sa ponuđenim saobraćajem i zahtijevanim kvalitetom servisa
- Proračun potrebne širine radio spektra potrebnog da se obezbijedi zahtijevani kapacitet sistema

ITU - R Izvještaj M.2290 " Procjena budućih zahtjeva spektra za terestrijalni IMT" predstavlja prognozu rasta ukupnog iznosa mobilnog saobraćaja u svijetu do 2020. godine te modele potražnje za spektrom za IMT usluge bazirane na prognozi saobraćaja [2]. Ovi modeli su relativno sofisticirani i dijele saobraćaj na različita servisna okruženja (npr. urban, suburban, rural, home, office, public) i klase usluga (npr. streaming, interactive, background). Ovaj model trenutno predviđa da će biti potrebni IMT spektar između 1340 (niže tržišne postavke) i 1960 MHz (više tržišne postavke) za zadovoljenje potreba za uslugama do 2020. godine.

Imajući u vidu veće zahtjeve ukupni procijenjeni zahtjevi su oko 1800MHz. Poredeći sa već do sad identificiranih 1172MHz, trebaće još oko 500MHz dodatnog spektra.

III. METODOLOGIJA PROCJENE RF SPEKTRA ZA MOBILNE SERVISE

Preporuka ITU-R M.1390 - Metodologija za proračun potreba za spektrom za zemaljski IMT-2000 sadrži metodologiju proračuna zahtijevanog spektra procijenjenog za IMT-2000. Obezbjeđuje sistematski pristup koji uzima u obzir geografski uticaj, uticaj tržišta i saobraćaja, tehničke i sistemske aspektre i konsolidaciju rezultata zahtjevanog spektra. Metodologija je primjenjiva na radio prenosne tehnologije bazirane na komutaciji krugova i komutaciji paketa i može objediniti servise koji su okarakterizirani asimetričnim protokom saobraćaja. Metodologija je dovoljno fleksibilna da obezbijedi globalno sagledavanje potreba za spektrom ili jedinstvene zahtjeve regionalnih tržišta [3].

Osnovni cilj metodologije je odrediti individualne zahtjeve za spektrom za sve kombinacije specifičnih okruženja i usluga u datom geografskom području i kombiniranje seta individualnih zahtjeva za spektrom zajedno sa ukupnom terestrijalnom komponentom koja se procjenjuje podrazumijevajući odgovarajuće faktore kao što su konkurentske usluge na datom geografskom području. Pored toga uzeti su u obzir uticaj više operatora, i dijeljenje spektra.

Obzirom da procjena ne daje konkretne rezultate, metodologija prije svega uzima u obzir primarne faktore za potrebni spektar.

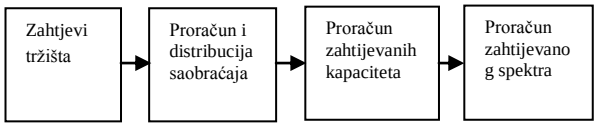
Ograničenja ove metodologije su što ona razmatra samo jedan tip mreže IMT-2000 bez mogućnosti razmatranja višestrukih mreža koje egzistiraju istovremeno. Također, M.1390 nije model mreža koje se sastoje od višestrukih tipova ćelija u istom geografskom području (kao što su makro ćelije i mikro ćelije za ukupno pokrivanje ili povećanje kapaciteta.

O metodologiji proračuna najviše odgovora je dato u okviru preporuke M.1768 koja je razvijena 2005 godine, a usvojena na Svjetskoj radio konferenciji WRC-07. Zbog ograničenja dotadašnjih metodologija i novih zahtjeva za spektrom na WRC-07 standardizirana je metoda ITU-R Recommendation M.1768, 2005 [4].

Nova verzija metodologije definisana preporukom ITU - R M.1768-1 (Metodologija proračuna potreba za spektrom za zemaljsku komponentu IMT) ima u vidu najnoviji napredak u IMT tehnologiji i razvoj IMT mreže: uvođenje dijeljenja spektra između makro i mikro ćelijskih slojeva (IMT-Advanced) i uvođenje novih parametara spektralne granulacije za IMT sisteme. Ažurirana metodologija sadrži zahtjeve za spektrom IMT u cjelini, a dijeli se između dvije pristupne radio tehnike (RATG-Radio Access Tehnic Group): RATG 1 : Pre - IMT sistem, IMT -2000 i njegova poboljšanja, RATG 2: IMT - Advanced. Metodologija procjene zahtjeva za spektrom slijedi tehnološki neutralan pristup i uzima u obzir tržište, tehnologije i razvoj informacija. [4]

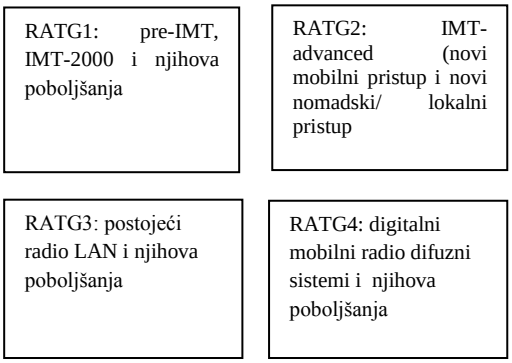
Dijagram toka metodologije

Opći dijagram toka metodologije iz Preporuke ITU-R M.1768-1 sastoji se od devet uzastopnih koraka koji su grupisani u četiri glavna bloka kao što je prikazano na slici 2. Metodologija počinje sa analizom podataka tržišta, prelazi na poračun i distribuciju saobraćaja, a zatim proračuna potrebnog kapaciteta, prije nego što se zaključi dijagram sa procjenom zahtijevanog spektra. Slika 2. prikazuje opći tok proračuna zahtijeva za spektrom.



Slika 2. Opći tok proračuna zahtijeva za spektrom

Definicije u metodologiji Preporuke ITU-R M.1768-1 radio pristupne tehnologije (prethodno tehnika) grupe (RATG) koriste definicije iz Izvještaja ITU-R M.2074 – Radio aspekti za zemaljsku komponentu IMT-2000, koji su sažeti u slici 3. Izvještaj ITU-R M.2074 također predstavlja opravdanje za korištenje RATG pristupa [5]. Treba napomenuti da metoda u Preporuci ITU-R M.1768-1 procjenjuje potražnju spektra samo za RATG 1 i 2, ali ne i RATG 3 (RLAN) i RATG 4. [4] Iako metoda uzima u obzir saobraćaj globalnih mobilnih telekomunikacije različitih mobilnih sistema kao što su IMT i RLAN, ipak zahtjevi spektra se računaju samo za IMT. Treba imati na umu da se ove metodologije mogu koristiti za procjenu ukupnih IMT zahtjeva za spektrom određene zemlje samo ako se sve trenutne vrijednosti ulaznih parametara zamjenjuju vrijednostima koje se primjenjuju na određene zemlje.



Slika 3. Radio pristupne tehnologije

U najnovijoj verziji preporuke ITU Rec M. 1768-1, objavljene u aprilu 2013. godine definiše se metodologija proračuna potrebnih količina RF spektra na način da se usvajaju novi trendovi koji vladaju na tržištu telekomunikacija, tehičko-tehnološki razvoj proizvoda i usluga te ostale promjene nastale u radio i servisnom okruženju. Komponenete metodologije su funkcionalno podijeljene u devet koraka i to:

1. Definisanje pojmova kategorije servisa (SC), servisnog okruženja (SE), radio okruženja (RE) i grupa pristupnih radio tehnologija (RATGs - Radio access technology groups);

2. Analiza prikupljenih podataka o tržištu koja je definisana u ITU izvještajima ITU-R M.2072 i ITU-R M.2243;

3. Izračunavanje potrebnog saobraćaja za određenu kategoriju servisa i okruženja;

4. Raspoređivanje saobraćaja na različite RATG-ove i radio okruženja unutar RATG-a;

5. Određivanje sposobnosti kapaciteta sistema da opsluži datu količinu saobraćaja;

6. Proračun potreba za spektrom za različite RATG1 i RATG2;

7. Usvajanje potrebnih podešavanja kao što su zaštitni opsezi, više operatora, primjena podešavanja uzimajući u obzir praktični razvoj mreže;

8. Proračun ukupnih potreba za RF spektrom;

9. U zadnjem koraku se raspoređuje odgovarajući resurs na odgovarajuće RATG-ove. [4]

Procjena potrebnog spektra proračunata je za RATG1 (pre-IMT, IMT-2000, i njihova poboljšanja) i RATG 2 (IMT-Advanced) za 2020. godinu. Kako za različite zemlje postoje razlike u procjeni i rastu podataka o mobilnim komunikacijama, tako su u okviru izvještaja ITU M.2290 date dvije ekstremne situacije. Naravno u nekim zemljama će doći do odstupanja i nacionalni

zahtjevi za spektrom mogu biti manji ili veći od predviđenih vrijednosti.

	Ukupni zahtjevi za spektrom za RATG 1	Ukupni zahtjevi za spektrom za RATG 2	Ukupni zahtjevi za spektrom za RATG 1 i 2
Situacija sa manjom gustinom korisnika	440 MHz	900 MHz	1340 MHz
Situacija sa većom gustinom korisnika	540 MHz	1420 MHz	1960 MHz

Tabela 4 - Ukupni zahtjevi za spektrom za oba RATG u 2020. godini prema izvještaju ITU-R M.2290-0

Ovaj izvještaj nudi rezultate istraživanja procjene zahtjeva za spektrom za zemaljski IMT [6].

Procijenjeni zahtjevi za spektrom se računaju pomoću metodologije definirane u Preporuci ITU-R M.1768-1 i odgovarajućih vrijednosti ulaznih parametara, uzimajući u obzir najnoviji napredak u tehnologijama i razvoj zemaljskih IMT mreža, kao i najnovija zbivanja na tržištima mobilnim telekomunikacija. Ukupni zahtjevi spektra za oba RATG 1 (tj pre-IMT, IMT-2000 i njegova poboljšanja) i RATG 2 (tj IMT-Advanced) u 2020. godini procjenjuju se pomoću dvije različite situacije kako bi se istakle razlike u tržištima i vremenske rasporede za rast mobilnih podataka u različitim zemljama. Procijenjeni ukupni zahtjevi spektra za RATGs 1 i 2 su 1340 MHz i 1960 MHz za postavke niže gustoće korisnika i postavke veće gustoće korisnika, respektivno.

Prema ITU modelu koji se zasniva na preporuci 1768-1 ukoliko se koristi poseban program za proračun spektra, ubacujući realne cifre dobije se da je potreba za spektrom puno manja nego ona proračunata ITU modelima. Tako se prema istraživanjima LS Telcom-a [7], EBU-a, ESOA došlo do rezulatata da su potrebe za spektrom 240 MHz u 2020. god.

IV. PRIMJENA U BIH

Procjena potrebnih RF kapaciteta za potrebe mobilnih mreža predstavlja veoma važan faktor u razvoju telekomunikacija koji treba da ponudi smjernice za budući razvoj tehnologije i usluga. U velikom procentu ovo je pitanje, kako je ranije prikazano u radu, determinisano zahtjevom tržišta odnosno odgovarajućim socio-ekonomskim faktorima koji preovladavaju u određenoj sredini. Zahtjev za proširenje kapaciteta mobilne mreže predstavlja stalni proces koji se postavlja pred telekom operatore. S druge strane operatori moraju planirati proširenja vodeći računa o stepenu iskorištenja trenutnih kapaciteta, troškovima uvođenja novih kapaciteta te napraviti procjenu koja će ponuditi izbalansirano rješenje koje će omogućiti kreiranje pozitivnog ambijenta za poslovanje kompanije. Obzirom da su frekvencijski resursi koje koristi mobilna mreža za pristup do krajnjeg korisnika veoma vrijedni planiranju mreže treba pristupiti studiozno i sa velikom pažnjom.

Primjena metodologije procjene predložene u ITU Rec.1786-1 u Bosni i Hercegovini se mora, zbog specifičnosti geopolitičke situacije u BiH, razmatrati sa više aspakata koji u određenoj mjeri odstupaju od principa tržišno orijentisane ekonomije čemu BiH teži. Zvanično definisanom strategijom razvoja telekomunikacija definisana su tri operatora mobilne telefonije što predstavlja administrativno ograničenje odnosno zabranu ulaska novih „ igrača“ na tržište telekomunikacija. Ova odrednica direktno utiče na razvoj konkurencije na tržištu te s tim u vezi i na razvoj novih proizvoda i usluga. „Atipično“ tržište mobilnih komunikacija u BiH se mora uzeti kao faktor koji nepovoljno utiče na razvoj, uvođenje i eksploataciju mobilnih mreža.

Sa aspekta iskorištenja RF spektra namijenjenog za potrebe mobilnih komunikacija, tradicionalno namijenjeni opsezi GSM-900 i DSC 1800 su podijeljeni između tri operatora. U novije vrijeme postoje zahtjevi za korištenje tehnologije koja na efikasniji način koristi navedene opsege tako da danas imamo korištenje LTE tehnologije u GSM 900 opsegu. Ovaj indikator pokazuje da su trenutni kapaciteti u dobroj mjeri iskorišteni. Što se tiče opsega koji su namijenjeni proširenju kapaciteta odnosno razvoju novih tehnologija u mobilnim mrežama, kao što su opsezi 700 MHz, 800 MHz, 2.6 GHz, administrativne pretpostavke za korištenje su ispunjeni ali ne postoje zahtjevi za komercijalno korištenje. LTE mreže egzistiraju u Bosni i Hercegovini samo u fazi testiranja. Ako se uzme u obzir da u regionu već egzistira LTE, može se odrediti odgovarajuće „kašnjenje“ za implemetacijom LTE mreža u BiH.

Preporuka ITU Rec.1786-1, koja je napravljenja za globalnu estimaciju potrebnih količina RF spektra, predstavlja dobru polaznu tačku za određivanje potrebnih količina RF spektra i u BiH. Analizom referentnih podataka datih u preporuci utvrđeno je da predimenzionisana vrijednost potrebnih resursa te da navedeni podaci ne korespondiraju sa realnim korištenjem ovih usluga (referenca LS Telcom). Ovdje se može lako doći do zaključka da je u izradi preporuke u velikoj mjeri učestvovala i industrija koja je jako zainteresirana za razvoj novih radiokomunikacijskih sistema.

U slučaju BiH, referentni podaci su uzeti iz zvaničnih podataka Agencije za statistiku BiH kao što su: broj korisnika mobilne telefonije, ukupan broj minuta u mobilnom saobraćaju, gustoća pretplatnika po km^2, pokrivenost signalom u zavisnosti od korištene tehnologije (2G, 3G, 4G) [8]. Analize pokazuju da najveći broj korisnika koristi veoma mali kapacitet prenosa manji od 16 kbit/s i to u 40 %. Multimedijalne sadržaje na brzinama do 144 kbit/s korsiti 30% korisnika, a ostatak korisnika se može svrstati u kategorije zahtijevnijih korisnika za srednjim i višim zahtjevom za kapacitetom prenosa podataka. Što se tiče mreža sa ćelijskim pokrivanjem u Bosni i Hercegovini, dominiraju makro i mikro ćelije ovisno o okruženju da li je urbano ili rualno. Hot spot mreže su realizovane u nokoliko gradskih središta, a piko ćelijie nisu zastupljene u značajnijem obimu.

V. ZAKLJUČAK

Količina potrebnog radiofrekvencijskog spektra u mobilnim komunikacijama važan je faktor prije svega zbog neprekidne potrebe za povećanjem mobilnog saobraćaja. Za procjenu zahtjeva za frekvencijama, novi modeli moraju biti razvijeni i primijenjeni na način da omogućavaju razmatranje prostornih i vremenskih korelacija između telekomunikacijskih usluga, uzimajući u obzir uslove na tržištu i scenarije mrežne implementacije. Metodologija predviđena ITU modelima analizira podatke tržišta, poračun i distribuciju saobraćaja, proračuna potrebnog kapaciteta da bi došla do procjene zahtijevanog spektra.

Uzimajući sve ranije navedeno u obzir možemo zaključiti da sadašnje korištenje RF spektra namijenjnog za mobilne komunikacije u BiH odgovara trenutnim i potrebama u bliskoj budućnosti (dvije godine). Trend povećanja obima saobraćaja proteklih godina je 5% u odnosu na predhodnu godinu. To znači da će za period do 2020. godine, u svakom sluačaju biti potrebno manje RF spektra nego što to predviđa ITU. Uzmemo li u obzir i povećanje obima saobraćaja koji se ostvaruje preko WiFi te korištenje manjih ćelija unutar makro ćelije kao metode povećanja efikasnosti RF spektra, proširenje na nove opsege će biti značajno manje od globalnih predviđanja.

Literatura:

[1] ITU World Radiocommunication Conference 2015 (WRC-15) Agenda and Relevant Resolutions, 2012

[2] Report ITU-R M.2290-0 (12/2013) Future spectrum requirements estimate for terrestrial IMT

[3] Rec. ITU-R M.1390. Methodology for the calculation of IMT-2000 terrestrial spectrum requirements, 1999

[4] Rec. ITU-R M.1768-1. Metodology for calculation of Spectrum Requirements for future developments of IMT-2000 and systems Beyond IMT-2000 (IMT Advanced), 2013.

[5] Report ITU-R M.2074. Radio aspects for the terrestrial component of IMT-2000 and systems beyond IMT-2000

[6] Estimated Future Spectrum Needs for IMT, Hiroyuki Atarashi, NTT Docomo, INC, 2014

[7] LS Telcom. Mobile spectrum requirement estimates: GETTING THE INPUTS RIGHT, 2014

[8] Agencija za statistiku BiH Telekomunikacijska opreme, mreže i usluge, Saopštenje. Saobraćaj, 2014

Uloga kvalitetnog dizajna poslovnih procesa u složenom okruženju telekomunikacijskog tržišta

Darko Markulin, D.Sc., Kristina Musa, M.Sc.
Odjel za razvoj i upravljanje integriranim sustavima operativne podrške
HT d.d.
Kruge 26, 10000 Zagreb, Hrvatska
Telefon: (385) 1-4912 721 Fax: (385) 1-4917 777
E-mail: darko.markulin@t.ht.hr, kristina.musa@t.ht.hr

Sažetak - Razvoj poslovnih procesa potrebnih za realizaciju i aktivaciju korisničkih usluga jedan je od osnovnih preduvjeta za uspješno korištenje usluga na području javne telekomunikacijske mreže. Kako se radi o velikom broju različitih i složenih poslovnih procesa vrlo je važna njihova precizna specifikacija i dokumentacija. Pri tome je vrlo bitno odabrati način i format prezentacije poslovnih procesa koji se uspješno može koristiti ne samo u specificiranju potrebnih promjena u sustavima za podršku potrebnim za upravljanje telekomunikacijskom mrežom, već i u svrhu dokumentiranja poslovnih procesa na način da se dokumentacija može koristiti u različite svrhe kao što je analiza poslovnih procesa ili razumijevanje poslovnih procesa od strane različitih sudionika kao što su IT stručnjaci ili eksperti iz poslovne domene. Vrlo je bitno da se dokumentacija može koristiti i u svrhu razvoja nove arhitekture sustava za podršku kao što je NGSSM (Next Generation Service and System Management).

U radu je prezentiran jedan od načina specificiranja i dokumentiranja poslovnih procesa za realizaciju telekomunikacijskih usluga koji omogućava primjenu specifikacija u različite svrhe.

Ključne riječi: poslovni procesi, sistemska operativna podrška, korisničke usluge, modeliranje, specifikacija

I. UVOD

Telekomunikacijsko tržište čini vrlo složeno i živo poslovno okruženje u sklopu kojeg je potrebno osigurati poslovne procese koji osiguravaju efikasnu realizaciju telekomunikacijskih usluga. Ovo nije nimalo jednostavno ostvariti jer uspješna realizacija usluga u tako složenom okruženju ovisi o mnogo čimbenika. Od osiguravanja potrebnih akcija za realizaciju usluge, preko koordinacije više poslovnih procesa za realizaciju složenih usluga pa do razmjene potrebnih podataka za konfiguraciju mrežnih elemenata i komunikaciju između sustava koji upravljaju telekomunikacijskom mrežom.

Da bi se osigurali efikasni poslovni procesi u tako složenom poslovnom okruženju jednu od ključnih usluga imaju specifikacije poslovnih procesa. Specifikacije koje se uspješno mogu koristiti u dizajniranju poslovnih procesa kako na području javne telekomunikacijske mreže, tako i u drugim poslovnim područjima, ne bi trebale imati samo jednostruku namjenu. Osim u svrhu specificiranja potrebnih promjena u sustavima za podršku koji djeluju u sklopu sustava za upravljanja telekomunikacijskom mrežom, specifikacije trebaju omogućiti i prikaz poslovnih procesa koji omogućava njihovu analizu te njihovo dokumentiranje koje se može koristiti u različite svrhe.

Jedno od vrlo bitnih svojstava specifikacija je i mogućnost njihovog korištenja u definiranju poslovnih procesa koje je potrebno podržati od strane novih arhitektura sustava za podršku. Način dizajna poslovnih procesa treba biti takav da se pojedini segmenti poslovnih procesa mogu uspješno preslikati u strukturu podržanu novom arhitekturom sustava.

U radu se razmatra i predlaže jedan od mogućih načina dizajniranja i dokumentiranja poslovnih procesa koji se koriste za implementaciju i aktivaciju usluga na području javne telekomunikacijske mreže.

II. PREDLOŽENI NAČIN SPECIFIKACIJE POSLOVNIH PROCESA

Da bi se uspješno modelirali poslovni procesi za realizaciju telekomunikacijskih usluga potrebno je definirati model poslovnih procesa kroz koji se uspješno mogu definirati svi elementi potrebni za specificiranje akcija koje je potrebno poduzeti u sklopu realizacije i aktivacije usluge. Osnovni preduvjet za modeliranje poslovnih procesa je korištenje grafičkog načina prikaza kao jednog od standardnog oblika prezentacije poslovnih procesa. U [2] su definirani osnovni elementi modela potrebni za modeliranje i specificiranje poslovnih procesa koji se koriste za implementaciju telekomunikacijskih usluga. Ti su elementi:

A. Početni i završni čvor

Ovi se elementi modela koriste za definiranje početka i kraja poslovnog procesa koji se modelira.

B. Ručni zadatak

Služi za definiranje akcija koje trebaju izvršiti akteri poslovnog procesa.

C. Odlazni i dolazni signal

Specificira se komunikacija između sustava koji sudjeluju u procesu realizacije telekomunikacijske usluge.

D. Podprocesi

Definiraju cjeline od kojih se sastoje poslovni procesi. Podprocese čine akcije koje se odnose na pojedinu uslugu

i potrebno ih je poduzimati u određenom dijelu poslovnog procesa.

E. Čvor odluke

Definira mjesto u poslovnom procesu kada je na temelju određenog kriterija potrebno odlučiti u kojem će se smjeru nastaviti tijek poslovnog procesa.

F. Čvor dijeljenja

Definira se mogućnost paralelnog odvijanja aktivnosti u poslovnom procesu.

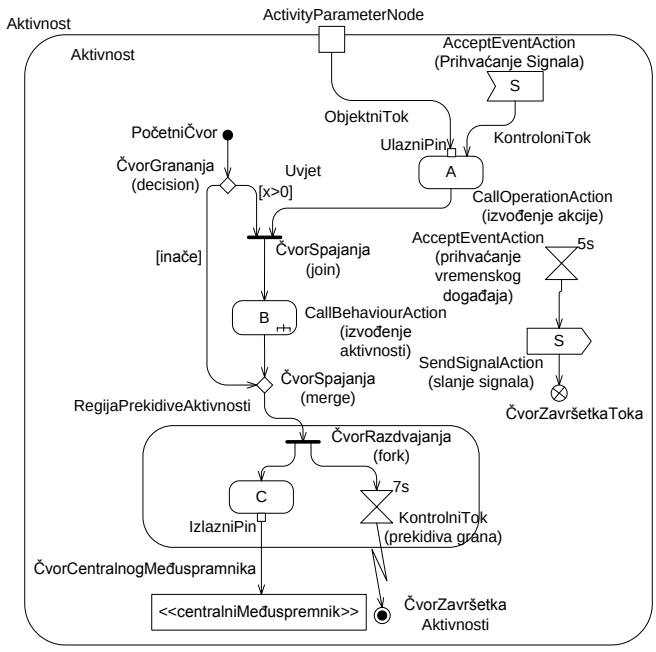
G. Vremenska kontrola

Definira zaustavljanje poslovnog procesa na određeno vrijeme.

H. Prekid procesa

Specificira se mogućnost prekida procesa na temelju definiranih kriterija kao što je tehnička nemogućnost realizacije usluge.

Na temelju definiranog modela poslovnog procesa bilo je potrebno odabrati alat za dizajniranje poslovnih procesa čija sintaksa podržava elemente modela. U [3] je predložen UML2 dijagram aktivnost čiji elementi odgovaraju elementima modela. Slika 1 prikazuje sintaksu UML2 dijagrama aktivnosti.



Slika 1: Sintaksa UML 2 dijagrama aktivnosti

UML (Unified Modeling Language) je standardizirani alat koji ima široku primjeru u grafičkom modeliranju softverskih rješenja. Alat koristi više različitih dijagrama koji modelirani sustav prikazuju iz različitih perspektiva. Dijagram aktivnosti definira dinamičko ponašanje sustava. To ga uz njegovu sintaksu čini vrlo korisnim alatom za dizajniranje poslovnih procesa i specificiranje potrebnih promjena u sustavima za operativnu podršku kojima se upravlja odvijanjem poslovnih procesa.

No, iako je odabrani dijagram aktivnosti alat kojim se poslovni procesi za realizaciju usluga u telekomunikacijskoj mreži mogu vrlo dobro dizajnirati njegovo korištenje samo po sebi nije dovoljno bez definiranog koncepta za modeliranje i specificiranje složenih poslovnih procesa. Naime, poslovni procesi postaju sve složeniji. Tu se ne radi samo o realizaciji jedne određene usluge, već se vrlo često radi o procesima kojima se istovremeno korisniku realizira više usluga. Tu su i procesi migracije usluga gdje se prilikom realizacije određene usluge pokreće i migracija postojećih usluge s jedna na drugu tehnologu. Na području telekomunikacija djeluje veći broj operatora te je potrebno osigurati sve veći broj različitih složenih poslovnih procesa koji upravljaju prelaskom korisnika između operatora.

S druge strane sustavi za operativnu podršku kojima se upravlja poslovnim procesima imaju svoju arhitekturu i to je potrebno uzimati u obzir prilikom specifikacije potrebnih promjena u sustavu. Zaista je veliki izazov modelirati poslovne procese i specificirati programsku podršku na način da su složeni poslovni procesi jasni sa stanovišta potrebnih akcija za realizaciju usluga s jedne strane i da jasno i precizno definiraju potrebne promjene u sustavima sistemske operativne podrške s druge strane.

Zato je bilo potrebno definirati koncept kojim se uspješno mogu modelirati složeni poslovni procesi i specificirati potrebne promjene u sustavima. Jedan od mogućih načina, koji se i predlaže u ovom radu, je korištenje scenarija i osnovnih podprocesa kojima se modeliraju poslovni procesi i dizajnira sistemska operativna podrška [5].

Svaki scenarij predstavlja poslovni proces koji se može pokrenuti prilikom realizacije ili korištenja telekomunikacijskih usluga. Složeni poslovni procesi na čije se modeliranje korištenjem koncepta scenarija želi naročito utjecati su na primjer migracija govorne usluge s PSTN (Public Switched Telephone Network) mreže na VoIP (Voice over Internet Protocol) i istovremeno uključenje pristupa Internetu i IPTV-a (Internet Protocol Television) ili istovremeno isključenje govornih i širokopojasnih usluga (pristup Interentu i IPTV) te portiranje numeracije i iznajmljivanje lokalne petlje (pretplatničke parice) drugom operatoru.

Scenarij se sastoji od osnovnih cjelina (podprocesa) kojima se definira skup akcija koje je potrebno odraditi u određenom periodu odvijanja poslovnog procesa. Primjer takvog podprocesa su konfiguracije na mrežnim elementima operatora koje je potrebno odraditi prije aktivnosti na području pristupne mreže ili aktivacije usluga kod korisnika. Podprocesi se, kao i osnovni procesi, modeliraju kroz dijagrame aktivnosti. U modeliranju podprocesa dalje se specificiraju akcije koje je potrebno odraditi u tom dijelu odvijanja poslovnog procesa.

Na ovakav se način postiže modeliranje poslovnih procesa kroz više razina složenosti. Prva razina definira scenarij i specificira osnovne cjeline (podprocese) od kojih se sastoji scenarij. Ovo je najniža razina složenosti i kroz nju se poslovni proces jasno modelira opisom osnovnih

cjelina od kojih se sastoji. Vrlo važan element modela su napomene koje se pridjeljuju svakoj cjelini i u kojima se opisuju osnovne aktivnosti koje je potrebno pokrenuti u sklopu cjeline. Ova razina modela vrlo je bitna u kontekstu analize E2E (s kraja na kraj) procesa kojeg operator treba provesti kako bi od korisnika preuzeo zahtjev, realizirao telekomunikacijsku uslugu i konačno naplatio svoje usluge.

Kroz ostale razine složenosti specificiraju se potrebne akcije za realizaciju telekomunikacijske usluge. Podprocesi kojima se modeliraju cjeline u sklopu scenarija mogu u sebi također sadržavati podprocese koje specificiraju proces kroz veću razinu složenosti. Na ovim razinama složenosti specificiraju se detalji potrebni za implementaciju potrebnih promjena u sustavima sistemske operativne podrške.

Ovakvim konceptom modeliranja poslovnih procesa za realizaciju telekomunikacijskih usluga postiže se jasno, strukturirano i precizno specificiranje potrebnih akcija u sklopu poslovnog procesa.

Drugi izazov je kako razvijeni model poslovnog procesa koristiti za specifikaciju potrebnih promjena u postojećoj arhitekturi sistemske operativne podrške, a da istovremeno model može poslužiti za definiranje poslovnih procesa koje je potrebno realizirati kroz novu arhitekturu sistemske operativne podrške.

Sami koncept scenarija i osnovnih cjelina od kojih se sastoji scenarij daje već dobru osnovu za rješenje ovog izazova. Scenarij opisuje procese na način da se definiraju osnovne akcije koje je potrebno poduzeti za realizaciju telekomunikacijske usluge bez obzira na arhitekturu sistemske programske podrške. I u ostalim razinama složenosti kojima se specificiraju detalji za implementaciju sistemske programske podrške elementi modela prvenstveno su usmjereni na ono što je potrebno realizirati, a ne kako. Način implementacije kroz postojeću arhitekturu sustava opisan je ili odvojeno od samog modela ili u sklopu modela, ali jasno odvojeno od ostalog opisa koji se prvenstveno odnosi na opis što je potrebno realizirati. Time se omogućava korištenje modela poslovnog procesa i za specifikaciju poslovnih procesa za realizaciju kroz drugu arhitekturu sistemske programske podrške ili barem definiranje osnovnih cjelina (podprocesa) i poslovnih pravila koje je potrebno osigurati kroz tu novu arhitekturu sustava.

III. PRIMJER SPECIFIKACIJE

Na slici 2 je prikazan primjer modeliranja scenarija. Model se odnosi na specificiranje poslovnih procesa kojima se upravlja kroz sustav za upravljanje tijekom poslova (workflow). Scenarij modelira složeni poslovni proces kojim se realizira migracija govorne usluge iz PSTN mreže u VoIP, portiranje (prijenos) numeracije u govornu mrežu drugog operatora te potrebna konfiguracija mrežnih elemenata kojom se omogućava prijenos govorne usluge preko ADSL pristupa iznajmljenog drugom operatoru. Ovo je vrlo složeni scenarij koji u sebi sadrži specifikaciju aktivnosti koje su potrebne za istovremenu realizaciju više usluga i više različitih procesa.

Poslovni proces se sastoji od više podprocesa koji definiraju cjeline unutar kojih je potrebno izvršiti određene aktivnosti u svrhu uspješnog odvijanja poslovnog procesa predstavljenog ovim scenarijem. Prva cjelina je ispitivanje tehničke mogućnosti u sklopu koje je potrebno ispitati postoje li potrebni mrežni kapaciteti za realizaciju usluga. Nakon toga se pokreću konfiguracije mrežnih elemenata potrebnih za realizaciju usluga i procesa. Kako se u ovom slučaju radi o poslovnom procesu koji se tiče i drugog operatora vrlo je bitno poštivati vremensku komponentu. Konfiguracije mrežnih elemenata koje imaju utjecaj na rad već postojećih funkcionalnosti potrebno je izvršiti u točno definirano vrijeme. To je vidljivo u napomeni kojom se opisuje podproces „Konfiguracija DSLAM-a“.

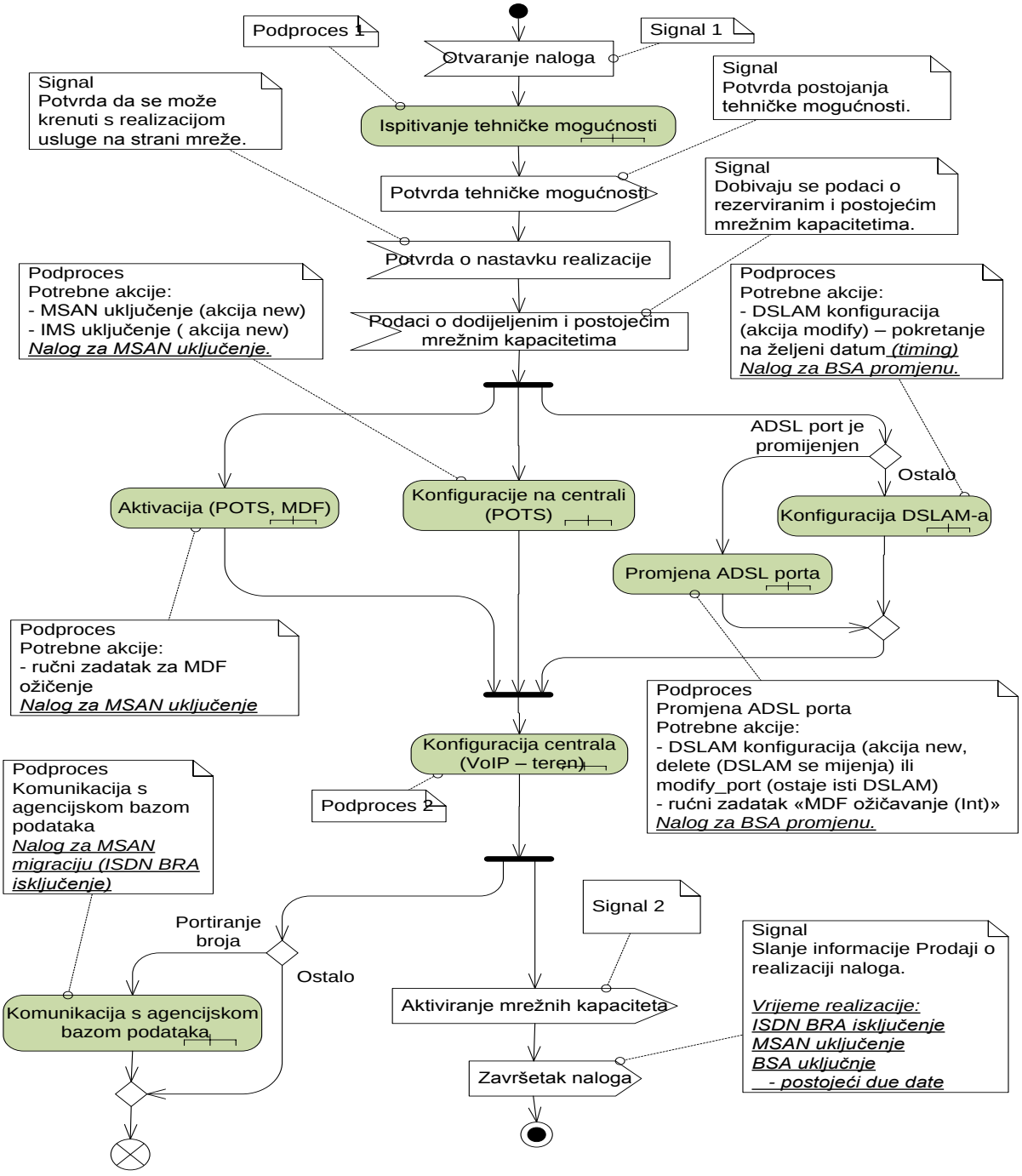
Nakon realizacije potrebnih konfiguracija mrežnih elemenata u definirano vrijeme može se pristupiti oslobađanju mrežnih elemenata koji više nisu potrebni nakon realizacije ovog scenarija, tj. složenog poslovnog procesa. To je definirano u podprocesu „Konfiguracija centrala (VoIP – teren)„. Ranijim pokretanjem ovog podprocesa dovelo bi se do prijevremenog prekida usluga što se svakako želi izbjeći.

Konačno, nakon provedenih aktivnosti pokreće se podproces kroz koji se obavještava baza podataka Agencije zadužene, između ostalog, i za definiranje odnosa između operatora da je portiranje brojeva između operatora izvršeno.

Model jasno opisuje osnovne cjeline i akcije koje je potrebno poduzeti, a i jasan je vremenski redoslijed i raspored akcija za uspješnu realizaciju složenog procesa. U modelu su specificirane i pojedinosti bitne za implementaciju poslovnog procesa kroz postojeću arhitekturu sistemske operativne podrške (označeno italic fontom). Te su pojedinosti jasno razgraničene od dijela modela kojim se modelira poslovni proces gledajući na aktivnosti u mreži potrebne za realizaciju usluge. Time se postiže mogućnost uspješnog korištenja modela za specificiranje potrebne sistemske podrške u postojećoj sistemskoj operativnoj podrški, a istovremeno i za modeliranje procesa koje je potrebno podržati kroz novu arhitekturu sustava buduće sistemske operativne podrške.

Na slici 3 prikazan je podproces kojim je opisana jedna od cjelina opisanog scenarija. Podprocesom se kroz dijagram aktivnosti opisuje dio procesa na način da se specificiraju potrebne promjene već do detalja potrebnih za implementaciju u sustavima za podršku. No, i na ovoj razini modela i dalje se zadržava princip da se specifikacije i opis elemenata modela odnose na potrebne aktivnosti u mreži, a ne i na postojeću arhitekturu sustava. Ukoliko je potrebno dodatno specificirati, recimo, neki od uvjeta kojima se kroz čvor odluke definira u koju granu će ići tijek poslovnog procesa to se radi uz dodatno specificiranje (na primjer tablica kroz koju se postojeći uvjet u modelu preslikava u uvjet koji u sebi sadrži tehničke detalje vezane uz arhitekturu i strukturu sustava). Naravno, postoje još i dodatni detalji koji se mogu specificirati u zasebnim dokumentima, a sami model ostaje „neopterećen“ detaljima vezanim uz postojeću strukturu i arhitekturu sustava.

Vrlo je bitno da model sadrži i podatke koje je potrebno razmjenjivati između sustava koji sudjeluju u upravljanju poslovnim procesima za realizaciju telekomunikacijske usluge te podatke potrebne za aktere koji sudjeluju u poslovnom procesu realizacije telekomunikacijske usluge.



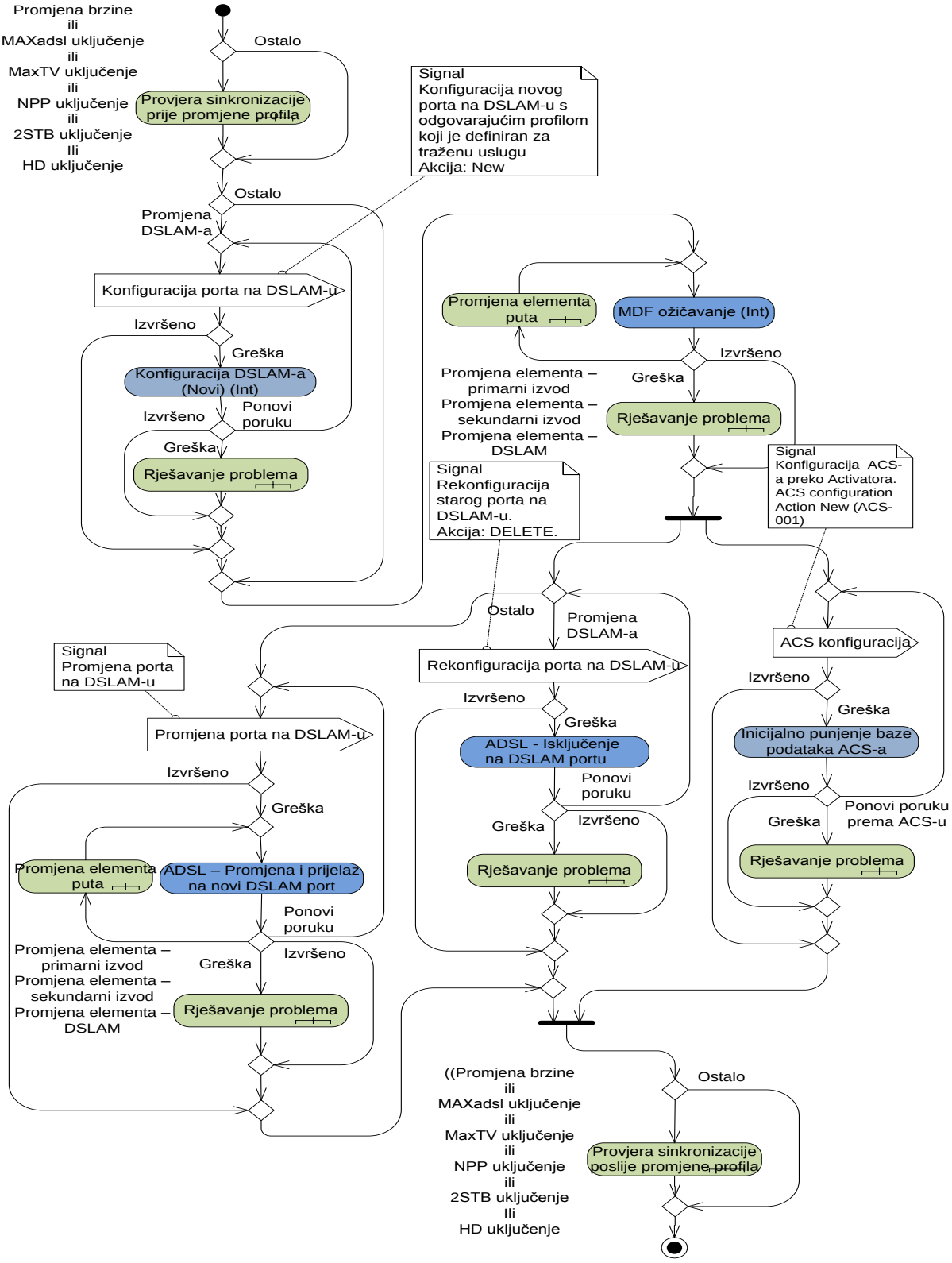
Signal 1 – Iniciranje otvaranje naloga na strani mreže
Tri naloga iz Donata (CRM sustav):- ISDN BRA isključenje (razlog «Migracija ISDN», produkt «Number portability»; MSAN uključenje (uključenje «Migracija ISDN»); BSA promjena
Kod scenarija: ISDN_TO_MSAN_BSA_CHANGE

Signal 2 – Aktiviranje mrežnih kapaciteta.
Koodinacija naloga prema DIS-u (sustav koji sadrži podatke o mrežnim kapacitetima): Prvot – ISDN BRA isključenje (akcija confirm); Drugo– MSAN uključenje (akcija confirm); Treće – BSA uključenje (akcija confirm)
Storniranje: Prvo – BSA uključenje (cancel); Drugo – MSAN uključenje (cancel); Treće – ISDN BRA isključenje (cancel)

Podproces 1 - Ispitivanje tehničke mogućnosti na strani mreže za ADSL i rezervacija potrebnih mrežnih kapaciteta
Koordinacija (čeka se na akciju PROCEED u prethodnom nalogu) prema DIS-u: Prvo – ISDN BRA isključenje (akcija start); Drugo - MSAN uključenje (akcija start); Treće – BSA change (akcija start)

Podproces 2- Konfiguracija IN centrale (akcija new; ako broj pripada PSTN numeraciji), PSTN centrale (ICS-8), IMS centrale (akcija UNLOCK); isključenje brojeva na ISDN BRA priključku (akcije disconnection numbering; ICS-8 add za broj koji migrira na VoIP; ICS-6 add za broj koji se portira drugom operatoru
Nalog za MSAN uključenje.

Slika 2: Scenarij složenog procesa



Slika 2: Podproces „Promjena ADSL porta“

IV. ZAKLJUČAK

Poslovni procesi koji se koriste na području telekomunikacijske mreže za realizaciju usluga vrlo su složeni što je rezultat vrlo složenog poslovnog okruženja. Da bi se uspješno upravljalo takvim procesima potrebna je upravljačka mreža kojom se osiguravaju potrebne aktivnosti za realizaciju usluge.

Ključnu uslugu u definiranju poslovnih procesa ima specifikacija. Specifikacija treba precizno definirati sve potrebne promjene u sustavima kako bi se osigurali potrebni poslovni procesi te jasno modelirati poslovne procese tako da se može jasno razaznati potrebni slijed aktivnosti u sklopu procesa. Koncept scenarija i prikaz scenarija kroz definirane cjeline kojima se opisuje određeni slijed aktivnosti kao dijela sveukupnog poslovnog procesa jedan je od načina kojim se pridonosi preciznosti i jasnoći specifikacije i modela poslovnog procesa. Na taj se način ne dolazi samo do uspješne specifikacije potrebnih promjena u postojećoj sistemskoj operativnoj podršci, već se model poslovnog procesa može uspješno koristiti i u definiranju i modeliranju poslovnih procesa koje je potrebno provesti kroz novu arhitekturu sistemske operativne podrške.

LITERATURA

[1] Musa K., Markulin D., "Realizacija složenih procesa uključenja kroz sustave za podršku", *Mipro 2009, HT d.d.*, 2009.

[2] Markulin D., Kunštić M., " Formal modeling business management of public telecommunication network", *Mipro 2010, HT d.d.*, 2010.

[3] Markulin D., Musa K., Kunštić M., "Using UML 2 activity diagram for visual business management modeling", *Mipro 2011, HT d.d.*, 2011.

[4] Markulin D., Musa K., " Proposal for specification of complex processes for telecommunication service provisioning in existing IT environment", *Mipro 2010, HT d.d.*, 2010.

[5] Markulin D., Musa K., "Proposal for business process modelling and development in condition of complex telecommunication market ", *Mipro 2012, HT d.d.*, 2012.

SERVICES

Semantic Approach to Accommodation & Booking Related Web Services

I. Pavković* and I. Štajduhar**
* Vienna University of Technology, Vienna, Austria
** Faculty of Engineering, Department of Computer Engineering, University of Rijeka, Croatia
ivan.pavkovic@student.tuwien.ac.at

Abstract - There are several types of information that tourists most often search for when planning a trip. Commonly, these are information about the location, travel, accommodation and nearby attractions. Almost all the knowledge that an average tourist needs already exists, but the current knowledge is highly decoupled and its aggregation is time consuming. Another problem with tourism-related Web services is that they do not correlate with each other. The information they share is often represented differently, is not connected and is sometimes even in contradiction. In this paper, we provide an overview of the differences in data representation between most popular Web tourist services, and introduce the *A&B* ontology that covers their data representation and enables the service discovery. Additionally, we propose a framework that is connected to the aforementioned Web services, translates the data to the *A&B* ontology, links the data and provides a single endpoint for multiple Web services.

I. INTRODUCTION

The rise in number of online accommodation and booking related Web services, provide the users huge amounts of data about different locations, accommodation, attractions, travel and similar information. In this paper we distinguish three different roles: (I) accommodation & booking related websites (Web service providers), (II) service providers (e.g. accommodation owners, attraction managers etc.) and (III) Clients.

A recent Google's travel study [1] discovered that 78% of the participants choose online sources for planning their travel activities. It is also worth noting that 85% of leisure travelers are undecided on a brand, when choosing their vacation packages. Moreover, the consultants of *Expedia*, one of the largest booking Web services, emphasised in their latest study [2] that travelers from UK and USA pay visits to 36.5 different travel websites in a 45 day period, prior to their booking. Additionally, there are several other problems regarding online travel services, e.g.:

- The data and knowledge provided by different service providers are decoupled, and its aggregation is time consuming.
- The data are redundant and inconsistent. Different services often provide the same data (e.g. one hotel is offered on multiple services), but they represent the same data differently.
- Web service providers are unable to distinguish themselves from one another, i.e. they are unable to clearly point out their pros and cons, which would help the clients choose their service.
- Accommodation provider's goal is to maximize the visibility of his objects, and possibly to offer them to as many booking-related Web service providers as possible. However, managing and updating data over a large number of various services is a tedious and repetitive task.

Accommodation-related Web services rarely share their data. Moreover, knowledge sharing rarely exists or is completely nonexistent. Therefore, the today's advanced Web 2.0 technologies and techniques, such as *Semantic Web*, service discovery and *Linked Open Data* (LOD) are rarely applied.

In this paper, we weigh the importance of *Semantic Web* technologies in tourism, travel and booking related Web-services domains. In knowledge management and *Semantic Web* research areas, ontologies are considered essential in order to describe various concepts and relationships between them [3]. Ontologies are formal specifications of a shared conceptualization of a domain [4]. They enable us to use shared vocabulary within a domain, thus standardizing the communication and solving the interoperability problem within a large heterogeneous network. These are, however, not of much use without the appropriate implementation within the domain. *Linked Data* is the usage of *Resource Description Framework (RDF)* and the *Hypertext Transfer Protocol (HTTP)*, for the purpose of publishing and connecting structured data on the Web, where different data sources point to each other.

García-Crespo et al. [3] state that tourism is a data rich domain, and depending on the purpose of the study, different approaches should be considered. This paper is organized as follows. In the following section, we compare the currently available ontologies in the tourism domain, and evaluate their applicability to our domain of accommodation and booking-related Web services. In section III, we analyse the data representation used by most popular services, describe the process and define a new universal ontology for tourism-related Web services. In section IV, we propose a system that is connected to

The work presented in this paper is partially supported by the University of Rijeka research grant 13.09.2.2.16.

TABLE I. ANALYSIS OF EXISTING TOURISM-RELATED ONTOLOGIES. THE COVERAGE OF RELEVANT TOURISM CONCEPTS AND THE APPLICABILITY TO ACCOMMODATION-RELATED WEB SERVICES SCORES WERE ESTIMATED.

Ontology	Main Concepts	Coverage of relevant tourism concepts [a]	Covered domains	Applicability to accommodation-related Web services [b]
E-Tourism	What, Where, When, Tourist	2	Activities, Locations, Accommodation, Weather	2
Harmonise	Events, Accommodation, Attractions, Gastronomy	3	Events, Gastronomy	3
QALL-ME	Country, Destination, Site, EventContent, Event, Transportation, Contact, CreditCard, Currency, Facility, Genre, Language, Location, Period, PersonOrganization, Price, Room	4	Events, Locations	3
Hontology	Accommodation, Facility, Room, Service/Staff, GuestType	4	Accommodation	4
ACCO	Accommodation Feature, BedDetails, Brand, DeliveryMethod, PriceSpecification, Warranty scope, Offering, Document…	2	Accommodation offers	2 (as extension only)
cDOTT	MeansOfTransport, Room, Accommodation, Attraction, FoodServiceOrganization, GeographicalRegion, Activity, Event, Booking, Reservation, Ticket, Voucher, CityPass, AttractionPass	5	Transport, Accommodation, Attractions, Locations, Calendar	4

a. range: from 1 (narrow) to 5 (broad)

b. range: from 1 (not applicable) to 5 (excellent applicability)

multiple accommodation-related Web services. Our system aggregates the data from multiple sources, translates the data, connects the data and represents it with our universal ontology, finally providing a single endpoint for multiple sources.

II. OVERVIEW OF THE EXISTING TOURISM-RELATED ONTOLOGIES

There are more than several ontologies that try to cover the aspects of e-Tourism and related areas. Many of the existing ontologies are specific to certain use-case or organization-specific scenarios. Some of the more popular ontologies, those more relevant to our domain of accommodation-related Web services, are described next. A brief analysis of the presented ontologies is summarised in Table I.

A. E-Tourism ontology

E-tourism ontology [5] was built upon the idea of further development of Web applications. It gives an answer to three questions: (I) What – What does a tourist see (activity); (II) Where – Where are the interesting places for a tourist to stay (location, accommodation); and (III) When – When can the tourist visit a particular place (timetable, weather). These three questions represent the basic concepts (classes), along with the forth concept, Tourist, where the tourist and his preferences are described. This ontology, however, covers a rather small domain, does not use standardised taxonomy, does not cover concepts widely used in tourism-related applications today (e.g. reviews, calendar), and does not recognize Web service providers, nor the concept of a service provider.

B. Harmonise ontology

Harmonise ontology [6] was derived from the EU project *Harmonise*, which was aimed at data-interchange standardisation. Because various tourist organisations use different data formats, the purpose of *Harmonise* was to exchange information without needing to change the data scheme. Several organisations participated in its development, e.g. *WTO (World Tourism Organisation)*, *eCTRL (e-Commerce and Tourism Research Laboratory)*, and several national tourist boards. It is represented with *RDF*, and covers events (conferences, performances and sports), accommodation (hotels and guest houses, disregarding private holiday apartments and camp sites), attractions and gastronomy domains. The project was, however, discontinued and the ontology is no longer available online for free use.

C. QALL-ME ontology

QALL-ME [7] is yet another EU-funded project that led to the development of a domain-specific ontology for multilingual and multimodal question answering in the tourism domain. It has seventeen classes: Country, Destination, Site (accommodation, gastro, attraction, and infrastructure), EventContent, Event, Transportation, Contact, CreditCard, Currency, Facility, Genre, Language, Location, Period, PersonOrganization, Price and Room. *QALL-ME* ontology covers a wide area that can be reused for the domain of accommodation-related Web services. On the other hand, it has a narrow definition of Accommodation (Bed and Breakfast, Campsite, Chalet, Cottage, Hostel, Hotel, Resort), Rooms (CinemaRoom, ConferenceRoom, Guestroom, RestaurantRoom), and the Facilities are not even defined (e.g. Pool, Air Conditioner). Understandably, taking into

account its initial purpose, the concept of a Web service provider does not exist.

D. Hontology

Hontology [8] is a multilingual ontology for the accommodation sector. It was built on concepts similar to *QALL-ME* ontology, and brought in alignment with structured content on *DBpedia.org* and *Schema.org*. *Hontology* was initially created, considering a scenario of applications, processing reviews about accommodation present in modern Web services. Its main concepts are Accommodation, Facility, Room, Service/Staff and GuestType. Unlike *QALL-ME* ontology, *Hontology* utilizes expanded concepts, e.g. Accommodation (IceHotel, CapsuleHotel, etc.), Facility (HeatingSystem, Parking, Wi-FiZone, etc.) and Rooms (SingleRoom, DoubleRoom, StudioApartment, etc.). The concept of a service provider and a Web service provider, however, do not exist. Moreover, a calendar implementation does not exist, also.

E. ACCO Accommodation Ontology

ACCO Accommodation Ontology serves as an extension to *GoodRelations* ontology [9, 10]. *GoodRelations* ontology is a lightweight ontology for annotating offerings and other aspects of e-commerce on the Web. It is the only *Web Ontology Language* (OWL) ontology, officially supported both by Google and Yahoo. *ACCO Accommodation Ontology* [11] is a Web vocabulary for hotels, vacation homes, camping sites, and other e-commerce accommodation offers. It mostly focuses on creating price features. For accommodation-related Web-services domain it can only serve as an extension, used for achieving better search-engine visibility.

F. cDOTT ontology

cDOTT ontology [12] integrates modularised ontologies such as *W3C Time* and *W3C Geo*, that center around their core domain ontology for the tourism sector. The motivation for the creation of such approach is that the process of developing and maintaining a single tourism ontology that covers the whole tourism market, including geographical, temporal, and user-related information, would be very tedious. This approach leads to a loose coupling of ontologies, while retaining flexibility and extensibility. Its core ontology covers a broad range of tourist concepts, also covering the temporal context which can be applied for reservation calendar purposes. It also distinguishes service providers for attractions, but not for accommodation service providers or Web service providers.

G. Summary of the existing, tourism-related, ontologies

Prantner et al. [13] suggest that the current state of domain ontologies is quite limited and unsatisfactory, due to the specific targeting to various different organisations, sectors and applications.

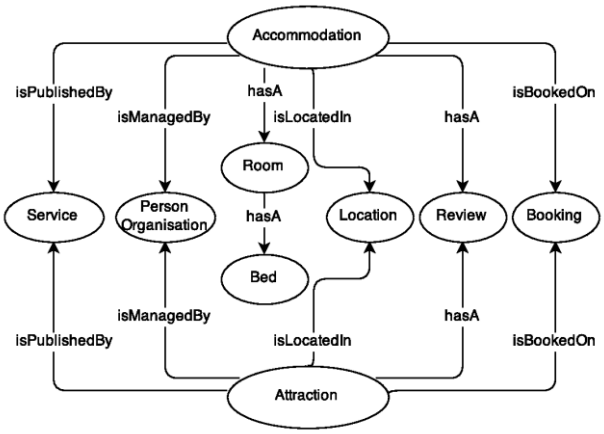


Figure 1. *A&B* ontology schema – basic concepts and data properties.

Summing it all up, most of the ontologies lack standardisation, agreed upon by multiple parties, and only a few offer a broad coverage of relevant tourism concepts. Moreover, none of the aforementioned ontologies offers a machine-accessible semantics, describing a network structure that could be applied to accommodation-related Web services. According to Ding et al. [14], this causes problems and affects content providers, intermediaries and content users.

III. ACCOMMODATION AND BOOKING (A&B) ONTOLOGY

The *A&B* ontology was developed, based on the *Ontology Development 101* method, proposed by Noy and McGuinness [15]. Our idea was to build an ontology that covers a wide range of accommodation and booking related Web services:

- Booking Web services – e.g. *Booking.com*, *Expedia.com* and *AirBnB.com*
- Hospitality exchange Web services – e.g. *Couchsurfing.com* and *BeWelcome.com*
- Accommodation swapping Web services – e.g. *HomeExchange.com* and *HomeForSwap.com*
- Travel-related Web services for reviews – e.g. *TripAdvisor.com*

During the ontology development, we focused our interests on three key roles in accommodation-related Web-services domain:

1. Accommodation & Booking Web service provider – Web services that provide certain data
2. Service provider – providers of real-world services, such as accommodation owners, hotel management, owners and managers of attractions
3. Clients – users of a Web service

We constructed a tourism-related *OWL* ontology. *OWL* is a vocabulary extension of *RDF*, and it consists of

classes (concepts within a domain), class-hierarchy (taxonomy of concepts), properties and property values, relations between classes (inheritance, disjointness, and equivalence), restrictions (cardinalities) and individuals.

In section II, we identified the similarities applicable to accommodation-related Web services with two existent ontologies, *Hontology* [8] and *cDOTT* [12], which were partially reused in our ontology. Semantic snippets, such as *Schema.org* and *Microformats.org* were also under consideration.

A. Main concepts

We defined the following main concepts in the *A&B* ontology, within the domain of accommodation-related Web services: Location, Attraction, Review, Room, Bed, PersonOrganization, Booking and Service. We describe the concepts and differences in concepts that differentiate our ontology, next.

Accommodation

The concept of accommodation is common to all tourism ontologies. We took a specific approach in order to define its subclasses and data properties. For defining subclasses (types of accommodation), we examined the data from various booking-related Web services (e.g. *AirBnB.com*, *Booking.com*, *Flipkey.com*), *Schema.org* and *DBpedia.org*. The overview of the accommodation types found is presented in Table II.

We added a broad range of accommodation types to our ontology. We chose the concepts that are more specific, e.g.:

- Watermill and Windmill on *FlipKey.com*, over Mill on *VRBO.com*
- Camper/Mobile Home and Camper/Van on *FlipKey.com* over a Camper/RV on *AirBnB.com*

Finally, the concept of accommodation has over fifty defined accommodation type subclasses.

In order to define data properties, we used a similar approach, where we compared the data representation over various Web services. There are over two hundred different data properties. Due to page limit, they are not described in detail in this paper. Regardless, the method we used can be roughly described using the following principles:

- If there were conflicts on data properties between different services, e.g. property Shower had a Boolean (has shower/does not have shower) value on one service, and an integer value on another service (has two showers), we used the version supported by more Web services;
- If a certain property was defined as string value (description) on one service, and as a Boolean value on another, we kept them both, since they semantically provide different meanings.

Service

In the *A&B* ontology, a specific accommodation or attraction is managed by a certain person or organisation, and that refers to a service provider role. Certain accommodation or attraction is, however, published by a certain service, where we refer to the role of a Web service provider, e.g. *Apartment in Vienna* is managed by *Jane*, and is published on *AirBnB.com*. Therefore, the primary purpose of the service concept is to identify the Web service, an instance belongs to.

Another purpose of the service concept is to resemble the *Universal Description, Discovery and Integration (UDDI)* [16] mechanism with Web services. *UDDI* is a standard by which Web services broadcast themselves on the Internet, as a mechanism to register and locate different services, in order to enable service composition.

TABLE II. OVERVIEW OF ACCOMMODATION TYPES ON POPULAR WEB SERVICES, SCHEMA.ORG AND DBPEDIA.ORG

URL	**Supported accommodation types**
AirBnB.com	Apartment, House, Bed&Breakfast, Loft, Cabin, Villa, Castle, Dorm, Treehouse, Boat, Plane, Camper/RV, Igloo, Lighthouse, Yurt, Tipi, Cave, Island, Chalet, Earth House, Hut, Train, Tent, Other
Booking.com	Apartment, Bed&Breakfast, Country House, Guesthouse, Homestay, Farm Stay, Lodge, Vacation Home, Villa, Chalet, Hotel, Motel, Inn, Capsule Hotel, Condo Hotel, Resort, Resort Village, Campground, Luxury Tent, Riad, Ryokan, Love Hotel, Boat
FlipKey.com	Apartment, Bed&Breakfast, Barn, Beach Hut, Boathouse, Bungalow, Camper/Mobile Home, Camper/Van, Castle, Cave House, Chalet, Chateau, Condo, Converted Chapel, Cottage, Farmhouse, Finka, Fort, Gite, Guesthouse, Hotel Apartment, House, Houseboat, Lighthouse, Lodge, Log Cabin, Manor House, Narrowboat, Penthouse, Riad, Shepherd's Hut, Ski Chalet, Studio, Tented Camp, Tipi, Tower, Townhouse, Treehouse, Trullo, Villa, Watermill, Windmill, Yacht, Yurt
HouseTrip.com	Apartment, Boat, Castle, House, Studio
VRBO.com	Apartment, Barn, Boat, Bungalow, Cabin, Castle, Chalet, Chateau/Countryhouse, Condo, Estate, Farmhouse, Hotel, House, Houseboat, Lodge, Mill, Studio, Tower, Townhome, Villa, Yacht
Wimdu.com	Holiday Home, Apartment, Private Room, Chalet Or Cabin, Castle, Farmhouse, Treehouse, Boat, Car
Schema.org	Bed&Breakfast, Hostel, Hotel, Motel, Lodging Business, Residence
DBpedia.org	Beach Hut, Gîte, Tourist Home, Sleep Box, Camping Coach, Hotel, Motel, Safari Lodge, Houseboat, Hospitality Service, Holiday Cottage, Vacation Rental, Lodging, Villa, Home Exchange, Coaching Inn, Homestay, Wilderness Hut, Mountain Huts, Resorts, Hostels, Campsites, Bed And Breakfasts, Vacation Rental, Villas

There are several different approaches to service discovery with semantic technologies [17,18,19]. Ours, however, can be seen as a *KISS* (keep it simple, stupid) approach within our small domain: every web service provider marks accommodation or attractions of his own. This approach assumes information sharing between different services.

The final purpose of the service concept is to serve as a policy negotiator between different Web services. As such, a Web service publishes terms, under which it shares internal data. E.g., a certain Web service allows reuse of its own accommodation units to another service, under the condition of taking a commission of three percent. Additionally, from a service provider's perspective, the accommodation manager can state the conditions under which he will allow publishing the information on his or hers accommodation unit on other Web services (e.g. apartment owner allows reuse on another Web services that take a commission lower than ten percent, and whose payment is upon the guest's arrival).

Booking

Booking concept defines a temporal aspect in the *A&B* ontology, i.e. covering the time-related aspects in domain of accommodation-related Web services. This concept is usable both for Accommodation and Attraction classes. For Accommodation, its purpose is to be used as a reservation calendar. With the assumption of knowledge sharing between Web services, this concept also eliminates the need of changing the availability periods on all Web services, i.e. a Web service can access the unavailable periods from other services and form a real-time availability calendar. For several attraction types, it is used to represent the booked periods, e.g. a masseur that is already booked on a certain date and time or specifying the time and date properties for a needed car-rental service. In order to define the Booking concept, we used the same approach as in the *cDOTT* ontology [12]. We integrated the *W3C Time* ontology, in order to model temporal concepts. *W3C Time* ontology is used for describing the temporal content of Web pages and the temporal properties of Web services.

Other concepts

The Attraction concept is used to describe attraction and nearby places in a certain location. It was modelled in alignment to *Schema.org*' CivilStructure, Landform, LandmarkHistoricalBuildings, LocalBusiness, TouristAttraction schemas and hListing on *Microformats.org*.

Bed is a basic capacity unit for a certain accommodation type. The Bed concept is used to broaden the range of Web services that can use the *A&B* ontology, e.g. the accommodation listing on *Couchsurfing.com* has a shared room that has a couch bed.

The PersonOrganization concept is used to define the owners or managers of a certain accommodation or attraction. It was modelled in alignment to *FOAF* ontology [20], *Schema.org* Person and Organization schemas and hCard on *Microformats.org*.

The Location concept, used to describe a certain location, has the following subclasses: Place, Region and Country. These subclasses are interlinked: a region can be located in a country and a country can be located in a region. However, due to the complexity of geolocations, this concept was not completely elaborated and will probably be remodelled and integrated with an external ontology.

Finally, the Review concept, used to evaluate a location or a service, was modelled in alignment to *Schema.org* Review and AggregateRating schema and hReview on *Microformats.org*.

IV. FUTURE WORK

In the future, we are planning to develop a framework that implements and comprises of all the aspects of the *A&B* ontology. We propose the following simplified framework's architecture:

- Knowledge aggregator – the component whose task is to aggregate the data from multiple accommodation-related Web services. The data is accessed with usage of various APIs, data dumps, provided *SPARQL* endpoints [21] or data extraction techniques. This component can be seen as a knowledge aggregation mashup composition of various Web services.
- Translator & Connector – the component that translates the aggregated data to the *A&B* ontology. After the translation, the component searches for similarities between individuals. In cases of high similarities that have the same data, the individuals are connected as Linked Data.
- Frontend – a Web frontend that enables users to query different Web services, enables service providers to add their properties and define policies for sharing, enables Web services to use SPARQL endpoints of this framework and enables them adding their SPARQL endpoints to the framework.

The benefits of this approach were discussed in previous sections of this paper.

V. CONCLUSION

In this paper, we study the currently available tourism-related ontologies and analyse their applicability to the domain of accommodation and booking-related Web services. Furthermore, we introduce the *A&B* ontology, an ontology aimed to cover the domain of accommodation and booking-related web services. The ontology was built in alignment to data represented on the most popular booking-related web services, *W3C Time* ontology and semantic snippets such as *Schema.org* and *Microformats.org*.

Furthermore, we proposed a framework that fully utilizes the benefits of the *A&B* ontology: one that connects to various accommodation-related web services, translates the data to the *A&B* ontology, connects it and provides a single endpoint, both for users and for other Web services. This approach solves several current problems related to accommodation and booking-related Web services, such as data dispersion and inconsistency, limited visibility of individuals and undistinguished Web services.

REFERENCES

[1] "Google Travel Study," Ipsos MediaCT, 2014, available online: https://www.thinkwithgoogle.com/research-studies/2014-travelers-road-to-decision.html

[2] "A Look into Consumer Traveler Behavior," Expedia Media Solutions, 2014, available online: http://info.advertising.expedia.com/travelerspathtopurchase

[3] A. García-Crespo, et al., "SPETA: Social pervasive e-Tourism advisor," Telematics and Informatics, vol. 26.3, 2009, pp. 306-315.

[4] N. F. Noy, and C. D. Hafner, "The state of the art in ontology design: A survey and comparative review," AI magazine, vol. 18.3, 1997, 53-74.

[5] J. Cardoso, "Developing an owl ontology for e-Tourism," Semantic Web services, processes and applications, Springer US, 2006, pp. 247-282.

[6] M. Dell'Erba, O. Fodor, F. Ricci, and H. Werthner, "Harmonise: a solution for data interoperability," Towards the Knowledge Society, Springer US, 2003, pp. 433-445.

[7] S. Ou, V. Pekar, C. Orasan, C. Spurk, and M. Negri, "Development and Alignment of a Domain-Specific Ontology for Question Answering," LREC, 2008.

[8] M. Chaves, L. Freitas, and R. Vieira, "Hontology: a multilingual ontology for the accommodation sector in the tourism industry," 2012.

[9] M. Hepp, "Goodrelations: An ontology for describing products and services offers on the Web," Knowledge Engineering: Practice and Patterns, Springer Berlin Heidelberg, 2008, pp. 329-346.

[10] M. Hepp, "GoodRelations," 2011, available online: http://semanticWeb.org/wiki/GoodRelations

[11] M. Hepp, "ACCO: The Accommodation Ontology," 2013, available online: http://ontologies.sti-innsbruck.at/acco/ns.html

[12] R. Barta, C. Feilmayr, B. Pröll, C. Grün, and H. Werthner, "Covering the semantic space of tourism: an approach based on modularized ontologies," in Proceedings of the 1st Workshop on Context, Information and Ontologies, ACM, 2009.

[13] K. Prantner, Y. Ding, M. Luger, Z. Yan, and C. Herzog, "Tourism ontology and semantic management system: state-of-the-arts analysis," 2007.

[14] Ding, Ying, et al. "OnTourism: Semantic etourism portal." Proceedings of the 2nd International Scientific Conference of the e-Business Forum-E-Business in Travel, Tourism and Hospitality, Athens, Greece. Vol. 3. 2008.

[15] N. F. Noy, and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," 2001.

[16] F. Curbera, et al. "Unraveling the Web services web: an introduction to SOAP, WSDL, and UDDI," IEEE Internet computing, vol. 6.2, 2002, pp. 86-93.

[17] B. Medjahed, A. Bouguettaya, and A. K. Elmagarmid, "Composing web services on the semantic web," The VLDB journal, vol. 12.4, 2003, pp. 333-351.

[18] M. Klusch, B. Fries, and K. Sycara, "Automated semantic web service discovery with OWLS-MX," in Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems, ACM, 2006.

[19] D. Chakraborty, F. Perich, S. Avancha, and A. Joshi, "Dreggie: Semantic service discovery for m-commerce applications," Workshop on Reliable and Secure Applications in Mobile Environment, 20th Symposium on Reliable Distributed Systems, 2001.

[20] D. Brickley, and L. Miller, "FOAF vocabulary specification 0.98," Namespace document 9, 2012.

[21] B. Quilitz, and U. Leser, "Querying distributed RDF data sources with SPARQL," Springer Berlin Heidelberg, 2008.

Voice over Wi-Fi

D. Bilajbegovic
Ericsson Nikola Tesla d.d., Zagreb, Croatia
damir.bilajbegovic@ericsson.com

Abstract – Voice over Wi-Fi, also known as Wi-Fi Calling, is a new service introduced by mobile operators just few months ago. Voice over Wi-Fi enables establishing "ordinary", Circuit Switch like, calls via Wi-Fi network. Unlike Over The Top Voice over IP calls which establish calls via any type of IP network, Voice over Wi-Fi enables seamless integration to existing telecommunication core offering same look and feel as well as same security level but without any intervention made by customer. This paper describes Voice over Wi-Fi integration in operator's core network, show reusing of the existing security solutions and influence of this feature on the Mobile Operator behavior. As this feature is new and very interesting so it is going to be very interesting to see how operators are going to adopt this feature as well as what technology changes it will push.

I. INTRODUCTION

Traditionally Mobile Network Operators made their brand and revenue by offering Circuit Switch voice services to their customers. And for the years this was the main service and main source of income. But with the rapid development of IP data communication, this position of undisputed provider of voice communication services is threatened with by the new Over The Top (OTT) service providers.

Voice over IP services was first introduced for personal computers. They offered a free alternative for long distance calls and solution for free text communication (chatting) compared to SMS texting. The most known and still the synonym for such solutions was Skype.

With Smartphones development and advanced mobile data infrastructure, Voice over IP services was offered to the customer via his phone. These services were now offered to the customers "all the time" making them available not only via traditionally mobile networks but also via these new OTT platforms. In parallel to development of traditional Mobile networks, Wi-Fi networks started to the widely available, not only at customer home and office but also on various free public hot spots.

Customer had for the first time alternative solution for establishing voice calls. They could use:

- Traditional CS mobile network to establish and receive voice calls
- OTT solutions for establishing and receiving voice calls via Mobile packet network or Wi-Fi network, for just a fraction of the price

Traditional MNO are now facing the situation where they are providing only "data pipe" for the OTT operators, and even worse, they are not any more the exclusive "data pipe" provider for specific customer.

II. USER PERSPECTIVE

With the fact that the Mobile network operators are threatened to be neglected and left to invest a lot of money in network infrastructure while their revenues are shrinking, Voice over Wi-Fi solution appeared as one of the services that could change the market back in the MNO favor.

Voice over Wi-Fi enables establishing Voice over IP calls via Wi-Fi network. At first this looks like something that already exists but the difference is in the way how this solution is implemented compared to the OTT solutions.

Voice over Wi-Fi was introduced with presentation of Apple iOS 8 platform in September 2014. The iOS 8 based phones iPhone 6 and 6 Plus offered Wi-Fi Calling service natively supported [5]. Also iPhone 5s could be upgraded to support Wi-Fi calling. At the same time Samsung offered the same feature in their flagship models of smartphones.

Figure 1. Voice over Wi-Fi configuration on mobile phone

Customer with such phone has only to enable Wi-Fi Calling feature in the phone setting and this is all the interaction that customer needs to do [7]. Example of this configuration is showed in the Figure 1. Unlike OTT applications, it does not need to install any application, configure it, configure its user profile, nor link or maintain separate phonebook. Their look and feel is the same as with normal phone calls.

Native build in support for Wi-Fi Calling (as it is often referred by smartphone manufactures) is important for quick adoption and fast market penetration for this service.

III. TECHNICAL IMPLEMENTATION

In the core of mobile network operator a lot of changes needed to be introduced and modified to support Voice over Wi-Fi. MNO needs to introduce the IP Multimedia Subsystem (IMS) with all the modifications to the MNO network. Beside implementation nodes, MNO must also modify the MNO Operating and Support System (OSS) and Business Support System (BSS). This could be a problem as the implementation could be costly from financial and man power usage point of view.

Basic foundation for simpler Voice over Wi-Fi introduction is that Mobile Network Operator already has IMS or Voice over LTE (VoLTE) implemented in its network. Support VoLTE is big benefit for easier Voice over Wi-Fi introduction. MNO that already have VoLTE needs only to modify Packet Data Network Gateway (PDN) or/and introduce evolved Packet Data Gateway (ePDG) in their network to support the customers connected via Non-3GPP network (Figure 2). Additional modifications (configurations) need to be implemented on Policy and Charging Rules Function (PCRF) node and on the Authentication, Authorization and Accounting (AAA) Node to support Wi-Fi users.

In the traditional mobile network defined by 3GPP standards Packet Data Network Gateway (PDN Gateway) need to be upgraded to support User Equipment (UE) access via Trusted Non 3GPP IP Access network [1]. Trusted Non 3GPP IP access networks are typically Wi-Fi networks whose Access Points are maintained by the same MNO and which has already implemented solution for authentication and authorization of such users. These users attach to the PDN Gateway and via IMS register themselves in MNO network.

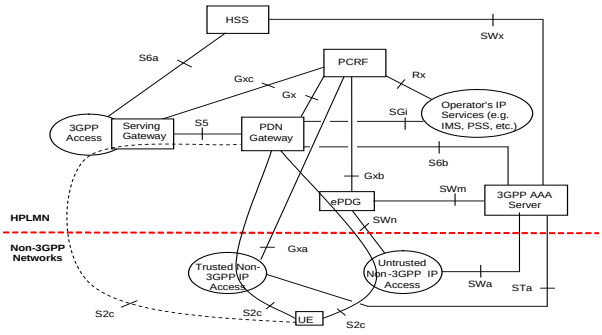


Figure 2. Voice over Wi-Fi configuration on mobile phone

For the users that are attached via Untrusted Non 3GPP IP Access networks, additionally evolved Packet Data Gateway (ePDG) node needs to be added to the network which enables establishing trusted connection between UE and PDN. ePDG perform authentication and authorization of the user and will create IP sec tunnel between ePDG and UE and GPRS Tunneling Protocol (GTP) Default tunnel between ePDG and PDN Gateway.

These users are typically connected via home, office or public free Wi-Fi network and represent majority of the typical Wi-Fi users and MNO does not have power to influence on the UE Access Point.

When performing Authentication and Authorization, the system uses UE already available security information stored in the UE SIM module and in the AAA/HSS server. What makes Voice over Wi-Fi really different from OTT solutions is the message exchange that is showed on Figure 3. Traditional OTT solutions will have only first two steps of message exchange. Voice over Wi-Fi solution, in step 3-8, uses the existing information about the UE (3), reuses it also during authentication procedure (4) and enforces it via already deployed network elements (5-8). Unlike the OTT solutions, MNO and UE do not need any additional security mechanism nor have any weak spots during initial configuration of the application (i.e. initial authentication and authorization of user). For the details please refer to the [3].

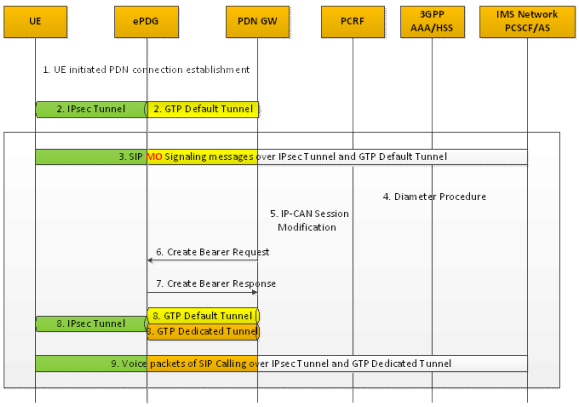


Figure 3. Voice over Wi-Fi SIP/IMS messagess exchange

After the user has been attached to the MNO network, UE starts usual IMS registration procedure. After successful registration, depending on the network configuration, UE can call or be called via Wi-Fi network or 2G, 3G or LTE network. When UE moves out of the Wi-Fi coverage, user will still be available via 2G, 3G and/or LTE network. If the user has already established Voice over Wi-Fi call, when loosing Wi-Fi coverage call can be handed over to the LTE (Voice over LTE call) and vice versa. Currently there isn't defined solution for the Wi-Fi call to be handed over from Wi-Fi to 2G / 3G network. As Voice over Wi-Fi is just introduced we could expect that the standardization of 2G/3G handover to Voice over Wi-Fi (and vice versa) is going to be the next step in standardization process and we will see it in following year or so.

The IMS itself needs only to be configured for additional type of users. As IMS is designed as access network independent system, adding support for Wi-Fi users is easy task. It might seem to be a complicated process to implement Voice over Wi-Fi but implementation is often done by simple software upgrade of existing nodes with some additional configuration. The most important and often time consuming task of every MNO - integration with OSS and BSS tools is minimal or non-existing. So MNO that wants to introduce Voice over Wi-Fi needs first to implement LTE network and then with easy modification of current network enable support for Voice over Wi-Fi.

IV. INFULANCE ON THE CUSTOMER AND MNO BEHAVIOUR

With implementation Voice over Wi-Fi solution, both customer and MNO are facing different behavior patterns.

MNO are now extending their network cover beyond their physical network infrastructure. It is as just as the MNO has turned up millions of base stations in just one day. MNO signal coverage can be now extended to the places where their mobile signal is very bad or it is not reachable at all. Typically in the city they are house basements, which are dig in the ground and the MNO signal is very weak. At the same time home Wi-Fi signal is typically stronger as it is only few meters above typical house basement. In the rural hilly areas, there are often some gaps where few houses are covered with weak or without MNO signal but if there is Wi-Fi access point, signal strength problem is solved.

These locations are typical for the places where MNO has network coverage. For instance T-mobile USA is with this feature providing higher network coverage, compensating lower 2G/3G/LTE network coverage compared to other Tier 1 operators in USA.

On the other hand, UE can access from the locations that are "roaming locations". When user is traveling abroad, income based roaming calls is split between home MNO and visiting (roaming) MNO. By offering Voice over Wi-Fi, MNO can reduce the price of such "roaming" calls and at the same time keep all the income based on such calls. Some MNO, like T-Mobile USA, priced these roaming calls the same was as it is local call, while other like France Telecom are treating this calls as ordinary roaming calls.

From the user perspective, Voice over Wi-Fi solution enabled "higher network signal" (MNO network coverage) and enabled lower roaming costs while not spending time in mobile phone configuration, maintaining different users profiles and different phone books. When the user is using Voice over Wi-Fi, it is going to be indicated by the icon and it will be only note that the customer is using Voice over Wi-Fi. User experience should be the same as it is traditional CS call.

Regarding availability of the smartphone devices, all the mayor phone manufactures implemented this feature in their flagship models and as the time of writing there are more than 30 models supporting Voice over Wi-Fi.

V. CONCLUSION

Introducing Voice over Wi-Fi enables MNO to extend their network coverage, increased availability of their services and extended potential number of customers.

Compared to the OTT operators, Voice over Wi-Fi enabled MNO to have the same or better network coverage and the same or better availability of their services. With "One Click" configuration procedure and uniform look and feel user experience MNO are in advantage position over the OTT operators.

From the standardization point of view, it is also going to be interesting to see how the 2G/3G handover to Voice over Wi-Fi (and vice versa) is going to be standardized as well as the policy of MNO how to handover the calls.

As Voice over Wi-Fi is launched just half a year ago it is going to be interesting to observe implementation of this solution at MNO network as well as the MNO price and marketing strategy for the customers.

REFERENCES

[1] Architecture enhancements for non-3GPP accesses, 3GPP TS 24.302, 3GPP, 2014, http://www.3gpp.org

[2] 3GPP Evolved Packet System (EPS), Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control plane (GTPv2-C), 3GPP TS 29.274, 3GPP, 2014, http://www.3gpp.org

[3] 3GPP System Architecture Evolution (SAE) – Security aspects of non-3GPP accesses, 3GPP TS 33.402, 3GPP, 2014, http://www.3gpp.org

[4] Evolved Packet System (EPS), 3GPP EPS AAA interfaces, 3GPP TS 29.273, 3GPP, http://www.3gpp.org

[5] Apple Live Special Event, Apple, September 17th 2014, http://www.apple.com/live/2014-sept-event/

[6] Wi-Fi calling web page by T-mobile USA, T-Mobile USA, 2015, http://www.t-mobile.com/offer/wifi-calling-wifi-extenders.html

[7] Make a call over Wi-Fi with iOS 8, Apple, 2015, http://support.apple.com/en-us/HT203032

An Example of Real Time QoE IPTV Service Estimator

N. Goran*, M. Hadžialić **
* BH Telecom JSC, Sarajevo, Bosnia and Herzegovina
** University of Sarajevo, Faculty of Electrical Engineering, Bosnia and Herzegovina
nermin.goran@bhtelecom.ba; mesud.hadzialic@etf.unsa.ba;

Abstract - this paper will consider an estimator which includes mathematical modelling of physical channel parameters as information carrier and the weakest links in the telecommunication chain of information transfer. It will also identify necessary physical layer parameters which influence the quality of multimedia service delivery or QoE (Quality of Experience). With the modelling of the above mentioned parameters, the relation between degradations will be defined which appear in the channel between the user and the central telecommunication equipment with domination of one media used for information transfer with certain error probability. Degradations in a physical channel can be noticed by observing the change in values of channel transfer function or the appearance of increased noise. Estimation of QoE IPTV (Internet Protocol Television) service is especially necessary during delivery of real time service. In that case the mentioned degradations may appear in any moment and cause a packet loss.

***Keywords* — Access network; Quality of Service; Quality of Experience; Quality of Physical Layer; Channel estimation;**

I. INTRODUCTION

Evaluation of video signal quality using different objective and/or subjective methods at application layer of OSI/ISO model has been tested in the point of presence of the tested signal or one of its components. During IPTV (Internet Protocol Television) service delivering there is no information about original signal on user's side; therefore QoE/QoS estimation in real time is a big challenge. If so, there are at least two possible scenarios. There can be feedback from users about service, such as crowdsourcing etc., which is used to warn network components about decrease of video signal quality as some authors researched in [1-5]. The second scenario is usage of available QoS indicators for quality estimation without necessary evaluation on the user's side as some authors researched in [6-10]. In the first case without already defined procedures that the user is expected to fulfil so the service provider can get a clear image of an upcoming problem, it cannot be unambiguously defined why or in what amount the service degradation occurred in particular time. It is also important to mention that this type of testing would take a lot of customer's time. In the second case estimator needs to be put between the user and service provider in order to give information about level of quality being delivered to the user. This refers to correlation between user subjective estimations and appearance of QoS (Quality of Service) indicator degradations in parts of telecommunication network. Thus, there are two benefits. First of all to get objective video signal testing, since video signal is estimated on the basis of objective measure parameters. Then to find the part of network which causes problem in distribution of IPTV service. Telecommunication network consists of three logical parts: service, core and access. Each of them has specific parameters that can be measured and used for estimating error number that appears during IPTV service delivery. At the same time, communication inside telecommunication network is carried out via protocols in accordance with OSI reference model. During a service delivery, MPEG-2 packets are transferred through OSI layers and certain parts of the packets e.g. the information from header can be used to ensure QoS. As we know, there is a huge number of transformations in message formed by OSI layer rules. The message passes through physical and data layer to network layer and so on. It is obvious that degradations in physical layer are propagated to user who uses and estimates the service by own subjective measure. This paper will consider influence of information carrier in access network as well as the weakest link in information carrier chain. Using estimation of the mentioned parameters, it is possible to find the relation between appearances of degradations on location user-telecommunication network with domination of one media for information transfer with a certain error probability. The problem is especially expressed during service delivery in real time with the possibility of appearance of the mentioned degradations which can cause packet loss and additionally complicate the estimation. The paper will also present the frame for mathematical modelling of transfer information loss through OSI reference model.

II. FRAME MODEL

Frame model begins at the service part of telecommunication network in the place where service is obtained by TV station. In that point, it is possible to define starting level of service quality using simple methods, due to having original signal characteristics necessary for distribution towards the user. Therefore, by

testing the parameters such as PSNR in particular frames of video signal or degradation caused by resolution decrease, different artefacts in video or blurring, system can enhance the quality of IPTV service. When service is prepared and code scheme chosen, the service is sent into network of numerous switches and routers. Degradation of network parameters can occur in any of them, such as packet losses in video signal, delays or delay variations. These parameters make errors that occur in core network and have a great influence on IPTV service i.e. they make video frame degradations or complete frame loss. The mentioned parameters are related to network layer of OSI/ISO reference model and they are often considered by a large number of authors who deal with this topic. After video packets are distributed to access node such as DSLAM (Digital Subscriber Line Access Multiplexer), they are supposed to be delivered to end user's LAN (Local Area Network) over a physical media. Thus, at the end point it is necessary to transfer video signal to user over physical channel with specific information capacity which can be obtained on that channel using different modulation and code techniques. For this purpose, different media and technologies are used. If QoE values are considered in real time, higher or lower QoE violation can occur with impairment of QoS parameters of network and/or application layer in OSI/ISO referent model. As it was mentioned in introduction, a great number of authors have researched the relation between proposed QoS parameters and QoE of the user of delivered service. However, in the case of consideration of physical telecommunication channel, which can be time varying, it can be concluded that a problem can appear as well as disappear in different times. In access network domain e.g. relation between access network and network of network operator, this is related to appearance of stationary and non-stationary noise. Therefore, it is necessary to observe QoS/QoE relation in real time e.g. in time of service delivery and its consummation.

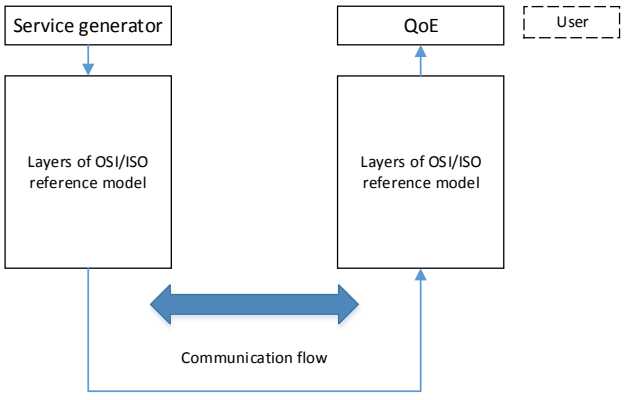


Figure 1. Communication flow of IPTV service

Fig. 1 depicts an example of communication flow between service provider and user during service delivery.

According to OSI/ISO reference model, communication flow passes through every layer which adds header in message of the previous layer. At the top is user's QoE which besides of numerous subjective parameters depends on QoS indicators. It's possible, due to the different manipulative data, that are provided by specific protocols, using monitoring (decoding) of these data, to have insight into how negative changes in lower layer influence data in higher layer of OSI/ISO reference model. This relation between different OSI layers of QoS indicators can be used for ME depicted in the Fig.2.

Considering the Fig.2, a connection can be found

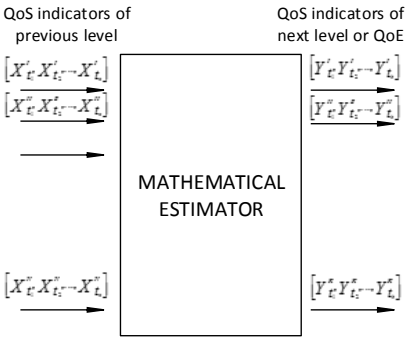


Figure 2. Example of mathematical estimator

between QoS indicators of the previous and the next OSI layers. For example, if we consider three vectors of input parameters, like delay X^D, jitter X^J or packet loss X^P and estimate output vectors that represent QoS indicators of next OSI layer or even QoE, we can find correlation between them. Considering that communication chain, for example, negative appearances in physical OSI layer will be propagated through data link layer, network layer and so on. On other hand, QoE is a perceived and subjective measure of services delivered through numerous telecommunication links. Appearance of impairment, perceived by user, in video services is a reflex of error appearances in any OSI/ISO layers. Monitoring QoS indicators can help to find link between parameters to forecast value of QoE in every telecommunication part, during exchange of information between different layers. As we can conclude, there are so many parameters and it is necessary to identify the essential one that is valuable for determination of QoS/QoE correlation and consequently QoE level.

III. MATHEMATICAL OBSERVATIONS

Observing the Fig. 2, it can also be noticed, there are time realizations of indicators in the previous OSI/ISO layer that cause incidental realizations in the next OSI/ISO layer. If vectors Ys represent vectors of time series parameters of interest for estimating incidents in the next layer of OSI/ISO model and Xs represent time series of incidental parameters, we can write the next equation:

$$Y^I{}_t = \sum_{K=1}^{N} \sum_{n=1}^{p} A_n X^K{}_{(t-p)} + E_t \qquad (1)$$

In equation (1), $X^{K}_{(t-p)}$ and Y^{I}_{t} represent different vectors of a particular time series while E vector represents stochastic variables. Those vectors represent only one realization of random process in specific time monitoring. Some kind of estimation of these particular processes has to be done to show the amount of influence of every observed variable on the processes in the next OSI level. To determine what has happened here we need to examine the nature of processes i.e. if particular time series are stationary or non-stationary or even ergodic. Also correlation between different X series i.e. series in the same layer should be examined. We can use consideration of autocorrelation and partial autocorrelation function for forecasting process nature and examination of next layer parameters. For the sake of simplicity, we have considered only one physical parameter i.e. noise margin parameter labelled as X. If process depends on its own previous values and its fluctuations, the next equation can be written as:

$$X_t = \sum_{j=1}^{p} \varphi_j X_{t-j} + \mathrm{e}_t - \sum_{i=1}^{q} \theta_i e_{t-i} \tag{2}$$

This equation represents known autoregressive moving average model. If there are some increasing or decreasing values of variable X, then the next equation can be defined as:

$$\varphi(\mathrm{L})\nabla^d X_t = \theta(\mathrm{L})\,\mathrm{e}_t \tag{3}$$

Equation (3) represents the same model including trends i.e. ARIMA model. Operator ∇^d signifies difference between operator order d, $\varphi(\mathrm{L})$ and $\theta(\mathrm{L})$ characteristic polynomial order L and e process of white noise. ARIMA model defines if process is stationary or non-stationary essentially. For analyzing the process behavior and its influence on QoS indicators of next layer, we have to determine time autocorrelation function of variable X obtained from equation (2) and (3) as well as ensemble autocorrelation function. These functions are defined by next two equations:

$$R_{tt} = \frac{\sum_{t=k+1}^{n} (X_t - \bar{X}) \bullet (X_{t-k} - \bar{X})}{\sum_{t=1}^{n} (X_t - \bar{X})^2} \tag{4}$$

$$R_{xx} = \frac{\sum_{x=s+1}^{n} (X_x - \bar{X}) \bullet (X_{x-s} - \bar{X})}{\sum_{x=1}^{n} (X_x - \bar{X})^2} \tag{5}$$

Considering equations (4) and (5) we can identify the nature of process of particular time series and it can be used for error propagation estimation through layers. We need to notice, if E(X) and Et(X) from equations (4) and (5) have the same values, the process is ergodic.

IV. Test Environments

In order to validate above mentioned model, we have chosen an experimental method in which SDTV (Standard-definition Television) IP service is delivered via MPEG-2 transport stream, UDP transport protocol, IP network protocol, ATM and ADSL2+ data link and physical layers. We have used noise margin [11] as parameter of physical layer and code violation CV and number of corrected errors as parameter of data layer, respectively. Packet loss was chosen as QoS indicator and we have correlated it with modified IQX hypothesis. We have used correlation between these losses and QoE for examination of video loss and compered our results and the results obtained by IQX hypothesis [7]. Physical channel was under the influence of stationary and non-stationary noises that cause negative effects in different OSI layers.

V. Results

Overview of a part of the results obtained from test environment is given in the Table I and Table II. Parameter X determines average changeability of noise margin that caused the appearance of frame loss; ΔX represents noise margin difference between different time intervals; FEC determines the number of corrected bits at the same time; E(X), Et(X), Var(X), Rxx, Rtt are expected values, variance and autocorrelation functions of ensemble and time, respectively. We have considered packet loss like QoS indicator because it can be correlated, based on IQX hypothesis, with QoE perceived by user. Mentioned parameters have been monitored in the same time intervals.

TABLE I: Measured parameters of Physical Layer in DSL environment

T(s)	X	ΔX	Et (X)	Var (X)	E (X)	Rxx	Rtt
t1	14	-0,1	10,93	4,38	10,36	-4,61	0,88
t2	14,1	0,1	10,91	4,79	10,36	-3,8	0,78
t3	14	0	10,83	4,79	10,36	-3,74	0,63
t4	14	0	10,93	4,64	10,36	-4,06	0,48
t5	14	0	10,89	4,91	10,36	-3,6	0,33
t6	14	2,4	10,77	5,14	10,36	-2,24	0,18
t7	11,6	-0,7	10,51	3,26	10,36	-5,98	0,09
t8	12,3	4,4	10,63	3,47	10,36	-1,5	-0,02
t9	7,9	0	9,94	3,83	10,36	-0,82	-0,03
t10	7,9	0	9,93	3,98	10,36	-0,75	-0,05
t11	7,9	0,1	9,93	3,91	10,36	-0,71	-0,06

After t5 there is appearance of non-stationary noise that caused negative effects in data and network layers. These effects produced video packet loss in IPTV service and consequently they disturbed a proper level of QoE. The

coefficients of autocorrelation function Rtt are shown in table I and the presence of non-stationary noise can be noticed especially in three correlation coefficients of Rtt.

TABLE II: MEASURED PARAMETERS OF DATA LINK LAYER, NETWORK LAYER AND QOE

T(s)	Frame loss	FEC	Packet loss	ARIMA	Mod QoE	IQX QoE
t1	0	0	0	14,01	4,07	4,07
t2	0	27	0	13,7	4,07	4,07
t3	0	0	0	13,18	4,07	4,07
t4	0	0	0	12,21	4,07	4,07
t5	0	0	0	10,68	1,90	1,06
t6	84	20165	28,70%	9,46	4,02	1,06
t7	1	927	0,30%	8,43	4,07	1,06
t8	0	222	0	7,87	4,07	1,06
t9	0	272	0	7,86	4,07	1,06
t10	0	0	0	7,85	4,07	1,06
t11	0	43	0	7,85	4,07	1,06

At the same time ARIMA model (eq. 3) forecasts values of variable X. Appearance of non-stationary noise has caused loss of the frames and consequently packet loss.

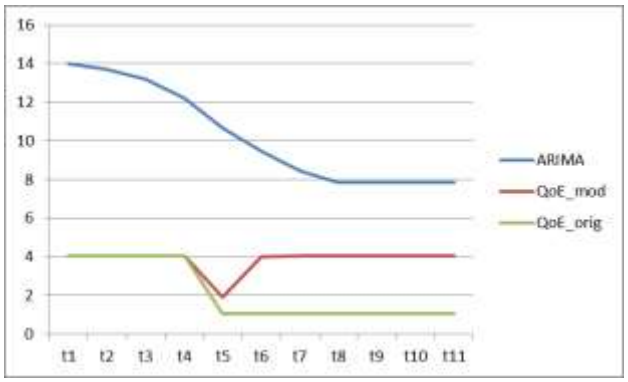


Figure 3. Model results

We have also calculated originate IQX QoE hypothesis [7] and modified one given in [11]. It can be noticed in Fig. 3, upper line defines ARIMA trends calculated by (3) and tracks changes of variable X caused by non-stationary noise while lower lines define values of originate and modified QoE. If we consider originate IQX hypothesis (QoE_orig), after appearance of packet loss, there is a fall of QoE level which lasts for all considered time. In contrary, that time dependency between QoS indicators (noise margin in this case) and QoE level is considered in our model. Although QoE_mod fall is noticed at the same time interval, in t8 interval after non-stationary noise disappears, we have a proper level of QoE again. Consequently, it cannot precisely define QoE level, because of time varying channel and its randomness i.e. without considering physical layer parameters.

VI. CONCLUSION

In this paper, the mathematical modelling of an estimator has been given. Also we have shown what the estimator needs to have for proper forecasting of QoS indicator time series which represents one realization of random process. This consideration refers to time and using the estimator QoE can be valued in a particular time interval. In the case of SLA (Service Level Agreement) offered by network provider, implementing this kind of solution can give the insight into how much of the service consumed time a customer had a proper level of QoE. Also, it can be used for feedback that helps system to repair itself without user intervention. For future work we plan to use more parameters in every OSI layer and find causality between them.

REFERENCES

[1] R. Schatz, T. Hoßfeld, L. Janowski, S. Egger, "From Packets to People: Quality of Experience as a New Measurement Challenge", in Data Traffic Monitoring and Analysis, Lecture Notes in Computer Science, vol. 7754, pp. 219-263, 2013

[2] M. Hirth, S. Scheuring, T. Hossfeld, Ch. Schwartz, Ph.Tran-Gia, "Predicting result quality in Crowdsourcing using application layer monitoring", in Proceedings of the 2014 IEEE Fifth International Conference on Communications and Electronics (ICCE), Da Nang, Vietnam, Aug. 2014, pp. 510-515

[3] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, Ch. Lorentze, "Initial Delay vs. Interruptions: Between the Devil and the Deep Blue Sea" in proceedings of 4th International Workshop on QoMEX 2012, Yarra Valley, Australia, Jul. 2012, pp. 1-6

[4] Fei Chen, Cong Zhang, Feng Wang, Jiangchuan Liu, "Crowdsourced Live Streaming over the Cloud", Cornell University Library, submitted on 23 Feb 2015

[5] Mu Mu, A. Mauthe, F. Garcia, "A Discrete perceptual impact evaluation quality assessment framework for IPTV services" in Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME), Suntec City, Singapore, Jul. 2010, pp. 1505-1510

[6] P. Calyam, P. Chandrasekaran, G. Trueb, N. Howes, R. Ramnath, D. Yu, Y. Liu, L. Xiong, D. Yang, "Multi-Resolution Multimedia QoE Models for IPTV Applications", in International Journal of Digital Multimedia Broadcasting, vol. 2012, Article ID 904072, 13 pages, Apr. 2011

[7] Markus Fiedler, Tobias Hoßfeld and Phuoc Tran-Gia, "A Generic Quantitative Relationship between Quality of Experience and Quality of Service", in IEEE Network, vol. 24, issue 2, pp 36-41, April 2010

[8] H. J. Kim and S. G. Choi, "A study on a QoS/QoE correlation model for QoE evaluation on IPTV service," in Proceedings of the 12th International Conference on Advanced Communication Technology, Phoenix Park, Korea, Feb. 2010, pp. 1377–1382

[9] Mu Mu, A. Mauthe, F. Garcia, "A Utility-Based QoS Model for Emerging Multimedia Applications" in Proceedings of the Second International Conference on the Next Generation Mobile Applications, Services and Technologies (NGMAST '08), Cardiff, UK, Sept. 2008, pp. 521-528

[10] J. Sen, S. Bhattacharya, "A Survey on Cross-Layer Design Frameworks for Multimedia Applications over Wireless Networks", in International Journal of Computer Science and Information Technology, vol. 1, no. 1, pp. 29-42, Jun. 2008,

[11] N. Goran, M. Hadzialic, "Real time assuring QoE in the lowest OSI/ISO layers during delivering of IPTV services", in Proceedings of 37th MIPRO, Opatija, Croatia, May 2014, pp. 532-535

Primjena mobilnih komunikacijskih tehnologija sa svrhom unapređenja usluga javnog prijevoza

Mićo Dujak[1], Zoran Civadelić[1], Krešimir Vidović[1], Pamela Miletić[2]
[1]Ericsson Nikola Tesla d. d., Krapinska 45, 10000 Zagreb, Hrvatska
[2]Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, Unska 3, 10000 Zagreb, Hrvatska
E-mails: { mico.dujak, zoran.civadelic, kresimir.vidovic}@ericsson.com; pamela.miletic@fer.hr

SAŽETAK: Transportna infrastruktura je važan segment u sektoru transporta i po svom značaju predstavlja glavnu podlogu za održivi razvoj zemlje. Kreiranje ekonomske politike neizostavno podrazumijeva kreiranje i razvoj transportne politike, odnosno razvoj infrastrukture. Poseban značaj pridjeljuje se kvalitetnijem usmjeravanju prometnih tokova, smanjenju broja prometnih nesreća i bržoj reakciji kada se one ipak dogode te kreiranju novih usluga primjenom informacijsko komunikacijskih tehnologija.

Integracija tehnologija iz područja mobilnih telefonskih mreža (GSM, 3G, HSDPA) s prometnom tehnikom omogućuje povećanu kvalitetu prometnih usluga i širi spektar njihovih funkcionalnosti. To je naročito primjetno u području informiranja putnika, pružanja turističkih usluga, akcija u slučaju prometnih nesreća te novih usluga iz područja inteligentne logistike te naplate prometnih usluga javnog gradskog prijevoza.

Potreba za sve većom pokretljivošću u suvremenim urbanim sredinama, kao i funkcioniranje gradskog prometa u sve složenijim uvjetima rezultirala je potrebom za uvođenjem naprednih prometnih usluga baziranih na informacijsko komunikacijskim tehnologijama i M2M komunikaciji. Među značajnije M2M aplikacije u prometnom kontekstu ubrajaju se domene umreženih vozila i unapređenje usluga javnog gradskog prijevoza.

U radu je opisana usluga spajanja vozila javnog gradskog prometa na širokopojasnu mrežu u svrhu naplate usluge javnog gradskog prijevoza u urbanim sredinama. Osnovna ideja uvođenja navedene usluge je uvesti dodatni kanal prodaje karata korištenjem mobilnih uređaja te na njima pripadajućih naprednih tehnologija za NFC i QR kôdove.

KLJUČNE RIJEČI

1. Javni prijevoz, 2.Naplata usluga, 3. Mobilne komunikacijske tehnologije, 4. NFC, 5. QR

I. UVOD

Potreba za sve većom pokretljivošću u suvremenim urbanim sredinama, kao i funkcioniranje gradskog prometa u sve složenijim uvjetima rezultirala je potrebom za uvođenjem naprednih prometnih usluga baziranih na informacijsko komunikacijskim tehnologijama i M2M komunikaciji. Očigledna ograničenja klasičnog pristupa razvoju prometnog sustava, u skladu s načelima znanstveno utemeljene politike gospodarenja i održivog razvoja, potakla su zahtjeve za novim usklađenim rješenjima u cestovnim i drugim granama prometa te njihovim sučeljima (lukama, kolodvorima itd.). Naplata usluga javnog gradskog prijevoza u urbanim sredinama se smatra jednim od značajnih prometnih problema s obzirom da troškovi za nenaplaćene usluge utječu kako na operativne tako i na administrativne troškove

Prema dosadašnjim istraživanjima [3] oko 60% usluga javnog gradskog prijevoza uspije se naplatiti, što znači da oko 40 posto korisnika ne plaća voznu kartu. Istraživanje o uslugama tvrtke Gradski prijevoz putnika d.o.o. u gradu Osijeku 2012, koju je izradila tvrtka Audeo [4] pokazuje da čak 60% korisnika smatra kako je cijena vozne karte previsoka. Uzme li se u obzir da se u Gradu Zagrebu preveze godišnje oko 300.000.000 putnika [2], može se zaključiti da prijevoznici javnog gradskog i prigradskog prijevoza trpe velike gubitke.

U ovom radu opisano je rješenje Smart Transport Sustav koje služi za naplatu usluga javnog prijevoza korištenjem mobilnih komunikacijskih tehnologija te platformi za mobilna plaćanja. Osnovna karakteristika ovog modela je da putnik plaća samo prijeđeni put, što uslugu čini atraktivnom za krajnje korisnike. U naletu novih tehnologija koje se koriste za plaćanje i sve većoj popularizaciji platformi za mobilna

plaćanja, Smart Transport je nova usluga koja predstavlja savršen spoj mogućnosti telekom operatera i potreba prijevoznika i krajnjih korisnika.

II. ZNAČAJ MOBILNIH KOMUNIKACIJSKIH TEHNOLOGIJA U JAVNOM PRIJEVOZU

Prosječna komercijalna brzinu vozila javnog prijevoza u gradovima u regiji iznosi do 20 km/h To znači da se na temelju iskustava može zaključiti kako bi uvođenje naprednih sustava upravljanja javnim prijevozom otvorio mnogo prostora za poboljšanja i povećanje zadovoljstva korisnika. Treba posebno istaći da primjena i implementacija inovativnih informacijsko-komunikacijskih sustava omogućuje povećanje protočnosti i ubrzanje javnoga i privatnoga prometa u gradu, a u konačnici se postiže i rasterećenje prometnica u gradu te na prilazima gradu.

Upravo mobilne komunikacijske tehnologije te širokopojasna rješenja igraju ključnu ulogu pri implementaciji naprednih usluga. Samim time mobilne komunikacijske tehnologije postaje sve važniji u transportnoj industriji budući da kompanijama u ovoj branši omogućava da posluju uspješnije, povećava sigurnost putnika i podržava zaštitu okoliša [1]. Ovako osmišljena inteligentna transportna rješenja otvaraju nova zanimljiva ICT tržišta velikih potencijala.

Povezivanje vozila javnog gradskog i prigradskog prijevoz na širokopojasnu mrežu može imati svrhu uvođenja dodatnog kanala prodaje i kontrole karata u javnom prijevozu. Integracija takvog sustava sa sustavom za satelitsko praćenje vozila omogućit će prijevoznim operaterima pristup širokom rasponu informacija o poziciji, brzini i smjeru kretanja autobusa, vremenu zaustavljanja, prijeđenoj kilometraži, vremenu polaska i odlaska s određenih lokacija, itd. Svi navedeni podaci mogu poslužiti pri uspostavi modela naplate usluge „plati koliko koristiš“ (eng. PAYG – Pay As You GO) [6].

Inovativna rješenja ovakvog tipa su temelj stvaranja novih usluga za planiranje kretanja gradom, pa i putovanja na međugradskim rutama. Budući da im sustav omogućuje mobilni širokopojasni pristup Internetu korisnici mogu jednostavno dobiti uvid u informacije o lokaciji vozila. Putnici tako štede vrijeme i imaju mogućnost kupnje karata na način koji im najviše odgovara. Operaterima javnoga prijevoza pak širokopojasni pristup omogućuje implementaciju aplikacija koje će podržati razvoj njihovoga poslovanja.

III. SMART TRANSPORT SUSTAV

Prema određenim procjenama i pokazateljima, smatra se da se samo oko 40% usluga javnog gradskog prijevoza uspije naplatiti. Prema tome može se zaključiti da prijevoznici javnog gradskog i prigradskog prijevoza trpe velike gubitke.

Osnovna ideja uvođenja Smart Transport sustava je uvesti dodatni kanal prodaje karata korištenjem mobilnih uređaja te na njima pripadajućih naprednih tehnologija za NFC i QR generatora. Primjenom Smart Transport sustava se želi postići sljedeće:

- Povećati broj novih putnika koji će nakon uvođenja usluge koristiti javni gradski prijevoz – u ovom kontekstu, radi se o putnicima koji su do sada koristili osobne automobile ili druge modove prijevoza te putnici koji imaju kraće rute za vožnju te im se nije isplatilo plaćati punu cijenu karte
- Povećati broj platiša – mnogi putnici koriste javni gradski prijevoz no ne plaćaju ga. Ovom uslugom bi stimulirali putnike da plaćaju prijevoz s obzirom da će na način koji je opisan tehničkim rješenjem imati mogućnost plaćanja po PAYG modelu.
- Unaprijediti sigurnost usluge mobilnog plaćanja putne karte, na način da se eliminiraju svi sigurnosni propusti koje su imali slični, ranije korišteni (i uglavnom odbačeni), sustavi.

Rješenje podrazumijeva da putnik plati samo onaj dio puta koliko se vozio u autobusu/tramvaju ili nekom drugom vozilu (eng. PAYG model – Pay As You Go). Treba naglasiti da se ovim rješenjem razmatra javni prijevoz, no uslugu je moguće modificirati i za druge oblike prijevoza (npr. taxi uslugu). Plaćanje, tj. računanje cijene, može biti definirano na temelju prijeđene kilometraže, vremena vožnje ili broja stanica između ulaska putnika u vozilo i izlaska putnika iz vozila. U ovom projektu cijena će se računati prema broju prijeđenih stanica.

IV. ARHITEKTURA I FUNKCIONALNOSTI SUSTAVA

Arhitekturalno gledano, sustav se sastoji od uređaja u vozilu (OBU) i Smart Transport aplikacije (STA) na platformi za mobilno plaćanje. Također, potrebna je i

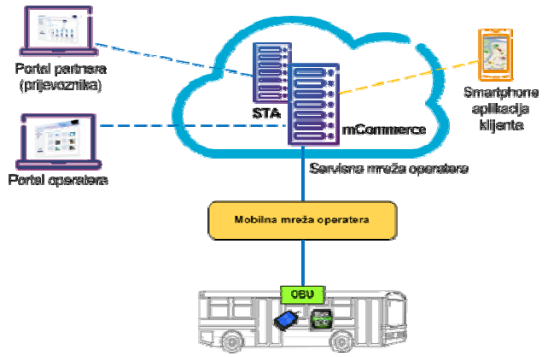
klijentska aplikacija na pametnom uređaju korisnika usluge.
U proces plaćanja uključene su slijedeće platforme telekom operatera i tehnološke cjeline:

- On-board unit (OBU) na vozilu, sa pripadajućim senzorima
- mCommerce platforma za mobilna plaćanja
- Smart Transport aplikacija
- Smartphone aplikacija sa NFC ili QR-kôd funkcionalnosti (za putnika i za kontrolora)

Smart Transport rješenje obuhvaća centralnu Smart Transport aplikaciju (STA) na platformi za mobilna plaćanja, te aplikaciju za pametne telefone putem koje se inicira plaćanje "putne karte" putem NFC kanala i QR kôda, i aplikaciju kojom se vrši kontrola karata u vozilu. Platforma za mobilno plaćanje poput mCommerce platforme u servisnoj mreži mobilnog operatera koristit će se kao platforma za plaćanje, na koju će se integrirati Smart Transport Aplikacija.

Koncept cijelog rješenja temelji se na komunikaciji u stvarnom vremenu između jedinice u vozilu (on-board unit) i centralne Smart Transport aplikacije povezane s mCommerce sustavom [5]. Infrastruktura treba biti osigurana od strane telekom operatera, a istovrememo telekom operater igra i važnu ulogu u cijelom procesu – operater pruža podršku za autentikaciju krajnjeg korisnika. Autorizacija svake transakcije vrši se na centralnom mCommerce sustavu (unutar mreže mobilnog operatera). Dakle, mCommerce platforma u mreži mobilnog operatera uz pripadajuće elemente mreže mobilnog operatera omogućava vrhunsku sigurnost svih transakcija, te nosi ulogu i centralne AAA platforme.

Dva su osnovna modela rješenja, ovisno o tehnologiji koju krajnji korisnik koristi prilikom ulaska u vozilo / izlaska iz vozila – rješenje bazirano na NFC tehnologiji i rješenje koje se koristi QR-kôd tehnologijom. Bez obzira koristi li se za identifikaciju i autentikaciju korisnika NFC ili QR-kôd, sam prometni model podataka („*call flow*“) je isti u oba slučaja. OBU na sebi ima integriran i NFC čitač i QR-kôd čitač, a koja će se tehnologija koristiti – ovisi o mogućnostima mobilnog uređaja korisnika. U ovom modelu svi podaci vezani uz transakciju prenose se do centralne aplikacije (STA) zaštićenim kanalom, preko OBU-a. Moguće je implementirati i alternativni QR model, koji ima drugačiji tok podataka (preko mobilnog uređaja korisnika).



Slika 1: Smart Transport aplikacija (NFC) – pregled sustava

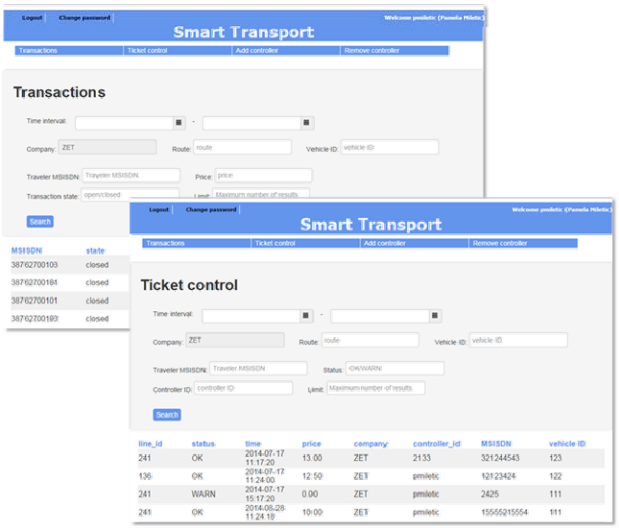
V. VALIDACIJA KONCEPTA

Validacija samog koncepta napravljena je razvojem i testiranjem ovog sustava u laboratorijskom okruženju, te testiranjem istog u stvarnim prometnim uvjetima. Komunikacija OBU i centralne aplikacije izvedena je putem GPRS/HSPA komunikacijskih kanala mobilnog operatera.

Smart Transport aplikacija pokrenuta je na Tomcat poslužitelju i direktno povezana sigurnim IP kanalom (VPN tunel) sa mCommerce platfromom u mreži mobilnog operatera. Backend STA aplikacije komunicira s jedne strane s mCommerce sustavom a s druge strane s jedinicama u vozilima (OBU). Komunikacija s OBU je zaštićena na više razina, kako bi sigurnost komunikacije i samih transakcija zadovoljila standarde tražene na tržištu.

Jedinica u vozilu (On Board Unit – OBU) je prilagođena smještanju u vozilo, te su sve postavke vezane uz napajanje, komunikaciju, sigurnost i izdržljivost uzeti u obzir i zadovoljeni prilikom dizajniranja iste. OBU prema krajnjem korisniku ima izložene NFC i QR ulazne kanale, a prema centralnoj aplikaciji je povezana ranije opisanim načinom. Fizički je u vozilu smještena odmah na ulaznim/izlaznim vratima.

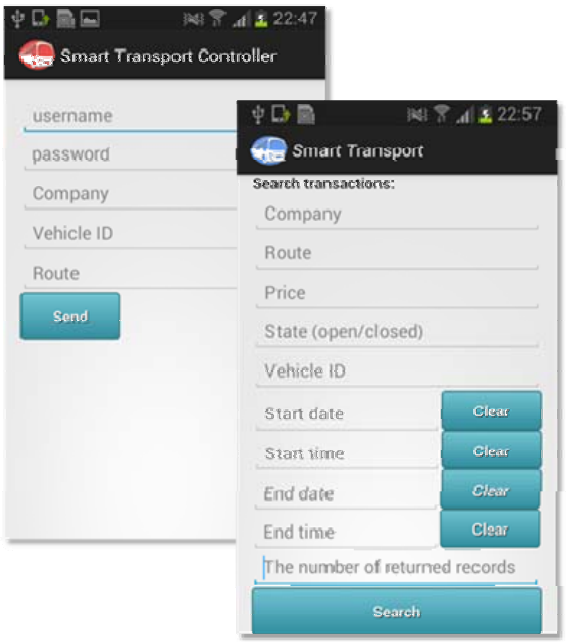
Sama centralna STA aplikacija u svom frotendu nudi sučelje za upravljanje jedinicama u vozilima, upravljanje samom uslugom, te administracijsko-reporting sučelje. Na slici ispod vidi se nekoliko ekrana sa GUI-ja centralne aplikacije.



Slika 2: Smart Transport aplikacija GUI – primjer sučelja

Smartphone aplikacija za korisnika realizirana je na Android platformi za Smartphone uređaje. Aplikacija omogućava krajnjem korisniku korištenje usluge putem NFC ili QR tehnologije (ovisno o mobilnom aparatu), te istovremeno i pregled svih svojih transakcija direktno na Smartphone aplikaciji mobilnog uređaja. Posebno je razvijena smartphone Aplikacija i za Kontrolora, koja koristi iste tehnologije, a prilagođena je kontroli karata u vozilu.

Na slijedećoj slici pokazani su primjeri grafičkog sučelja testne smartphone aplikacije za korisnika i aplikacije za kontrolora.



Slika 3: STA Smartphone App. – primjeri sučelja

VI. ZAKLJUČAK

Sveprisutnost pokretnih mreža, niske cijene sklopovlja za bežičnu komunikaciju i dostupnost uređaja sa podrškom za umrežavanje stvorilo je uvjete za razvoj kako platformi za M2M komunikaciju, tako i specifičnih usluga i aplikacija zasnovanih na M2M komunikaciji.
Uslugom spajanja vozila javnog gradskog prijevoza s mobilnom širokopojasnom mrežom putem modula za mobilni širokopojasni pristup visokih brzina te postavljanjem senzora na vozila, unapređuje se usluga elektroničke naplate karata i satelitskog praćenja autobusa a operaterima se omogućuje i pristup širokom rasponu informacija o poziciji, brzini i smjeru kretanja autobusa, vremenu zaustavljanja, prijeđenoj kilometraži, vremenu polaska i odlaska s određenih lokacija, te uvjetima u okolini (emisija CO, CO2, O2, temperatura i vlaga…) itd.

VII. REFERENCES

[1] Z Kljaic. M Dujak, K Meic. (2011). Prediktivni sustavi za povećanje sigurnosti cestovnog prometa temeljeni na statističkim geoprostorno organiziranim prometnim podacima. Mipro 2011, 34th international convention on information and communication technology, electronics and microelectronics. Opatija, Croatia.

[2] www.zet.hr, 13.01.2015.godine

[3] ZET za pet, Za Grad, 2013.

[4] Istraživanje o uslugama tvrtke Gradski prijevoz putnika d.o.o. u gradu Osijeku 2012., Audeo, Osijek, 2012

[5] Internal Ericsson mCommerce project documentation

[6] M. Mezgahani. (2008). Study on electronic ticketing in public transport, Dinal Report, EMTA

SIMULATIONS AND MEASUREMENTS

Computer Modeling and Simulation of Teletraffic Models for 5G Networks

E.P. Ivanova*, T. Iliev*, Gr. Mihaylov*, I. Stoyanov, * F. Tsvetanov**, E. Otsetova*** and D. Radev ***
* University of Ruse, Ruse, Bulgaria
** Southwest University, Blagoevgrad, Bulgaria
*** College of Telecommunications and Posts, Sofia, Bulgaria
epivanova@uni-ruse.bg, tiliev@uni-ruse.bg, gmihaylov@uni-ruse.bg, stoyanov@uni-ruse.bg, ftsvetanov@swu.bg, eotsetova@hctp.acad.bg, dradev@abv.bg

Abstract - Advances in the new mobile networks are steadily increasing the demand on network resources, and new standards must ensure that the networks are managed effectively in order to ensure that users of the network remain satisfied with their Quality of Service (QoS) and Quality of Experience (QoE). The simulations opportunities represent a chance to reduce the digital divide and to ensure the new 5G networks with minimum cost of time and financials. The simulation of teletraffic models allows the modeling of resource management issues, admission of system users with different parameters.

I. INTRODUCTION

Mobile networks constantly evolve. In the 1980s, when wireless technology began to go mainstream, the 1G analog networks supported the bulky, brick-like mobile phones from that era. The 90s saw the introduction of 2G digital networks, when Code Division Multiple Access moved mobile phones beyond voice transmission into data and text messaging. Around 2001, third generation devices appeared, first in Japan, offering the high-speed data transfer and Internet connection used in the earliest smartphones. The first 4G LTE and WiMAX networks appeared in 2011 [1,7]. Given the natural decade-long development cycle, fifth generation (*5G*) networks are due around 2020.

Simulations represent a good opportunity to reduce the digital divide and to ensure the new 5G networks. Many universities and researchers are currently involved in simulation modeling to provide better results. This paper presents an analysis and development of the teletraffic models used in three of the Bulgarian universities: University of Ruse, College of Telecommunications and Post in Sofia and Southwest University of Blagoevgrad.

The implementation of simulation program is made with .NET. The teletraffic models with geometrical and Pareto arrival distributions, which are versatile, possessing a pseudo-memoryless property which makes the solution of many GE-type queuing systems and networks analytically tractable. The new communication systems are interested in the modeling and performance evaluation of single communication links, but in the modeling of more devices in the whole network. Typically, the performance of the whole network containing many links working in parallel and interfering with each other is of primary interest.

This type of simulation allows the modeling of resource management issues, admission of system users with different priority, interaction and network loading by several types of traffic sources, etc. The aim of such modeling is network capacity evaluation and probabilities of failure.

Project № BG051PO001-3.3.06-0008 "Supporting Academic Development of Scientific Personnel in Engineering and Information Science and Technologies"

II. TRENDS OF 5G MOBILE NETWORKS

The trends of 5G mobile access is to be built upon new radio access technologies, evolving 4G, LTE, HSPA, GSM and WiFi. The new requirements are widely varying. The services, devices, number of clients are in super wide variety.

The new communication needs of a massive deployment of various devices, such as with low complexity parameters to more advanced parameters. The industrial and medicine devices must be lifesaving. The variety of services must be with very low latency and probability of blocking.

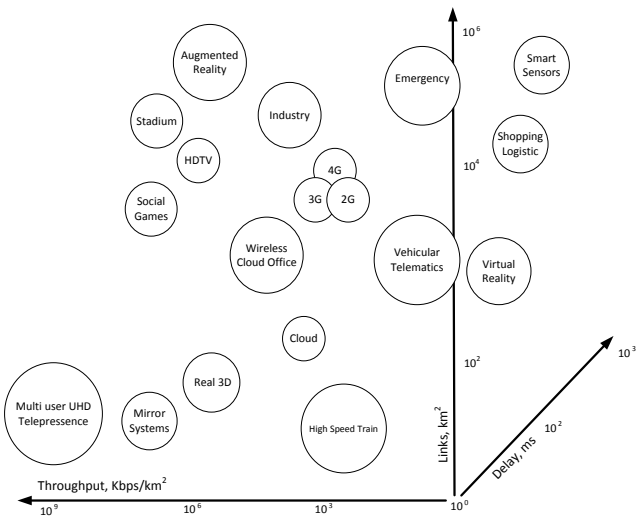


Figure 1. Services of 5G mobile network

The services must be amazingly fast in very dense crowds of users [8]. The fast connections is a key factor for success of development of so important services as health and security sectors, fire brigade, high speed trains and cars (Fig.1). The services are:

- Shopping logistic;
- Vehicular telematics;
- Virtual reality;
- Emergency;
- Stadium, crowd of people;
- Wireless cloud offices;
- High speed trains;
- Multiuser UHD telepresence;
- HDTV;
- Mirror Systems;
- Social games;
- 2G, 3G and 4G services;
- Smart sensors.

The Quality of Experience requires fully connected users not only for static users. The expectation of connectivity is no matter where or how one is moving [8]. Highly mobile devices, e. g. trains and cars, are communicating machines for this scenario, but also sensors related to widely varying applications, e.g. moving components in industries, plants, other vehicles.

III. SCENARIOS FOR TELETRAFFIC MODELS FOR 5G MOBILE NETWORKS

The heavy traffic models must be simulated, because the various expectations must be fulfilled on very high level. Some scenarios could be described:

- ***Emergency communications***; ***traffic efficiency*** and ***safety*** must be with the highest priority of all services and scenarios. When occurs an event as an accident, the devices and the network must be able to switch to "emergency mode", with the highest priority, and the battery life must be surviving for a long time.
- ***Teleprotection*** in smart grid network will ensure the people with very important messages, which will be sent between substations for prevention of some kind of accident (for example as car crash).
- ***Blind spots*** are the next scenario. The wireless systems have to provide ubiquitous coverage in rural and heaving shadowed areas. The real time remote computing for mobile terminals must ensure the connection.
- ***Virtual reality office*** is the communication point with the necessary of high data rates. The telepresence and communication between the offices of one company must work without problems. Applications with high data rates must be with high security.
- ***Shopping mall***, ***traffic jam***, ***concerts***, ***stadium*** or ***open festivals*** are places with very high density of users and devices. The expectations are for huge amount of aggregated data traffic about up to 900 Gb/km^2. The users inside the stadium or concert will generate high quality videos or could transmit outside the HD movies. The density of users in the shopping mall is high, but the access of broadband communications must be fulfilled.

The necessity of transmitting a huge amounts of data in very short times, very fast connections and priorities of different services is exact description of arrivals and departures of different clients in described scenarios in 5G. It could be described with Geometrical and Pareto distributions with heavy load. The high data rates, the demand and availability of bandwidth is raising at a much faster rate [2,3,5]. The emergency communication must have very low probability of blocking and high reliability. Therefore, the priorities in the model are dynamic.

The teletraffic delay system with two parallel inputs with Geometrical and Pareto distribution, and one service output with Geometrical distribution, with finite capacity N was applied with the scenarios of 5G networks (Fig.2).

The GE is of the form:

$$F(t) = P(W \le t) = 1 - \tau e^{-\sigma t}, \quad t \ge 0, \\ \text{where} \quad \tau = 2/(C^2 + 1) \quad \text{and} \quad \sigma = \tau v, \tag{1}$$

where W is a mixed-time random variable of the inter event – time, C^2 is the squared coefficient of variation of W [4].

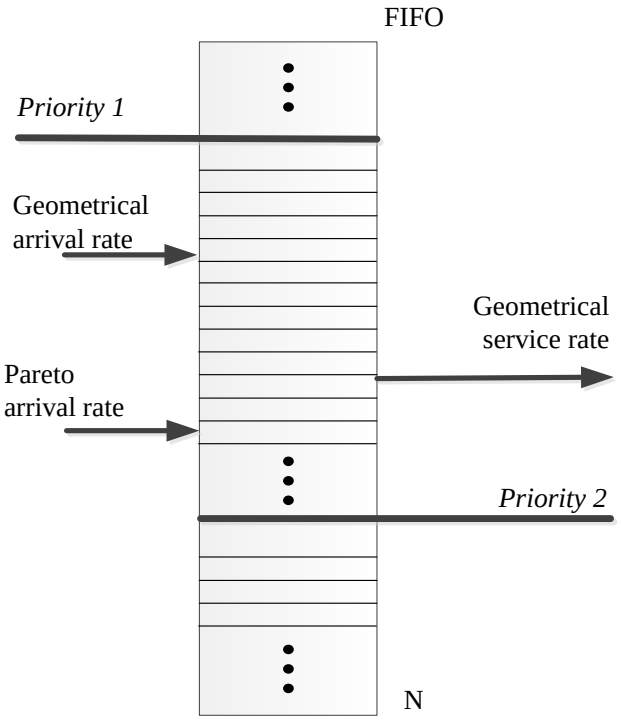


Figure 2. Teletraffic system for 5G mobile network

For $C^2<1$, the GE distribution model cannot be physically interpreted as a stochastic model.

For $C^2>1$, the GE model is a mixed-time probability distribution and can be interpreted as bulk time distribution with underlying counting process equivalent to compound Poisson process with parameter $2\nu/(C^2+1)$ and geometrically distributed bulk size with mean $(C^2+1)/2$ and squared coefficient of variation $(C^2-1)/(C^2+1)$.

The approximate universal closed form expression for GE for blocking probabilities π_i, i=1,2,…R, can be determined by using GE-type probabilistic arguments and is described by (2):

$$\pi_i = \sum_{\upsilon=0}^{N_1} \delta_i(\upsilon)(1-\sigma_i)^{[N_i-\upsilon]^+} P(\upsilon), \qquad (2)$$

where $\delta_i(\upsilon) = \delta_i$, if $\upsilon = 0$ or otherwise, $\sigma_i = (C_{ai}^2+1)/2$, $\delta_i = r_i / (r_i(1 - \sigma_i) + \sigma_i)$, and $r_i = (C_{si}^2+1)/2$.

The queuing system has for each class i (i=1,2,…,R, $R>1$) the arrival rate $1/\lambda_i, C_{si}^2$ and the service rate $1/\mu_i, C_{si}^2$.

The blocking probability is π_i that an arrival of class I will find in the system at the full capacity [7].

The blocking probability is determined by arrival sequence, independently by the rule of service FCFS (*First Come First Serve*), LCFS (*Last Come First Serve*) or HOL (*Head of Line*) static users.

IV. Simulation of Teletraffic Models for 5G Mobile Networks

The program for Simulation of Rare events in teletraffic systems with single queue is made with the .NET Framework, because it is the way that programming will be done on a Microsoft machine from now. And not just on a Microsoft machine. The ADO.NET is used for creating web site, and for manipulating databases. This framework create applications for mobile phones and PDA's with .NET, this is one of the main reasons for using it.

The .NET Framework is a whole lot of Classes (called Namespaces) and the technology to get those Classes to work [6]. The main component is called the Common Language Runtime. With .NET, more than 15 different programming languages can use the Common Language Runtime. One of these languages is, of course Visual Basic .NET. Another is C#.

The simulation of the three main types of queuing systems is made of the AMD Athlon(tm) II X2 250 Processor 3.00 GHz and 3 GB RAM, running 32-bits operation system is Windows 7 Professional.

First simulations are made with M/M/1/N queuing sytems with the .Net tools, because it was necessary to compare analytical and simulation results, to prove and validate the new results. The simulation of the queuing system is made for different number of observations n. The increasing of arrival rate leads to insignificant increasing of blocking events. We choose arrival and service rate, to receive appropriate result. First simulations were made with the disciplines of service. The amount of rare events in system with disciplines of service: LIFO and FIFO are with insignificant difference.

TABLE I. Teletraffic system with N=60

Service rate	n	Arrival rate 1	Arrival rate 2	Dynamic priority 1 and 2		Blocking events
0.5	100000	0.5	Pareto	30	30	4
	30000	Pareto	0.5	30	30	1
	50000		0.5	30	30	1
	10000	0.2	Pareto	50	40	2
	10000	0.4		30	60	2
0.48	10000	0.3	Pareto	30	60	3
0.44	10000	0.2	Pareto	40	60	2
	10000	0.25		60	60	3
0.4	10000	0.2	Pareto	50	40	2
	30000			50	40	4
	50000			50	40	5
	10000			60	50	1
	10000			40	60	5

The priorities in the modeled system are dynamic; the service rates of GE are with values 0.2, 0.3, 0.4, 0.44, 0.48 and 0.5. The number of examples are 10 000, 30 000, 50 000 and 100 000. The GE service rate is with values from 0.2 to 0.5. Table 1 shows simulation results with N=60 places for service in the system, the relative error is 10%, the received blocking events. The results of teletraffic system with N=90 are described in Table 2.

TABLE II. Teletraffic system with N=90

Service rate	n	Arrival rate 1	Arrival rate 2	Dynamic priority 1 and 2		Blocking events
0.5	100000	Pareto	0.35	60	60	5
	10000		0.35	50	60	1
	50000		0.35	60	60	2
	10000		0.35	40	90	3
	30000		0.35	40	90	4
	50000		0.35	40	90	4
	100000		0.35	40	90	5
	10000		0,35	80	60	1
0.4	10000	Pareto	0.1	90	70	5
0.4	30000	Pareto	0.1	90	70	2
0.4	10000	0.5	Pareto	50	80	6
0.3	30000	0.4	Pareto	45	45	3
0.3	50000			45	45	4
0.3	10000			45	45	1
0.2	10000	Pareto	0.1	90	70	1

Service rate	n	Arrival rate 1	Arrival rate 2	Dynamic priority 1 and 2		Blocking events
0.2			0.1	90	40	1
0.2		0.4	Pareto	45	45	1

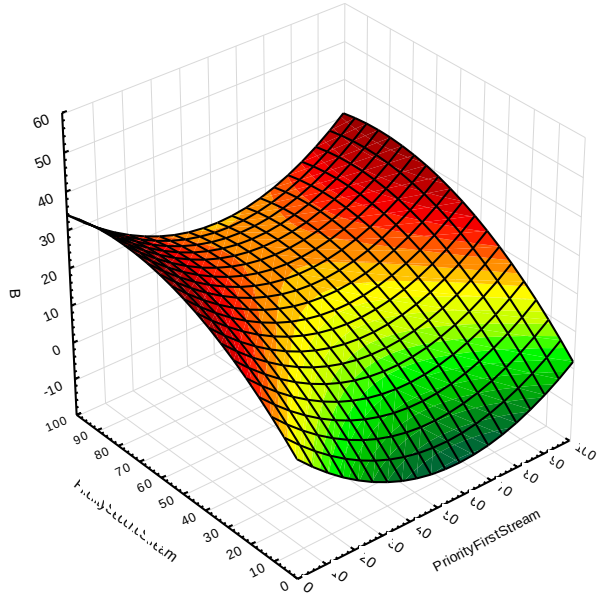


Figure 3. The relationship between dynamic prorities and blocking events B

On Fig. 3 and Fig.4 are shown the dependences between system parameters as first and second priorities, arrival and service rates and blocking events *B*.

The simulation experiments show that the blocking events in the system decrease with the length of the queues and increase if the dynamic priorities are increasing.

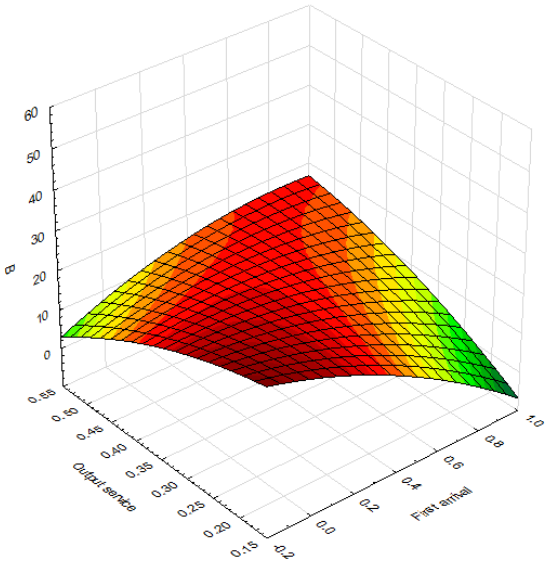


Figure 4. The relationship between arrival rate and service rate and blocking events

The potential users of the simulation product and software could be working groups in the area of the next generation mobile networks.

V. CONLUSION

In this paper we have reviewed, presented, validated and discussed the scenarios for teletraffic models for 5G mobile networks. Secondly, this paper presents an implementation of teletraffic models developed in .NET for estimation of blocking events.

The simulation results are received for queuing system with Geometrical and Pareto arrivals with two dynamic priorities, the service rate is implemented with Geometrical distribution. The service is organized with FIFO and a finite buffer size *N*.

The simulation results show that the blocking events in the system decrease with the length of the queues and increase if the dynamic priorities are increasing.

The future work is the study for the blocking probability with RESTART method for with *c* servers and finite capacity *N*.

ACKNOWLEDGMENT

The present document has been produced with the financial assistance of the European Social Fund under Operational Programme "Human Resources Development". The contents of this document are the sole responsibility of "Angel Kanchev" University of Ruse and can under no circumstances be regarded as reflecting the position of the European Union or the Ministry of Education and Science of Republic of Bulgaria:

Project № BG051PO001-3.3.06-0008 "Supporting Academic Development of Scientific Personnel in Engineering and Information Science and Technologies".

REFERENCES

[1] T. Iliev and Gr. Mihaylov. "3GPP LTE system model analisys and simulation of video transmission". Proceedings in Advanced Research in Scientific Areas, Slovak Republic, 2012, pp. 2016-2021.

[2] D. Radev, "Modelling of Rare Events in Broadband Digital Networks", Kolbis, Sofia 2006. (in bulgarian)

[3] R.Raev, E. Ivanova, Ek. Dudin and D. Radev. Educational Platform for Teletraffic Engineering. IN: EDULEARN, 2014, Barcelona, 2014, ISBN ISSN 978-84-617-0

[4] E. Ivanova, R. Raev and D. Radev. Simulation of Rare Events in Teletraffic Systems with Single Queue. IN: ICEST, Veliko Tarnovo, 2012

[5] D. Radev, "Teletraffic Engineering", University publishing center Ruse 2012. (in bulgarian)

[6] C. Sells, "Windows Forms Programming in C#", Addison-Wesley Professional, 2003.

[7] D. Koleva and S. Sadinov, "Comparative analysis on the state and development of telecommunication Services, International Scientific Conference UNITECH 2014, Bulgaria, Gabrovo, 2014, pp. 141 - 146

[8] M. Fallgren, B. Timus, P. Popovski, and V. Braun, Deliverable D1.1 "Scenarios, Requirements and KPIs for 5G mobile and wireless system", 2013

Simulation of a Quasigroup Error-Detecting Linear Code

N. Ilievska*, D. Gligoroski **
* Faculty of Computer Science and Engineering, "Ss.Cyril and Methodius" University, Skopje, Republic of Macedonia, Email: natasa.ilievska@finki.ukim.mk
** Department of Telematics, Norwegian University of Science and Technology, Trondheim, Norway, Email: danilog@item.ntnu.no

Abstract - We analyze an error-detecting code with a fixed length redundancy, based on quasigroups. We obtain experimental results for the probability of undetected errors when quasigroups of different order are used for coding and conclude how the probability of undetected errors depends on the order of the quasigroup used for coding and the parameter of the code.

I. INTRODUCTION

In [1] we have defined a model of error detecting code based on quasigroups. An important parameter for every error-detecting code is the probability of undetected errors. In this paper we will analyze this probability for our error-detecting code.

In Section II are given the basic mathematical definitions needed for the code, while in Section III is given the definition of our error-detecting linear code. In Section IV are given the probabilities of undetected errors when for coding are used quasigroups of order four, order eight and order sixteen for different lengths of the input block. These probabilities are obtained using simulation process, which is also explained in this section. Also, in this section we analyze how the probability of undetected errors depends on the length of the input block and the parameter of the code for all three quasigroups used for coding. Finally, in Section V we analyze how the order of the quasigroups used for coding affects on the probability of undetected errors. At the end, in Section VI are given the conclusions of the paper.

II. MATHEMATICAL PRELIMINARIES

Since the considered model of error-detecting codes is based on linear quasigroups, in this section we will give a definition of this algebraic structure.

Definition 1: Quasigroup is algebraic structure $(Q,*)$ such that

$$(\forall\, u, v\in Q)(\exists!\, x, y\in Q)(x*u=v \wedge u*y=v) \qquad (1)$$

Definition 2: The quasigroup $(Q,*)$ of order 2^q is linear if there are non-singular binary matrices A and B of order $q\times q$ and a binary matrix C of order $1\times q$, such that

$$(\forall u, v\in Q)\ x*y=z \Leftrightarrow \boldsymbol{z}=\boldsymbol{x}A+\boldsymbol{y}B+C \qquad (2)$$

where $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ are binary representations of x, y and z as vectors of order $1\times q$ and + is a binary addition.

In what follows when we say that $(Q,*)$ is a quasigroups of order 2^q, then we take $Q=\{0, 1, 2,\ldots, 2^q-1\}$.

Due to the Definition 2, linear quasigroups instead to be presented with the quasigroup operation * can be presented with a binary matrices A, B and C. We will do so in the future in this paper.

III. THE DEFINITION OF THE QUASIGROUP ERROR-DETECTING LINEAR CODE

The model of error-detecting linear code based on linear quasigroups is defined in a following way: Let $(Q,*)$ be a linear quasigroup and let k be a number such that $1\le k\le n-1$. Each input block $a_0a_1\ldots a_{n-1}$, where all $a_i\in Q$, is extended into a block $a_0a_1\ldots a_{n-1}d_0d_1\ldots d_k$. The redundant characters are defined as follows:

$$d_i=a_i*a_{i+1}*a_{i+2}*\ldots*a_{i+n-2},\ \ i=0, 1, \ldots, k \qquad (3)$$

where all operations in indexes are modulo n. The coded block $a_0a_1\ldots a_{n-1}d_0d_1\ldots d_k$, turned into binary form, is transmitted through the binary symmetric channel.

Restriction $k\ge 1$ ensures that all information characters $a_0, a_1,\ldots, a_{n-1}$ are controlled by at least one redundant character. The second constraint ($k\le n-1$) ensures that there will be no two identically defined redundant characters.

As we can see from the definition of the model, this code always adds fixed number of redundant characters, regardless of the length of the input block. More precisely, for a given k, this model always adds $k+1$ redundant characters on the input block.

Since, there are noises in the channel, some of the characters may not be correctly transmitted. In order to check whether the received block is correct, the receiver checks whether equations (3) are satisfied for the received block. If there is some i for which the equation (3) is not satisfied, the receiver concludes that there are errors in transmission and it asks the sender to send the block once again. But, since the redundant characters are transmitted through the binary symmetric channel, it is possible that they are incorrectly transmitted, too. This can result with a situation in which (3) is satisfied for all $i\in\{0,1,\ldots,k\}$, although some of the information characters $a_0, a_1,\ldots, a_{n-1}$ are incorrectly transmitted. This means that it is possible to have undetected errors in transmission. For this reason

it is important to obtain the probability of undetected errors. From the definition of the model, it is obvious that besides from the bit-error p, it depends on the length of the input block n, the parameter k, but also it depends on the order of the quasigroup used for coding.

IV. The Probability Of Undetected Errors

In this section we will briefly explain the process of simulation (subsection *A*). Using this process, we will obtain experimental results for the probability of undetected errors when for coding are used quasigroups of order four (subsection *B*), order eight (subsection *C*) and order sixteen (subsection *D*). For each of these cases, we will consider two choices of the parameter k in the error-detecting model defined with (3): one in which the redundancy is twelve bits and the other in which the redundancy is sixteen bits (i.e. fifteen bits for quasigroup of order eight). For each of these cases, the probability of undetected errors is obtained for length of the input block from k+1 to 20. Additionally, we obtain the probability of undetected errors for input blocks of length 64 and some other block lengths, which we will need in Section V. These probabilities are obtained for probability of bit-error p in the binary symmetric channel from 0.01 to 0.5 with step 0.01.

A. The simulation process

In this subsection we will give short description on the process of simulation. The probability of undetected errors for a block with length n, transmitted through a binary symmetric channel with probability of bit-error p is obtained in the following way:

First we generate an input message of length s which is the largest integer smaller than 1000000 which is devisable with n. The generated message is over the alphabet Q. Then, this message is separated into blocks with length n and each block is coded separately. After that, this coded message, which contains all the coded blocks, is turned into binary form. We generate the noise in the binary symmetric channel, which is a binary string with length equal to the length of the coded input message turned into binary form. The noise string is generated such that the probability of occurrence of one in this string is p. The positions in the noise string to which there is one are the positions on the coded input message in a binary form that are incorrectly transmitted. The output message is calculated with simply adding the coded input message in binary form and the noise string per modulo 2. Now, for every incorrectly transmitted block we check if equations (3) are satisfied. If the equation is not satisfied for some i, the error in transmission is detected. But, if all equations are satisfied, the error will not be detected. We count the number of incorrectly transmitted blocks in which the error is not detected. The probability of undetected errors is calculated by simply dividing this number with the total number of blocks in the input message.

B. The probability of undetected errors when linear quasigroup of order 4 is used for coding

For coding is used the linear quasigroup of order four represented with the following binary matrices:

$$A=\begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix},\ B=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \text{ and } C=\begin{bmatrix}0 & 0\end{bmatrix}.$$

Using the simulation process described in subsection *A*, we obtained the probability of undetected errors for different values of the block length n when for coding is used the above linear quasigroup of order 4. The maximums of this probability are given in Table I.

In the first column of Table I is given the length of the input block, in the second column are given maximums of the probability of undetected errors when k=5, and in the third column are given maximums of the probability of undetected errors when k=7. In the last column is given the quotient of the maximums of the probability of undetected errors in the case when k=5 and k=7.

Recall that the number of redundant characters is k+1. This means that when k=5, there are 6 redundant characters, while when k=7, there are 8 redundant characters. Since when a quasigroup of order 4 is used for coding, every character is represented with two bits in a binary form, follows that when k=5 there are 12 redundant bits, while when k=7 there are 16 redundant bits.

TABLE I. MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS WHEN FOR CODING IS USED QUASIGROUP OF ORDER 4

n	k=5	k=7	(k=5)/(k=7)
6	0.000330		
7	0.000392		
8	0.001216	0.000072	16.888889
9	0.003123	0.000153	20.411765
10	0.005740	0.000650	8.8307692
11	0.008327	0.001947	4.2768413
12	0.011700	0.003660	3.1967126
13	0.015249	0.005135	2.9696145
14	0.019306	0.007826	2.4669118
15	0.022080	0.009855	2.2404846
16	0.024880	0.012992	1.9150246
17	0.029750	0.015810	1.8817275
18	0.031320	0.017712	1.7682897
19	0.033763	0.020691	1.6317758
20	0.038900	0.022420	1.7350580
24	0.049441	0.032112	1.5396123
30	0.067411	0.045450	1.4831674
36	0.076682	0.059006	1.2995711
42	0.087236	0.069259	1.2595514
48	0.100034	0.081745	1.2237278
54	0.104439	0.088508	1.1799884
60	0.111784	0.097684	1.1443442
64	0.122080	0.099000	1.2331313
66	0.124876	0.104284	1.1974608
72	0.129248	0.108655	1.1895265
78	0.131903	0.112480	1.1726796
84	0.131903	0.115423	1.1427792
90	0.133831	0.112771	1.1867501
96	0.138441	0.119624	1.1573012
102	0.141385	0.125370	1.1277419
108	0.145000	0.136700	1.0607169
114	0.149900	0.139000	1.0784173
120	0.155000	0.141400	1.0961810

The empty fields in Table I are due to the constraint $k \le n-1$, which implies $n \ge k+1$. For this reason when $k=7$, block length n must be greater that or equal to eight.

As we can see from Table I, for a given k, when the block length n increases, the probability of undetected errors increases, too. This is expected result. Namely, when block length n increases, the number of characters that these k redundant characters need to control increases, thus increasing the chance that some incorrectly transmitted character will pass unnoticed. For this reason, from the aspect of the probability of undetected errors it is better to divide the messages in to blocks with as small as possible lengths.

From Table I can be seen that for small values of the block length n the probability of undetected errors in the case $k=7$ is much smaller than the probability of undetected errors in the case $k=5$. But, as the block length n increases, the quotient decreases and for larger values of the block length n, the two considered cases have almost equal probabilities of undetected errors. This means that for smaller values of the block length n it is better to use larger values of the parameter k, but as block length increases it loses the advantage since the probability of undetected errors is not so smaller any more, but computations are more complicated and larger value of k adds more redundant characters.

C. *The probability of undetected errors when linear quasigroup of order 8 is used for coding*

In this case for coding is used the linear quasigroup of order eight represented with the following matrices:

$$A=\begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 1\\ 1 & 1 & 1\end{bmatrix},\ B=\begin{bmatrix}0 & 1 & 1\\ 1 & 1 & 1\\ 1 & 0 & 1\end{bmatrix} \text{ and } C=\begin{bmatrix}0 & 0 & 0\end{bmatrix}$$

The values for the parameter k for which we obtained the probability of undetected errors are chosen in the following way. When quasigroup of order 8 is used for coding, binary representation of each character contains three bits. For this reason, in order to have 12 bits redundancy, which is 4 characters of the quasigroup, the parameter k must be 3. Since 16 is not divisible by 3, it is impossible to have 16 bits redundancy when quasigroup of order 8 is used for coding. The nearest natural number to 16 which is divisible by 3 is 15. In order the redundancy to be 15 bits, which is 5 characters from the used quasigroup, k must be 4.

The maximums of the probability of undetected errors when the above quasigroup of order 8 is used for coding in the cases when $k=3$ and $k=4$ are given in Table II.

As in the case when quasigroup of order 4 is used for coding, for a given k, when block length n increases, the probability of undetected errors increases, too. Also, for small values of the block length n the probability of undetected errors in the case $k=4$ is smaller than the probability of undetected errors in the case $k=3$. But, as the block length n increases, the quotient decreases and for larger values of the block length n, the two considered cases have almost equal probabilities of undetected errors, which can be seen from Table II.

TABLE II. MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS WHEN FOR CODING IS USED QUASIGROUP OF ORDER 8

n	$k=3$	$k=4$	$(k=3)/(k=4)$
4	0.00034		
5	0.00030	0.00007	4.285714
6	0.00032	0.00015	2.133333
7	0.00067	0.00010	6.700000
8	0.00193	0.00030	6.433333
9	0.00353	0.00135	2.614815
10	0.00532	0.00234	2.273504
11	0.00716	0.00374	1.914438
12	0.00958	0.00536	1.787313
13	0.01111	0.00688	1.614826
14	0.01336	0.00810	1.649383
15	0.01478	0.00961	1.537981
16	0.01679	0.01091	1.538955
17	0.01791	0.01310	1.367175
18	0.01952	0.01391	1.403307
19	0.02208	0.01577	1.400127
20	0.02321	0.01745	1.330086
24	0.03072	0.02396	1.282137
28	0.03598	0.02876	1.250967
32	0.03970	0.03366	1.179454
36	0.04500	0.03697	1.217312
40	0.04652	0.04116	1.130223
44	0.05336	0.04440	1.201802
48	0.05668	0.04695	1.207242
52	0.05885	0.04960	1.186492
56	0.06280	0.05390	1.165121
60	0.06645	0.05835	1.138817
64	0.06664	0.05834	1.142269
68	0.06970	0.05907	1.180022
72	0.07167	0.06113	1.172302
76	0.06587	0.05987	1.100223
80	0.06773	0.06447	1.050671

D. *The probability of undetected errors when linear quasigroup of order 16 is used for coding*

In this case we used the quasigroup of order sixteen represented with the following binary matrices:

$$A=\begin{bmatrix}0 & 1 & 1 & 1\\ 1 & 0 & 1 & 1\\ 1 & 1 & 0 & 1\\ 1 & 1 & 1 & 1\end{bmatrix},\ B=\begin{bmatrix}1 & 0 & 1 & 1\\ 1 & 1 & 0 & 1\\ 1 & 1 & 1 & 1\\ 0 & 1 & 1 & 1\end{bmatrix} \text{ and } C=\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}$$

The maximums of the probability of undetected errors when this quasigroup of order 16 is used for coding are given in Table III. We consider two cases: $k=2$ (given in the second column) and $k=3$ (given in the third column). In the last column is given their quotient. When quasigroup of order 16 is used for coding, binary representation of each character contains four bits. Therefore, when $k=2$, there are 3 redundant characters, i.e. 12 bits, while when $k=3$, there are 4 redundant characters, i.e. 16 bits.

As can be seen from Table III, the probability of undetected errors has exact the same properties explained in the case when quasigroups of order 4 and order 8 are used for coding. For this reason in all three cases are valid conclusions explained in Subsection *B* of this section.

TABLE III. MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS WHEN FOR CODING IS USED QUASIGROUP OF ORDER 16

n	k=2	k=3	(k=2)/(k=3)
3	0.00043		
4	0.00031	0.00006	5.166667
5	0.00034	0.00005	6.800000
6	0.00041	0.00006	6.833333
7	0.00086	0.00015	5.733333
8	0.00220	0.00040	5.500000
9	0.00359	0.00135	2.659259
10	0.00445	0.00216	2.060185
11	0.00563	0.00301	1.870432
12	0.00670	0.00392	1.709184
13	0.00774	0.00454	1.704846
14	0.00853	0.00531	1.606403
15	0.01020	0.00662	1.540785
16	0.01144	0.00707	1.618105
17	0.01228	0.00855	1.436257
18	0.01341	0.00931	1.440387
19	0.01394	0.01021	1.365328
20	0.01480	0.01099	1.346679
21	0.01582	0.01136	1.392606
24	0.01922	0.01454	1.321871
27	0.02024	0.01614	1.254027
30	0.02333	0.01803	1.294116
33	0.02543	0.02160	1.177250
36	0.02720	0.02067	1.316127
39	0.02768	0.02227	1.243112
42	0.03020	0.02448	1.233660
45	0.03070	0.02440	1.258197
48	0.03260	0.02720	1.198529
51	0.03355	0.02705	1.240296
54	0.03105	0.02850	1.089474
57	0.03320	0.02930	1.133106
60	0.03420	0.03020	1.132450
64	0.03646	0.03170	1.150157

V. IMPACT OF THE QUASIGROUP ORDER ON THE PROBABILITY OF UNDETECTED ERRORS

It is reasonable to compare the probabilities of undetected errors when lengths of the input blocks expressed in bits are equal and the numbers of redundant bits are also equal. For this reason we will separately analyze the two cases of the redundancy length: 12 redundant bits (Subsection *A*) and 16 redundant bits (Subsection *B*). In both of them we will compare the probability of undetected errors for input blocks with equal lengths expressed in bits, when for coding is used quasigroup of order 4, order 8 and order 16. Comparison is made using Table IV - Table VII in which the block length n is expressed in bits.

A. Twelve Bits Redundancy

In the first column of Table IV is given the block length in bits, in the second column is given the maximum of the probability of undetected errors when for coding is used quasigroup of order 4 (k=5), while in the third column is given the maximum when for coding is used quasigroup of order 8 (k=3). In the last column is given the quotient of these two probabilities.

TABLE IV. MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS WHEN REDUNDANCY IS 12 BITS AND FOR CODING ARE USED QUASIGROUPS OF ORDER 4 AND ORDER 8 AND THEIR QUOTIENT. BLOCK LENGTH IS EXPRESSED IN BITS

n	order 4	order 8	(order 4)/(order 8)
24	0.01170	0.00193	6.06217617
36	0.03132	0.00958	3.26934238
48	0.04944	0.01679	2.94465753
60	0.06741	0.02321	2.90438173
72	0.07668	0.03072	2.49616211
84	0.08724	0.03598	2.42471468
96	0.10003	0.03970	2.51993853
108	0.10444	0.04500	2.32086667
120	0.11178	0.04652	2.40292347
132	0.12488	0.05336	2.34025487
144	0.12929	0.05668	2.28031051
156	0.13190	0.05885	2.24134237
168	0.13190	0.06280	2.10036624
180	0.13383	0.06645	2.01401053
192	0.13844	0.06664	2.07744598
204	0.14138	0.06970	2.02847920
216	0.14500	0.07167	2.02325487
228	0.14990	0.06587	2.27580856
240	0.15500	0.06773	2.28838695

In the first column of Table V is given the block length in bits, in the second column is given the maximum of the probability of undetected errors when for coding is used quasigroup of order 8 (k=3), while in the third column is given the maximum when for coding is used quasigroup of order 16 (k=2). In the last column is given the quotient of these two probabilities.

The data for the probabilities of undetected errors in Table IV and Table V are extracted from Table I - Table III by selecting appropriate k and calculating the block length expressed in characters from the appropriate quasigroup. In order to obtain the block length expressed in characters from the quasigroup of order 4, order 8 and order 16, the block length n expressed in bits should be divided by 2, 3 and 4, respectively.

As we can see from Table IV and Table V, using quasigroup of higher order provides smaller probabilities of undetected errors. For smaller values of the block length n, the reduction of the probability of undetected errors is larger. For larger values of the block length n, the probability of undetected errors is a little bit more than twice smaller when quasigroup of order 8 is used than when quasigroup of order 4 is used. Also, for larger values of the block length n quasigroup of order 16 gives approximately twice smaller probability of undetected errors than the quasigroup of order 8. From Table IV and Table V can be obtained the result how many times the quasigroup of order 16 gives smaller probability of undetected errors than the quasigroup of order 4. Namely, it can be easily verified that if for a given block length n, the quasigroup of order 8 gives t_1 times smaller probability of undetected errors than the quasigroup of order 4 and for the same n, the quasigroup of order 16 gives t_2 times smaller probability of undetected errors than the quasigroup of order 8, than for that n, the quasigroup of order 16 gives t_1t_2 times smaller probability of undetected errors than the quasigroup of order 4.

TABLE V. MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS WHEN REDUNDANCY IS 12 BITS AND FOR CODING ARE USED QUASIGROUPS OF ORDER 8 AND ORDER 16 AND THEIR QUOTIENT. BLOCK LENGTH IS EXPRESSED IN BITS

n	order 8	order 16	(order 8)/(order 16)
24	0.00193	0.00041	4.70731707
36	0.00958	0.00359	2.66852368
48	0.01679	0.00670	2.50597015
60	0.02321	0.01020	2.27549020
72	0.03072	0.01341	2.29082774
84	0.03598	0.01582	2.27419722
96	0.03970	0.01922	2.06540062
108	0.04500	0.02024	2.22332016
120	0.04652	0.02333	1.99371713
132	0.05336	0.02543	2.09842461
144	0.05668	0.02720	2.08382353
156	0.05885	0.02768	2.12608381
168	0.06280	0.03020	2.07947020
180	0.06645	0.03070	2.16449511
192	0.06664	0.03260	2.04417178
204	0.06970	0.03355	2.07749627
216	0.07167	0.03105	2.30810628
228	0.06587	0.03320	1.98393675
240	0.06773	0.03420	1.98050585

For illustration of the reduction of the probability of undetected errors for all values of the bit-error p in the binary symmetric channel and the shape of this function (as function of the bit-error p) in Fig. 1 are given probabilities of undetected errors for smaller block length, while in Fig. 2 are given probabilities of undetected errors for larger block length for all three used quasigroups.

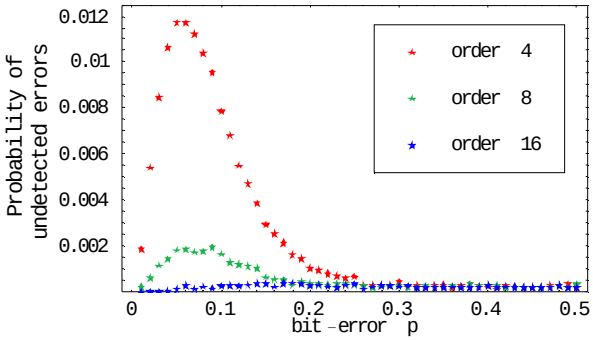


Figure 1. The probability of undetected errors when block length is 24 bits and redundancy is 12 bits.

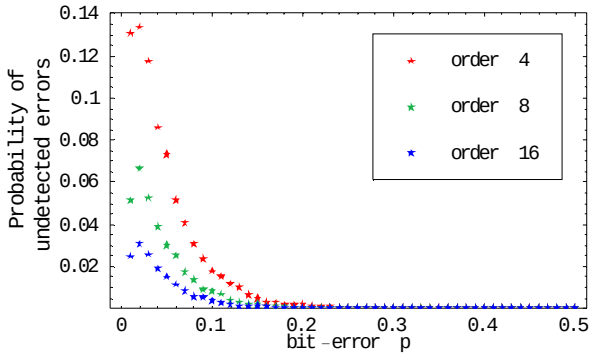


Figure 2. The probability of undetected errors when block length is 180 bits and redundancy is 12 bits.

B. *Sixteen Bits Redundancy*

In Table VI are given the maximums of the probabilities of undetected errors when the quasigroup of order 4 (k=7) and the quasigroup of order 8 (k=4) are used for coding, while in Table VII are given the maximums of the probabilities of undetected errors when the quasigroup of order 8 (k=4) and the quasigroup of order 16 (k=3) are used for coding. The values for these two tables are obtained from Table I – Table III in a same manner as described for Table IV and Table V.

From Table VI and Table VII we see that the same conclusions that we made in the case when there are 12 redundant bits are valid in this case, too.

For illustration in Fig. 3 are given probabilities of undetected errors for smaller block length, while in Fig. 4 are given probabilities of undetected errors for larger block length for all three used quasigroups.

TABLE VI. MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS WHEN REDUNDANCY IS 16 BITS (I.E. 15 BITS FOR THE QUASIGROUP OF ORDER 8) AND FOR CODING ARE USED QUASIGROUPS OF ORDER 4 AND ORDER 8 AND THEIR QUOTIENT. BLOCK LENGTH IS EXPRESSED IN BITS

bits	order 4	order 8	(order 4)/(order 8)
24	0.00366	0.00030	12.20003333
36	0.01771	0.00536	3.30451492
48	0.03211	0.01091	2.94340055
60	0.04545	0.01745	2.60461318
72	0.05901	0.02396	2.46267529
84	0.06926	0.02876	2.40818846
96	0.08174	0.03366	2.42876837
108	0.08851	0.03697	2.39427647
120	0.09768	0.04116	2.37327259
132	0.10428	0.04440	2.34873874
144	0.10865	0.04695	2.31427050
156	0.11248	0.04960	2.26774193
168	0.11542	0.05390	2.14142857
180	0.11277	0.05835	1.93266495
192	0.11962	0.05834	2.05046280
204	0.12537	0.05907	2.12251573
216	0.13670	0.06113	2.23609718
228	0.13900	0.05987	2.32182499
240	0.14140	0.06447	2.19338046

.

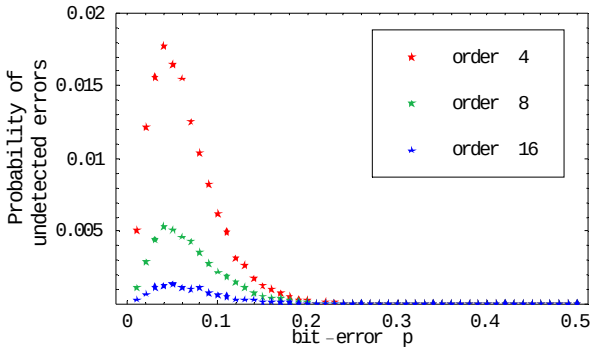


Figure 3. The probability of undetected errors when block length is 36 bits and redundancy is 16 bits.

TABLE VII. MAXIMUMS OF THE PROBABILITY OF UNDETECTED ERRORS WHEN REDUNDANCY IS 16 BITS (I.E. 15 BITS FOR THE QUASIGROUP OF ORDER 8) AND FOR CODING ARE USED QUASIGROUPS OF ORDER 8 AND ORDER 16 AND THEIR QUOTIENT. BLOCK LENGTH IS EXPRESSED IN BITS

n	order 8	order 16	(order 8)/(order 16)
24	0.00030	0.00006	5.00000000
36	0.00536	0.00135	3.97037037
48	0.01091	0.00392	2.78316326
60	0.01745	0.00662	2.63595166
72	0.02396	0.00931	2.57357680
84	0.02876	0.01136	2.53169014
96	0.03367	0.01454	2.31479367
108	0.03697	0.01614	2.29037794
120	0.04116	0.01803	2.28282391
132	0.04440	0.02160	2.05555556
144	0.04695	0.02067	2.27177053
156	0.04960	0.02227	2.22754158
168	0.05390	0.02448	2.20179739
180	0.05835	0.02440	2.39139344
192	0.05834	0.02720	2.14485294
204	0.05907	0.02705	2.18361183
216	0.06113	0.02850	2.14502807
228	0.05987	0.02930	2.04323208
240	0.06447	0.03020	2.13465894

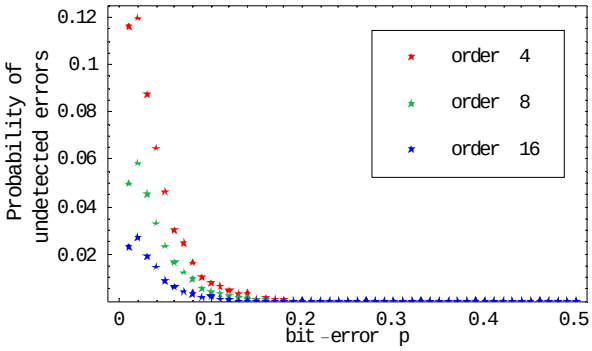


Figure 4. The probability of undetected errors when block length is 192 bits and redundancy is 16 bits.

From Fig.3 and Fig. 4 can be seen the proportion of the functions of the probability of undetected errors for one smaller and one larger input block. We give one figure with smaller and one with larger block length since in the above analyses we concluded that in these two cases the proportion of the probabilities are different.

VI. CONCLUSION

Using simulation we analyze a model of error-detecting linear code based on quasigroups. We considered the cases when linear quasigroups of order 4, order 8 and order 16 are used for coding. For each of these cases are analyzed two cases for the parameter *k* of the error-detecting model: one which gives 12 redundant bits and the second which gives 16 redundant bits when quasigroup of order 4 and order 16 is used, i.e. 15 redundant bits when quasigroup of order 8 is used.

From obtained results we conclude that for a given quasigroup and parameter *k*, when block length *n* increases, the probability of undetected errors increases, too. Also, for small values of the block length *n*, the larger value of the parameter *k* gives smaller probability of undetected errors. But, as the block length *n* increases larger *k* loses the advantage since for larger values of the block length *n* the two considered cases for the parameter *k* have almost equal probabilities of undetected errors. From this we conclude that for smaller values of the block length *n* it is better to use larger value of the parameter *k* (more redundant characters), but for larger values of the block length *n* it is better to use smaller values of the parameter *k* since probability of undetected errors is not so larger any more, but computations are simpler and are added less redundant characters. Overall, from the aspect of the probability of undetected errors it is best to divide the messages in to as small as possible blocks and in the same time to choose larger values for the parameter *k*. These conclusions are valid for all considered orders of the quasigroup used for coding.

Also, we conclude that the multiplication of the order of the quasigroup used for coding by two results with approximately twice smaller probability of undetected errors for larger values of the block length *n*. For smaller values of the block length, using quasigroup of higher order for coding results with larger reduction of the probability of undetected errors.

Due to the simple linear construction of the quasigroups, the codes that we defined and analyzed in this paper are suitable for use in light-weight hardware devices.

REFERENCES

[1] N. Ilievska, D. Gligoroski, "Simulation of some new models of error-detecting codes," Proc. of the 22nd Telecommunications Forum Telfor, Belgrade, Serbia, pp. 395-398, 2014.

[2] V. Bakeva, N. Ilievska, "A probabilistic model of error-detecting codes based on quasigroups," Quasigroups and Related Systems, vol. 17, no. 2, pp. 135–148, 2009.

[3] N. Ilievska, D. Gligoroski, "Error-detecting code using linear quasigroups," ICT Innovations 2014, Advances in Intelligent Systems and Computing Vol. 311, Springer International Publishing, pp. 309-318, 2015.

[4] N. Ilievska, "Proving the probability of undetected errors for an error-detecting code based on quasigroups," Quasigroups and Related Systems, vol. 22, no. 2, pp. 223-246, 2014.

[5] N. Ilievska, D. Gligoroski, "An error-detecting code based on linear quasigroups," Proc. 11th Int. Conf. on Informatics and Information Technologies, R.Macedonia, 2014, in press.

[6] N. Ilievska, V. Bakeva, "A model of error-detecting codes based on quasigroups of order 4," Proc. 6th Int. Conf. on Informatics and Information Technologies, R.Macedonia, pp. 7-11, 2008.

[7] D. Gligoroski, V. Dimitrova, S. Markovski, "Quasigroups as Boolean functions, their equation systems and Grobner bases," short-note for RISC Book Series, Springer, "Groebner, Coding and Cryptography", Ed.T. Mora, L.Perret, S.Sakata, M.Sala and C.Traverso, pp. 415-420, 2009.

[8] B. D. McKay, I. M. Wanless, "On the number of Latin squares," Annals of Combinatorics, vol. 9, pp. 335-344, 2005.

An integrated remotely controlled oceanographic measuring station

Kalinić, H.; Jelavić, D.; Ivanković D.; Bašić, J.; Muslim, S.; Marasović, S; Dadić, V.

Institute of Oceanography and Fisheries, Croatia

hrvoje.kalinic@izor.hr

Abstract – In this paper describes an oceanographic measuring station as a part of automated system for oceanographic measurement with remote control and data visualization. Oceanographic measuring station consists of operator's computer and a number of different measuring devices that produce data continuously. Each device has its own software interface that allows user interaction with the data. Such interface suffice when operator work with only few measuring devices and when the type (manufacturer) of a measuring device is not frequently changing, but it becomes cumbersome as more and more various devices became available for oceanographic data acquisition. In our setup the operator's computer is replaced by embedded computer dedicated to serve as a middle layer between sensor and database server with whom the operator communicates. The embedded Linux box run a set of daemons which acquire and archive data and transfer it to database, and a daemon that supervise the device and the entire process. A part of the acquired data is publicly available through visualization on Institute's website (www.izor.hr). The dynamic visualization of the data is done in javascript.

I. INTRODUCTION

The Mediterranean is one of the busiest seas in the world in terms of traffic. Even if it covers only 0.7% of the total sea surface in the world, it hosts 30% of the overall international maritime traffic. In order to monitor it and measure the rate of change and human influence on such important oceanographic system many information from both simulations and in-situ data acquisition is required.

In oceanography a vast amount of information is produced by many oceanographic sensors that measure physical[1], biological[2], chemical[3] parameters which are scattered across the shore, on the sea surface or in the sea. The type of data acquired by various oceanographic sensors ranges from meteorological data like air temperature, pressure, wind speed, insolation, humidity, rain accumulation (rain rate), ultraviolet radiation etc. to oceanographic data like sea temperature, conductivity (a proxy for salinity), chlorophyll, scattered matter, CDOM, sea current velocity, wave size, depth (measured with pressure) of the water etc. [4,5]

An important and interesting kind of oceanographic data acquisition happens when multiple informations are acquired from various sensors at various locations (or depths). This necessary leads to large amount of data, provided by the large amount of various sensors. Managing multiple data sources easily becomes cumbersome. Thus, an automated system for ocanographic research was a long term goal of many ocenographers [6,7]. Remote control of these systems are preferable, as they can be sparsely distributed over a large area or might need a frequent sampling at the specific area. Moreover, centralization of this access would contribute for easier data integration, research and could allow sensors to interact to one another and to adapt to environmental changes.

II. PROBLEM DESCRIPTION

As large amount of sensors produce data continuously the need for a more systematic approach for data load and storage becomes necessary. Many measuring devices come with heterogeneous proprietary client based software which makes them suitable for on-situ operations, but less suitable for automated, remote or even centralized control. In a typical setup, measuring device is connected to an operator's personal computer via serial data interface using RS232, RS485 or sometimes even RS422. This, as a consequence, puts the requirement for the operator to be within in a close range of the sensor in order to access and manipulate data. The close range of is defined by the RS232 standard that states the maximum capacitive load of 2 500pF. Although low-capacitance cables can be used for serial connection and thus allow larger distance, for an ordinary cable, this means that at the most a few tens of meters of cable should be used for a reliable serial connection. Therefore, for a longer distance, other standards are better suited.

In oceanography, often a satellite communication is used to transfer data to a distant location. However, this is an expensive form of communication. Luckily, the Croatian side of Adriatic is largely covered by the signal from commercial mobile operator due to the largely indented coast. Thus, we aim to exploit this advantage and use a general packet radio service (GPRS) for data transfer.

The measuring devices come have vendor specific communications protocols. This is usually solved with a dedicated user applications provided by the manufacturer, and is quite suitable for manual data acquisition. However, in order to automate the process, an automated method is preferred. Moreover, for a centralized data access all the sensor data needs to be stored in a uniform way.

All of this can be done with a single device that (i) translate the data from the sensor and store it in a file and (ii) uses GPRS to transfer it to a file server. For the first part the communication with each vendor specific device needs to be solved, and in the second case it is preferable to reduce power consumption by switching the GPRS

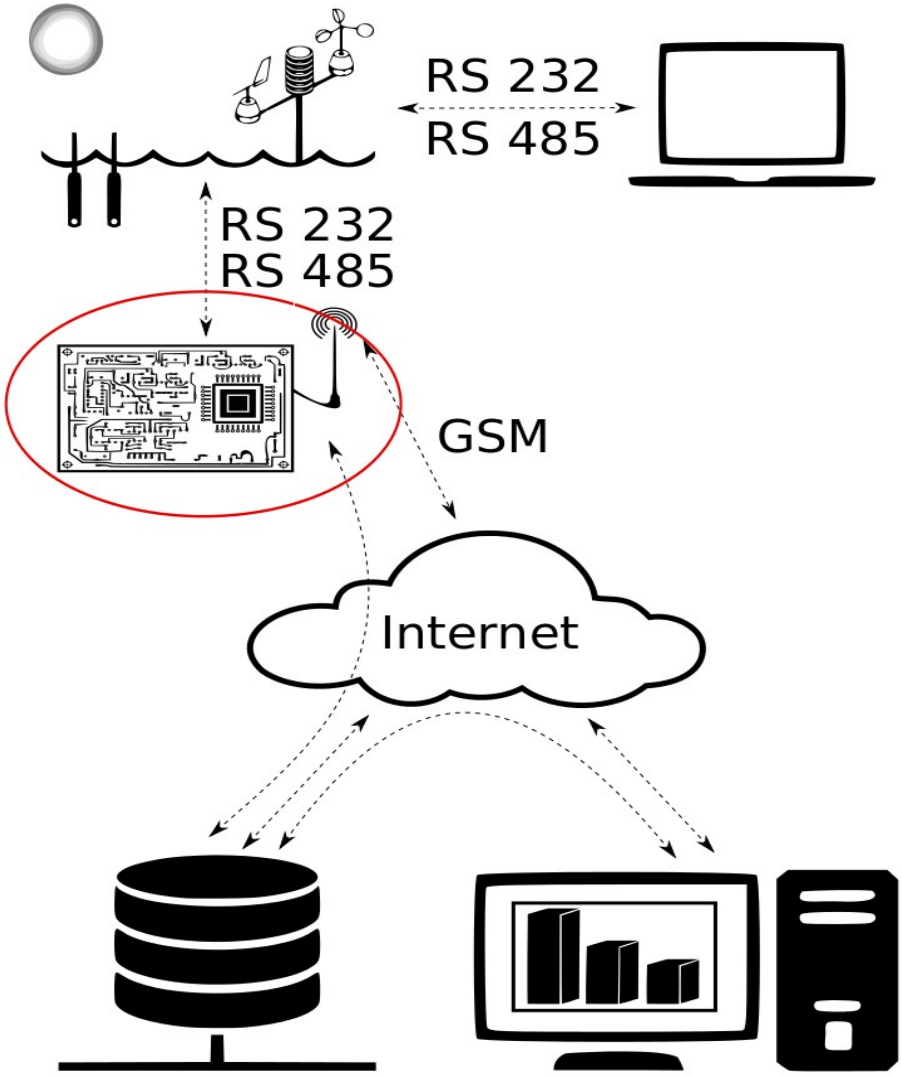


Figure 1. A schematic overview of an automated system for oceanographic measurement with remote control and data visualization. The first row depicts a typical setup where measuring device is connected to a computer (via RS232) through which the operator access and manipulate the data. This setup is replaced with the embedded computer and the database server with whom the operator communicate. The embedded computer is marked in red.

module on and off depending on the amount of acquired data. As addition, a watch-dog or a supervisor that contributes to larger reliability of the system is desirable. Finally, once the data is available on the server, it should be visualized for a data inspection or dissemination.

Both the classical approach and the proposed solution to the data acquisition are described in the schematic overview given in Figure 1. In the top of the Figure 1, a typical setup where measuring device is connected to a computer through which the operator access and manipulate the data. This setup is replaced with the embedded computer and the database server with whom the operator communicate.

Notice how the proposed architecture allows automated near real-time oceanographic data acquisition and does not require operator presence on the site to control the device or visualize the data. In the next section we will describe the embedded computer in more details

III. PROPOSED SOLUTION

Our proposed solution extends the idea of a measuring system by replacing the local personal computer of the measuring station with the embedded computer. In this way, the embedded computer can be utilized as the middle layer between the sensor and a server that stores the data. In this setup the data is transferred via Internet to a file server using FTP protocol. The data is then transferred to a database which permanently stores the data and metadata (e.g. the measuring system configuration). Once the data is in the database the operator can access and manipulate the data, or even change the configuration of the measuring system from remote location.

In order to automate this process a set of specially designed applications which run as a daemons on a Linux [8] embedded computer are initialized to gather the data from a device and to write the data to a file which is later transferred to the database server. Although all daemons that gather the data communicate with the device via serial data (RS232, RS422 or RS485) interface, each daemon is device specific and can be considered as a wrappers which unify the interface for a second stage of the process, which is file transfer. We will refer to such a program as data acquisition program and it its functionality can broadly be described with steps depicted in flow diagram shown in Figure 2. Naturally, the program is enclosed with opening of the serial port, sensor wake-up and reset of the watch-dog counter and it repeats for each cycle of the sensor data acquisition.

The gathering of the data is done in cycles where in each cycle each sensor produce a number of measurements that are stored in separate temporary files. Once the measurement cycle is complete the temporary files are marked as new, and ready for sending. Thus, the files are compressed and sent to folder named "2send" which acts as a pipe. Each of these files are appended to a file in the archive folder which is compressed once a day. If there is an Internet connection available, the files are sent, otherwise they are appended to a sensor specific file, which is scheduled to be sent once the connection is restored. The "archive" folder is acting like a large circular buffer, and any data from this folder can be sent to server upon request. The process is depicted in the chart given by Figure 3, and most of the depicted operations happen in the daemon called file manager.

In particular, file manager operates on the newest files once the sensor data acquisition finishes its cycle and crates a new file from a temporary file. In most basic implementation file manager operates with FTP protocol, and apart form sending or scheduling files for a later transfer, the file manager can respond to a request form the server, and resend the files form the archive if they are

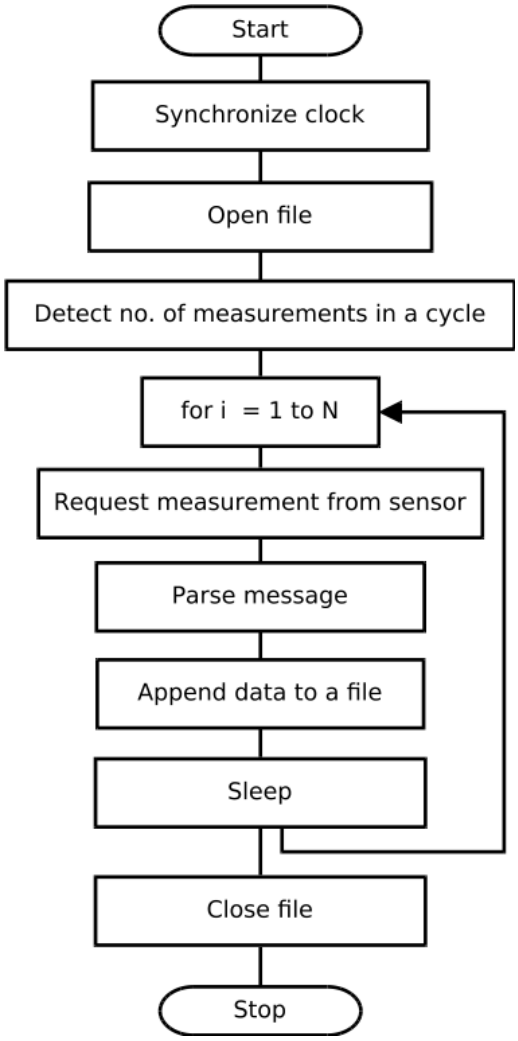


Figure 2. Flow diagram for data acquisition

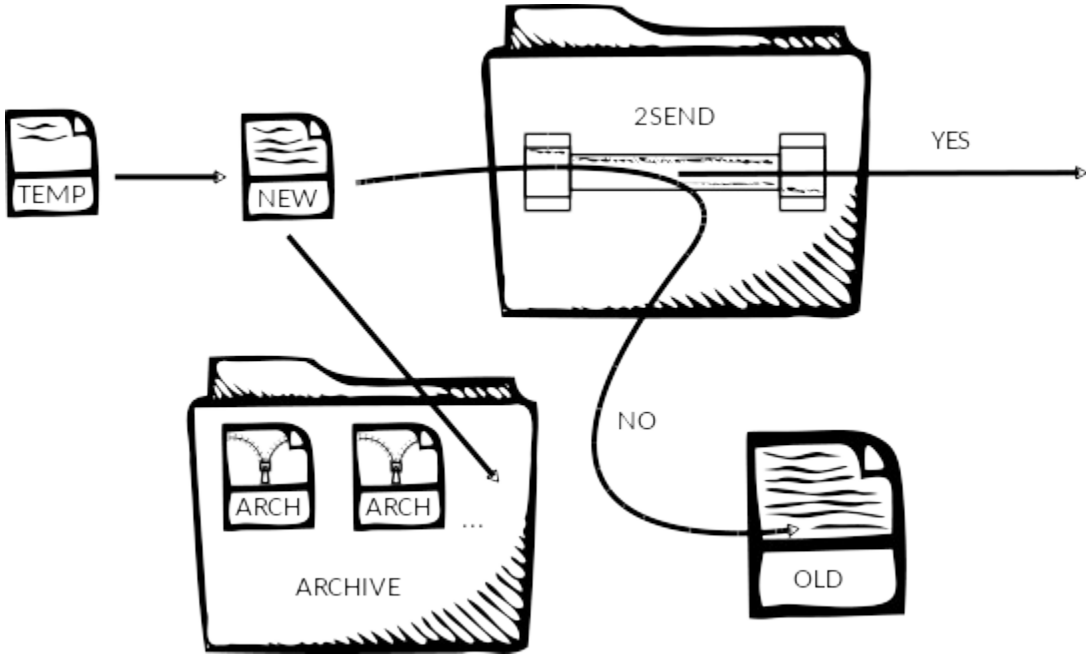


Figure 3. Process file procedure within the file management daemon

corrupted, inconsistent or missing. An overview of a file manager implementation is given by the following pseudo-code:

```
read configuration
reset watch-dog counter
download new configuration files (if any)
process files
sleep
```

As can be noticed form the pseudo-code, apart form data management, a file manager is used to download configuration download as well. The configuration files downloaded by the file manager can be used to set-up file manager, as well as the sensor or the data acquisition process, or some other part of the system. The notification that the configuration files are downloaded file server receives through standard process file procedure, i.e. the configuration file (once loaded) copies to the "2send" folder. In this way, the server will receive the station configuration once the network is available. If the station configuration is as the server requested it to be, it will stop sending the configuration files.

The daemons are supervised by a separate program – the supervisor – that assures that no data is lost due to the temporary power failure, unresponsive devices or programs, broken processes, etc. Thus, both data acquisition daemon and the file manager daemon are expected to report their proper functionality, or face restart. In addition to basic functionality of the watchdog (if timer is not reset, reset the daemon) the supervisor also inspects the overall health of the system and logs the information in a separate file. This file can be sent to the server as any other file. Similarly, the configuration file that describes which part of the system are monitored can be downloaded from a file server using the file manager as well.

The final part of the system is on the server side, where received data has been unpacked and passed to database. Database input is done through external table, which apart from data input, also validates the data, (if necessary, depending on the sensor type) calculate the mean value and in case of error or inconsistency send warnings.

IV. System performance

The embedded device we are using as the Linux box has a 192 Mhz RISC CPU and 32 MB SDRAM with up to four RS232 ports. In the experimental setup the larger number of measuring devices than the number of RS232 ports can be attached if a ETH to RS232 conversion or a multiplexer is used. Each device have multiple sensors and sometimes measure 20 or more parameters.

The system was tested in near real time data acquisition, where measuring cycle of devices was reduced to one minute with data transmission cycle reduced to a minute as well. The number of generated files depend on the number of the attached devices, and the size of each file depends on the number of sensors within each devices, or more precisely, on the number monitored parameters. Naturally, if the transmission cycle is not the same as the measuring cycle the size of the file will depend on that parameter as well.

Each manufacturer equips the measuring device with different sensors and various additional parameters can be retrieved from it. Typical parameters are: temperature humidity, pressure, conductivity, wind, waves, oxygen concentration, rain or hail intensity or duration etc. Some parameters can be measured outside or inside, some (as pressure or temperature) are measured for both sea and air, and some are measured on different depths, while some or vector values whose velocity and direction are reported as independent paremeters.

Consider a typical setup the number of monitored parameters easily surpass few dozens since 4 measuring devices can be attached to the embedded Linux box. If the system is operating in near real time data acquisition each device each minute create a file between 100B and 1kB in size, which for one day totals to 1440 files or 5760 files from all four devices. The size of the archived files, after compression, typically ranges between 13 and 300 kB per device.

V. Acquired data visualization

For data visualization we used pure JavaScript charting library - Highcharts. Along with numerous chart types and data manipulation techniques, Highcharts supports many functionalities that increases interactivity, such as: tooltip labeling, zooming, exporting and printing images. It also supports multiple axes if you want to compare variables that are not the same scale - for example air temperature versus Humidity. This Web page uses JSON (lightweight data-interchange format) adapted to read raw data fetched from the database and to draw these values into the chart. This method of formatting data allow us to load JSON file externally and draw dynamic charts that are automatically updated each minute.

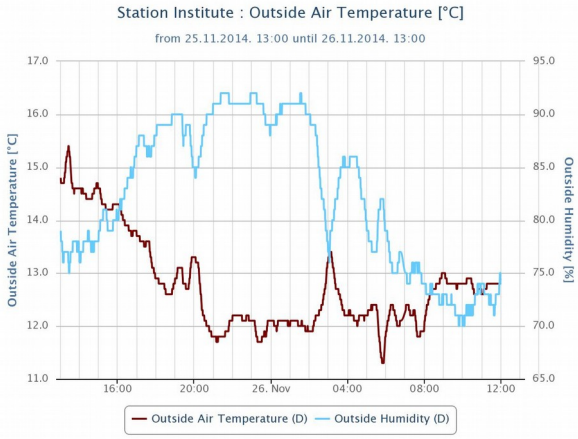


Figure 4. 24h graph for air temperature and humidity

It can be defined one or two measuring parameters to display on chart. Visualization procedure also enables to define date and time (rounded to the minute) depending on the last value that should be shown on a chart, and timespan (defined in hours) preceding that time. If these two variables are not user defined, procedure draws

measured data of the last 24 hours. Data density is rounded by the minute and is displayed in line chart. An example of such chart from the webpage is given in Figure 4. With Highcharts other visualization form sensors are possible, as well, such as an example showing a wind rose graph in Figure 5.

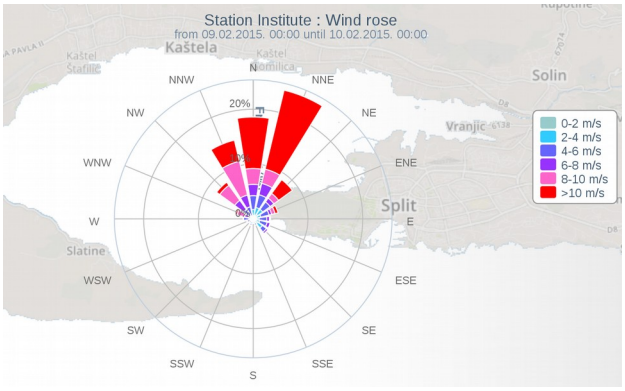


Figure 5. A wind rose form the station at the shore near Institute

VI. Conclusion

The proposed solution extends the idea of a measuring system by replacing the local personal computer of the measuring station with the embedded computer. On the embedded computer three daemons operate: (i) data gathering daemon – which is device specific and responsible for communication with the device, (ii) file manager daemon – which is is responsible for unified communication with the server, and (iii) the supervisor that assures the first two daemons work correctly and inspects the overall health of the system.

The proposed architecture allows automated near real-time oceanographic data acquisition and does not require operator presence on the site to control the device or visualize the data. The system is already functional and part of the acquired data is publicly available through Institute's website (www.izor.hr). The dynamic visualization of the data is done with Highchart visualization engine and visualization engine developed by Institute.

VII. Bibliography

[1] H. Mihanović, S. Cosoli, I. Vilibić, D. Ivanković, V. Dadić, and M. Gačić, “Surface current patterns in the northern Adriatic extracted from high-frequency radar data using self-organizing map analysis”. Journal of geophysical research 2011, pp. 8033-8033

[2] M. Rixen, J.M. Beckers, C. Maillard, M. Fichaut, G. Maudire, C. Coatanoan, E. Balopoulos, A. Iona, A. Lykiardopoulos, Karagevrekis, M.J. Garcia, B. Manca, A. Giorgetti, A. Mosetti, N. Mikhailov, E. Vyazilov, A. Kuznetsov, A. Puzova, R. Boukortt, B. Boudjellal, N. Eddalia, D. Dooley, A. Drago, S. El-Agami, G. Kortchev, I. Gertman, Y. Tsehtik, S. Lakkis, G. Manzella, I. Oliounine, A. Orbi, J. Larissi, S. Zizah, M. Ozyalvac, F. Berkay, N. Pinardi, M. Zavatarelli, A. Suvorov, A. Khaliulin, G. Zodiatis, K. Bilashvili, Z. Savaneli, V. Dadic, V. Diaconu, R. Gelfeld, and S. Sammari, “A hydrographic and bio-chemical climatology of the mediterranea and Black Sea: a technical note on the use of coastal data”. Bollettno di Gofisicia teorica ed applicata. Vol 4., 2005, pp. 319-327

[3] J.C. Orr, V.J. Fabry, O. Aumont, L. Bopp, S.C. Doney, R.A. Feely, A. Gnanadesikan, N. Gruber, A. Ishida, F. Joos, R.M. Key, K. Lindsay, E. Maier-Reimer, R. Matear, P. Monfray, A. Mouchet, R.G. Najjar, G.K. Plattner, K.B. Rodgers, C.L. Sabine, J.L. Sarmiento, R. Schlitzer, R.D. Slater, I.J. Totterdell, M.F. Weirig, Y. Yamanaka, and A. Yool, “Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms”, Nature, 2005, pp. 681-686 . doi: 10.1038/nature04095

[4] C.W. Anderson, D.A. Kitchin, K.S. Decker, “Oceanographic Sensor System Development”, Johns Hopkins APL Tehnical Digest, 2002, vol 23, no. 4

[5] George L Pickard and William J. Emery. “Descriptive Physical Oceanography, An introduction”. 5th ed., Butterworth-Heineman, Elsevier Science, 1990

[6] W. Yu, L. Yingjian, and G. Zhongwen, “Three-Dimensional Ocean Sensor Networks: A Survey”, J. Ocean Univ. China, 2012, vol 11, no. 4

[7] E. Crisalfi, F. Azzaro, G. Zappala, and G. Magazzu, “Integrated Automatic systems for oceanographic research: some applications”, OCEANS '94. 'Oceans Engineering for Today's Technology and Tomorrow's Preservation.' Proceedings IEEE, 1994, vol 3

[8] E. Siever, S. Figgins, R. Love, and A. Robbins, Linux in a Nutshell, 6th edition, 2009, O'Riley Media

Role of Performance Measurements in Process of Migration to IP Mobile Backhaul in 3G Networks

D. Ilišević * and N. B. Ćurguz*
* Telekomunikacije RS a.d., Banja Luka, Bosnia and Herzegovina
dijana.ilisevic@mtel.ba, natasa.banovic-curguz@mtel.ba

Abstract - New 3G and Long Term Evolution (LTE) mobile network has again made mobile backhaul a key area, especially regarding current network, and considering LTE growing traffic requirements. Network planning is undergoing major changes in terms of the technologies and methodologies. Smooth evolution to LTE networks will enable transformation from circuit-switched (CS) to all-IP architecture and ultimately to network function virtualization (NFV) to facilitate ICT transformation. This paper presents use of performance measurements to improve quality in process of migration to IP mobile backhaul in 3G networks (LTE license is not available in B&H).

I. INTRODUCTION

Demand for mobile backhaul (MBH) capacity continues to rise as consumers using more data and video services with their mobile handsets. The constant increase in data traffic is also forcing Mobile Service Providers (MSP) to invest in their infrastructure capacity in all segments of the networks, from the core to the access and transport. Ensuring multiple network technologies increases operational and engineering complexity, reinforcing the need for new ways to evaluate quality.

While transitioning to an IP infrastructure help MSPs address their current bandwidth demands; it does introduces some new level of management complexity that can impact service quality and often leads to strong OPEX increase. At the same time, as the scope and nature of mobile services continue to evolve (from 2G Base Transceiver Station (BTS) to 3G+ NodeB (NodeB is a term used in UMTS equivalent to the BTS) , 4G LTE and now LTE Advance), efficient utilization of MBH in migration process requires using a mix of pseudowire, native and hybrid deployments, and each require separate end-to-end management.

II. IP MOBILE BACKHAUL

After the deployment of HSPA (High Speed Packet Access), the volume of mobile packet data traffic has exploded to the point where it has exceeded circuit-switched traffic. Furthermore, with the deployment of LTE and 4G mobile systems, it is expected to increase even more [1]. To cope with the changed traffic composition, operators need to migrate their legacy Time Division Multiplexing (TDM) networks to packet-switched backhaul networks capable of supporting high volumes of packet traffic (while maintaining low OPEX).

Mobile backhaul refers to the network between base station sites and network controller/ gateway sites for all generations of mobile technologies (Base Station Controller – BSC in 2G networks and Radio Network Controller-RNC in 3G networks). As MSPs are moving to all – IP (Internet Protocol) mobile world, transport network are undergoing a dramatic migration to support a large bandwidth data and video services. At the core of the mobile network transformation is the migration from voice-based technologies TDM to multimedia communications based on the IP standards. This migration reflects the fact that consumers are moving from voice-centric services to more data-intensive and mobile services such as real-time video streaming. MSPs that utilize IP and the related SIP (Session Initiation Protocol) standards can deliver mobile data services at a higher quality and lower cost.

Current network transformation involves much more than replacing TDM-based equipment with IP-based equipment. Transformation is a holistic process that typically happens in phases, depending on the unique business plans that a MSP is facing. Analysis of the different migration options shows that there is no unique solution or single migration path that fits all types of backhaul networks. Irrespective of the level of Ethernet support, the migration path involves an intermediate step of a TDM/ATM/Packet hybrid backhaul. The hybrid nature of the mobile backhaul will depend upon using interface between the BTS/NodeB and BSC/RNC. Hybrid approaches of tomorrow will reverse the hierarchy by carrying TDM traffic encapsulated over packet based transport. Existence of 2G/3G/4G over a common wireless access network besides new traffic demands gives us a lot of challenges, usually to support equipment configuration of different generation [2]. Main goal is to fulfill the inexpensive and rapid deployment of the following:

- Convergence of 2G/3G/4G over a common mobile backhaul (provide comprehensive support for current 2G and 3G traffic, and provide an easy migration path to LTE).
- Scalable solutions (sufficient E1 capacity and allocate enough Ethernet capacity to fulfill current and future traffic growth).

- QoS ensured through traffic classification, policing (rate limiting), scheduling to queues and shaping of egress traffic to a specific line rate.
- Synchronization (ensure backhaul nodes capable of effectively interworking between TDM, SyncE and Precision Time Protocol)

To satisfy such escalating and changing demands, MSPs need to be able to deploy and expand the new systems more easily, and also have the means to monitor and optimize system performance automatically. Therefore, such demands give a lot of pressure to transport part of network to make them to become self – tuning especially in 2G/3G/4G networks. New Radio Access Network (RAN) architecture such as HSPA, HSPA+ and LTE has led to an increasing demand for backhaul capacity. BTS and NodeB interfaces A_{bis} and I_{ub} could be realized as shown in Table I.

TABLE I. A_{BIS} AND I_{UB} INTERFACES

2G	Interface	3G	Interface
A_{bis}	FE/GE L2	I_{ub}	FE/GE L2
A_{bis}	FE / GE L3	I_{ub}	FE / GE L3
A_{bis}	E1	I_{ub}	E1

Regarding possible scenarios, it is imperative to choose best solution of interfaces, layers (L2 or L3) to ensure network capacity demands regarding investments, legacy and traffic growth and following possible advantages and disadvantages of L2/ L3 networking.

Network planning is undergoing major changes in terms of the technologies and methodologies. As MSP industry becomes increasingly competitive, planning become ever more important. To ensure LTE ready IP MBH is possible even in 3G networks. In process of modernization of 2G/3G network new RAN equipment has to be swapped and new deployments are LTE ready. With new advanced transport modules it is possible to BTS/NodeB will be directly connected to optical or electrical FE interface on site using WMPT module (ALL IP sites). It should be L3 networking and optical/ electrical FE port on WMPT is used for co-transmission for 2G and 3G base station as given in Figure 1. WMPT (WCDMA Main Processing and Transmission Unit) board is an IP gateway for GTMU boards (GSM Transmission& Timing&Managament Unit) and WMPT acts as a router.

There are four options for VLAN usage:

- 2G and 3G use only one VLAN: 2G/3G service, signaling and NodeB O&M will be in same VLAN.
- 2G and 3G use separate VLAN: 2G is one VLAN and 3G is in another.
- 2G and 3G use separate VLAN: 2G and 3G uses one VLAN for service, other VLAN for 3G O&M (recommend this option).
- 2G and 3G use three VLANs: 2G and 3G use separate VLAN. NodeB O&M is assigned third VLAN.

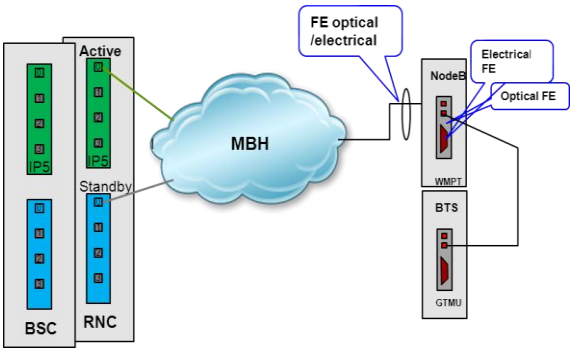


Figure1. ALL IP 2G/3G sites.

Use of L3 is proposed, and basically is questioning what level IP (LTE) MBH is mature [3]. Understanding L3 advantages requires understanding additional transport capabilities to build up LTE ready MBH. This approach allows MSPs to make a strategy to support future LTE network and easy to predict future capacity requirements. Deployment of LTE MBH also assume besides transition from TDM to IP but also reviewing the network topology. In case of microwave MBH, this assume move from 1+1 to 1+0 and ring topology as enabling more capacity and availability (reducing Capex and Opex).

III. PERFORMANCE MEASURMENTS – IMPROVING NETWORK QUALITY

Process of standard telco metrics were initially designed to measure part of network and with new customer oriented approach misses the point. Key Performance Indicator (KPI) is a new standard to provide service oriented approach in performance measurements. KPI measurements provide monitoring the entire End-to-End (E2E) mobile infrastructure from packet or circuit switched core elements to radio access network nodes, and detailed visibility to the underlying transport entities along the entire path of service delivery.

This allows MSPs to achieve numerous efficiencies in operating and engineering their networks and also standardize their performance management needs. Migration process in MBH shown in Figure 2 is followed by total transformation of mobile industry and business model.

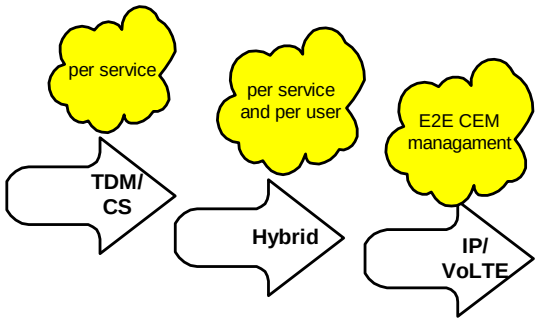


Figure 2. Migration process in MSPs approach.

New LTE networks is data oriented technology and enables proactive operation and moves focus on improving E2E customer experience through core, radio, and IP planning and optimization The circuit-switched (CS) and future VoLTE network will coexist for a long time and coordination between the two is crucial for seamless Customer Experience Management (CEM) and whole mobile network evolution.

Mobile network performance measurements per service may not be the best approach for systematic troubleshooting when degradations are detected on customer experience. The real challenge comes in diagnostic network problem that impact customer experience. These problems may be specific to a particular cell, device, core network element or application. Relevant KPIs shall be able to set alarming thresholds for different services. Furthermore, it is important to be able to detect relevant node/nodes and improve service delivery and loss of income – customer churn. Introducing performance measurements in MSP`s organization shows how to improve whole network efficiency including network design, network deployment, dynamic scaling, service and capacity forecast, and troubleshooting [4],[5]. New mobile network deployments are unprecedented in its complexity and represent a great challenge to MSPs and telco business transformation. Transformation process is very complex and covers cultural, engineering and business transformation [6]. The main goal is to move from just engineering point of view to customer oriented approach in process of maintaining the strength in core business while unlocking other areas like data analytics, music distribution and payment channels.

IV. MEASURMENTS RESULTS IN MTEL NETWORK

In this paper, a part of real Mtel mobile 2G/3G radio access network is analyzed after modernization of network elements in access and transport network (replacement of all existing 2G and 3G base stations by new LTE ready SingleRAN stations) and transport network (replacement of all existing TDM equipment with hybrid or IP equipment). For LTE ready backhaul solution L3 is recommended, and Mtel has viewed as more flexible to optimization and supports easier additions and changes (VLAN option 3 is used).

Rise of new application and services, together with competitive market has led to complex scenario in which quality of service plays a major role. When new IP MBH is introduced, MSPs require complete coverage from a quality perspective, but at the same time the operators is very sensitive to CEM impacts. KPI oriented approach has taken center stage as the performance platform offers significant advantages over traditional network monitoring systems and troubleshooting tools. A key challenge in this migration process has been how to ensure high quality for both voice and data services. Synchronization hybrid scheme, consisting of SyncE, PTP and TDM components, is needed to address synchronization requirements in a practical deployment scenario.

The general concept of performance indicators is that network elements and probes, which are used as service resource instances, are placed at certain nodes of the network infrastructure to pick up performance-related data, e.g. cumulative counters of protocol events. In constant time intervals or in near real time this performance-related data is transferred to higher level service assurance and performance management systems. Figure 3 shows the performance management process of the M2000. It includes the following procedures: performance measurement setting, performance data import, performance data display, performance data synchronization, and performance data northbound export.

There are different kinds of input for a KPI formula: cumulative counter values, constant values, timer values seem to be the most important ones. Also KPI values that have been already computed are often seen in new KPI formulas. In this paper, all chosen KPIs use standard counters and formulas in accordance to 3GPP and vendor recommendation [7], [8]. They are given in the equations (1), (2), (3), (4), (5), (6), (7) and (8). All of them are computed, measured and defined with its relevant counters, using network management software iManager M2000, Huawei [7].

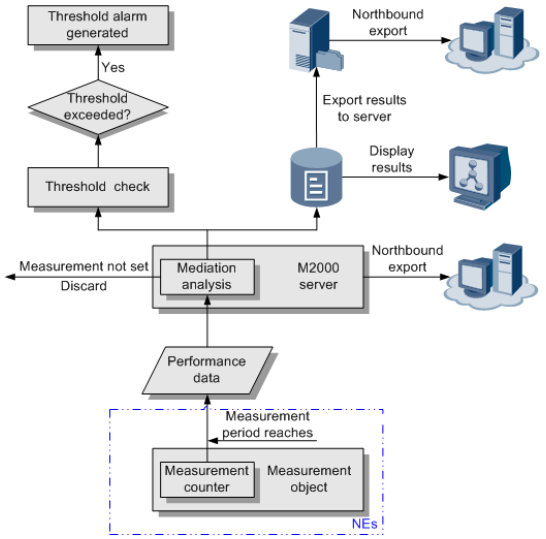


Figure 3. Performance management process, [7].

RRCSRservice = *([RRCSuccConn[Stab]/[RRCAttConn[Stab])*{100}* (1)

AMR RAB SR =*VS.RAB.SuccEstab.AMR]/ [VS.RAB.AttEstab.AMR]*{100}* (2)

PS RAB Setup SR = *([VS.RAB.SuccEstabPS.Conv]+ [VS.RAB.SuccEstabPS.Str]+[VS.RAB.SuccEstabPS.Inter]+[VS.RAB.SuccEstabPS.Bkg])/([VS.RAB.AttEstabPS.Conv]+[VS.RAB.AttEstabPS.Str]+[VS.RAB.AttEstabPS.Inter]+ [VS.RAB.AttEstabPS.Bkg])*{100}* (3)

AMR Call Drop Ratio =*VS.RAB.Loss.CS.AMR]/ ([VS.RAB.Loss.CS.AMR] +[VS.RAB.Loss.CS.Norm.AMR])*{100}* (4)

PS Service Drop Ratio= *([VS.RAB.Loss.PS.RF]+ [VS.RAB.Loss.PS.Abnorm])/([VS.RAB.Loss.PS.RF]+ [VS.RAB.Loss.PS.Abnorm]+[VS.RAB.Loss.PS.Norm])*{100}* (5)

Soft Handover SR (Cell)= *([SHO.SuccRLAddUESide]+ [SHO.SuccRLDelUESide])/([SHO.AttRLAddUESide]+ [SHO.AttRLDelUESide])*{100}* (6)

CS Inter-RAT Handover SR (WCDMA to GSM)(Cell)=IRATHO.SuccOutCS]/[IRATHO.AttOutCS]*{100} (7)

PS Inter-RAT Handover SR (WCDMA to GPRS)(Cell)= *[IRATHO.SuccOutPSUTRAN]/ [IRATHO.AttOutPSUTRAN]*{100}* (8)

In Tables II and III are given measurement results (KPIs) for

- call completion - *RRC Setup Success Rate*, *AMR RAB Setup Success Rate* and *PS RAB Setup Success Rate*,
- retainability - *AMR Call Drop Ratio* and *PS Service Drop Ratio,* and
- mobility management – *Soft Handover Success Rate*, *CS Inter RAT Handover Success Rate* and *PS Inter RAT Handover Success Rate* [8], [9].

KPI`s target values are defined in accordance to Communications Regulatory Agency BH [10]. Recommendations values for KPIs can be changed depending on the local network and its environment, growth and customer requirements.

Implementation of mobile backhaul using IP based solution provides results under expected thresholds. We used highly efficient Performance Management tool - iManager M2000, Huawei. We have analyzed average values of chosen KPIs for six sites with different number of cells (1, 2 or 3) in Banja Luka region, after network modernization in comparison to referent KPI target value. The measurements have been made in busy hours for 40 samples. Sites are chosen with MBH realized with different transmission mediums (microwave and fiber).

Table II shows that call completion KPIs, for all sites after swap of NEs, are better than 98%.

TABLE II KPIS FOR CALL COMPLETION

U Cell	RRC SR service (%)	AMR RAB SR (%)	PS RAB SR (%)
KPI target	**98%**	**98%**	**98%**
Site1_Cell1	99.57	100.00	100.00
Site2_Cell1	99.88	100.00	100.00
Site3_Cell1	99.79	100.00	100.00
Site3_Cell2	99.67	100.00	100.00
Site3_Cell3	99.92	100.00	99.96
Site4_Cell1	99.88	99.76	99.95
Site4_Cell2	99.77	99.76	99.83
Site4_Cell3	99.87	100.00	99.95
Site5_Cell1	99.79	96.67	99.81
Site5_Cell2	100.00	100.00	100.00
Site5_Cell3	99.59	100.00	98.61
Site6_Cell1	99.96	100.00	99.87
Site6_Cell2	99.65	100.00	100.00
Site6_Cell3	100.00	100.00	99.90

Table III shows that retainability and mobility management KPIs, for all sites after swap of NEs are better than KPIs target values.

TABLE III KPIS FOR RETAINABILITY AND MOBILITY MANAGAMENT

U Cell	AMR Call Drop Rate (%)	PS Call Drop (%)	Soft Handover SR (%)	CS ISHO SR (%)	PS ISHO SR (%)
KPI target	**2%**	**2%**	**98%**	**95%**	**95%**
Site1_ Cell1	0.00	0.00	100.00	100.00	100.00
Site2_ Cell1	0.00	0.34	100.00	100.00	100.00
Site3_ Cell1	0.00	0.04	100.00	100.00	100.00
Site3_ Cell2	0.00	1.08	99.90	100.00	100.00
Site3_ Cell3	0.00	0.20	100.00	100.00	97.50
Site4_ Cell1	0.00	0.15	99.98	95.22	100.00
Site4_ Cell2	0.18	0.56	99.93	98.89	100.00
Site4_ Cell3	0.16	0.27	99.98	99.55	100.00
Site5_ Cell1	0.00	1.55	100.00	100.00	100.00
Site5_ Cell2	0.00	0.67	100.00	100.00	100.00
Site5_ Cell3	0.00	0.92	100.00	100.00	100.00
Site6_ Cell1	0.00	0.11	99.93	100.00	100.00
Site6_ Cell2	0.00	0.85	99.86	95.83	100.00
Site6_ Cell3	0.00	0.17	100.06	100.00	100.00

Through every mobile network upgrades or changes, MSP aim to increase capacities and improve network performance to attract more subscribers and reduce subscriber churn through increased customer satisfaction, at the lowest possible cost. To achieve these goals, MSPs must carefully evaluate the capacity and performance capabilities of the products they are considering for deployment. As new hardware, firmware, or services are introduced, they should be carefully evaluated for performance and it must be verified that they do not negatively impact the performance of existing network services.

Measurments values in Table II and III shows that even in 3G MBH it is possible to start transitioning to an IP infrastructure, and minimize difficulty in future process of migration to IP MBH and make current operations more efficient (scalable solution that provide capacity management in large traffic dynamics).

V. CONCSLUSION

Results given in this article shows that migration to all IP transport ensure that KPI`s of existed services are under expected thresholds values and also using this approach is promising for future development of mobile networks (migration to all IP transport networks). Using KPIs validation is proven way to ensure quality improvements and provide quality monitoring for every element in mobile network and increase overall quality of network. This IP transport evolution in MBH for 3G networks ensure connecting future and present generation of mobile networks.

Smooth evolution to LTE networks will enable transformation from circuit-switched (CS) to all-IP architecture and ultimately to network function virtualization (NFV) to facilitate ICT transformation. This is very important task for customer oriented approach that every MSP trying to accomplished.

VI. REFERENCES

[1] Matthew Baker, "From LTE Advanced to the Future", *IEEE Magazine., Communication.* vol. 50, no. 2, pp. 116-120, 2012.

[2] http://metroethernetforum.org/carrier-ethernet

[3] Humair Raza, "A Brief Survey of radio Access Network Backhaul Evolution: Part II", *IEEE Magazine., Communication*, vol. 51, no. 5, pp. 170-177, 2013.

[4] Dijana Ilisevic, Natasa Banovic-Curguz, **"**Synchronization in mobile backhaul using hybrid microwave links**",**Telfor Belgrade, pp. 208-211, November 2013.

[5] Dijana Ilisevic, Natasa Banovic-Curguz, "Analysis and implementation of mobile backhaul in Mtel network using hybrid microwave links **",** ETRAN Vrnjacka Banja**,** pp. TE 1.7 1-4, June 2014.

[6] D. Ilisevic and N. B. Curguz, "Changing Face of Mobile Service Providers in Bosnia and Herzegovina", Telfor Belgrade, pp. 51-54, November 2014.

[7] Mobile Element Management System manager M2000, Huawei, V200R011C01SPC240.

[8] ETSI TS 132 410, Technical Specification

[9] ITU-T Recommendation E.800

[10] Communications Regulatory Agency BH, www.rak.ba

Comparative Analysis of GSM Coverage Prediction with Measurement Results for Urban Areas Using Statistical Nonparametric Mapping

Haris Lučkin*, Emina Lučkin*and Mirko Škrbić**
*BH Telecom, Sarajevo, Bosnia and Herzegovina
**University of Sarajevo, Bosnia and Herzegovina
haris.luckin@bhtelecom.ba

Abstract - **Planning, building and optimization process of radio access networks is a complex dynamical activity, which requires a lot of planning effort and time. The final goal is to build a radio network, which consists of minimum infrastructural components, and offers the best coverage and service quality. In order to consider all the possibilities of GSM signal coverage during planning time, and before the installation of equipment on certain location, it is very important to do software simulation of the GSM signal coverage of the observed area. Consequently, it is important to use an optimal predictive model which, using the appropriate digital maps, will offer the most realistic picture of the future radio access network. Taking this into account, subsequent optimization process and re-analysis would be reduced to a minimum.**

In this paper, comparative analysis of the software prediction GSM coverage area signal with the results obtained by measurements after the implementation of planned radio infrastructure at the proposed base station, will be presented. In order to assess the credibility of the software predicted GSM signal coverage area, measurements of GSM signal in urban area were performed. Final results were verified using statistical nonparametric chi square test. Further, the software prediction relation in regard to the measurement signal after the implementation of system is defined, and potential causes are treated.

Keywords—GSM, prediction, measurements, Chi-square test.

I. INTRODUCTION

One of the primary objectives of any mobile services provider is well planned radio-access network, which is essential in minimizing additional costs. Access network is a key part of any network in a sense of providing appropriate availability and quality of service to the end user. Therefore, mobile signal coverage, number of potential users, and traffic demands represent major input parameters in planning of all wireless networks. On the other hand, the optimization of mobile networks can be observed as a process of improving overall quality in network, that can be perceived by end-user [1]. By doing this, operators enable an efficient using of network resources.

Further, the regular functioning of one radio access network greatly depends of initial assumptions and software simulation which predict future coverage of the area with mobile network signal. The important step during this process represents the selection of appropriate prediction model which, using convenient digital maps, have to be capable to give, as much as possible, the realistic idea of future radio network coverage area.

In this paper, the results of conducted drive-test measurement of GSM signal for urban network case, are presented, with the ultimate goal of checking credibility and correctness of coverage area of this signal predicted using appropriate software. Further, comparative analysis of the values of GSM signal levels obtained by these procedures, e.g. checking and behavior of software prediction in regard to measurements. This analysis is conducted using nonparametric statistical chi-square mapping, which treats the level of reliance on the certain hypothesis acceptance or rejection. Lastly, results and potential causes for such results are interpreted.

II. PRACTICAL ANALYSIS OF GSM SIGNAL COVERAGE

Below, the analysis of GSM signal coverage for the certain area will be described. The final goal is to evaluate precision of software prediction of GSM signal coverage for urban area, without using objects elevation maps. Firstly, the selected cell with all parameters will be described, then the software prediction and drive test measurements will be conducted in the planed cell coverage area. Adjustable features of the “real" base station and base station software created are identical.

A. Selecting the base station and area to be analyzed

When choosing base station, it was primarily important to select an urban area, to prove efficiency of prediction model. Due to urbanity demand, the area of city Sarajevo was chosen. Also, it was necessary to take into account some limitations of GPS device, and limitations of software for GSM signal level measuring. Because of need for four satellites to be visible, it wasn’t favorably to consider the town center area of Sarajevo. The limitations of measuring tool are possibility of limitating the measurements on one carrier frequency, so it was necessary to choose BCCH channel frequency of the base station with a minimum number of adjacent cells with the same BCCH channel frequency. During analysis, base station Merkur, cell number one was considered, located in Otoka, Sarajevo, with frequency 1800 MHz.

The reason why the 1800 MHz cell was chosen is because various characteristics of spreading radio wave on different frequencies. Because of less attenuation of signal on lower frequencies, GSM 900 signal will at any point will be better than GSM 1800 (in case of cells co-located on one site). On the other side, to unload GSM 900 cell, a part of overalll capacity should take over GSM 1800 cell, and that is done using hierarchical structure of cells (HCS content).

B. Base station coverage prediction

In order to predict signal coverage for 1800 MHz cell (MERKUR18), TEMS software tool was used [3]. In this tool, all the data are loaded in form of layers. For this coverage prediction, a few layers were used:

- Georeferenced orthophoto map of the area.
- Database with data for all base stations for chosen mobile service provider.
- Elevation map, with relief data for chosen area, not including information about area type (urban, suburban, rural).
- Roads and vegetations information.

For GSM signal coverage prediction Ericsson 9999 propagation model is used [3], because other models need map of objects in order to analyze coverage, which was not available at the time of research. Prediction model tuning was not conducted.

In Figure 1 prediction of composite GSM signal coverage for chosen cell is shown (Best Server-composite), using selected layers and overall base station's EIRP.

For predicting cell coverage, 2000 meters is taken as cell radius, with cell in the center of the area, for getting a clearer overview of area with smaller signal levels, and therefore better analysis and comparation with results obtained by measurements.

C. Base station signal level measurements results

Analysis of base station coverage is done using Nemo Outdoor and Nemo Analyze software tools [6]. Measurements which can be conducted when analyzing GSM signal area coverage are divided into three categories:

- IDLE mode measurements. Idle mode means that GSM signal is measured without connection establishment, and this type of measuring is used to determine GSM signal coverage level of certain area.
- Measurements during connection establishment.
- Measurements with connection established (DEDICATED mode). Dedicated mode is used for determining the quality of signal, handover procedure success rate and determining continuity of established call (measuring with connection established-connected mode).

Measuring of signal level in idle mode are conducted for this paper for the reason in dedicated mode connection on radio link is established and functions „Power Control" and „Discontinuous Transmission" on link are activated. This functions, in order to keep level of network interference low and to extend battery lifetime, limit emitted power of base station and mobile phone to minimum needed value [2]. For this reason, level of signal in dedicated mode is 5 to 10 dBm lower than real level of signal, measured in idle mode.

Measuring signal levels in the coverage area of cell MERKUR18 are conducted in idle mode. These results are compared with signal power levels results obtained by prediction. On the cell coverage area three measurements of signal level are conducted, in order to average the results, and observe variations of signal level in different periods of time.

One of the results is shown on Figure 2. Observing the conducted measurements, it can be concluded that signal levels measured through different time periods slightly differ, and it is consider normal taking multipath fading into account.

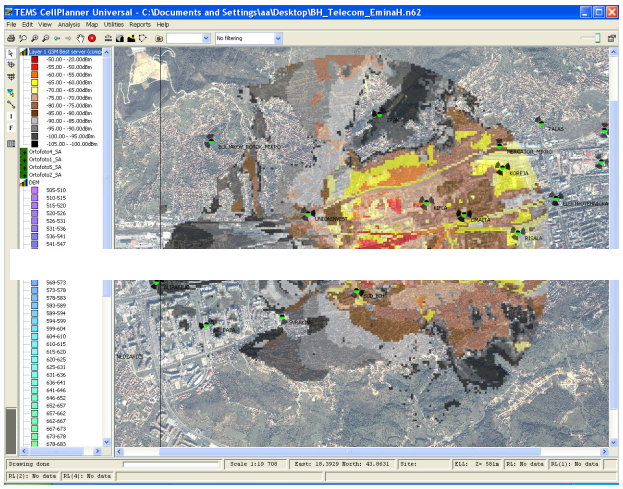


Figure 1. Coverage prediction map (composite) for chosen GSM cell

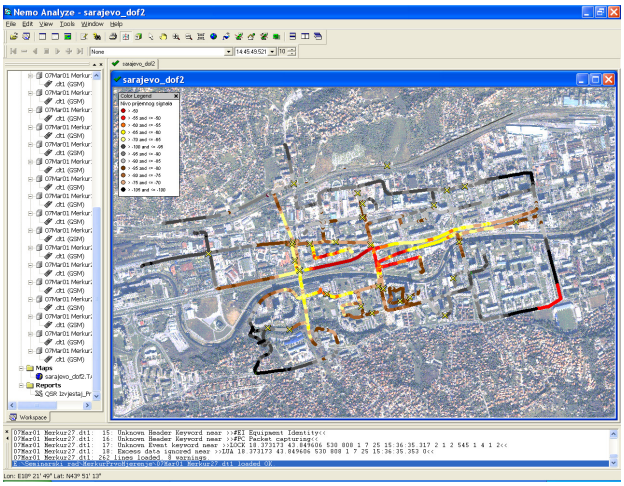


Figure 2. Measurement results showing GSM signal strength for chosen cell (first reading)

Figure 3 represents GSM signal strength as a function of distance from base stations antenna, values were taken from measuring files. In Figure 4, the same diagram is shown, while approximated diagrams for path loss and slow fading are added.

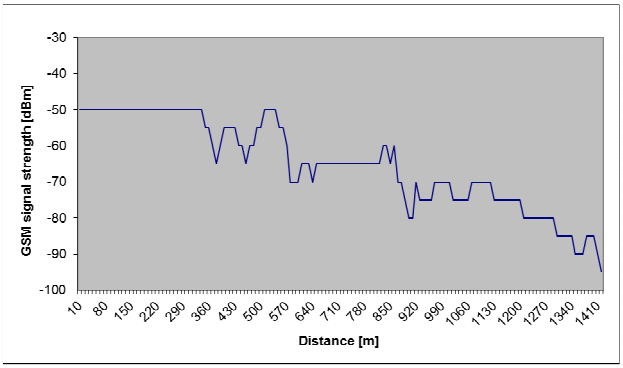


Figure 3. GSM signal strength as a function of distance (based on measurements results)

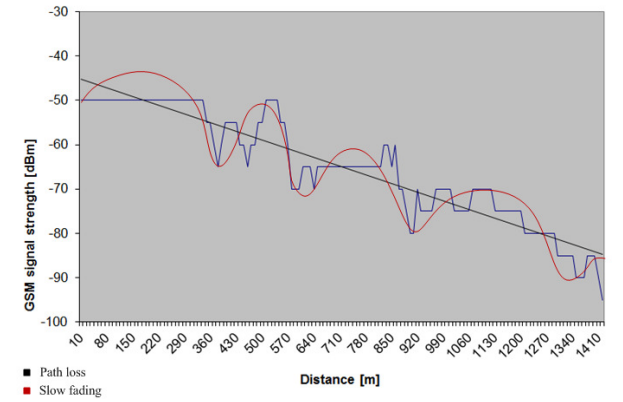


Figure 4. Path loss and slow fading shown using measurement results

III. STATISTICAL SIGNIFICANCE TEST

After obtaining the results of an area GSM signal coverage level, and representing the results in format suitable for further analysis, it is necessary to determine the relationship between the values of GSM signal strength obtained by both procedures, and to assess credibility of predicted GSM signal coverage using software tool propagation model. In order to do that, it is necessary to conduct certain statistical significance test, which will statistically compare values obtained by both procedures. Any statistical significance test which is correctly performed, has conclusions about confidence grade with which certain hypothesis could be accepted or rejected. In this paper, χ^2 test is chosen.

Nonparametric χ^2 statistical test is statistical significance test for the tabelar analysis with two variables. Typically, hypothesis being tested with χ^2 distribution is: whether or not are to different samples different enough in some characteristics or aspects of their behavior, to conclude that these samples' population of origin are also different or not different, in certain features or aspects of their behavior.

Nonparametric χ^2 test is rough estimation of credibility; it accepts weaker, less accurate data as input, unlike parametric tests (e.g. t-tests and variance analysis), so it has lower status among other statistical tests. However, it's limitations are it's power, because χ^2 distribution is tolerant of input data accuracy, so it can be applied to wide research area.

χ^2 test is usually used to test statistical significance of the results presented in tables with two variables. Interpretation of tables with two variables is a part of χ^2 test results interpretation, so the analysis will be conducted using tabular analysis with two variables [4].

A. *Conducting the χ^2 test*

As we mention before, χ^2 is nonparametric test. It does not require that the sampled data are more ore less arranged by normal distribution (like t-test do), although it is based on assumption that variable is normally distributed in the population from which samples were taken.

Requirements of the χ^2 test are as follows:

- The sample should be taken randomly from the population.
- The data should be presented as frequency of occurrence (not a percentage).
- Measured variables should be independent.
- The observed frequencies of occurrence should not be too low.

In this work, it will be assumed that the values that we collected from GSM signal coverage correspond to practically measured values.

In carrying out χ^2 test is necessary to provide sufficient number of samples that will be used for testing. In order to achieve reliable results, we will take the number of samples in the amount of 1.000.

TABLE I. GATHERED RESULTS FOR PREDICTION AND MEASUREMENTS

GSM signal strength of the sample	Result obtained by prediction	Result obtained by measurements
-105<x<-100	30	99
-100<x<-95	25	181
-95<x<-90	33	147
-90<x<-85	68	142
-85<x<-80	119	64
-80<x<-75	155	60
-75<x<-70	106	70
-70<x<-65	117	56
-65<x<-60	222	43
-60<x<-55	41	33
-55<x<-50	48	52
-50<x	36	53

TABLE II. PARAMETER VALUES BEFORE AND AFTER CORRECTION

GSM signal strength of the sample	χ^2 value	χ^2 value after correction
-105<x<-100	158,70	1,375
-100<x<-95	973,44	0,090
-95<x<-90	393,81	4,266
-90<x<-85	80,53	0,007
-85<x<-80	25,42	0,690
-80<x<-75	58,22	3,368
-75<x<-70	12,22	2,647
-70<x<-65	31,80	2,173
-65<x<-60	144,32	0,734
-60<x<-55	1,56	0,923
-55<x<-50	0,33	0,285
-50<x	8,02	1,422

These samples were taken randomly based on their coordinates (longitude and latitude). For every sample obtained by prediction, the sample with same coordinates obtained by measurements is taken for comparison purposes.

Measured signal strength value will be collected, then the points with the same coordinates will be found on the GSM signal prediction coverage map, and value of the signal strength in these points will be collected, too. For each range of values of the GSM signal strength, for measured and predicted values is necessary to provide the sum of samples with a given signal strength and record it for further analysis.

Furthermore, it is necessary to determine the desired error tolerance. We will assume error probability Pe=0,05. For further analysis is also necessary to know the number of degrees of freedom ν. This number is defined as:

$$\nu = (r-1)(c-1) \tag{1}$$

where r is the number of rows in the table with data to be analyzed, and c is the number of columns in the same table.

After sampling, it is necessary to ”measure” the difference between the pairs of observed and expected frequencies of occurrence. Observed values in this case are values obtained by measuring the GSM signal strength, and expected values are values obtained by software prediction tool. Further, it is necessary to calculate the difference between the observed and expected frequencies of occurrence for each column (for each range of values of the signal strength), square the difference, and finally divide the resulting number by expected frequency of occurrence.

The resulting value is called χ^2 value for given signal strength values range in our case.

$$\chi^2 = \frac{(O-E)^2}{E} \tag{2}$$

Where :

χ^2– constant by which the accuracy of assumptions is evaluated;

O – measured value;

E – predicted value.

After collected data, values shown in Table I, χ^2 values shown in Table II are calculated. Summing the values from Table II we get the overall $\chi 2$ value, which in our case is 1.888,37.

The resulting number should be compared with the values in the table of upper critical values of χ^2 distribution with ν degrees of freedom. For Pe=0,05, ν=11, critical value of χ^2 distribution is 19,675. The tested value is obviously much higher than the critical value χ^2 distribution, indicating that the initial hypothesis is incorrect.

B. Correction of the results

Since the results obtained by measurements deviate significantly from the results obtained by the software prediction, we started the analysis of the results obtained, in order to establish a correlation between these two procedures.

Given that there is no real possibility of correction practical measurements (intervention on the antenna system, base station, which refers to the modification of azimuth, tilt and height of the antenna), the only to affect the correlation of these values is to correct prediction software parameters.

Corrections made to the parameters of the base station (antenna system) used during software prediction, were reflected in the modifications to the azimuth, tilt, and height of the antenna system.

So, after a detailed analysis and comparison of prediction and measurement results, it was found that the optimal alignment of these results is obtained using the parameters given in Table III.

Base station’s antenna system radiated power (EIRP) has not been modified after the correction (base station’s combiner type remained the same - Gc, and output power of the base station cabinet 2206). The result after correction made using above parameters is shown in Figure 5.

TABLE III. BASE STATION PARAMETERS BEFORE AND AFTER CORRECTION

Value type	Antena height	Azimuth	Downtilt
Before correction	23m	75°	3°
After correction	22m	80°	2°

TABLE IV. GATHERED RESULTS FOR PREDICTION AND MEASUREMENTS AFTER CORRECTION

GSM signal strength of the sample after correction	Result obtained by prediction	Result obtained by measurements
-105<x<-100	88	99
-100<x<-95	177	181
-95<x<-90	124	147
-90<x<-85	141	142
-85<x<-80	71	64
-80<x<-75	76	60
-75<x<-70	85	70
-70<x<-65	46	56
-65<x<-60	49	43
-60<x<-55	39	33
-55<x<-50	56	52
-50<x	45	53

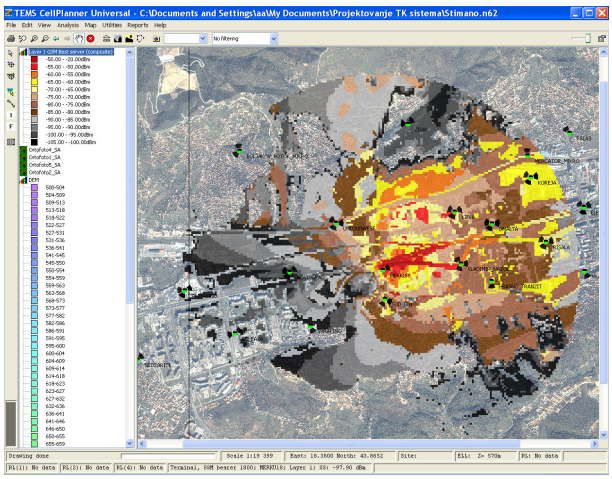


Figure 5. Coverage prediction map after correction

After correction of the overall sum of the number of samples, the values for the $\chi 2$ parameter are presented in Table IV.

By summing above values a corrected $\chi 2$ value 17.98 is obtained.

The resulting number should be compared with the values in the table with the upper critical values of $\chi 2$ listed. For values of Pe = 0.05, ν = 11, critical value of $\chi 2$ distribution is 19.675. The test value is in this case less than the critical value of $\chi 2$ distribution, indicating that the initial hypothesis is true.

IV. CONCLUSION

The basic step in the development of each mobile radio network is good planning. Today different software tools are used to plan coverage of the radio network. After software coverage planning, and commissioning a certain transmitter, it is necessary to control the inspect signal strength value, in the way of measuring the GSM signal strength. During this process, conclusion that the measured value more or less deviate from the values obtained by prediction is often. The reasons are different, and one of the key reasons is a lack of quality elevation terrain maps, that prediction model uses to predict the signal strength.

In order to test the accuracy of prediction using the appropriate model intended for urban and suburban areas, applied to an urban area, measuring the signal strength in the service zone for cell MERKUR18 is performed. Values obtained by measurements and prediction were compared by nonparametric $\chi 2$ test. Results of this test indicated the large deviations of the measured values compared to the values obtained by prediction. From the values that were used during analysis it can be concluded that measured values deviations from predicted values are higher in the weaker signal strength samples, and lower for areas with good signal coverage area. Areas with higher quality signal are areas near the antenna system, i.e. areas where there is line of sight to the base station (without objects, and therefore the reflection). From this we can conclude that the main reason for the poor results the lack of high-quality maps, more precisely, the lack of folders to the present elevations, as the terrain, and objects (folders with facilities were not available for analysis). Another potential reason could be used predictive model, however, as all the prerequisites for the correct operation of the model are not provided used, valid comparison of the different models wasn't done, and it could be an objective of future researches.

In order to perform the correlation of the measurement results and prediction software results, only the possible corrections for this case were done: corrections of parameters of the antenna system. Thus, after a number of changes optimal parameters were identified, and by applying the $\chi 2$ test on the obtained results it was found that these results were very well correlated.

REFERENCES

[1] A. R. Mishra, "Fundamentals of Cellular Network Planning and Optimisation", John Wiley & Sons, Ltd, 2005.

[2] J. Lempiainen and M. Manninen, "Radio Interface System Planning for GSM/GPRS/UMTS", Kluwer Academic Publishers, 2002.

[3] Tems CellPlanner Universal User's Guide, Ericsson, 2006.

[4] Milan Merkle, "Verovatnoća i statistika za inženjere i studente tehnike", Akademska misao, Beograd, 2002.

[5] www.georgetown.edu, last access date 10.03.2014.

[6] www.nemotechnologies.com, last access date 10.03.2014.

SOFTWARE ENGINEERING & DATA MANAGEMENT

Software Defined Networking Demands on Software Technologies

T. Galinac Grbac*, C.M. Caba**, J. Soler **
* University of Rijeka/Faculty of Engineering, Rijeka, Croatia
** Technical University of Denmark/Networks Technology and Service Platforms, DTU Fotonik, Lyngby, Denmark
tihana.galinac@riteh.hr, {cosm, joss}fotonik.dtu.dk

Abstract - Software Defined Networking (SDN) is a networking approach based on a centralized control plane architecture with standardised interfaces between control and data planes. SDN enables fast configuration and reconfiguration of the network to enhance resource utilization and service performances. This new approach enables a more dynamic and flexible network, which may adapt to user needs and application requirements. To this end, systemized solutions must be implemented in network software, aiming to provide secure network services that meet the required service performance levels. In this paper, we review this new approach to networking from an architectural point of view, and identify and discuss some critical quality issues that require new developments in software technologies. These issues we discuss along with use case scenarios. Here in this paper we aim to identify challenges for further evolution of software technologies in addressing problems of network evolution. Our view is based on the need for strong integration of network and software technologies, and therefore we establish the main focus of the paper on discussing SDN demands on software technologies. The main contribution is in categorization of open research problems and presenting ideas and opening new opportunities for future research.

I. INTRODUCTION

Software Defined Networking (SDN) is a paradigm that is related to idea of offering the network resources to end users as a service (NaaS) over an open Application Programming Interface (API). In this paradigm, from the network operator perspective, the key differential element is to replace industrial standard Command Line Interface (CLI) for managing directly network devices with an open interface, which can be used to programmatically perform device management. Today, we have two protocols NETCONF [1], proposed by IETF, which provides mechanisms to install, manipulate, and delete the configuration of network devices, and Open Flow[2], proposed by ONF, as communications interface between the control and forwarding layers of an Software Defined Network (SDN) architecture. Introduction of open protocols enables faster evolution and interoperability among equipment of different vendors. Thus, leads to an idea of separating switching hardware and control logic as is presented in Fig. 1.

In OSI layer two and layer three implementations of the modern transmission networks there is a plethora of control protocols allowing for autonomous switching and routing of Internet traffic. Autonomous control is achieved by their distributed communication relying on continuous tracking of surrounding network topology. Thus significant processing power of network equipment is spent on spreading common topology view among these network nodes (according to some research studies more then 30%). Also, transmission overhead and link utilization may represent also significant issue. With the continuous trend of network scale, creating larger transmission chains, this current autonomous control approach becomes more and more inefficient and leading to unacceptable performances and convergence time exceeding conventional constraints. As network grows, more and more centralized approaches to network management are needed. The main problems identified are with **responsiveness**, **reliability** and **scalability**. A solution that is getting more and more attention nowadays is introduction of aforementioned layering approach with centralization of control logic. The aim of separating control logic from the switching logic into two distinct network planes is to decouple the forwarding and device operating functions from the network control logic. This new approach increases the efficiency in handling a dynamic network, and provides better adaptation to changing traffic patterns.

The idea of fully integrated network solution with separation of control logic into separate layer from pure hardware switching layer without logic, just based on dynamically changing data table, is far away from realization. Although, it looks simple it requires significant virtualization of network management activities and procedures that are not so simple in the current state of the art. Therefore, an endeavor has to be invested into simplification of existing network management operations providing backward compatible, easy to scale and performance wise solutions.

The aim of this paper is to introduce into concepts of Software Defined Networking and explore and discuss open research problems on SDN software technology. For this purposes we survey work on SDN software technologies and on concrete use cases we identify and explain main open research problems in that area. Finally, main contributions of the paper are in introducing new ideas on future research that is to be considered in developing of SDN software technologies.

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

II. SDN ARCHITECTURE

The basic key elements of SDN are separation of network control logic from network hardware operation and simplification of network device logic. This enables automatization of network management processes and logic above the control layer, centralization of network control logic, and openness of network resources through open standards and to end users.

Figure 1 captures the overall SDN architecture, which consists of three planes (layers): data plane, control plane and application plane. The data plane comprises the forwarding devices. At the control plane, the logic is implemented in a software platform termed SDN controller (SDNC). The SDNC's interface towards Data Plane is realized through the southbound API (*Application Platform Interface*) [3], see Fig 1. The SDNC platform contains a logical representation of the underlying network resources, and several control functions that implement operations and functions for the network. The capabilities built into the network control plane are exposed to the application plane through the Northbound API that is SDNC's interface for communication between Application and Network Control Plane. The applications, located in the application plane, leverage the northbound interface to implement higher layer service logic.

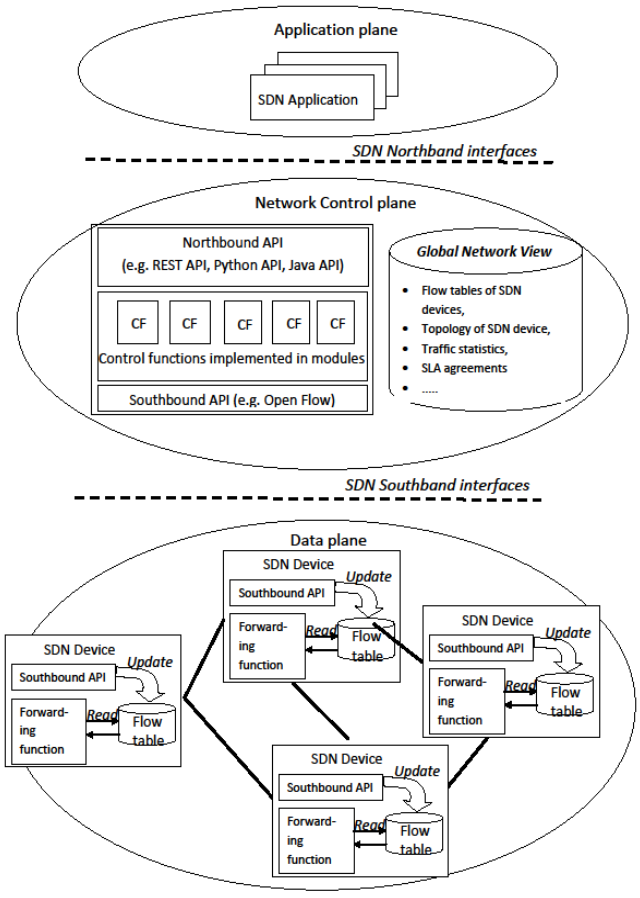


Figure 1. Overview of Software Defined Network Architecture

A. SDN Northbound Interfaces

The SDN northbound interfaces, as is defined in [3], specifies the functions exposed by the SDN control plane, for programming of SDN applications and services and it is implemented as an Application Programing Interface (API). There can be various implementations for the northbound API, depending on the network control plane and the requirements of the network services built on top of it. Figure 1 shows some possibilities for such implementations (e.g. REST, Java, etc.). These implementations fall under three main categories:

- Controller APIs that use a general purpose language (e.g. Java, C, Python).
- APIs based on web services (e.g. REST).
- APIs based on Domain Specific Languages (DSL).

The first category contains the most common and used APIs. They allow the developer to write new service modules integrated in the controller platform (Figure 1). The variety of languages depends on the implementation language of the controller platform. Some examples of controllers are: NOX in C++, POX in Python, Floodlight, ONOS and OpenDaylight in [4], [5], [6], [7], [8]. The northbound APIs in this category lead to tight coupling of the services (control functions) with the controller platform.

The second category of APIs is based on web services. Some controllers offer a REST API through which applications can get information about the state of the network, and send control messages to the switches. This type of APIs emphasizes simplicity and modularity: the network services may be programmed in any language as long as they comply with the REST interface. They are well suited for integration with external platforms, using service oriented architectures. The main disadvantage is that it may be difficult to expose the full set of capabilities of the control plane as REST APIs. While the control plane and the southbound interface are generally event-driven (network generated events), it is harder to implement an event driven REST API thus to preserve the same programming model towards the SDN applications. There are efforts in the open source community to provide event notifications through the northbound interface [6], although this is challenging. The RESTCONF protocol [9] tries to establish a standardized method for exposing the control plane resources (e.g. devices, flows, etc.) to the SDN applications using an event-driven paradigm (including event notifications and remote procedure calls).

The third category of APIs is based on domain specific languages. Their main purpose is to ease the task of expressing network policies by using high level abstractions. Examples of research work in this area are Netcore, Frenetic, and Pyretic [10], [11]. A network control plane that exposes such an API, takes as input a policy describing the expected network behavior, and compiles the policy into low level rules (e.g. Open Flow entries) that are installed in the forwarding devices. This programming model makes application development easier by moving the complexity from the applications to

the control plane policy compiler. The policy compiler ensures that the behavior enforced in the data plane is correct and does not lead to inconsistent traffic treatment (due to possible conflicting policies).

One of the challenges regarding the definition of northbound APIs is choosing the correct level of abstraction for the network resources exposed through the APIs. If the abstractions are very low level, then the application development becomes increasingly complex. On the other hand, high level abstractions make development easier but they may become less powerful (i.e. decrease visibility and control over the resources).

B. Southbound Interface

The southbound interface defines the communication between the network control plane and the data plane (Figure 1). The OpenFlow (OF) protocol has been standardized in 2009 and it is currently the most used protocol implemented on the SDN southbound interface. OF is based on abstraction of forwarding logic using flow tables. Inside a forwarding device, the forwarding function is performed based on the OF entries in one or several flow tables (Figure 1). An OF entry is a *match* and *action* tuple. The *match* part defines the headers to match against the packets arriving at the forwarding device, while the *action* part defines the actions to apply to the matching packets. In its first version, OF had limited capabilities, covering mostly Ethernet, IP, and transport headers [2]. In later versions several features have been defined, making OF more versatile:

- Pipe-line processing of packets using multiple flow tables. This mechanism enables applying a series of actions for a packet.
- Tunneling support using Multi-Protocol Label Switching (MPLS)[12] and Provider Backbone Bridge (PBB) [13].
- Support for IPv6.

Several other extensions have been defined for the OF protocol in order to support circuit switched and optical networks. While OF enables controlling the forwarding behavior in the data plane and collecting statistics about the network, it does not make it possible to configure the devices. Examples of protocols that enable configuration and management of devices are NETCONF, SNMP, and OF-Config [1] [14] [15]. These protocols enable for example configuration of operating parameters of the devices, such as queues, ports, etc.

III. USE CASE SCENARIOS

Two exemplary case scenarios are chosen here to illustrate the advantages provided by SDN in managing current network services and infrastructures.

1. Network Slicing / Virtualization

A network infrastructure under the supervision of an SDN controller (SDNC) can be "shared" by multiple users without mutual knowledge or impact, creating de-facto "virtual networks". This is a likely solution for "multitenant" operations where network infrastructure resources can be dynamically allocated and modified among the different tenants via the centralized and programmatically control of the SDNC. While this type of network "slicing" can be achieved by the physical separation and control of ports and queues in the forwarding devices, a more scalable and flexible solution can be achieved if the SDNC becomes the manager for tag-based (overlay) conventional virtual networks, i.e. by managing VLAN (Ethernet) tags or MPLS labels.

2. Data Center & Virtual Infrastructure

This is very much exploited currently in different commercial solutions targeting Datacenter (DC) networks, where computer resources virtualization has become the de-facto standard mechanism to maximize the utilization of equipment and distributing its cost. DC network topologies are based on massive tree structures of servers rooted by common switches in share physical collocated racks (i.e. TORs: Top of the Rack switches). TORS are as well interconnected in tree architecture towards gateways connecting to external networks. While traditionally "leaf" nodes in these tree topologies were mere hardware servers, in today virtualized environments they are virtual machines (VMs), instantiated when and where needed by virtualization hypervisors. These VMs, in order to enable connectivity within the virtualized network are linked via virtual switches, which can be controlled by an external SDNC. This enables to create overlay networks, as the ones mentioned previously, within the DC to represent different tenants (customers) or applications. DC network management challenges, such as dynamic traffic migration (i.e. when copying or migrating a VM from one physical server to another) or inclusion-deletion to-from different virtual (overlay) networks within the DC, become extraordinarily simplified when based on a centralized-programmatic SDNC operation, rather than a manual configuration-reconfiguration from a human (fail-prone) manager.

IV. SOFTWARE TECHNOLOGIES

The introduction of network virtualization enabled further development and flexibility of networking in cloud environments. However, SDN aims to centralize network control, thus it aims to *softwarify* network engineering theory into network services offered to its users leading to easier and more efficient network management. However, numerous problems exists and still more work has to be done to define network engineering theory that would be ready for network softwarisation.

Here we aim to introduce some of the problems and open areas for research but we are not aiming to review all the state of the art in this field.

V. OPEN PROBLEMS

Movement to the new SDN paradigm, aiming to *softwerize* the network, is full of challenges. There is a growing body of research in that direction. Some well established top journals have devoted a special issue to SDN. There have been three HotSDN conferences with

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

continuously increasing number of papers submitted. A new conference started on Software Defined 5G Networks. Workshops are addressing particular topics in SDN such is for example Workshop on Architectures and programming paradigms for emerging 5G Networks. Numerous papers have already been written and published and number is continuously increasing. One of the biggest concerns is how can software technologies help in this migration and where we need their help. Main challenges with network virtualization are addressed per each identified use case scenario in [16]. Here in this paper we aim to identify challenges for further evolution of software technologies in addressing problems of network evolution. We survey problems addressed by numerous authors on SDN software technologies and discuss them here in several categories. Also, important to note is that we are not aiming to review all the existing literature on challenges and provided solutions. Here we just describe open problems and occasionally provide reference for better illustration of problem.

First problem that we address here is the consistency problem between the logical representation of the network resources at the network control plane, and the physical resources in data plane. This consistency problem is associated with liveness properties. This problem is twofold. If we strive to keep consistency then we will sacrifice responsiveness and vice versa. A consistent network model at the control plane requires that well defined synchronization procedures are executed frequently in order to update the network control logical model with information from the data plane. This leads to increased resource consumption at the SDNC side and introduces delays on the control channel, thus decreases the overall responsiveness at the control plane. On the other hand, optimizations may improve responsiveness but may also introduce uncoordinated behavior that in the case of large network complexities, is leading to a non-reliable system, with routing loops and black holes. Optimizations as shortcuts are not welcome solution in such complex systems. System scaling is limited with distributed relationship between logical and physical resources and maintaining their mutual relationship is not trivial [17]. Software technology research aims to identify set of representative relations, that are easy to manage and maintain, investigate specific trade-offs that are dependent on controller implementations and to find alternative mechanisms on delays or reliability. Consistency effect is evaluated from different perspectives and vast amount of software technologies has been proposed to address this problem and still this is a hot research area widely involving computer science and complex network theory.

To achieve high scalability, flexibility and availability it is identified as mandatory to implement redundancy at control plane. In [18] experimenting with different topologies it is identified that one controller location is often sufficient to responsiveness requirements but not for fault tolerance requirements. Some researchers has come with statement that the only possible way to secure highly available but reliable platform is to introduce redundancy and distribution of control logic. However, in that scenario we introduce completely new branch of problems dealing with synchronization and concurrency issues while handling concurrent policy updates [19]. Here we face with full complexity of programming network control function with high reliability and availability demands while capable for executing numerous concurrent processes for variety of different users. Majority of knowledge in this domain has already been collected within industry such is for example Ericsson knowledge base developed in more than 40 years of evolution of AXE based telephone exchange that in its *AXE 810* version that implements control logic as standalone node that is central in next generation networks for handling control (signaling) of end user traffic in resource network layer. This switch is central to many core network implementations of Next Generation Networks (see 3GPP standards) and is reliable, robust and easy to scale solution. However, most of the solutions are not implemented to be used at runtime as is required in new network evolution phase (5G networks). The synchronization and concurrency issues are big concern of these networks. As communication is a central element of these networks and aiming to *softwarify* network engineering theory we may need to control harmonies execution of numerous industries specific hardware related policies and numerous interacting industry standards. This is actually one of the biggest concerns of further network evolution. For that reason we may need new abstractions for structured communication - centered programming and to specify and implement program communication safe software. One promising example is behavioural type theory that introduces new behaviuoral types for describing interactions involving multiple peers and that abstracts these interactions as a global scenario [20], [21]. Message exchange among concurrently running multiple peers are verified at global level thus enabling independent certification of each party's local implementation if it is in correspondence to global behaviour. Based on that theory a lot of interesting approaches has been presented at workshop on *Programming Language Approaches to Communication and Concurrency cEntric Systems (PLACES)*. One good example is Scribble programming language [22] that allows certification of global protocol interaction and projection onto local protocol implementation. Except the tools for editing, verifying and projecting, there is work in progress of development of numerous libraries that allow its integration with some general purpose languages such are java or python.

In centrally controlled network architecture that is continuously growing, its management becomes extremely complex and security vulnerabilities becomes more complex and error prone. Data integrity and confidentiality for information exchange in a network with dynamic policies require innovation in the area of cryptographic algorithms or authentication certificates. Some cyber security controls, that would be implemented as part of software tehnology for next generation systems

such as access control, network isolation or monitoring are areas under investigation.

Another problem is with central management of global network view. This task requires huge amount of network data to be effectively stored into database, restored and up to date. Problem of storing of huge amount of data as well as effective information capturing from this big data volumes is a huge issue. One approach is usage of hierarchy abstraction. In [23] proposal is to use hierarchical policies that show positive performance results.

Expression of configuration protocols such as Open Flow and Netconf is limited. A numerous studies question their flexibility to different configuration needs, and their expression abilities in relation to the performance requirements [24]. Here we have tight interaction between telecommunication and computer science profession in defining right abstractions that would provide expressive protocol and its implementation language for efficient, reliable but secure communication between abstract network notation and physical resources. There is continuous work in progress in standardization of these protocols and its implementations as well as identification of new avenues for further evolution in configuration protocols and languages.

Modeling abstractions, software organization at control but also at device level may have significant influence of service performance offered by SDN network. This is especially interesting to observe in relation to system potential to scale. For example, an earlier study show that splitting of architecture in Mobile Switching Centre has significantly degrade service performances and scaling abilities, [25]. Software modelling methods have to be revised with software dynamics in mind. A software engineering modeling tool that aims to model how software application with intensive dynamic scaling requirements has to be distributed in respect optimized resourse costs has been proposed by [26]. An object oriented programming approach implementing type system aiming to optimize resource usage costs by governing software distribution within multicore systems has been proposed in [27].

Architectures that rely on open standards and open source code community benefit in terms of progressive development and prototyping. Moreover, an integrated view that is maintained among various industries through open community turns out with positive interoperability characteristics. However, there are numerous drawbacks. Open community is good choice for simple prototyping solutions without strict and formal rules, based on easy and general purpose languages. But, in case of complex system there are numerous studies that proved that human factor should not be neglected. For example, programming language *Erlang* contains some interesting concepts that are developed assuming unreliable software. The right balance between advanced software technologies that human should consume while developing software and software technologies that accepts not perfect world has to be achieved. One way to deal with this issue is to develop certification technology for software packages that are to be deployed into complex network that would be based on using built in typecheckers.

The way how system evolves very much influence its further ability to scale, further develop, maintain required reliability levels, etc. Learing from existing complex systems that evolved over decades may bring some fruitful knowledge. Identification of software behaviour and identification of autonomous modeling techniques that could be easily implemented into network or development environemnt may be used to introduce discipline in SDN network in open source environment.

VI. CONLUSION

Software Defined Networks have been developed and invented as new paradigm in telecommunication world but it is highly based on computer science theories. Numerous efforts that were invested within computer science field developing theories in formal languages and verification tools have finally show their full practical benefit. Developing complex software systems such a modern telecommunication networks become extremely challenging task. Add hoc shortcutting engineering solutions is not desirable any more but we should strive on systematized solutions. The best known systematization technology is governed by sound formalisms that are implemented within technology that humans use to engineer complex systems. In the SDN world the need for systematization of existing network and traffic engineering knowledge is evident. And this is still not enough; we need this systemized knowledge packed in nice formal packages that will be simple and easy to integrate in complex system that we develop and use. For this future step there is need for diverse skills and competences, a huge experience and practice base of engineering existing networks, theoretical sense for knowledge systematization and representation in meaningful way and again engineering knowledge to apply systematized solutions in practice. Moreover, expertise from different knowledge domains are identified as needed. Just to mention a few, network and traffic engineers, reliability engineers, computer scientist, mathematicians, and complex network theoreticians. Software technologies just have to be developed for new networking paradigm used for developing complex software systems and lot of efforts has to be invested prior final solution. Here in this paper we survey key SDN problems, demonstrate details on real use case examples and based on that we identify main categories of existing research in SDN software technology. Furthermore, we open some new avenues by introducing some ideas for future research in SDN software technology.

ACKNOWLEDGMENT

The work presented in this paper is supported by COST

action 1304 Autonomous Control for a Reliable Internet of Services (ACROSS), COST action 1201 Behavioural Types for Reliable Large Scale Systems (BETTY), the research grant 13.09.2.2.16. from the University of Rijeka, Croatia, and the research project project "SDN: applicability and service possibilities" at DTU.

REFERENCES

[1] IETF NETCONF Network Configuration protocol, RFC 6241, http://tools.ietf.org/html/rfc6241, retrieved on January 2015.

[2] ONF, Open Flow Switch Specification v 1.05. https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-switch-v1.5.0.noipr.pdf retrieved on January 2015.

[3] ONF SDN Architecture, Issue 1 (June 2014). https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR_SDN_ARCH_1.0_06062014.pdf, retrieved on January 2015.

[4] http://www.noxrepo.org/ , accessed on January 2015.

[5] http://www.projectfloodlight.org/floodlight/ , accessed January 2015.

[6] http://www.opendaylight.org/ , accessed January 2015

[7] P. Berde, et al., "ONOS: Towards and Open, Distributed SDN OS", Proc. HotSDN , 2014.

[8] http://onosproject.org/, accesed on January 2015

[9] Bierman et alt., ETF RESTCONF Protocol, http://www.ietf.org/archive/id/draft-bierman-netconf-restconf-04.txt, retrieved on January 2015.

[10] http://www.frenetic-lang.org/ , accessed January 2015,

[11] C. Monsanto, N. Foster, R. Harrison, D. Walker, "A compiler and run-time system for network programming languages", ACM SIGPLAN Notices - POPL '12 , January 2012.

[12] Rosen et alt. IETF RFC 3031- Multiprotocol Label Switching Architecture. http://tools.ietf.org/html/rfc3031 , retrieved on January 2015.

[13] IEEE802.1Qay- Provider Backbone Bridge Traffic Engineering http://www.ieee802.org/1/pages/802.1ay.html , accessed on January 2015

[14] Harrington et. al. IETF RFC 3411 – An Architecture for Describing Sinple Network Management Protocol Management Frameworks http://tools.ietf.org/html/rfc3411, retrieved on January 2015.

[15] ONF OF-Config v1.2. OpenFlow Management and Configuration Protocol.https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow-config/of-config-1.2.pdf , retrieved on January 2015.

[16] Network Functions Virtualization (NFV) Use Cases, ETSI GS NFV 001 v1.1.1 (2013-10). http://www.etsi.org/deliver/etsi_gs/NFV/001_099/001/01.01.01_60/gs_NFV001v010101p.pdf , retrieved on January 2015.

[17] D. Levin, A. Wundsam, B. Heller, N. Handigol, and A. Feldmann. "Logically centralized?: state distribution trade-offs in software defined networks," In *Proc. of the first workshop on Hot topics in software defined networks* (HotSDN '12), ACM, 2012, pp. 1-6, 2012.

[18] B. Heller, R. Sherwood, and N. McKeown. "The controller placement problem", In *Proceedings of the first workshop on Hot topics in software defined networks* (HotSDN '12). ACM, New York, NY, USA, pp. 7-12, 2012.

[19] M. Canini, P. Kuznetsov, D. Levin, and S. Schmid, "Software transactional networking: concurrent and consistent policy composition," In *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking* (HotSDN '13), ACM, New York, NY, USA, pp. 1-6, 2013.

[20] Kohei Honda, Nobuko Yoshida, and Marco Carbone, "Multiparty asynchronous session types", In *Proceedings of the 35th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages* (POPL '08). ACM, New York, NY, USA, pp. 273-284, 2008.

[21] Lorenzo Bettini, et al. Global Progress in Dynamically Interleaved Multiparty Sessions. In Proc. of *Conference on Concurrency Theory* (CONCUR 2008), volume 5201 of LNCS, pp. 418–433. Springer, 2008.

[22] Scribble programing languague, http://www.scribble.org/, Accessed Feb 2015.

[23] A. D. Ferguson, A. Guha, C. Liang, R. Fonseca, and S. Krishnamurthi. Hierarchical policies for software defined networks. In *Proc. of the first workshop on Hot topics in software defined networks* (HotSDN '12). ACM, New York, USA, 37-42, 2012.

[24] A.Voellmy, H. Kim, and N. Feamster. "Procera: a language for high-level reactive network control," In *Proceedings of the first workshop on Hot topics in software defined networks* (HotSDN '12). ACM, New York, NY, USA, pp. 43-48, 2012.

[25] Galinac, Tihana, Huljenić, Darko; Influence on basic call set-up by integration of packet and circuit switched Core Network *In Proceedings of the International Conference on Computers in Telecommunications of the 25th International Convention* (MIPRO 2002) Opatija, Croatia, 2002.

[26] Tanković, Nikola, Galinac Grbac, Tihana; Truong, Hong-Linh; Dustdar, Schahram, Transforming vertical web applications into elastic cloud applications *In Proc. of International Conference on Cloud Engineering* 2015, Phoneix, USA. 2015.

[27] Franco, Juliana, Drossopoulou Sophia; Behavioural types for non-uniform memory accesses, *In Proc. Of Programming Language Approaches to Communication and Concurrency cEntric System, 2015, London, UK, 2015.*

Data Collection for Software Defect Prediction – an Exploratory Case Study of Open Source Software Projects

Goran Mauša*, Tihana Galinac Grbac * and Bojana Dalbelo Bašić**
* Faculty of Engineering, University of Rijeka, Rijeka, Croatia
** Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
goran.mausa@riteh.hr, tihana.galinac@riteh.hr and bojana.dalbelo@fer.hr

Abstract – Software Defect Prediction (SDP) empirical studies are highly biased with the quality of data and widely suffer from limited generalizations. The main reasons are the lack of data and its systematic data collection procedures. Our research aims at producing the first systematically defined data collection procedure for SDP datasets that are obtained by linking separate development repositories. This paper is the first step to achieving that objective, performing an exploratory study. We review the existing literature on approaches and tools used in the collection of SDP datasets, derive a detailed collection procedure and test it in this exploratory study. We quantify the bias that may be caused by the issues we identified and we review 35 tools for software product metrics collection. The most critical issues are many-to-many relation between bug-file links, duplicated bug-file links and the issue of untraceable bugs. Our research provides more detailed, experience based data collection procedure, crucial for further development of SDP body of knowledge. Furthermore, our findings enabled us to develop the automatic data collection tool.

I. INTRODUCTION

The research community is in need of a systematically defined data collection procedure for Software Defect Prediction (SDP) studies. In order to achieve this goal, we need to be aware of all the issues that could lead to data bias. Data collection from freely available open source projects is becoming increasingly popular, so we decided to reveal as many issues that are present in their development repositories.

A great number of empirical studies tried to indentify the best approach to SDP and most of them are collected in several systematic review [1,2,3,4]. However, there is still work to be done on reproducibility of empirical studies based on datasets retrieved from development repositories [5]. Detailed and replicable data collection procedures have to accompany the datasets in order to achieve greater generalization of SDP studies [1]. Furthermore, it is known that SDP repositories often suffer from data quality issues like: outliers, missing values, redundant or irrelevant attributes and instances, overlapping classes, contradictory samples, data shifting, imbalance of classes or replicability [6], [7]. Thus, we need to start systematically build a knowledge base on reusable research datasets. In this paper, we want to stress on role of human introduced bias in data collection procedure because it is developers who make entries into development repositories. We also recognize the importance of reusable tools for automated data collection that could play significant role in systematically building of uniformly obtained datasets. Freely available reusable tools, that have open source software with clearly defined data collection mechanism implemented, could form a base for wider discussion of data collection issues and its relations to bias. It is important that all research in the area of SDP collects the data according to a standardized procedure in order to perform repeatable and verifiable case studies that will benefit the body of knowledge. Thus, we aim to develop a rigorous systematic data collection plan that could be reused by other researchers. We also plan to quantify the main sources of bias. That kind of information would improve comparability between studies and generalization ability. Building a detailed data collection procedure and addressing all the possible issues is the first and crucial step in that process. Moreover, we believe the procedure should also be explained clearly enough to be applicable in industrial research as well as in academic that most often rely on open source software [8].

The paper is structured according to the guidelines given by Runeson and Höst [9] and Kitchenham et al. [10]. Section II puts our research into context and presents the related work. Section III explains in detail experimental planning where we organized the objective of our research into several research questions, presented the subject of the experiment and the means to perform it, defined the tasks that are going to be executed and methods how to do it, described the experiment design and prepared the analyzes and validation procedures that will enable us to answer the research questions. Section IV describes the execution of experiment. Finally, we present the results and answers to each research question in Section V and conclude the paper in Section VII.

II. BACKGROUND AND RELATED WORK

The importance of building a verifiable, repeatable and accurate data collection plan was recognized and put into several steps by Basili and Weiss [11]. Our research goal is to create most often used type of SDP datasets

using such a data collection plan. We analyzed all the 36 papers that were selected for final analysis in [1] because those were the ones that contained description of data collection procedure. However, we identified there is a lack of systematic unified mechanism in that area. There is a number of SDP studies that contain a section describing their data collection. However, there are no guidelines for reporting the crucial details of data collection process. Hence, it is not not performed systematically and one could not compare and quantify bias introduced with the data collection process.

Nevertheless, there were some attempts in that direction. A very good example of planned validation procedure can be found in [12]. They analyzed possible threats to validity, prepared and performed a manual check of a subset of potentially flawed data and evaluated the soundness and completeness of their data collection procedure. A detailed plan of analysis that should confirm the appropriateness of a novel approach for the data collection procedure in the phase of linking bugs to changes is present in [13]. They prepared a sort of benchmarking data set, manually collected and validated using observer triangulation. They named it "*the ground truth*" and used it to compare their novel approach to the traditional one in terms of precision, recall and F-measure evaluation metrics. We believe this approach should be used more often when introducing novel concepts because with the use of statistical tests of significance it could accelerate the diversification of true improvements. Confirmation to our statements comes from research presented in [14] that used the same "ground truth" data set and in [12] that even improved and enlarged such a data set. One commonly identified issue is the *bug linking* issue when collecting the data from two separated repositories of bugs and source code changes is addresed in [15]. The paper [16] identified the bug-feature bias and commit feature bias in defect datasets.

To the best of our knowledge, this is the first research intended to find all other issues present in the data collection procedure that could lead to biased software defect prediction datasets. We want to identify all the critical information that may be useful in comparison of different studies. It is important to be systematic because otherwise we could unintentionally include a source of bias and not be able to evaluate its impact. In our study we used students to manually collect the datasets and we draw conclusions from their data collection process. The similar approach is mentioned as in [16] but no systematic approach for the analysis of manually collected datasets is proposed. In this paper we conduct a systematic experiment which aims to identify all the traps in data collection process that lead to bias. In the remainder of the paper we refer to them as **issues**.

III. EXPERIMENTAL PLANNING

A. Research Goal

Our research goal is to create a systematic data collection procedure for SDP data sets that will enable us to develop automated data collection tool. To achieve this goal, we follow the process of data collection procedure development presented in Figure 1. This paper presents the first step of that process. Its purpose is to reveal and quantify all the issues that stand in the way of creating the data collection procedure, and to find the appropriate software metrics tool.

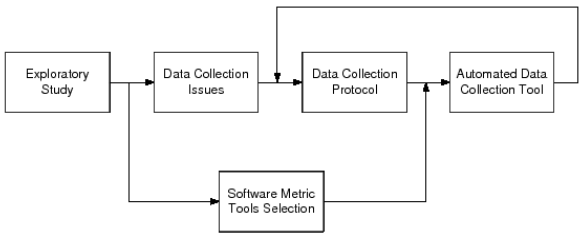


Figure 1. Research plan

B. Research Questions

Our research goal is divided into two research questions that will guide our exploratory study:

RQ1. Which issues affect the data collection procedure and how could they be overcome?

RQ2. Which software metrics tools are best suited for automatic collection?

Giving answer to **RQ1** is the primary purpose of this study. We hope to acquire as much experience and insight into the issues of data collection process as possible. Answering **RQ2** will help us develop an automated data collection tool. There is a number of software metrics (SM) tools present on the Internet and in related work so we need to review them and implement the ones that satisfy certain criteria.

C. Experimental Subjects

The participants that conduct the data collection are 12 students of the Faculty of Engineering at the University of Rijeka, Croatia, enlisted in the course of Software Engineering Management, obligatory in the first year of graduate Computer Science study program. This is an appropriate course because it teaches the students the importance of proper and accurate data in the decision making processes in the evolution of a large and complex software product. The participants had no previous experience with SM tools, little experience with bug repository use and mixed experience with source code management repositories. That is why they were all introduced to the concepts of these repositories in preparation exercises and seminars. Meanwhile, we omitted the fact that this is an exploratory study and that the observer triangulation is used.

D. Experimetnal Material

The main experimental material are: software development repositories, SM tools, and data collection forms. There are two **development repositories** crucial to the SDP data collection: the bug tracking (BT) repository and the source code management (SCM) repository. The Eclipse open source community is the most often used subject in SDP data collection so we used its largest projects in our experiment: JDT, PDE, Platform, BIRT and Mylyn. We use the term large projects for those that contain houndreds or thousands of

source code files and have been developed through at least a decade of evolution. The Eclipse community uses Bugzilla as its BT and GIT as its SCM repository. Bugzilla is a Web-based general-purpose bug tracker that allows developers to keep track of both defects that cause loss of functionality and requirements' requests. GIT is a distributed version control and source code management system with complete history and full version tracking capabilities. The **SM tools** are required in the second phase of our experiment for calculation of various product metrics. Further details are given in section IV-A. We use five **data collection forms** in the data collection procedure. Each form is prepared and explained in details for each of the tasks demanded from the participants to solve:

Form 1 is designed to collect the pieces of information from the BT repository: Bug ID, Product, Version, Summary (Full), Priority, Assignee, Number of Comments, Opened, Changed.

Form 2 is designed to document the bug-commit linking output: Bug ID, Found (yes/no), Commit ID.

Form 3 is designed to collect the pieces of information related to the linked files from the SCM repository: File path, Bug ID, Insertion, Deletion

Form 4 is designed to collect all the metrics obtained by SM tools both before and after the bug fix: File path, Product Metrics (m_1, m_2, ..., m_n).

Form 5 is designed as the final dataset form that containes all the metrics and the number of bugs for each file: File path, Metrics (m_1, m_2, ..., m_n), Number of bugs, Release.

E. Tasks and Methods

We need to perform several **tasks** to answer our research questions:

1. Examine the appropriateness of SM tools for our data collection procedure
2. Execute the data collection procedure manually and automatically
3. Analyze the collected data and examine the sources of bias

The first task gives us the answer to the **RQ2**. It needs to be done prior to data collection because it is an investigation of its own and the results are implemented in the second task. The manual execution of the data collection reduces initial researcher bias and gives insights into the issues that need to be addressed. Making the data collection precess automatic could reveal some additional issues. Furthermore, the manual and automatic data collection provide us with valuable data for evaluation of the issues and for answering the **RQ1**. In order to conduct all of these tasks, we use the following **methods**:

1. **Seminars** for reviewing the SM tools
2. **Written Instructions with Examples** for each step of the data collection procedure
3. **Data analysis and validation methods** of the results obtained during the experiment

The participants are given a preselected subset of the SM tools for reviewing and they are involved in the execution of data collection procedure. The authors of this paper perform the literature and Internet survey to locate all the SM tools, form a list of the most representative range of product metrics, find the candidate tools for further examination through seminars, plan the data collect procedure, prepare the written instructions and examples for each step of the data collection procedure, and determine the data analysis and validation techniques that we explain in more details in the following section.

F. Experimental Design

The manual collection is executed to identify all the issues present in our data collection process. The task scheduling is performed so that it allows a validation of the performance of different SM tools and the data itself early enough to enable timely corrective action on data collection procedure. Since participants could be the source of bias as well, we assign each project and each tool to more than one participant, i.e. following the principle of observer triangulation [13]. The participants are given personalized tasks in 4 exercises with 52 steps and explained by 21 screen-shot figures. The data collection procedure is divided into several phases. Each phase, along with its issues is presented in Figure 2. We describe them in more details and give a proposition how to solve them in the remainder of this subsection.

SM Tool Selection: The initial phase is to select the SM tools for collection of software metrics that represent the independent variables in SDP. Since we decided to analyze Eclipse projects and collect the most often used software metrics, the tool needs to be able to analyze large projects written in java and give reports for calculated software product metrics on file or class level of granularity.

Issue 1.1 The SM tools requirements are:

1. **Availability of tool** - for in-use testing
2. **Programming language** – java
3. **Level of granularity** – file / class level
4. **Software metrics** – product metrics
5. **Usability - input** – capable for large projects
6. **Usability - output** – *csv*, *html* or *xml* reports

Project Selection: The initial phase also needs to select the open source project that will be the source of SDP data set. The issues that affect this initial phase are:

Issue 2.1 The project needs to be present both in the BT and in the SCM repository

Issue 2.2 The project's releases from BT must be present as release tags in SCM repository

The presented issues are rather self-explanatory. We need to be able to access the source code, the bugs and the corresponding releases in both repositories. We evaluate these issues by choosing a great number of projects from the chosen Eclipse community, and analyze which projects contain release designations in both repositrories.

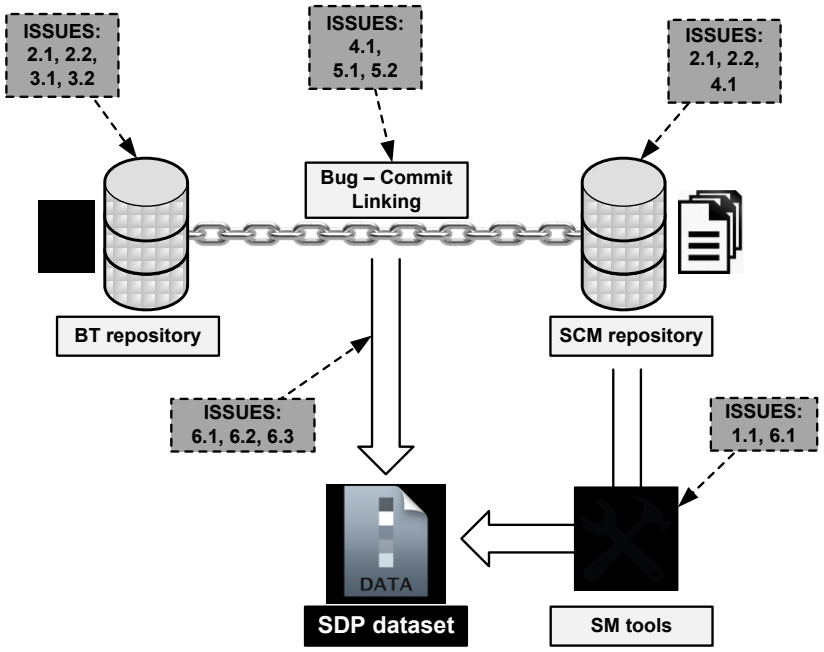


Figure 2. The impact of identified issues on each data collection phase

Bug Repository Collection: The first data collection phase is to find defects from the Bugzilla repository and collect all the valuable pieces of information.

Issue 3.1 The bug must be relevant to SDP
Issue 3.2 The bug should be solved

BT repositories typically store bugs and enhancements. Furhteremore, bugs may relate to minor typographical errors. The bugs relevant to SDP are the ones that represent a loss of functionality which is indicated by severity level above trivial. Since we need to linking the bug to source code, the bugs need to be solved in order to be present in the commits from the SCM repository. A BT tabular report, as presented in **From 1**, for all releases of each project is required with following parameters:

- Bug Status: resolved, verified and closed
- Bug Resolution: fixed
- Bug Severity: blocker, critical, major, normal and minor

This way we bypass all the unconfirmed, invalid or duplicated defects, defects that are not solved or that could not be solved and requirement requests.

We evaluate the impact of these issues by comparing the number of bugs that satisfy our parameters to the number of bugs that satisfy opposite parameters (Bug Status: unconfirmed, new, assigned and reopened; Bug resolition: invalid, wontfix, duplicate, worksforme, moved and not_eclipse; Bug severity: enhancement, and trivial).

SCM Repository Collection: The second data collection phase is to find the SCM repository for the chosen project in GIT. We need stable releases of the project for future analysis of software metrics. If issues 2.1 and 2.2 are both satisfied, there remains only one issue, related mostly to large projects and we evaluate it by counting the number of such projects.

Issue 4.1 The project may be divided into several sub-repositories within the foundation's top level GIT repository

Bug - Commit Linking: The most important data collection phase is to link the BT and SCM repositories because it yields the number of defects per software module. Every source code change is stored within commits and each commit has the developers description of the change within a message. The linking is done by searching the Bug ID in the commit message. We use **Form 2** to record successful linking.

Issue 5.1 The Bug ID number may be a part of larger block of characters
Issue 5.2 The Bug ID may be present in more than one commit

The issue of varying length of the Bug ID presents an obstacle to automatic data collection. It is important to search for the exact match of the Bug ID. We evaluate its impact by calculating the percentage of the Bug IDs we can link with the simple search and the strict search we mentioned. The second issue is because of the nature of complex bugs that require a lot of change, possibly in different subrepositories, and because of the loose development rules in the community. We evaluate it by comparing the number of linked commits and the number of bugs.

Source Code Analysis: This phase consists of calculating the product metrics of all the source code files in each release of the chosen project. The SM tools are used for that purpose. All the product metrics obtained by these tools are collected in **Form 4**.

SDP Dataset Generation: The final phase of our data collection procedure is to construct the final SDP dataset in the **Form 5**, combining the output of the previous two phases.

Issue 6.1 The file identifier needs to be unique for the *Source Code Analysis* and the *Bug - Commit Linking*
Issue 6.2 File – Bug cardinality is many-to-many
Issue 6.3 Single file can be changed in multiple commits that are linked to one bug

The unique identifier of any file present in one release is its *File Path*. In order to cope with issue 6.1, the file path was extracted from commits in the same way it was extracted in the output of SM tools. Neglecting this issue would make it very difficult to connect the previous two phases. Due to issue 6.2 and 6.3, we count the number of distinct *File Path* and *Bug ID* (FB) ordered pairs in **Form 3**. To evaluate the potential impact of issue 6.2, we compare the number of FB pairs with the number of linked files and bugs. The impact of issue 6.3 will be evaluated through the number of FB duplicates.

IV. RESULTS

The literature and Internet survey found 35 SM tools (12 from [8] and additional 23 from the Internet search) and we analyzed them all. The tools were compared against demands presented in subsection III-F. In the first stage, we analyzed the tool descriptions. In the second stage, we anylzed the remaining tools more thoroughly in usage on a sample project through participants' seminars. In the third stage, we tested the tools in large scale manual data collection. 19 tools passed the first review stage, 5 tools passed the second and only 2 remained in the end. Here is a full list of tools we analyzed: iPLASMA, inFusion, inCode, CodeCover, JavaNCSS, ckjm, Understand, Borland Together, CodeSonar Static Analysis Tool, CodePRO Analytix, Metrics 1.3.8, Moose, JavaMetrics, Testwell CMT ++ (CMTJava), JDepend, Dependency Finder, JarAnalizer, CCCC, Source Code Metrics, Classycle, Sonar, Resource Standard Metrics, Jhawk, Jtest, SonarGraph, PMD, JDiff, CodeCount, LocMetrics, CodeAnalyzer, JMT, Jmetric, ES2, Xradar, Essential Metrics. Table 1 presents 7 categories of disqualifying factors we applied and the number of tools reject due to each of them.

The tools that were tested in use are: CodePRO Analytix, Metrics 1.3.8, Source Code Metrics, Jhawk and LocMetrics. Metrics 1.3.8. required projects to be built and Source Code Metrics required the restart of NetBeans for each following analysis. A great number of analyses dictated to exclude these two tools. We did not manage to automate the Code PRO Analytix and since it was incapable of analyzing whole project releases it was discarted as well. The LOC Metrics[1] proved to be a fast tool offering a sample of 10 simple product metrics and the JHawk[2] enriched the set of metrics with additional 40. The complete set of software product metrics provided by these two tools is wider than in any other related study.

[1] http://www.locmetrics.com/
[2] http://www.virtualmachinery.com/jhawkprod.htm

TABLE I. SM TOOLS REVIEW - ISSUE 1.1

Requirement	*Important characteristics*	*Rejection rate*
1	Availability of tool	5
2	Programming language - java	0
3	Level of granularity - files	7
4	Software metrics - product - too few metrics	4 4
5	Usability - large input	3
6	Usability - output reports	1
7	Other - older version of another Other - not a tool, a platform Other - trial version too limited	2 2 2

The evaluation of data collection issues is presented in tables 2, 3, 4, 5 and 6. Table 2 reveals that not all of the Eclipse projects are present in the GIT repository. We analyzed 85 different Eclipse projects from BT repository. We found 76 of them in SCM repository, indicating that the issue 2.1 is rarely present. Furtheremore, for 51 projects we managed to find release tags in SCM that correspond to release numbers from BT. It shows the naming convention of releases are not always consistent and that issue 2.2 could impact greater number of projects. Finally, we found that 37 projects are distributed in more than one sub-repository, showing that it is important to be aware of issue 4.1. Table 3 presents the importance of choosing the appropriate parameters while searching for bugs in the BT repository. We compare the number of bugs collected by Our parameters (subsection III-F-Bug repository collection) and by other parameters. We can expect a completely different SDP dataset if we do not limit the *Bug Resolution* to fixed only. On the other hand, bugs with incomplete data (expected by **Form 1**) are rare and could cause insignificant bias. The inclusion of bugs of opposite severity may cause from 10% - 15% bias in the number of bugs. For the Mylyn project, the bias rises to 45%. Table 4 demonstrates the need for a precise search criterion for establishing the bug - commit links. A great number of incorrectly linked bugs may occur if we simply search for the Bug ID as a series of digits (Simple Search) and not defining its surrounding characters (Strict search). The bias in the linking rate from 7% - 40% proves issue 5.1 to be very important. Table 5 reveals that a single bug ID may be linked to more than one commit and that it is highly important to be aware of issue 5.2. Table 6 reveals that the number of FB pairs is much greater than the number of files that were linked to bugs, proving that the majority of bugs is located in a smaller part of the software. We also notice that the number of linked files can be both greater (PDE, BIRT, Mylyn) or lower (JDT, Platform) than the number of linked bugs, proving that issue 6.2 needs to be taken into account. The number of FB pair duplicates is presented in table 7. We performed additional manual root cause analysis of a subset of FB duplicates. We discovered that some FB pairs even occurred more than twice and their cause was:

TABLE II. EVALUATION - ISSUES 2.1, 2.2 AND 4.1

Demand	*Important characteristics*	*Acceptance rate*
-	Total projects analyzed	85
1	Present in BR and SCM repository	76
2	Corresponding release tags	51
3	>1 SCM subrepository	37

TABLE III. EVALUATION - ISSUES 3.1, AND 3.2

Bug requirement	**Number of bugs per project**				
	JDT	***PDE***	***Platform***	***BIRT***	***Mylyn***
Our paramaters	20353	7318	38511	13071	3492
Opposite severity	2875	895	4604	1243	1581
Opposite status	0	0	0	0	0
Opposite resolution	16488	3570	32582	3059	1909
Incomplete bug data	64	26	108	7	1

TABLE IV. EVALUATION - ISSUE 5.1

Linking Technique	**Linking rate per project**				
	JDT	***PDE***	***Platform***	***BIRT***	***Mylyn***
Bugs:	18696	6822	34641	8101	2739
Simple Search:	80.5%	65.8%	80.7%	59.8%	52.3%
Strict Search:	72.0%	58.6%	71.2%	18.9%	16.1%

• There are both a bug test and a bug fix reported for the same FB pair

• There are 2 bug fixing commits reported in the SCM for the same FB pair

• There had been reported one or two temporary bug fixes before the final bug fix was reported

• There are both a bug fix and a revert bug fix reported for the same FB pair

All the reasons we found show us that only the final bug fixing commit should be regarded as relevant

V. THREATS TO VALIDITY

There are two major threats to validity to our experiment: the specificity of data (only several largest Eclipse projects written in java stored in Bugzilla and GIT) and the possibility of human error from the participants. However, observer triangulation and the partial automation of the process reduced this the human error influence and most recent experiments showed the procedure can be also successfully used for other projects from other foundations, like Mozilla or Apache. In contrast to usual data collection procedures, we welcomed the differences in collected data from different participants because those were indicators of issues we need to address.

VI. CONLUSION

The goal of this exploratory research was to identify as many issues that may cause bias in the SDP data collection procedure. Our survey of related research and other sources in this area available on the Internet resulted in a preliminary data collection protocol. That protocol, along with detailed descriptions was given to 12 participants of our experiment and they were assigned 5 major Eclipse projects. Their output data was then analyzed and validated through a number of metrics and manual inspections.

TABLE V. EVALUATION - ISSUE 5.2

Number of	**Multiple bug-commit links per project**				
	JDT	***PDE***	***Platform***	***BIRT***	***Mylyn***
Bugs linked:	13468	3995	24663	1532	441
Commits linked:	17667	7573	29669	2127	3002

TABLE VI. EVALUATION - ISSUE 6.2

Number of	**Multiple file-bug links per project**				
	JDT	***PDE***	***Platform***	***BIRT***	***Mylyn***
Bugs linked:	13468	3995	24663	1532	441
Files linked:	12080	6629	15084	4273	975
Bug-file links:	25476	16595	23784	6663	1258

TABLE VII. EVALUATION - ISSUE 6.3

Number of	**Duplicated file-bug links per project**				
	JDT	***PDE***	***Platform***	***BIRT***	***Mylyn***
FB duplicates:	768	1233	903	4	0

As an answer to our **RQ1,** we identified and quantified several important issues that may become a significant source of bias in the SDP dataset. The collection starts with the choice of proper open source projects that contain their data in the BT and the SCM repositories. Some larger projects contain several SCM subrepositories. Relevant bugs are the resolved, verified and closed ones with the resolution fixed and severity level above trivial because SDP requires only those that caused a loss in functionality. Linking between bugs and commits must be done in a carefully defined strict manner because otherwise we might obtain bias in linking rate. Multiple links between bugs and commits (i.e. files) and duplicated file-bug links can be expected. We also answered our **RQ2**. After we thoroughly analyzed 35 SM tools, we identified *LOC Metrics* and *JHawk* as the most appropriate for our automatic data collection tool [17]. These two SM tools provide us with a respectable subset of 50 product metrics and are suitable for automating the data collection process.

Finally we find evidence that human error is widely present in development repositories. Allthough the Bug ID is a unique identifier of a software defect, it cannot be always linked to the commit that made it fixed. Therefore, a certain number of bugs remains untraceable and becomes an issue in the phase of determining the non-fault-prone files as well as the fault-prone ones. However, there are indications that some versions suffer a lot less from this problem and a more detailed analysis is intended for our future work.

ACKNOWLEDGMENT

The work presented in this paper is supported by the University of Rijeka research grant Grant 13.09.2.2.16.

REFERENCES

[1] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell. A systematic literature review on fault prediction performance in software engineering. *IEEE Trans. Softw. Eng.*, 38(6):1276–1304, Nov. 2012.

[2] E. Arisholm, L. C. Briand, and E. B. Johannessen. A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. *J. Syst. Softw.*, 83(1):2–17, Jan. 2010.

[3] T. M. Khoshgoftaar and N. Seliya. Comparative assessment of software quality classification techniques: An empirical case study. *Empirical Software Engineering*, 9(3):229–257, 2004.

[4] M. D'Ambros, M. Lanza, and R. Robbes. Evaluating defect prediction approaches: A benchmark and an extensive comparison. *Empirical Softw. Engg.*, 17(4-5):531–577, Aug. 2012.

[5] J. M. González-Barahona and G. Robles. On the reproducibility of empirical software engineering studies based on data retrieved from development repositories. *Empirical Softw. Engg.*, 17(1-2):75–89, Feb. 2012.

[6] D. Rodriguez, I. Herraiz, and R. Harrison. On software engineering repositories and their open problems. In *Proceedings of RAISE '12*, pages 52–56, 2012.

[7] M. J. Shepperd, Q. Song, Z. Sun, and C. Mair. Data quality: Some comments on the nasa software defect datasets. *IEEE Trans. Software Eng.*, 39(9):1208–1215, 2013.

[8] T. Galinac Grbac, P. Runeson, and D. Huljenic. A second replicated quantitative analysis of fault distributions in complex software systems. *IEEE Trans. Softw. Eng.*, 39(4):462–476, Apr. 2013.

[9] P. Runeson and M. Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical Softw. Engg.*, 14(2):131–164, Apr. 2009.

[10] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. E. Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Trans. Softw. Eng.*, 28(8):721–734, Aug. 2002.

[11] V. R. Basili and D. Weiss. A methodology for collecting valid software engineering data. *IEEE Computer Society Trans. Software Engineering*, 10(6):728–738, 1984.

[12] T. F. Bissyande, F. Thung, S. Wang, D. Lo, L. Jiang, and L. Reveillere. Empirical evaluation of bug linking. In *Proceedings of CSMR '13*, pages 89–98, Washington, DC, USA, 2013. IEEE Computer Society.

[13] R. Wu, H. Zhang, S. Kim, and S.-C. Cheung. Relink: Recovering links between bugs and changes. In *Proceedings of ESEC/FSE '11*, pages 15–25, New York, NY, USA, 2011. ACM.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[14] A. T. Nguyen, T. T. Nguyen, H. A. Nguyen, and T. N. Nguyen. Multi-layered approach for recovering links between bug reports and fixes. In *Proceedings of FSE '12*, pages 63:1–63:11, New York, NY, USA, 2012. ACM.

[15] A. Sureka, S. Lal, and L. Agarwal. Applying fellegi-sunter (fs) model for traceability link recovery between bug databases and version archives. In *Proceedings of APSEC '11*, pages 146–153, Washington, DC, USA, 2011. IEEE Computer Society.

[16] C. Bird, A. Bachmann, E. Aune, J. Duffy, A. Bernstein, V. Filkov, and P. Devanbu. Fair and balanced?: Bias in bug-fix datasets. In *Proceedings of ESEC/FSE '09*, pages 121–130, New York, NY, USA, 2009. ACM.

[17] G. Mauša, T. Galinac Grbac, and B. Dalbelo Bašić. Software defect prediction with bug-code analyzer - a data collection tool demo. In Proc. of SoftCOM '14, 2014.

Trendovi u izgradnji podatkovnih centara u Hrvatskoj

Marko Galac*, Davorin Valenčić **
* Combis d.o.o./Sektor mrežne infrastrukture i povezivanja, Zagreb, Hrvatska
** Veleučilište Velika Gorica/Velika Gorica, Hrvatska
Marko.galac@combis.hr , davorin_valencic@yahoo.com

Sažetak – Izgradnja standardnih podatkovnih centra (engl. Data Center), gdje smatramo podatkovne centre u fiksnim objektima postala je na nekim način standard današnjeg vremena. U posljednjih nekoliko godina pojavljuje se trend izgradnje mobilnih ili modularnih podatkovnih centara. To su podatkovni centri (najčešće u kontejnerskoj izvedbi) gdje korisnik u jednoj „kutiji“ dobije kompletan podatkovni centar s ugrađenim svim potrebnim sustavima . Upravo iz razloga što je podatkovni centar malih dimenzija, vrijeme planiranja i izgradnje je svedeno na minimum te postoji mogućnost preseljenja na novu lokaciju kroz period eksploatacije. Uz sve ove prednosti, ovaj tip podatkovnih centara u potpunosti teži smjeru „zelene“ tehnološke infrastrukture. U ovom članku su prikazani trendovi u izgradnji podatkovnih centara, s osvrtom na trenutno stanje u Hrvatskoj..

I. UVOD

Računala u današnjim vremenima više nisu nešto što je dobro imati i koristiti, već su postala potreba. Aplikacije i korisnički programi koje koriste gotovo sve organizacije za vođenje svog poslovanja ili pružanje usluga korisnicima, svakim danom su sve zahtjevniji po pitanju računalnih resursa najviše u smislu procesorske snage i memorije. Također pojavljuje se potreba za maksimalnom mogućom dostupnosti računala, odnosno računalnog sustava, kako ne bi došlo do narušavanja produktivnosti organizacije. Nedostupnost računalnog sustava, ne samo kod profitabilnih organizacija, može uzrokovati velike gubitke. Primjerice nedostupnost sustava neke od službi za pomoć unesrećenima u nekoj od situacija može uzrokovati velike materijalne a i ljudske gubitke. Kod profitabilnih organizacija, poput banaka, pružatelja Internet servisa, telekom operatera gubitci se mjere u velikim svotama novca. Sukladno gore navedenim rizicima većina organizacija odlučuje investirati u dodatne mehanizme koji štite njihove računalne sustave od rušenja odnosno nedostupnosti istih.

Jedan od mehanizama za osiguranje i podizanje razine dostupnosti sustava je izgradnja podatkovnog centra. Podatkovni centar je prostor sa kontroliranim parametrima u koje se smještaju računalni sustavi, sustavi za pohranu podataka i ostala ključna komunikacijska oprema. Prostor podatkovnog centra u objedinjuje nekoliko sustava kako bi se postiglo optimalno i u potpunosti kontrolirano okruženje za rad računalnih sustava. Neke organizacije zbog svojih potreba, investiraju u izgradnju redundantnog podatkovnog centra, što je zapravo najčešće replikacija primarnog podatkovnog centra smještena na nekoj drugoj lokaciji.

U posljednje vrijeme pored izgradnje klasičnih podatkovnih centara, koji se izgrađuju u postojećim prostorima ili se grade zasebni objekti za podatkovni centar, sve veću primjenu nalaze mobilni i modularni podatkovni centri koji su svoju primjenu do sada nalazili najčešće u vojnim organizacijama ili u organizacijama za upravljanje u kriznim uvjetima.

A. *Prednosti novih trendova*

Prednost pristupa izgradnje modularnih i/ili mobilnih podatkovnih centara je prvenstveno rok za izgradnje. Kod dosada, standardnog pristupa, izgradnji, potrebno je ishođenje velike količine dokumentacije za što je u često bilo potrebno više vremena no što je potrebno za samu izgradnju podatkovnog centra.

Pored kratkog roka izgradnje, predefinirana i unaprijed standardizirana rješenja modularnih/mobilnih podatkovnih centara omogućuju organizaciji kvalitetno planiranje, financiranje i izračun povrata investicije takvog projekta.

II. POJAM PODATKOVNOG CENTRA

Pod pojmom podatkovni centar (engl. Data center), kako je već u uvodu spomenuto smatramo posebno izgrađen objekt ili posebno uređenu prostoriju unutar objekta koja je namijenjena za smještaj ključnih računalnih sustava, uređaja za pohranu podataka (eng. storage) te ključne komunikacijske opreme, poput preklopnika, usmjerivača (eng. router) te sličnih uređaja. Unutar podatkovnog centra oprema se smješta u za to posebno dizajnirane ormare (eng. rack), i to prema ranije definiranom rasporedu kako bi se osigurao dovoljan kapacitet hlađenja i napajanja električnom energijom. Za postizanje visokog nivoa dostupnosti podatkovnog centra sustavi hlađenja i napajanja , kao jedni od ključnih komponenata, izvode se u redundantnoj opciji kako bi se osigurao kontinuitet rada u slučaju kvara ili redovitog servisa jednog od sustava.

Podatkovni centar možemo podijeliti na nekoliko osnovnih elemenata [1]:

- Izvori napajanja;
- Sustavi hlađenja;

Ovaj rad je podržan od Veleučilišta Velika Gorica, Velika Gorica, Hrvatska i tvrtke Combis d.o.o., Zagreb, Hrvatska.

- Komunikacijska instalacija (komunikacijski kabeli);
- Zaštita prostora (sustavi automatskog gašenja, sustavi vatrodojave, sustavi tehničke zaštite, sigurne sobe – IT kavezi);
- Sustav nadzora.

Svaki od ovih sustava dodatno možemo dodatno podijeliti na vrste i tipove.

A. Povijest podatkovnih centara

Podatkovni centri počeli su se razvijati još u rano doba računalne industrije. Tadašnja računala (mainframe) bila su velikih dimenzija te je za njihov smještaj bilo potrebno osigurati zasebne prostorije. Paralelno sa time računala su zahtijevala posebne uvjete za rad, poput rashladnih sustava i velike količine električne energije. Povezivanje tadašnjih računalnih komponenti zahtijevalo je velike količine kabela. Zbog takvih zahtjeva pojavila se potreba za razvijem metoda organizacije smještaja računalnih komponenti unutar tog posebno namijenjenog prostora. Tada se počinju dizajnirati postolja za smještaj računalnih komponenti, prethodnica današnjih serverskih ormara, dizajnirani su podignuti podovi, danas poznati kao tehnički pod, kako bi se omogućilo polaganje kabela ispod razine poda te su razvijani posebni sustavi napajanja i hlađenja koji su mogli zadovoljiti zahtjeve tadašnjih računala [2].

Kroz godine računala su se počela razvijati, postajala su dostupnija te je upotreba postaja sve raširenija, posebno u većim organizacijama koje su ih koristile za svoje poslovanje. Razvojem novih tehnologija računalni sustavi postajali su sve složeniji i sve veći, te su i organizacije počele pridavati sve veći značaj kontroliranoj okolini u kojima ti sustavi rade.

Osamdesetih i devedesetih godina u primjeni se sve više pojavljuju klijent-poslužitelj arhitektura. Poslužitelji polako zamjenjuju stara računala u tim prostorima. Tada se počinje koristiti pojam podatkovni centar, za posebno namijenjen prostor sa zadovoljavajućom infrastrukturom u kojoj je smještena oprema za centraliziranu obradu i pohranu podataka [2].

Devedesetih godina počinje ubrzan razvoj informatičkih tehnologija, raste potreba za brzim pristupom podatcima te pristupu Internetu. Organizacije koriste računalne sustave za upravljanje poslovanjem, pružanje usluga svojim korisnicima. Sve to zahtjeva stabilan računalni sustav jer on postaje veoma važna komponenta koja osigurava sigurno poslovanje organizacija. S obzirom da uređenje takvih prostora i investicija u snažne računalne sustave zahtjeva znatnu financijski izdatak, manje organizacije nisu bile u mogućnosti izgraditi vlastite podatkovne centre, i tada se pojavljuje se pojam kolokacija (engl. collocation), što podrazumijeva najam prostora u podatkovnom centru. Kolokacijom smatramo najam određenog dijela prostora u podatkovnom centru za smještaj korisničke opreme (serverski ormari u koje se potom smješta oprema) ili čak najam prostora unutar serverskog ormara za smještaj jednog ili više servera korisnika. U današnje vrijeme veće organizacije grade podatkovne centre koji isključivo imaju takvu namjenu. U Hrvatskoj postoji nekoliko takvih podatkovnih centara a neki od primjera su HT-Hrvatski Telekom DC Selska, podatkovni centar Križ - Data Cross u Jastrebarskom pored Zagreba te Altus IT.

B. Standardi i klasifikacija podatkovnih centara

Jedan od najčešće korištenih standarda za podatkovne centre je standard TIA-942 kojim su specificirani osnovni zahtjevi u dizanju podatkovnih centara. Standard je donesen od strane TIA – *Telecommunications Industry Association*. TIA organizacija akreditirana je od strane američkog nacionalnog instituta za standarde (eng. *ANSI - American National Standards Institute*) za razvoj standarda koji se temelje na industrijskim standardima, a za široku paletu informacijskih i komunikacijskih tehnologija (ICT tehnologija). Uvođenjem standarda za dizajn i izgradnju podatkovnih centara uvelike je olakšan posao projektantima koji su prije standardizacije nailazili na probleme poput nekompatibilnosti opreme zbog različitih pristupa ili zahtjeva pojedinih investitora.

Sam standard definira ključne elemente podatkovnog centra:

- kvalitativne detalje prostora i dizajn objekta/prostora,
- kabelsku (komunikacijsku) infrastrukturu,
- energiju (sustave napajanja i sustave hlađenja) i
- utjecaj na okoliš.

Paralelno sa TIA-942 standardima podatkovni centri klasificiraju se i prema Tier klasifikaciji, a zavisno o arhitekturi izvedbe i dostupnosti podatkovnog centra. Uptime Institute definirao je Tier Classification System kako bi konstantno mogao kvalificirati podatkovne centre u smislu instaliranih performansi pojedinih sustava, odnosno infrastrukture, kao i njihovu dostupnost [3].

Osim spomenute Tier certifikacije, Uptime Institut dodjeljuje i takozvani certifikat operativne održivosti čiji je cilj upravljanje, poslovanje i održavanje podatkovnog centra a nije temeljen na njegovom dizajnu sustava. Operativna održivost utvrđuje ponašanja i rizike izvan sustava Tier klasifikacije (I, II, III, IV) koji utječu na dugoročnu učinkovitost podatkovnog centra. U pojednostavljenom smislu to je poveznica između upravljanja i Tier funkcionalnosti infrastrukture podatkovnog centra. Tri glavna elementa certifikacije operativne održivosti su [3]:

- Upravljanje i operacije (npr. prisutnost osoblja na lokaciji, kadrovska osposobljenost, programi održavanja);
- Karakteristike izgradnje (npr. dodatni sustav neprekidnog napajanja - UPS);
- Lokacija podatkovnog centra (npr. opasnost od poplave, transportni koridori);

Certifikacija operativne održivosti izdaje se u tri razine [3]:

- Gold – iskorišten potpun potencijal za neprekidan rad infrastrukture ili su čak i pojačani pojedini dijelovi infrastrukture;
- Silver – postoje mogućnosti za poboljšanje da bi se postigao puni potencijal instalirane infrastrukture;
- Bronze – postoje značajne mogućnosti za poboljšanje da bi se postigao puni potencijal instalirane infrastrukture.

Samim time certifikat operativne održivosti postaje dodatak na Tier klasifikaciju podatkovnog centra.

C. Tier klasifikacija

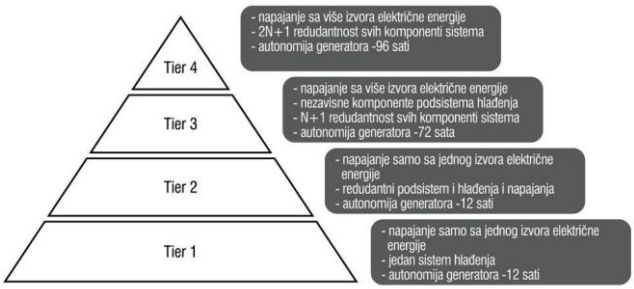
Tier klasifikacija je standardizirana metodologija koja određuje dostupnost (engl. availability) podatkovnog centra, a ujedno omogućuje investitoru da na temelju nje izrađuje proračune povrata investicije te performansi podatkovnog centra. Postoje dvije metodologije standardizacije koje se koriste, a to su [3]:

- Tier certifikacija izvedbene dokumentacije i
- Tier certifikacija izgrađenog podatkovnog centra.

Može doći do pitanja zašto postoje dvije standardizacijske forme po kojima se vrši certifikacija, ako se podatkovni centar gradi sukladno izvedbenoj dokumentaciji. Odgovor je veoma jednostavan i praktičan. Prilikom izrade izvedbene dokumentacije nemamo uvid u sve potencijalne izazove s kojima se možemo sresti na lokaciji izgradnje podatkovnog centra. Ukoliko se pojavi potreba zbog nekog ranije nepoznatog izazova, potrebno je izvršiti preinake na sustavima prilikom izgradnje i tada nastaje odstupanje od izvedbene dokumentacije i procjena na temelju izvedbene dokumentacije više nije vjerodostojna.

Tier certifikacija koja se odnosi topologiju infrastrukture podatkovnog centra podijeljena je u četiri razine, odnosno četiri nivoa dostupnosti [3]:

- Tier 1 – dostupnost iznosi 99,671% na godišnjoj razini, odnosno do maksimalno 28,8 sati godišnje neraspoloživosti, i ne postoji redundancija sustava.
- Tier 2 – dostupnost iznosi 99.741% na godišnjoj razini, odnosno do maksimalno 22 sata godišnje neraspoloživosti, a redundancija sustava je N+1.
- Tier 3 – dostupnost iznosi 99,982% na godišnjoj razini, odnosno do maksimalno 1,6 sati godišnje neraspoloživosti, a redundancija sustava je N+1 ili N+2.
- Tier 4 – dostupnost iznosi 99,995% na godišnjoj razini, odnosno sustav ne smije biti neraspoloživ duže od 0,4 sata godišnje, a redundancija sustava je 2N+1; 2N+2, …



Slika 1. Tier klasifikacija [9]

Pojmom redundancije (engl. redundancy) sustava smatramo sustave koji se instaliraju pored tj. paralelno sa primarnim sustavom a u cilju osiguranja dostupnosti tog sustava kao dijela infrastrukture podatkovnog centra. Postoji nekoliko modela redundancije od koji se u podatkovnim centrima koriste najčešće modeli [3]:

- 2N – redundantni model koji osigurava da svaka komponenta sustava ima rezervu (engl. backup) kako bi se postigao efekt što manje vjerojatnosti pada sustava, odnosno nema SPoF (engl. no single point of failure) točke.
- Need+1 odnosno koncept redundancije gdje je kapacitet konfiguriran da uključuje kapacitet koji se koristi, ali da postoji i rezervni koji se uključuje u slučaju pada primarnog kapaciteta.
- Need+2 odnosno koncept redundancije gdje je kapacitet konfiguriran da uključuje kapacitet koji se koristi, ali da postoje 2 rezervna koji se uključujuća u slučaju pada primarnog kapaciteta ili se treći uključuje u slučaju pada prva dva

U praksi se još mogu pronaći konfiguracije sustava koje se izvode kombinacijom nekih od gore navedenih modela.

III. OSNOVNI ELEMENTI PODATKOVNOG CENTRA

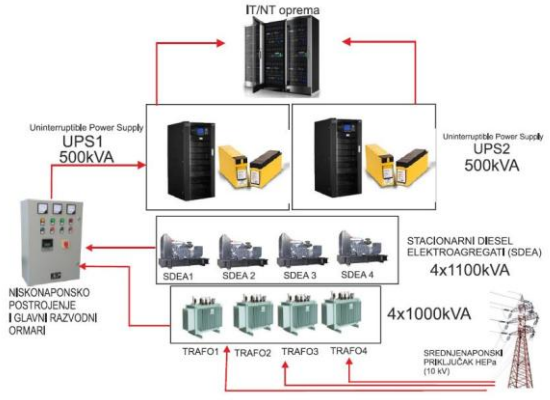
A. Sustavi napajanja

Sustav napajanja jedan je od najkritičnijih elemenata podatkovnog centra. Ispad sustava napajanja trenutačno uzrokuje prekid rada svih sustava podatkovnog centra što ima za posljedicu nedostupnost podataka, korisničkih servisa te samim time uzrokuje velike gubite u financijskom smislu. Kako bi se izbjegao takav scenarij sustavi napajanja se izvode na način da se osigura primarni i sekundarni izvor napajanja.

Primarni izvori napajanja su sustavi napajanja preko javne energetske mreže. Sukladno Tier standardima za Tier 3 i Tier 4 klasifikaciju potrebno je osigurati minimalno dva neovisna izvora napajanja sa javne energetske mreže. Primarni izvor napajanja sastoji se od srednje naponskog priključka, transformatorskog postrojenja, nisko naponskog postrojenja te glavnog razvodnog ormara u kojem se spajaju primarni i sekundarni izvori napajanja te se iz istog izvodi energetski razvod unutar podatkovnog centra.

Sekundarni izvori napajanja su sustavi napajanja neovisni o javnoj energetskoj mreži i služe kao izvori napajanja u slučaju pada odnosno nedostupnosti primarnog izvora napajanja. U sekundarne izvore napajanja svrstavamo diesel elektro agregate te potencijalno neke od alternativnih izvora napajanja. Neprekidni izvori napajanja (engl. uninterruptable power supply - UPS) također mogu biti uvršteni u sekundarne izvore napajanja iako u pravilu oni služe kao premosnica prilikom ispada primarnog izvora napajanja do uključenja sekundarnog izvora napajanja i tu imaju ključnu ulogu. U pojednostavljenoj verziji procesa, prilikom ispada primarnog izvora napajanja, sustav dojavljuje nadzornom sustavu alarm koji potom uključuje diesel eletro agregate kojima je potrebno neko određeno vrijeme da bi se sinkronizirali i pustili napajanje u energetsku mrežu podatkovnog centra. Upravu u tom trenutku neprekidno napajanje ima ključnu funkciju da premosti taj period tj. da oprema radi na izvoru napajanja iz uređaja za neprekidno napajanje. Kada se sinkroniziraju agregati i puste napajanje u mrežu, neprekidno napajanje prima signal da se pojavio izvor napajanja na njegovom ulazu i vraća se u čekanje i puni baterije za eventualni novi ispad.

Snaga diesel elektro agregata i sustava neprekidnog napajanje definira se sukladno maksimalnom očekivanom kapacitetu podatkovnog centra. Jedan od najčešćih pristupa u izgradnji podatkovnog centra da se izgradnja ovih sustava radi u dvije faze, gdje se u prvoj fazi instalira potreban kapacitet koji će biti instaliran u nekom skorijem periodu, a druga faza se pokreće kada je kapacitet ugrađen u prvoj fazi na 75 – 80 % iskorištenosti. U Tier klasifikaciji definirano je i vrijeme autonomije rada na sekundarnom izvoru napajanja. Prema Tier standardima spremnici goriva za agregate moraju imati spremnike goriva za minimalno 12 sati rada pod punim opterećenjem.



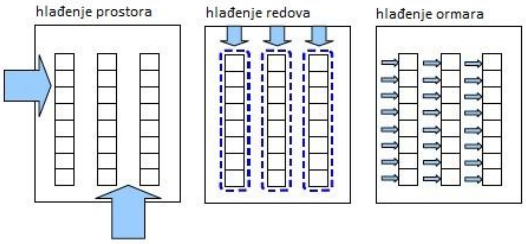
Slika 2. Blok shema sustava napajanja podatkovnog centra [7]

B. *Sustavi hlađenja*

Sustavi hlađenja podatkovnog centra također su jedan kritičnih dijelova infrastrukture podatkovnog centra. Namjena im je održavanje optimalnih uvjeta za rad računalne i komunikacijske opreme unutar podatkovnog centa. Optimalni uvjeti za rad opreme su između 16º C i 24ºC te relativne vlažnosti zraka između 40% i 55 %. U slučaju ispada sustava hlađenja dolazi do brzog porasta temperature u prostoru gdje je smještena oprema te postoji opasnost od pregrijavanja opreme i samim time kvara na istoj. Drugi parametar koji može uzrokovati kvarove na opremi je relativna vlažnost zraka, jer u slučaju prevelike vlažnosti zraka pojavljuje se kondenzat koji može uzrokovati kratki spoj na opremi, a u slučaju preniske vlažnosti zraka moguća je pojava statičkog elektriciteta koji također može uništiti ili značajno oštetiti opremu [2]. Svi ti potencijalni rizici moraju biti svedeni na najmanju moguću mjeru, pa se sukladno tome pridaje velika pažnja sustavima hlađenja. Kako bi bili sigurni da će sustav hlađenja biti dostupan u svakome trenutku, preporuka je da se spoje na način da imaju napajanje sa primanog i sekundarnog izvora napajanja. Naravno da redundancija ovog sustava nije opcija nego nešto je obavezno.

Sustave hlađenja možemo pojednostavnjeno podijeliti na vanjske i unutarnje jedinice , a međusobno spojene čine jedan sustav. Vanjske jedinice su najčešće veliki hladnjaci (engl. chiller), što su uređaji koji služe za hlađenje rashladnog medija koji se potom šalje kroz sustav cijevi u unutarnje jedinice koje hlade zrak u prostoru. Rashladni medij može biti voda, glikol ili plin, najčešće freon.

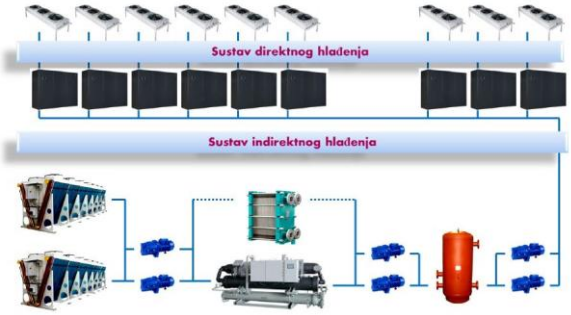
Unutrašnje hlađenje podatkovnog centra može se izvesti na nekoliko načina. U praksi najčešći pristupi u izvedbi su hlađenje prostora (eng. room based cooling), hlađenje između redova ormara u kojima je smještena oprema (eng. row based cooling) ili hlađenje pojedinog ormara (eng. rack based cooling)



Slika 3. Načini hlađenja prostora podatkovnog centra [4]

Unutarnje jedinice također mogu biti izvedene u nekoliko oblika, a najčešće su to CRAC sustav ili InLine/InRow sustav. CRAC sustav upuhuje hladan zrak direktno u prostor ili u prostor ispod tehničkog poda u podatkovnom centru, te na taj način stvara tlačnu komoru i hladan zrak izlazi ispred ormara u kojima je smještena oprema kroz izstrujne ploče u podu. InLine/InRow izvodi se na način da se klima jedinice smještaju između ormara u kojima je oprema te hladan zrak ispuhuju ispred same opreme. Kod ovog sustava postoji nekoliko opcija načina ispuhivanja hladnog zraka prema opremi ali generalno princip rada je isti.

Za postizanje veće efikasnost rada sustava oprema se zatvara u takozvane hladne zone, čime se smanjuje volumen prostora koji je potrebno hladiti.



Slika 4. Blok shema sustava hlađenja podatkovnog centra [7]

C. *Komunikacijska infrastruktura*

Komunikacijska infrastruktura obuhvaća sve potrebne bakrene i svjetlovodne kabele koji su potrebni za komunikaciju između opreme, senzora, mjerača ili drugih elemenata unutar podatkovnog centra. Kabelsku infrastrukturu možemo podijeliti na pristupne komunikacijske sustave te sustave vertikalnog i horizontalnog kabliranja.

Pristupna kabelska infrastruktura obuhvaća, najčešće mono modne svjetlovodne kabele, koji se koriste za povezivanje podatkovnog centra sa udaljenom lokacijom korisnika, redundantnim podatkovnim centrom (DR site[1]) ili za pristup Internetu. Za povezivanje udaljenih lokacija najčešće se koristi dark fiber, odnosno iznajmljena ili vlastita svjetlovodna veza koja se koristi isključivo za promet podataka. Za povezivanje na Internet infrastrukturu podatkovni centar se spaja na mrežu pružatelja usluge ili direktno na neko od čvorišta u Europi.

Sustavi vertikalnog i horizontalnog kabliranja obuhvaćaju lokanu pasivnu komunikacijsku infrastrukturu koja je namijenjena komunikaciji unutar podatkovnog centra, komunikaciji između ormara (međuveze) i povezivanje korisničkih računala. Za ovaj sustav najčešće se koriste više modna optička vlakna OM3 ili OM4 karakteristika, te bakreni kabeli (UTP ili STP) kategorije 6A ili 7.

OM3 i OM4 karakteristika svjetlovodnih kabela je standardom definirana karakteristika svjetlovodnog kabela a odnosi se na brzinu prijenosa podataka i propusnost (eng. bandwith). Konkretno za OM3 brzina prijenosa je do 10 Gbps na udaljenosti do 600 m. Za povezivanje komunikacijskih ormara (vertikalno kabliranje) i horizontalno kabliranje koriste se kabeli kategorije 6A ili kategorije 7. Brzina prijenosa je do 10 Gbps propusnost 500 MHz za Cat.6A, 600MHz za Cat.7. Kategorija 8 za sada je u povojima a omogućavati će brzine 40 Gbps. Identična stvar je i sa bakrenim kabelima gdje su standardom definirane kategorije koje određuju brzinu prijenosa i propusnost kabela [9].

D. *Zaštita prostora*

Zaštita prostora obuhvaća sustave zaštite od pojave požara, neautoriziranog pristupa podatkovnom centru. Ugrađuju se sustavi vatrodojave, sustavi gašenja požara te sustav tehničke zaštite koji obuhvaća video nadzor i kontrolu pristupa.

Sustavi vatrodojave pomoću senzora otkrivaju prisutnost čestica gorenja u zraku ili rade na principu prepoznavanja dima. Centralna jedinica sustava vatrodojave spaja se na centralnu jedinicu sustava gašenja te se obično spaja i na centralu koja automatizirano šalje obavijest o požaru lokalnoj vatrogasnoj postaji i odgovornoj osobi podatkovnog centra. U slučaju pojave požara , vatrodojavna centrala aktivira sustav gašenja, isključuje napajanje podatkovnog centra te šalje obavijest o požaru nadležnim službama i/ili osobama.

Sustav gašenja požara u podatkovnom centru aktivira se na osnovi dojave vatrodojavne centrale ili ručnog uključivanja sustava. Sustav gašenja izvodi se plinom , kako bi šteta nakon gašenja bila što manja te da ne bi došlo do ugroze osoba koji su se eventualno zatekle u prostoru za vrijeme požara , odnosno da oprema koja nije bila zahvaćena požarom ne bi bila oštećena djelovanjem medija za gašenje. U tu svrhu koriste se plinovi FM200 ili Novec 1230.

Sustav tehničke zaštite služi za nadzor prostora i perimetra podatkovnog centra te kontrolu ulaska i izlaska iz prostora. Sustav se sastoji od video kamera koje se raspoređuju sukladno definiranim standardima te čitača magnetskih kartica, otiska prsta ili slične metode prepoznavanja identiteta osobe.

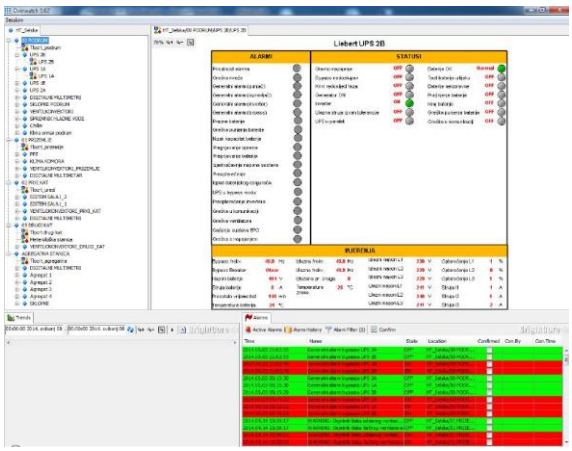
Sustavi zaštite prostora spajaju se na centralni nadzorni sustav objekta ili samog podatkovnog centra kako bi korisnik imao u svakom trenutku dostupnu informaciju o događajima u podatkovnom centru.

E. *Centralni nadzorni sustav*

Centralni nadzorni sustav, CNUS[2], omogućava kontrolu, prikupljanje informacija o radu svih ugrađenih sustava i njihovih elemenata, analizu podataka te udaljeno upravljanje. Daljinsko upravljanje ima značajnu ulogu u upravljanju DR podatkovnim centrima ili u slučaju pružatelja komunikacijskih usluga, upravljanje PoP čvorištima. .Implementacijom centralnog nadzornog sustava na temelju prikupljenih podataka i detaljne analize moguće je ostvariti znatne uštede u pogledu preventivnog tj. redovnog održavanja podatkovnog centra. Pored ušteda centralni nadzorni sustavi mogu preprogramiranom logikom izvršiti određene procese, poput gašenja određenih sustava ili automatskog podešavanja parametara na sustavima kako bi se izbjegao prekid u radu.

[1] DR Site - *Disaster recovery site* – podatkovni centar u kojem je smještena oprema na koju se repliciraju podatci iz primarnog podatkovnog centra. Najčešće se grade po principu zrcala, dakle identični kao primarni podatkovni centar.

[2] CNUS- Centralizirani nadzorni i upravljački sustavi



Slika 5. SNUPI - Sustav nadzora i upravljanja procesima infrastrukture tehnoloških prostora [7]

IV. TRENDOVI U IZGRADNJI PODATKOVNIH CENTARA

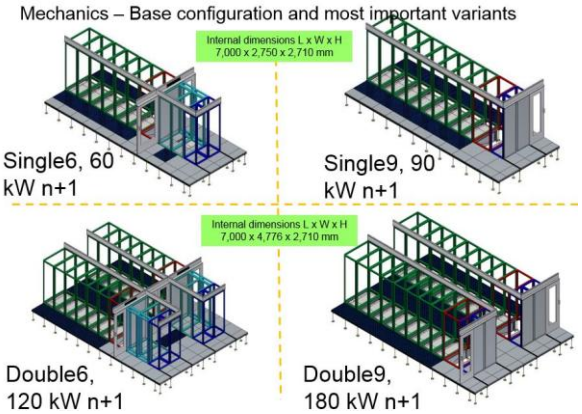
Izgradnja standardnih podatkovnih centra, gdje smatramo podatkovne centre u fiksnim objektima postala je na nekim način standard današnjeg vremena. U posljednjih nekoliko godina pojavio se trend izgradnje mobilnih ili modularnih podatkovnih centara. To su podatkovni centri, najčešće, u kontejnerskoj izvedbi standardnih dimenzija, gdje korisnik u jednoj „kutiji" dobije kompletan podatkovni centar sa ugrađenim svim sustavima. Dodatno je potrebno osigurati eventualno vanjsku rashladnu jedinicu i priključak napajanja koji su također dostupni u kontejnerskoj izvedbi. Jasno je da zbog zadanih dimenzija, podatkovni centar ne može imati kapacitet poput standardnog podatkovnog centra pa se stoga koristi metoda modula, koji se potom spajaju u jednu cjelinu kako bi se postigao potreban kapacitet. Ova izvedba podatkovnih centara posebno je zanimljiva organizacijama koje nisu izgradile podatkovni centar u vlastitim prostorima te postoji mogućnost da će kroz period eksploatacije doći do potrebe preseljenja podatkovnog centra na neku novu lokaciju Upravo iz razloga što je podatkovni centar sam za sebe ili su moduli standardnih dimenzija, preseljenje cijelog sustava ne predstavlja problem.

Ovakvim pristupom planiranja, izračuna povrata investicije i ostalih parametara investitor može detaljno isplanirati buduće korake vezane za proširenje kapaciteta sukladno potrebama. Neki od vodećih svjetskih kompanija za pružanje Internet usluga ili komunikacijskih usluga počele su koristiti ovakav pristup izgradnji svojih podatkovnih centara.

Iako i Hrvatskoj postoje tvrtke koje nude takva rješenja, investitori nisu još prepoznali prednosti novih trendova . Za primjer, možemo uzeti jednu od učestalih situacija gdje investitor, recimo tijelo javne uprave, investira u podatkovni centar, u prostoru koji nije u vlasništvu javne uprave već ima najam na određeni broj godina. Investicijom od nekoliko milijuna, javna uprava postaje ovisna o toj lokaciji bez obzira koliko troškovi najma porasli u određenom periodu. Upravo u takvim situacijama, organizacije pristupom izgradnje mobilnog ili modularnog podatkovnog centra štite svoju investiciju kroz cijeli period eksploatacije i samim time ostvaruju velike uštede.

Hrvatske tvrtke svoja rješenja plasiraju na vanjsko tržište gdje svojim korisnicima isporučuju visoko sofisticirana rješenja.

U Hrvatskoj i užoj regiji prepoznata je prednost MDC[3] rješenja koja na nivou jednog ili nekoliko spojenih serverskih ili komunikacijskih ormara osiguravaju sve zahtjeve željenog Tier standarda. MDC rješenja također su modularnog i mobilnog tipa. Ovakav tip rješenja najčešću primjenu nalazi u financijskim organizacijama koje svoje podatke štite najvišim stupnjem. Najviši stupanj zaštite osigurava neometan rad opreme unutar MDC-a sukladno normi vatrootpornosti F90, otpornosti od prodora vlage te svih ostalih parametara zaštite koji su usklađeni s EN normama.



Slika 6. Primjer standardiziranog modularnog rješenja za podatkovni centar [7]

[3] MDC-Micro Data Center – Sustav koji omogućava sve uvjete koje zadovoljava klasični podatkovni centar ali na nivou jednog ili više serverskih/komunikacijskih ormara

V. PODATKOVNI CENTRI U HRVATSKOJ

	Data Cross	***HT Selska***	***Altus IT***
Namjena	*Kolokacija*	*Kolokacija*	*Kolokacija*
Klasifikacija	*Tier 3*	*Tier 3/4*	*Tier 2*
Površina DC-a	*600 m2 (proširivo do 1300)*	*450 m2 (proširivo do 1000)*	*400 m2*
Primarno napajanje	*4 MW**	*4MW*	*nema podataka*
Broj petlji napajanja	*2*	*3*	*2*
Neovisni izvori	*2x 10 kV*	*3 x 10 kV*	*nema podataka*
Agregatsko napajanje	*4 x 1,6MW (2+2)**	*4 x 1,1 MW (3+1)*	*2*
Autonomija agregata	*72 sasta (3 dana)*	*120 sati (5 dana)*	*nema podataka*
UPS	*nema podataka*	*2x500 kVA*	*da*
Hlađenje	*Cold Aisle, pod*	*Cold Aisle, pod*	*Cold Aisle,pod*
Medij	*Glikol*	*Glikol/Freon*	*Glikol*
Gašenje	*Novec*	*Novec*	*FM200*
Godina izgradnje	*2012*	*2013*	*2005*
Investicija	*40 mil. Kn*	*65 mil. kn*	*nema podataka*

Tabela 1 Usporedba podatkovnih centara u Hrvatskoj

U gornjoj tabeli možemo vidjeti usporedbu izgrađenih podatkovnih centara u Republici Hrvatskoj, na temelju javno dostupnih podataka. Svi navedeni podatkovni centri izgrađeni kao podatkovni centri za kalkulaciju, dakle za najam korisnicima. Podatkovne centre tvrtki koje su gradile podatkovne centre za vlastite potrebe nisu uzete za usporedbu jer je većina podataka koje uspoređujemo poslovna tajna ili je potrebna posebna suglasnost istih da se ti podatci javno objavljuju. Iz usporedbe možemo vidjeti da su se investitori prilikom izgradnje podatkovnih centara vodili standardom te je usporedba veoma jednostavna i razumljiva. Također se može vidjeti da se kroz godine mijenjaju trendovi u pojedinim sustavima, posebno u ovom našem slučaju, na sustavu gašenja. FM 200 zamijenio je Novec , koji je ekološki prihvatljiviji. Na sustavima hlađenja, svi navedeni, zadržali su se na CRAC sustavu hlađenja sa zatvaranjem u hladnu zonu, kako bi što više smanjili trošak energije i održavanja sustava hlađenja.

Kako je ranije spomenuto vidljiv i je i pristup izgradnje u fazama, sukladno popunjavanju kapaciteta. Data Cross i T Com završili su prvu fazu izgradnje, napravljena je priprema za drugu fazu koja će biti pokrenuta čim kapaciteti prve faze budu gotovo iskorišteni [5]. Altus IT je imao nešto drugačiji pristup, možda iz razloga što je izgradnja u fazama jedan od novijih trendova, ta je odmah instalirao svu planiranu snagu, odnosno kapacitete [6].

VI. ZAKLJUČAK

Kako bi mogli lakše prikazati kolika je u stvarnosti korisnost i isplativost ulaganja u izgradnju podatkovnih centara, pokazuje studija koju je napravio Ponemon Institut, pod pokroviteljstvom Emerson Network Power. Neplanirani ispadi podatkovnog centra su skupi, a troškovi zastoja su u porastu, navodi se u studiji. Prosječna cijena po minuti neplaniranog zastoja je sada 7900 $, što je 41 % posto više od cijene iz 2010. godine koja je bila 5600 $ po minuti [1]. Iznosi koji su dobiveni nisu nikako mali, te se brzom analizom može doći do zaključka kako investicija u podatkovni centar je veoma isplativa a i povratak investicije je, u odnosu na druge projekte, veoma kratak.

	Unajmljeni podatkovni centar	Vlastiti podatkovni centar
Početni troškovi	+	-
Troškovi održavanja	+	-
Fleksibilnost nadogradnje	-	+
Nadzor prometa	-	+
Sigurnost	-	+

Tabela 2 . Usporedba izgradnje podatkovnog centra i kolokacije [2]

Iz ove usporedbe može se do odgovora da li investirati u vlastiti podatkovni centar ili koristiti uslugu kolokacije. Svaka organizacija bi trebala napraviti detaljnu analizu svojih potreba i zahtjeva za dostupnosti te na temelju toga donijeti konačnu odluku [2]. Ukoliko se radi o organizaciji sa manjim kapacitetom računalnog sustava, izgradnja vlastitog podatkovnog centra vjerojatno neće biti isplativa jer će povrat investicije biti veoma dug. Kako bi skratili taj rok, organizacije smanjuju investiciju na štetu dostupnosti sustava, pa u konačnici dobiju za uloženi novac nešto što neće ispuniti njihova očekivanja u smislu dostupnosti i pouzdanosti. Podatkovni centi koji su namjenski građeni za kolokacije uvijek su bolji odabir u takvim situacijama iz razloga što su građeni sukladno najvišim standardima i garantiraju određen nivo dostupnosti, te korisnik nema potrebu vršenja redovnih održavanja sustava i usklađivanja sa zakonskim normama.

Velike organizacije, posebno one koje djeluju na globalnom tržištu najvjerojatnije će za ishod analize imati rezultat da je investicija u izgradnju vlastitog podatkovnog centra isplativija te će shodno tome pokrenuti projekt izgradnje podatkovnog centra.

LITERATURA

[1] 2013 Cost of Data Center Outages, Independently conducted by Ponemon Institute LLC, prosinac 2013.

[2] Laboratorij za sustave i signale, Fakulteta elektrotehnike i računarstva, Sveučilišta u Zagrebu: „Redundantni podatkovni centri“, siječanj 2011.

[3] Uptime institute, http://uptimeinstitute.com/TierCertification/ , učitano 29.01.2015.

[4] Choices for data center cooling architectures http://www.electronicproducts.com/Thermal_Management/Fans_and_Blowers/Choices_for_data_center_cooling_architectures.aspx, učitano 30.01.2015.

[5] Wikipedia:Datacenter, http://en.wikipedia.org/wiki/Data_center , učitano 29.01.2015.
Web stranice Data Cross, www.pck.hr , učitano 30.01.2015.

[6] Web stranice Altus IT,

www.altus-it.hr , učitano 29.01.2015.

[7] Interna prezentacija Hrvatski Telekom d.d./Combis d.o.o. „Podatkovni centar Selska“

[8] Rittal GMbH – „RimatixS“ tehnička dokumentacija

Telekomunikacije – stručno naučni časopis republička agencije za telekomunikacije – Online izdanje, učitano 29.01.2015.

http://www.telekomunikacije.rs

[9] Krunoslav Kotarski, NOBIUM & FER - ZESOI – LSS, Generičko kabliranje, ISBN 978-953-6970-90-2

Data Storage and Analysis System for Conducting Biotechnological Experiments

Damir Spahić*, Goran Mauša*, Sandra Kraljević Pavelić** and Tihana Galinac Grbac*
* Faculty of Engineering/University of Rijeka, Rijeka, Croatia
** Department of biotechnology / University of Rijeka, Rijeka, Croatia
dspahic@riteh.hr, gmausa@riteh.hr, sandrakp@biotech.uniri.hr and tgalinac@riteh.hr

Abstract – Biomedical research, such as analyses of antiproliferative properties of certain compounds designed to target various human cancer cells, often requires extensive preclinical evaluation to obtain conclusive evidence. The first step in biological evaluation is typically very rigorous in comparing different compounds' concentrations on a number of target cells *in vitro*. Therefore, the experiments yield a set of data that needs to be properly documented, classified, stored and analyzed. This paper presents the Data Storage and Analysis System designed to make that process simpler, faster and less prone to human errors. The importance of such research and its implications lead to high sensitivity to bias. Our system minimizes the chance for human error in the process of data transfer by making the process as automatic as possible and allowing the posterior validation of data by the lead researchers. The analysis most often requires both numerical calculations, such as the IC_{50} value, and graphical visualization of compounds' performances. The system offers to researchers a simple interface for simultaneous comparison of results by choosing the appropriate combination of a compound, target cells and date of experiments.

Keywords: data storage, data analysis, visualization

I. INTRODUCTION

Antiproliferative evaluation of novel potential or existing anticancer compounds is a standard part of preclinical evaluation of great chemical and pharmacological importance. Those compounds that posses the ability to target specific signalling pathways, *i.e.* in cancer could be used in combined treatment approaches to improve cytotoxicity of drugs and clinical response of patients.

Extensive *in vitro* experiments need to be performed to achieve conclusive first evidence in such studies. For example, a study performed by Gazivoda, Kraljević et al. [1] tested 23 different compounds for antiproliferative activities on 7 cell lines (5 tumour cell lines and 2 human fibroblasts) at various micromolar concentrations. The results showed that five compounds exhibited antiproliferative effects and the one with the strongest concentration-dependent effect was chosen for further biological evaluation. The authors compared its effect with a commercial kinase inhibitor used as standard treatment of Philadelphia positive leukaemia. The new compound presented by the authors, exhibited promising results as it achieved greater antiproliferative effects *in vitro*. Indeed, *in vitro* evaluation of biological data coupled to bioinformatics analyses represents the 'golden standard' in the modern drug development process.

Having in mind a high number of potential compounds' combinations and tested concentrations, as well as requirement to perform the tests on diverse target cells, the amount of data may increase rapidly [2]. The need for a systematic data storage and analysis approach is evident [3] and this paper presents a system designed precisely for that purpose. The system's database is structured to allow multiple users – researchers to upload the data obtained in a standard form by any microplate absorbance reader. The process is made as automatic as possible to avoid human errors, which are typically the dominant reason of data bias. The researchers are required to fill the date related to compounds' names, tested concentrations and names of employed cell lines. The entering of data into the system is a two-step process with the first step based on initial values 'Day zero' and 'Blank' and the second step on final results template load. Since these steps might be slightly prone to human error, the researcher administrator is given the privilege to check and correct the entered data. The system also allows the researcher to automatically calculate the most important parameters like: Values for controls (untreated cells), percentage of growth cells (PG) treated with certain compounds at various concentrations and inhibitory concentration 50 (IC_{50}) and to visualize the concentration-response curves for each tested compound and cell lines. In this paper we present *data **S**torage and **A**nalysis for conducting **B**iotechnological **E**xperiments **S**ystem*, SABES. The SABES system was developed by students within the Software Engineering and Information Processing (SEIP) Lab, at the University of Rijeka [4], where we have had the experience of working on other ICT platforms like data collection tool for software defect prediction [5] and ambient assisted living environment [6]. The SABES system accelerates and improves the experimental workflow in early evaluation of potential drugs or biologically active compounds *in vitro*. In particular, it allows comparison of any compound's effect on any cell type *in vitro* by ensuring a simple and systematic storage within the system.

The structure of the paper is as follows: Section II presents the purpose of SABES system and gives a detailed design in five subsections; Section III presents the conclusion and future work intentions.

II. SABES SYSTEM

The usual workflow of *in vitro* experiments is presented in Fig 1. *Groups* of *Experimenters* are formed based on common research interest in research projects and one experimenter may be a part of more than one group. A group is analyzing the growth of a certain number of *Cells* when interacting with certain *Compounds*. In each experiment, the compound effect is analyzed in several levels of *Concentration*. The experimenter may perform a number of experiments using microplate absorbance reader that produces a standardized output table in .xls format for each experiment (*Experiment Table)*. There is one experiment table at the beginning of the experiment and it is called *Day* zero and the experimental end-point table that we refer to as the *Measurement*. Both experiment tables include *Blind Test* values and *Control value* only in the Measurement file. The blind test values refer to absorbance values of test reagents used for assessment of cell growth and the control value stays for untreated cells. These values are important for proper analysis of measurement results. Finally, besides the experimenters, we also distinguish the Role of researchers that perform these analyses. They can view, modify and combine data from several experiments.

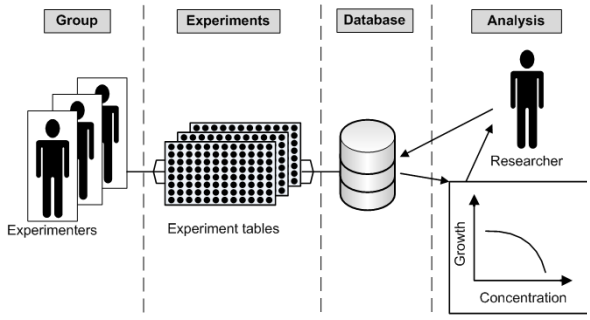


Figure 1. SABES Experiment

Here presented experiment storage and analyses is time-consuming and in particular, major problems are encountered when a need arises for localization of raw data - experimental files upon a certain length of time, or when finding only those files related to effects of the same compound or its behaviour for the same target cells, or even when a need is to additionally compare and group results from various experiments. Furthermore, the experiments may be run by students where unintentional mistakes occur at a higher rate in transfer or analysis of raw data. The previous modality of work included a manual copy-paste activity of raw data into semi-formatted Excel sheets prepared for data analyses. That is why the first task of SABES system was to provide for a controlled import of raw experimental data in a user friendly fashion and less prone to errors. With the SABES system, all the data is stored in a single database, and the user is guided through the data import process, without the need to be an expert in the field. Stored data are easily accessible through the interface. Furthermore, the system allows the supervision of main researchers or mentors and the possibility of additional manual checking of errors. The second task we achieved with the SABES system is comparison of previously performed experiments by the following filters: (1) the person who made the experiments, (2) by tested compound, or by (3) type of cells. The user may choose the data, make a graphical plot, calculate IC_{50} values, and even analyze mean values for a certain experimental point and standard deviations on a graph. In the following section we present the system architecture and technologies employed for its implementation where a detailed explanation of its most important parts is described.

A. Structure

The SABES data storage and analysis system is implemented using standard web development technology tools PHP, MySQL and jQuery. PHP is a server-side scripting language that is used for web development for a long time. Thus, it is supported both in Windows and Linux OS. It contains built-in modules for accessing many of the most popular database servers, making it easy-to-implement. One of the main advantages of PHP is its ability to run the scripts directly on a server, without previous compiling. The system's data storage functionality needs to support large amounts of raw data and that is why we used a relational database designed in MySQL, the most widely used open-source relational database management system. AJAX (Asynchronous JavaScript And XML) is a group of techniques used on the client-side, which allows the development of fast dynamic web applications because it sends and retrieves data asynchronously (in the backend). The jQuery framework is a cross-platform library designed to simplify the client-side scripting. We used it to simplify the implementation of AJAX queries, *i.e.* to achieve the asynchronous communication with server-side, which makes the application resemble to a desktop application from the user perspective [7]. The SABES system architecture is presented in Fig 2. The technologies we used removed the web browser incompatibility so the system can be run in any browser without any design adjustments.

B. Database

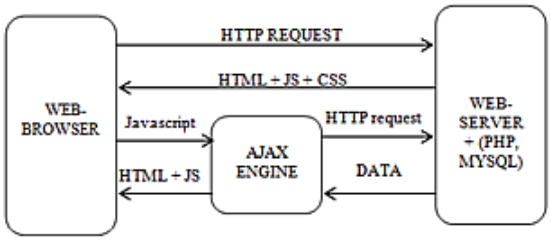


Figure 2. SABES system architecture

The database has two functions in the SABES system, the administrative function of assigning roles to the users and the data storage function. The Entity-Relationship (ER) diagram of the database is presented in Fig 3. Authentication of users is done based on the data from table User. It consists of the username, password and the

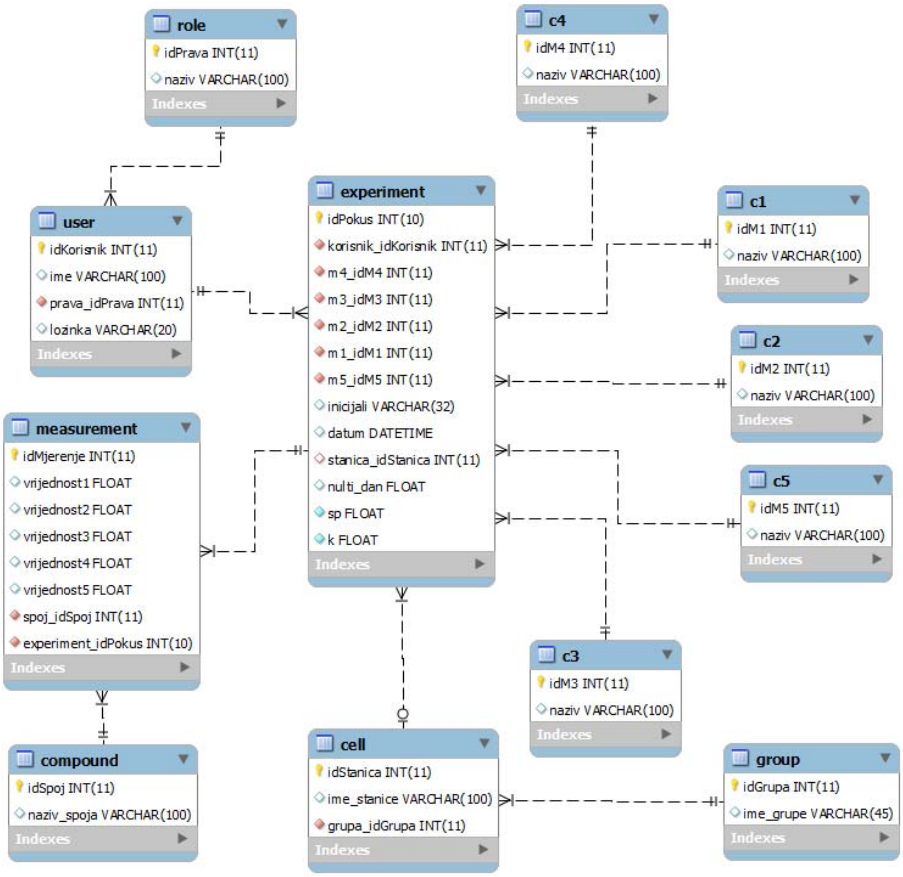


Figure 3. SABES system database

role assigned to each user by the administrators. The Role table divides the users into two groups: experimenter and researcher, as presented in Fig 1.

Experimenters are the users that conduct experiments, import the data into the system and can view and edit only their own data. Researchers have the administrative role and they can import, view and edit any data stored within the system.

The data import is done through the *Experiment table* obtained at the output of the device for cellular growth measurement, so it has the major influence on database structure. General pieces of information are stored within the table named Experiment, the most important table that consists of information about the experimenter, date of experiment, target cell (and indirectly about the group that conducted the experiment through table Group), the day zero, the blind test and the control value for an experiment. The measurement values obtained from the *experiment table* are stored within table Measurement. One *experiment table* can contain measurements for up to 4 compounds on 1 target cell. Each compound, stored within table Compound, has its own row in table Measurement. Each compound can be applied in up to 5 different concentrations. Thus, each compound concentration is stored in adequate tables (c1, c2, c3, c4, c5). Since the compound concentrations usually have recurring values, we avoided duplicated entries by separating them into such, additional tables. The mere form of *experiment table* does not allow the entry of more than 4 measurements and that limitation was not needed in the database design. The integrity of the data obtained for each research project (named *Group* in Fig 1) is achieved with table Group. One group can analyze more than one target cells. If more groups decide to analyze the same target cell, they should name it differently. In that way, no unintentional overlap of research between different groups will occur. Furthermore, the presented ER diagram supports both demands given by researchers-users:

- View the experiment results for a single Cell from a defined Group
- View the experiment results for all Cells tested with one Compound

C. Data Import

The user interface contains the data import form, presented in Fig 4. The *experiment table* obtained as output of an experiment is structured in a standardized xls format, which allowed us to create an automatic data import module. It is designed to make the automatic data import process less prone to human error than the previously accustomed manual process. Thus, it guides the user through the process by several steps. It is also designed to be easily understandable and usable by experimenters that are not experts in this field. The only restriction is not to change the default structure of the *experiment table*.

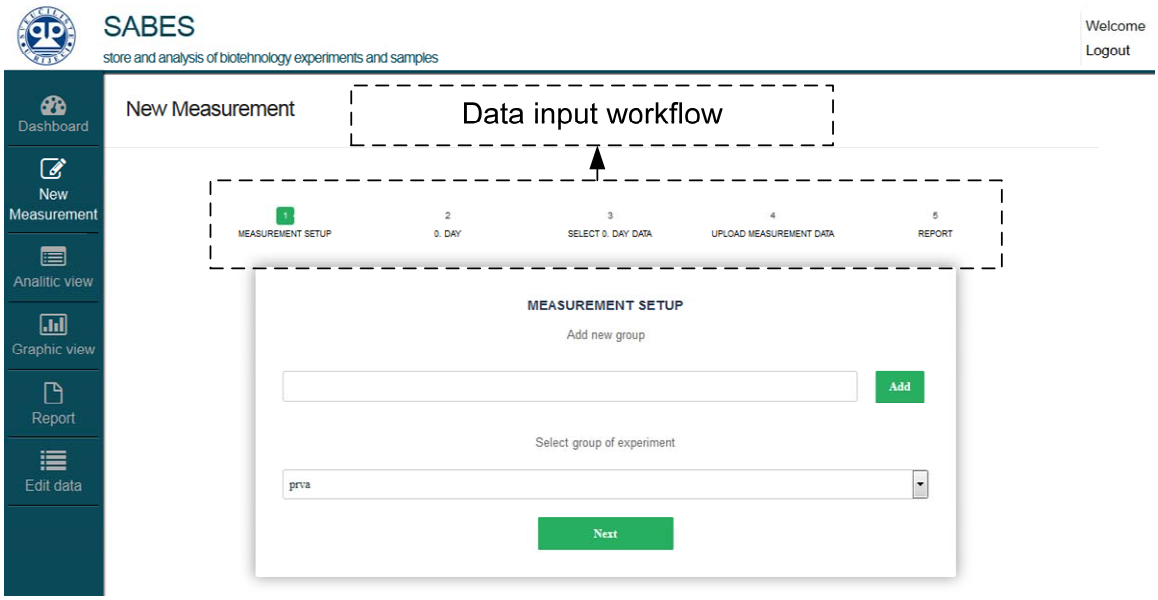


Figure 4. SABES system

The data import process initiates with the selection of the *Group*. The system offers the user to select from all the previously entered Group names. If the Group is a new one, the user is given the option to enter its name for the first time. The second step is to browse the location of the two *experiment tables* in xls format that are stored locally on the computer. The first *experiment table* that the user imports should contain the 0^{th} *day* values. The *experiment table* is then sent to server for processing. The relevant data are then structured and returned to the user for examination. The user then needs to select the column that represents the 0^{th} *day* values for the cell that is currently under examination. The second *experiment table* for import is the one that contains the *measurement* values. All the data are then stored in the database and the user is presented with the data import status. If there was an error in data import process, the user is informed that he should repeat the process. The complete data import sequence diagram is presented in Fig 5. Its steps are:

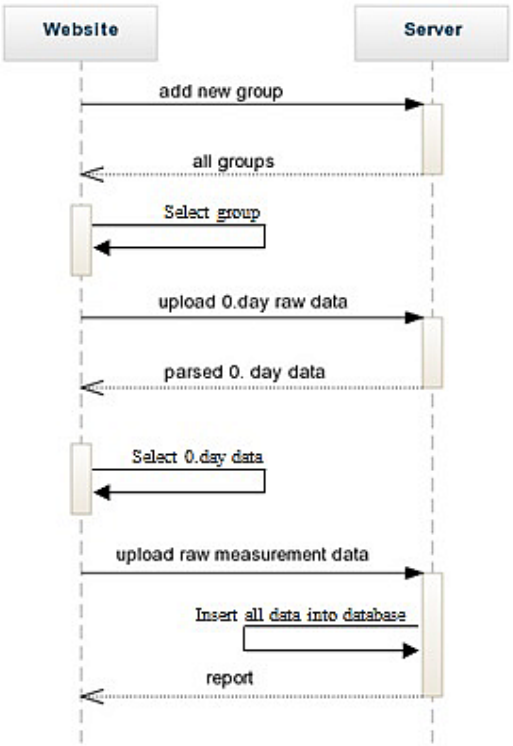


Figure 5. Data import sequence diagram

1. Selecting an existing *Group* or entering the name of a new one
2. Browse the *experiment table* with 0^{th} *day* values
3. Select the column of the *cell* under examination
4. Browse the *experiment table* with *measurement* values
5. Report the user of the data import success

D. Data Analysis

The SABES system automatically performs data analysis of imported data. Measures that are important for data analysis are maximum, minumum, mean and standard deviation values for each *Compound-Cell-Concentration* (*CCC*) combination and the IC_{50} value. The *CCC* combination represents a set of values that analyze the impact of the same *Concentration* of a *Compound* on a certain *Cell*. One *experiment table* contains four values for one *CCC* combination and their mean value and standard deviation for further analysis and visualisation must be calculated. The IC_{50} value represents the most important calculation within the experiment. It represents the *half-maximal inhibitory concentration*, *i.e.* the level of *Compound Concentration* that achieves 50% inhibition of the maximum cellular growth for the analyzed *Cell*.

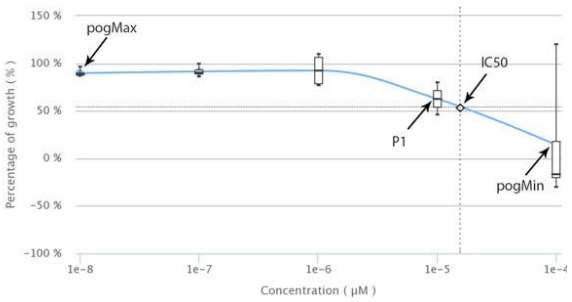


Figure 6. Određivanje intervala za izračun IC_{50}

The estimation of the IC_{50} value is done with linear approximation between two neighbouring levels of concentration that inhibit the cellular growth above and beyond 50%, as presented in Fig 6. The maximum percentage of cellular growth for one *Compound-Cell* combination is presented with the *pogMax* value and the mimum percentage is presented with the *pogMin* value. The 50% cellular growth (pog50) is then obtained by formula (1):

$$pog50 = pogMax - 0.5 \cdot (pogMax - pogMin) \quad (1)$$

The pog50 value is then used to find the two neighbouring values of concentrations that are used in the experiment and stored in the database using the general expression (2):

$$\dots pog_{n-1} < pog_n < pog_{50} < pog_{n+1} < pog_{n+2} \dots \quad (2)$$

The values of pog_n and pog_{n+1} are the two neighbouring values we are looking for. For the example given in Fig 6, the pogMin value is the pog_{n+1} value. The IC_{50} value is then calculated using a standard linear equation between two points given by Equation (3):

$$y = y_a + (x - x_a)\frac{y_b - y_a}{x_b - x_a} \quad (3)$$

where y_a is the pog_n value, y_b is the pog_{n+1} value, x_a is the *Compound concentration* for pog_n value and x_b is the *Compound concentration* for pog_{n+1} value. IC_{50} value is equal to x, when pog50 is inserted as y in (3).

E. Data Selection and Visualisation

The SABES system also offers the graphic view after analyzing the imported data. There are two possible ways to select the data for analysis:

6. Filter the data by *Group* and by *Cell*
7. Filter the data by *Compound*

The graphical user interface of the graphic view window is presented in Fig 7. Filtering the data by *Group* and by *Cell* can be done using the filter elements above the cellular growth graph and filtering the data by *Compound* can be done using the filter elements of the left side of the graph. Sequence diagram of filtering the data by *Group* and by *Cell* is presented in Fig 8. The *Group*

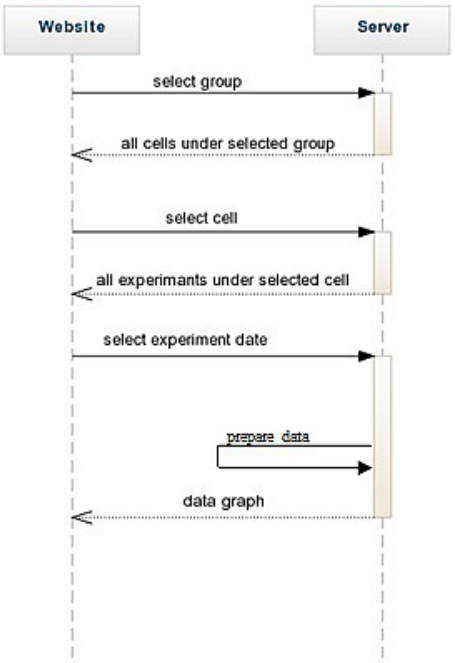


Figure 7. Data selection sequence diagram - filter by *Group* and *Cell*

filter offers a drop-down menu that is automatically filled with all the *Group* names within the database. The server then returns only the names of *Cells* that were analyzed by the chosen *Group*. After selecting the name of *Cell*, the server gives a list of all the experiments that were conducted for selected *Group* and *Cell*. The user then selects the experiment he wants to analyze and the graph is plotted in the cellular growth graph area, as presented in Fig 6. Each graph generated with the SABES system exhibits the effect of four different *Compounds* used in several levels of concentration with a target *Cell*. The x-axis of the graph represents the levels of concentration and the y-axis of the graph represents the percentage of cellular growth. Filtering the data by *Compound* is simpler and hence we omit its sequence diagram. The data filtering starts with the selection of the *Compound* and then the server returns the names of all *Cells* that were analyzed with it in *Cell* filter. The number of *Cells* that will be plotted on the same graph is arbitrary and left to user's decision.

There are two curve plotting styles implemented in the SABES system. The first one is based on plotting a continuous "*spline*" curve that is calculated based on mean values for a certain *CCC* combination. This is the standard presentation style of presented *in vitro* experiments. The other style is based on drawing of a box & whisker plot. The box represents the inter-quartile range, the whiskers represent the complete range of data, and the middle line presents the median. As we mentioned earlier, each experiment contains at least four measurements for the same *CCC* combination and these values are the basis for the box & whisker plot. If the data are filtered by *Compound*, then we can have a large number of experiments that have the same *CCC* combinations and the box & whisker plot is done for all of them. This style is optional and it can be omitted from the graph.

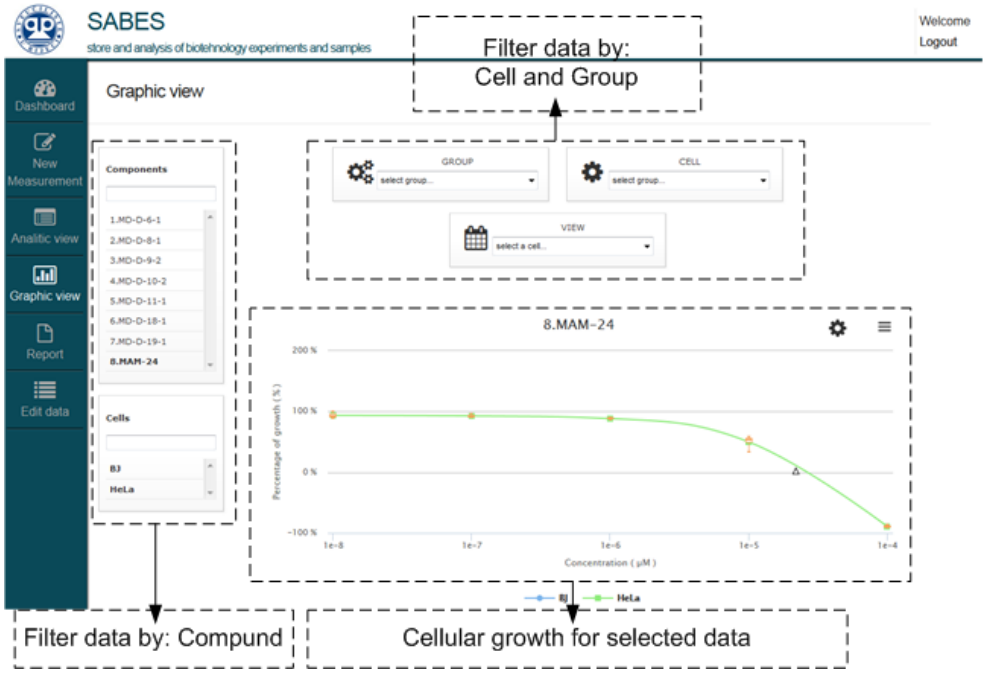


Figure 8. Graphic view for filtering and plotting the experiment data

III. CONLUSION AND FUTURE WORK

One of the biggest obstacles in advancements of any scientific field is limited data storage and analysis capacities. Although, there are vast technologies ready for data generation that are very well supported with powerful and relatively cheap processing equipment for data analysis, there is still a large area of intervention in smaller, specialized domains, *i.e.* biological experimental research where automatization and optimisation of processes might be designed in line with customers/users' needs. Importantly, scientists should understand deeply the phenomena under investigation on large number of data samples to generate generalising theories. And here the key obstacles remain data analysis, data presentation and systematization. This gap might be filled through interdisciplinary collaboration among researches from natural science and computer science. Their interaction indeed, may be of vital importance for future evolution in both disciplines. This paper presents an endeavour in that direction. Here we present a simple IT based solution to error prone process of manual data collection, data aggregation and data analysis. We focused on issues how to simplify the data storage, data aggregation and data analysis process and remove as much as possible human bias from datasets in analysis by taking care of an easy to use approach through the IT system called SABES system that is jointly developed by students and researchers from the University of Rijeka.

One of the major concerns of data management systems today is in their ability to deal with large data sets. Structured data collection process simplifies data accumulation. Data aggregation and analysis rules and policies that may be developed per each individual research experiment may be fault prone and computational intensive task. Future work should consider development of new technologies and programming abstractions that would be able to cope with these issues. We started to create a data management platform for scientific research purposes and future work will be on future evolution of SABES system by integrating more data structures, developing an environment for collaborative and interdisciplinary research as well as extending its abilities to support big data.

ACKNOWLEDGMENT

The work presented in this paper is supported by the University of Rijeka, Croatia research grants (numbers 13.09.2.2.16 and 511-21).

REFERENCES

[1] T. Gazivoda Kraljević, N. Ilić, V. Stepanić, L. Sappe, J. Petranović, S. Kraljević Pavelić, and S. Raić-Malić, "Synthesis and in vitro antiproliferative evaluation of novel N-alkylated 6-isobutyl- and propyl pyrimidine derivatives", Bioorganic & Medicinal Chemistry Letters, Vol. 24, No. 13, 2014, pp. 2913-2917

[2] E. G. Stephan, K. R. Klicker, M. Singhal and H. J. Sofia, "Problem Solving Environment Approach to Integrating Diverse Biological Data Sources", *in* 'CSB Workshops', IEEE Computer Society, 2005, pp. 47-50.

[3] C. S. Ang, "Data Management and Analysis Architecture for a More Efficient and Productive Bioinformatics Environment", *Proceedings of HICSS '05.*, 2005 pp.279 - 279.

[4] SEIP Lab: http://www.seiplab.riteh.uniri.hr/

[5] G. Mauša, T. Galinac Grbac and B. Dalbelo Bašić, "Software defect prediction with bug-code analyzer - a data collection tool demo", In: Proceedings of SoftCOM '14, Split, Croatia, 2014

[6] A. Grguric, G. Brestovac, D. Marin, T. Oroz, A. Vidović, S. Bozóki, M. Mošmondor, T. Galinac Grbac, "Ambient orchestration in assisted environment", Engineering Review, May 30, 2014

[7] J. Li and C. Peng, "jQuery-based Ajax general interactive architecture," In Proceedings of Software Engineering and Service Science (ICSESS), 2012, pp.304 - 306

CLOUD SYSTEMS

Automatizacija postavki mreže u *Cloud* sustavima preko samouslužnih portala

Mladen Maras*, Tomislav Kosanović**

* Combis d.o.o./Sektor razvoja i integracije aplikativnih rješenja, 10000 Zagreb, Hrvatska

** Hrvatski Telekom d.d. / Odjel za dizajn i razvoj usluga, 10110 Zagreb

Tomislav.Kosanovic@t.ht.hr, Mladen.Maras@combis.hr

***Sažetak* - Korištenje javnih usluga u oblaku (eng. cloud) od strane masovnih pružatelja usluga za smještaj poslovnih podataka i sustava još je uvijek prilično kontroverzna tema. Razlog tome su neutemeljeni strahovi u pogledu sigurnosti, nedostatak povjerenja u pružatelje usluga, njihovu mogućnost brzih promjena na sustavima, fleksibilnost i na kraju cijena.**

Usluge u oblaku vrlo su ovisne o mrežnim elementima u rješenju te se obično sastoje od virtualiziranih otoka u klasičnom mrežnom okruženju. Iako SDN (eng. Software Defined Network) još nije došao u pravom smislu do okruženja u oblacima, te ne postoji javno dostupna industrijska standardizacija, trenutno dostupne tehnologije omogućavaju dizajn i izgradnju dovoljno pouzdanih, fleksibilnih, a opet robusnih sustava za automatizaciju i mreže i korisničku kontrolu mrežnih postavki.

***Ključne riječi - cloud*, mreža, virtualizacija, automatizacija, portali, korisničko iskustvo**

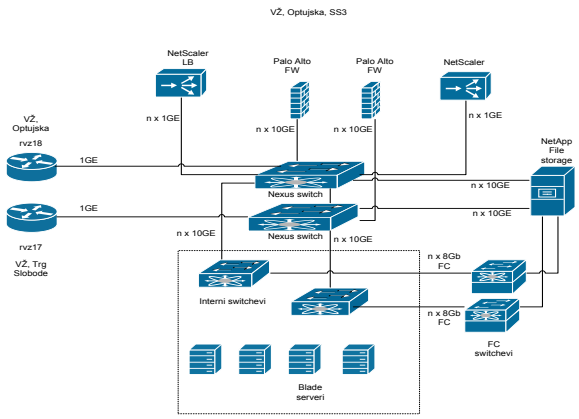
I. UVOD

Prema standardu NIST (National Institute for Standards and Technology), da bi se rješenje moglo smatrati *javnom uslugom u oblaku* (u daljnjem tekstu *Cloud*), ono nužno mora podržavati sljedećih 5 karakteristika:

- Samouslužnost,
- Širokopojasnu mrežu,
- Lokacijsku neovisnost,
- Elastične resurse,
- Mjerljive usluge.

Podatkovni centri i infrastruktura koja podržava *Cloud* sa sistemske strane prilično je kompleksna, kao što se može vidjeti na slici 1 (primjer za uslugu Virtualno Računalo u HT Cloud-u).

Mreže operatora i pružatelja *Cloud* usluga sastoje se od velikog broja specifičnih uređaja. Da bi se pokrenula nova usluga često je izazov pronalaženje prostora za prihvat ovih uređaja, održavanje postaje sve teže, velika složenost takvih okruženja, povećanje troškova energije, kapitalnih ulaganja, nalaženje vještina potrebnih za dizajn na tržištu rada zatim integracija i upravljanje svime time postaju gotovo nemogući. Štoviše, rješenja bazirana na vlastitim fizičkim resursima vrlo brzo dolaze do kraja isplativosti održavanja te zahtijevaju puno od osoba koje ih implementiraju. Što je još gore, životni vijek usluga postaje sve kraći, kako se tehnologija i inovacije ubrzavaju. Virtualizacija funkcije mreže ima za cilj rješavanje tih problema iskorištavanjem standardne IT tehnologije virtualizacije za konsolidaciju mnogo vrsta mrežne opreme. Ovaj rad se temelji na projektu koji je trebao dokazati da je virtualizacija mreže primjenjiva za bilo koju funkciju unutar fiksne i mobilne mrežne infrastrukture.



Slika 1. Arhitektura rješenja

Cilj

Cilj je uspostaviti potpunu automatizaciju mreže te omogućiti korisniku potpunu kontrolu nad mrežnim postavkama koje se odnose na korisnikove *Cloud* sustave a koji uključuju virtualne poslužitelje, virtualna računala, kolokacijske poslužitelje i lokalna korisnička računala spojena brzom vezom (*ClearPath* platforma).

Mrežni pristup trebao bi biti transparentan te bi korisnik morao moći mijenjati način mrežnog pristupa neovisno o *Cloud* usluzi koju koristi. Mijenjanje mrežnih postavki mora unutar *Cloud*-a biti moguće preko samouslužnih portala:

- Uspostava automatske aktivacije mrežnih elemenata:
  - *Nexus* komutatora na koje je spojen javni oblak,
  - *Palo Alto* vatrozid,
- Uspostava automatske aktivacije *ClearPath* platforme radi njezine integracije s uslugom virtualnih poslužitelja i kolokacijskom uslugom.

- Uspostava korisničkih portala koji će omogućiti korisniku upravljanje mrežnim resursima unutar njegovog zakupljenog prostora na javnom oblaku:
    - vatrozid između korisnikovih mreža,
    - balansiranje opterećenja korisnikovih usluga,
    - kontrola adresne translacije prema javnoj mreži za sve korisnikove privatne adrese,
    - uspostavu dinamičke translacije za korisnikove resurse unutar oblaka prema Internetu,
- Sve aktivacije unutar *Cloud* platforme moraju se izvršiti orkestracijom.

Ključ pojednostavljenju usluga, a time i smanjenju operativnih troškova i mogućnosti povećanja usluga je u apstrakciji usluga odnosno apstrakciji kompleksnosti usluga, aktiviranja i orkestracije.

Primjerice, ako izaberemo virtualni poslužitelj kojem povećavamo resurse (primjerice procesorsku snagu), na klasičan način prvo ćemo izabrati akciju, potom s liste virtualni poslužitelj na kojem vršimo promjenu. Pristup koji smo izabrali i koji se pokazao vrlo uspješnim jest da okrenemo uloge i da prvo izaberemo subjekt, a potom na subjektu vidimo listu akcija koje je moguće provesti nad odabranim subjektom u tom trenutku, te izaberemo odgovarajuću akciju. Dakle, subjekt – predikat pristup.

U konstelaciji samouslužnih portala i cilja da portali moraju pružiti maksimum informacija i pristup do željene informacije/produkta s minimalnim brojem korisničkih odabira dizajnirali smo samouslužne portale na bazi *dodataka* (u daljnjem tekstu *widget*).

Widgeti se dijele na:

- Informativne - bazirane na *live-tile* ili graf funkcionalnostima,
- Entitete – odnosno subjekte.

Informativni *widgeti* prikazuju informacije koje administratoru ili korisniku sustava (osobi koja kupuje, bilo tvrtki ili pojedincu) pruža informacije koje najčešće imaju potrebu pratiti.

Izgled ekrana (kontrolne ploče) administratora VPS portala sa svim relevantnim podacima, prikazan je na slici 2.

Entiteti su revolucionirali portale, grafička sučelja i korisničko iskustvo te način na koji su usluge prezentirane krajnjem korisniku. Ta apstrakcija obuhvaća razdvajanje aplikacija i podataka od fizičkih ograničenja, dok istovremeno pruža mogućnost upravljanja i zaštite.

U nastavku ćemo obraditi pristup *Cloud* uslugama kroz apstrakciju aplikativnim entitetima, te apstrakciju kompleksnosti mrežnih usluga kroz kontekst SDN-a, odnosno tehničku izvedbu koja treba simulirati SDN.

II. HT *CLOUD* SAMOUSLUŽNI PORTALI

HT *Cloud* infrastrukturni portali sastoje se od:

- VPS – Virtualnih privatnih poslužitelja (eng. Virtual Private Server) modula,
- VDi – Virtualnih radnih površina (eng. Virtual desktop) modula,
- Hex – Hosted exchange modula,
- SSO – Single Sign On modula.

Cloud samouslužni portali prošli su nekoliko stupnjeva evolucije da bi konačno došli do vrhunskog izgleda i stanja. Dizajnirani su za najbolje moguće korisničko iskustvo (eng. user-experience) i funkcionalnost.

Započeli smo dizajniranjem *klasičnih* portala – s listama, stranicama, velikim zaglavljem i podnožjem, listama i stranicama za unos korisničkih podataka, stranicama za prikaz izvještaja i slično.

Kada smo analizirali web stranice na osnovu predikata i subjekta, došli smo do toga da klasične web stranice funkcioniraju na osnovu izbora akcije (predikata) iz liste, te potom izbora subjekta na kojem će se izvršiti odabrana akcija. U HT *Cloud* portalima smo te postavke okrenuli.

Entiteti su *widgeti* koji predstavljaju apstrakciju neke usluge. U kontekstu VPS-a, virtualni poslužitelji su entiteti i prikazani su na slikama 3 i 4.

Svaki entitet zna koje sve akcije se na njemu mogu provesti. Entitet ima sve potrebne parametre, informacije i naredbe koje ga definiraju i koje su neophodne za upravljanje funkcionalnošću. Akcije postavljaju entitet u promijenjeno stanje u nekom vremenu i okruženju.

Možemo ići tako daleko i reći da su portali u ovom kontekstu samo spremnici za entitete i služe im za stvaranje novih entiteta i pokretanje drugih naredbi koje su iznad entiteta.

Tako dizajnirani portali, *widgeti* i entiteti su inherentno pripremljeni za responzivan web – da se prezentiraju na tabletima ili pametnim telefonima.

Svaka akcija koje ne može biti direktno izvršena klikom na entitet podiže *lightbox* (JS modal dialog), koji je također responzivan.

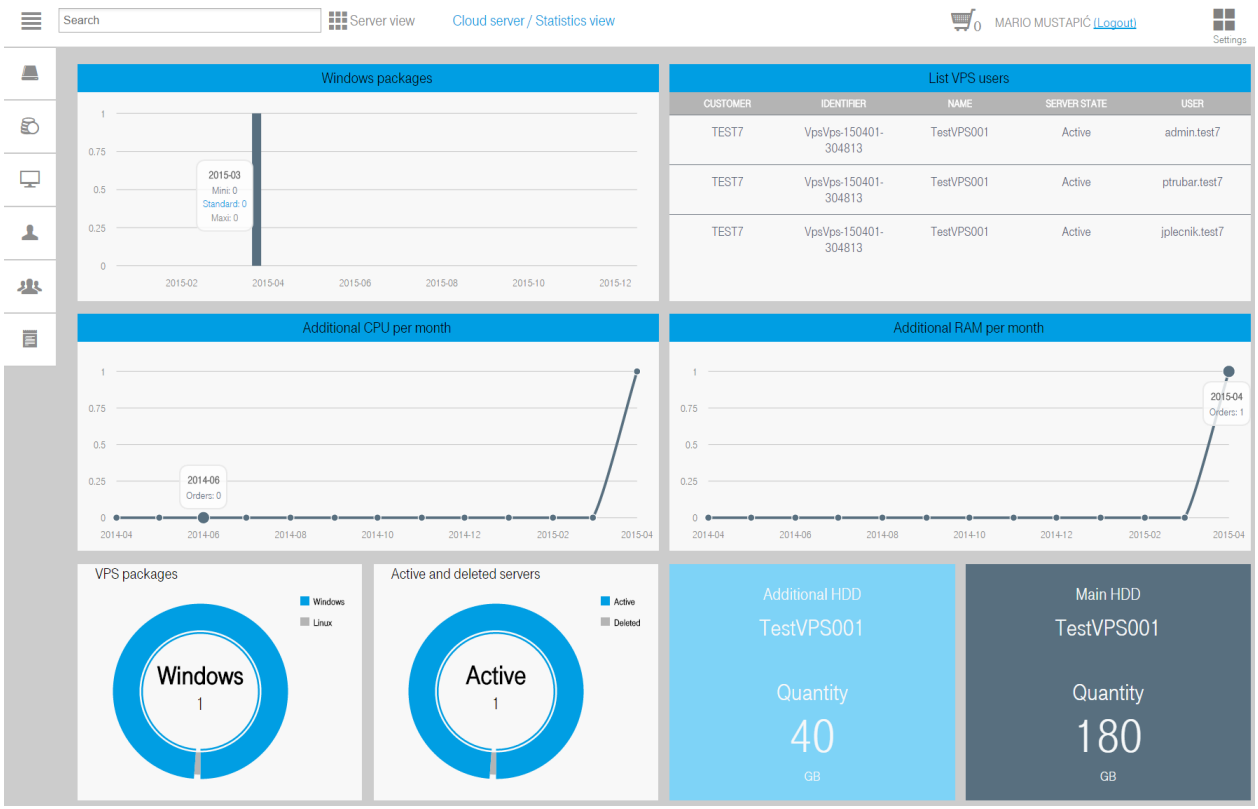
III. IMPLEMENTIRANO RJEŠENJE

Korisnik se spaja isključivo na Portale putem kojih definira i izvršava postavke na mrežnom nivou za sustave u svom vlasništvu.

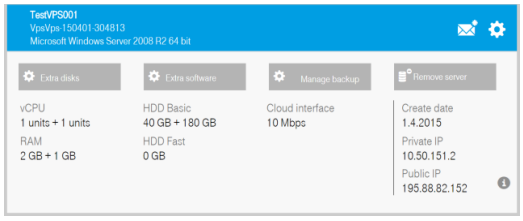
Portali korisničke akcije izvršavaju isključivo pozivanjem metoda na orkestratoru koji zatim iste izvršava na infrastrukturi. Sve postavke se spremaju u zasebnu bazu podataka kojoj pristupa *NetProv* portal.

Za potrebe ove usluge dodatno se konfigurira sljedeće:

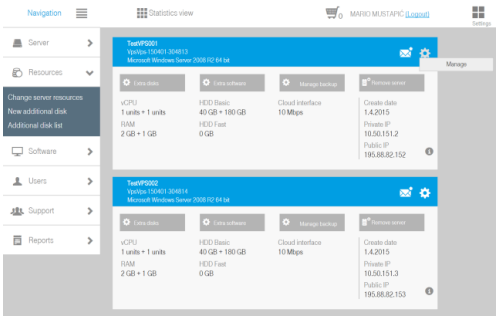
- Na PE uređajima rvz17 i rvz18 dodaje se nova logička veza i podmreža u VRF-u *CLOUD*_SME (podmreža ne smije biti iz 10.0.0.0/8). Preporuka je bar /28 podmreža kako bi ostalo dovoljno IP adresa za HSRP između rvz17 i rvz18 te IP adrese koje će služiti za NAT/PAT potrebe na PAN VATROZID.



Slika 2 – Kontrolna ploča



Slika 2: Entiteti (Widget-i)



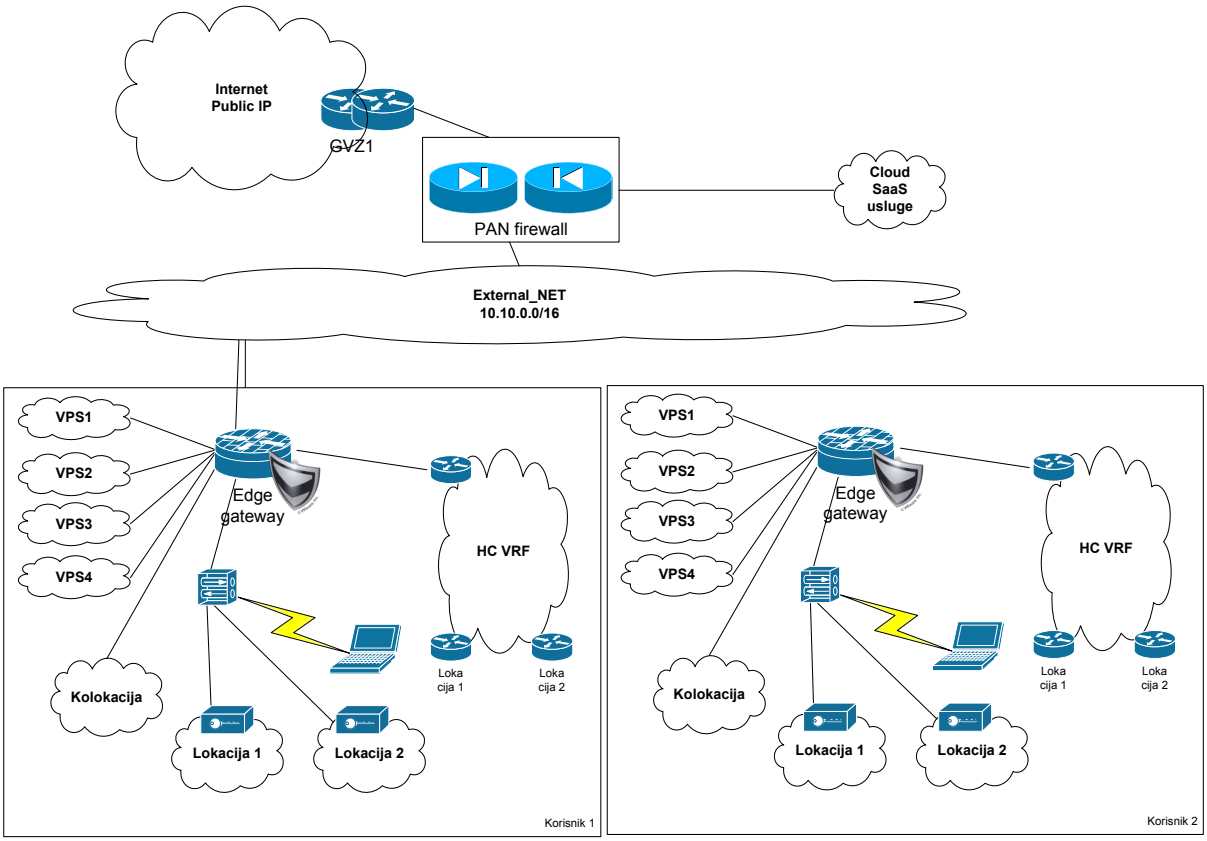
Slika 3: Entiteti opcije

- VLAN se konfigurira na *Nexus* komutatoru te pušta prema konektorima na koje je spojen PAN VATROZID
- Na PAN VATROZID-u se konfigurira novi virtualni usmjeritelj radi odjeljivanja *Cloud* pristupa od postojećeg Internet pristupa i nova IP adresa na podmreži prema PE.
- Na PAN VATROZID-u se dodaje NAT konfiguracija prema *Netscaler*-u (koristi se postojeća mreža PAN – Netscaler). IP adresa za pristup u VDI se upisuje na DNS poslužitelj.
- Na PAN VATROZID-u i *Netscaler*-u se dodaju odgovarajuća usmjerenja prema korisničkim podmrežama (radi povratnog prometa).
- Na PAN VATROZID-u se dodatno konfigurira NAT od strane virtualnih radnih površina prema korisničkim podmrežama. Za potrebe NAT-a na PAN VATROZID-u se stoga koristi jedna ili više adresa iz pristupnog podmreža. Virtualne radne površine mogu komunicirati s lokalnim poslužiteljem kod korisnika pri čemu vezu uvijek pokreće virtualna radna površina.
- Na PAN VATROZID-u treba postaviti odgovarajuće per AD user ACL koje pojedinom korisniku dopuštaju pristup samo njegovom LAN segmentu.
- Na PAN VATROZID treba podesiti *rate-limit in/out* prometa s *Cloud* infrastrukture prema korisničkim lokacijama (ograničenje je suma brzina svih lokacija)
- Za kolokacijsku mrežnu aktivaciju dodatno je potrebno na PAN-u konfigurirati ACL koji dopušta nekom korisniku *Cloud* računala pristup samo do njegovog kolokacijskog podmreža.

- Za pristup *Cloud* računala do virtualnih poslužitelja na PAN-u treba podesiti NAT prema podmrežama virtualnih poslužitelja, usmjeravanje prema podmrežama virtualnih poslužitelja i odgovarajući ACL.

Ciljano mrežno sučelje je prikazano na slici 5

Samouslužni portali koji entitetima podržavaju implementirano upravljanje ciljanim mrežnim sustavom prikazani su na slici 6.



Slika 5 Ciljano mrežno sučelje

Slika 6: Korisnički portali

IV. ZAKLJUČAK

Virtualizacija mrežnih funkcija već se događa pa za nekoliko godina možemo očekivati da će telekomunikacijska industrija izgledati vrlo slično IT industriji. Tu će biti široki spektar poslovnih modela koji više odgovaraju modelima softverske industrije. Složenost okruženja na apstraktnom nivou će biti daleko manja, a nivo samostalne kontrole će biti drastično veći. Kao što je navedeno u ovom dokumentu, virtualizacija mrežnih funkcija će donijeti mnoge prednosti za mrežne operatore te njihove partnere i klijente, istovremeno pružajući priliku za stvaranje novih vrsta eko-sustava (uz tradicionalne modele nabave na temelju povlaštenih cijena sa strateškim partnerima) koji će poticati i podržavati brze inovacija sa smanjenim troškovima i smanjenim rizicima. U ovome radu opisanim pristupom postigla se razina željene konfiguracije i apstrakcije virtualiziranih elemenata uz podržavanje svih traženih karakteristika i marketinških zahtjeva.

REFERENCE

[1] M. Goršeta, „HT „Cloud platforma“, HT Interna dokumentacija, veljača 2014.

[2] S. Bailey, D. Bansal, L. Dunbar, D. Hood, Z. L. Kis, and other, SDN Architecture Overview, https://www.opennetworking.org, prosinac 2013.

[3] M. Chiosi, D. Clarke, P. Willis, J. Feger and other, Network Functions Virtualisation, http://portal.etsi.org/NFV/NFV_White_Paper.pdf, listopad 2012.

Data Storage and Synchronization in Private Cloud

D. Hodanić, N.Vrkić and M. Tomić
University of Rijeka, Faculty of Engineering, Department of Computer Engineering, Rijeka, Croatia
{dhodanic, nvrkic, mtomic}@riteh.hr

Abstract - Usage of cloud systems for data storage has many advantages over the traditional approaches. It is already widely used and its popularity is still fast-growing. The systems must be implemented and maintained in a way that not only satisfies the performance and resource availability requirements, but also fully addresses the questions of security, privacy and data ownership. However, concerns related to those questions very often lead to considerations of a private cloud implementation.

In this paper, we explore a private cloud implementation suitable for small to medium businesses. We introduce main types of cloud computing as basic service models and analyze private cloud systems features. Advantages and disadvantages in comparison to public cloud services are considered. Implementation of private cloud solutions in a lab environment allowed us to examine the ease of the setup and maintenance as well as the usability of the chosen solutions and their applicability for the target user group.

I. INTRODUCTION

Cloud technologies have infiltrated society in many aspects. Our culture is currently built around mobile devices, social media, cloud platforms, intelligent systems and many more. There is a lot of personal and business information that has to be stored. Cloud storage and synchronization are highly popular for personal use. In contrast, organizations are still cautious with moving their businesses into the cloud. Privacy and security are the main concerns when considering cloud solutions. Most of the popular cloud services offer public cloud service model. Public cloud platforms opened the cloud concept to personal users. There is no doubt that these systems earned their popularity, but for some user groups more private solutions are potentially a more viable and acceptable option.

Although cloud platforms extended possibilities in terms of storage and synchronization for organizations, they also produced some uncertainties. With enterprises and cloud services, trust issues were bound to come up. Trust is depicted as a complex factor formed of control over assets, data ownership, failure prevention and security. Khan et al. showed in [1] that building trust in cloud services comes with time, good performances, security, more client control and transparency. Therefore, private clouds found their way into many organizations and application areas where trust is of great concern. For instance, Doelitzscher et al. in [2] presented a case for building a private cloud for collaboration and e-learning services in HFU University. Architecture and software aspects were analyzed in detail, with conclusions about private cloud benefits at the university and its IT department. A more general comparison of several cloud computing platforms has been given in [3] to provide better understanding and useful details for choosing the best cloud solution given a specific case. The authors focused on private and hybrid clouds, namely OpenNebula and Nimbus and considered OpenNebula solution to be suitable for building a highly scalable cloud computing environment. In [4], the Eucalyptus was considered in details. Eucalyptus is an open source private and hybrid cloud software representative, suitable for the use in enterprises. It is compatible with Amazon Web Services (AWS) cloud interface, allowing interaction between a private and a public cloud. Atefi et al. presented in [5] the use of open source private cloud for digital forensics. Seafile was chosen due to its advanced features for preserving privacy, file syncing, and collaboration. In a prototype research from Mościcki et al. [6], a private cloud solution for internal use by the CERN employees has been analyzed in depth, as an alternative to the public cloud services. The authors chose ownCloud as a viable platform for an open source private cloud solution for file sharing and synchronization on a very large scale. OwnCloud competitors, such as Pydio, Seafile, SparkleShare and Syncany were discarded because of feature incompleteness, usability issues or failing to support necessary file formats. They found that only ownCloud satisfied most of their demands for building an open source cloud solution which can compare well to commercial solutions.

Private cloud systems are highly adjustable to user's needs and offer great flexibility in terms of the number of active users and feature implementations. There are several solutions, but also many different concepts for custom usage within organizations. In this paper, we explore how private cloud solutions may benefit small to medium businesses (SMB). A deployment of an open source private cloud system with software solutions such as ownCloud, provides business users with internal control of privacy and sole data ownership. We also consider questions of security, scalability and cost.

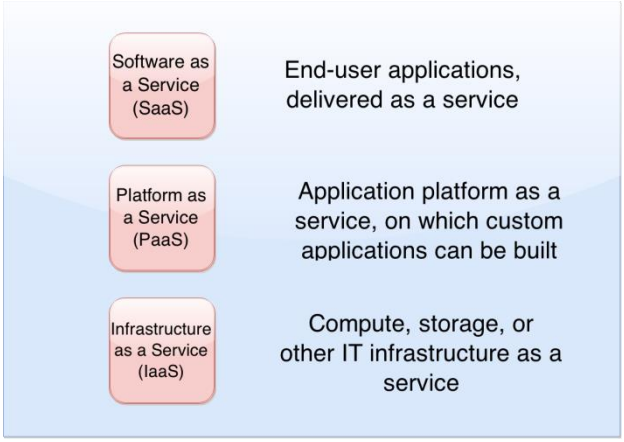


Figure 1. Basic cloud computing service models

II. CLOUD COMPUTING CONCEPTS AND CLOUD STORAGE MODELS

A. *Definition of the cloud*

There is more than one definition of cloud computing, but most of them have a similar meaning – using computer resources on demand. This is a model for delivering information technology services where resources are retrieved from the Internet using web-based tools. Applications can communicate with servers where all data and software packages are stored. Cloud computing model allows access to information as long as a client's device has access to the web. Companies may find that the cloud computing allows them to reduce the cost of information management, since they are not required to use their own infrastructure, but instead lease it from service providers [7].

B. *Cloud service models*

There are three basic cloud computing service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS), as shown in Fig. 1.

Infrastructure as a Service (IaaS) is a service model where cloud computing service providers share their storage, network, and other computing resources with their clients. Clients are able to run their own software. A user does not operate or control fundamental physical infrastructure of a cloud but has control over operating systems, storage, deployed applications, and possibly limited control of selected networking components. Service providers give their users access to multiple virtual server instances, so different users access the same infrastructure at the same time. They have flexible access, meaning they can reach the infrastructure from any location and any device that has a network connection. With regards to security problems, many IaaS providers offer a private cloud environment for cloud hosting [8].

Platform as a Service (PaaS) is a service model similar to IaaS but it provides an additional level of functionality. It is a model for running applications without the need to maintain the hardware and the software infrastructure at organizations. PaaS service providers allow customers to deploy, run or develop applications on their cloud infrastructure. A user does not manage nor control the underlying cloud infrastructure (including network, servers, operating systems), nor storage, but has control over the deployed applications and possibly over configuration settings for the application-hosting environment. PaaS offers software developers several advantages including cost savings, reduced technical maintenance and increased mobility. Costs for purchasing servers, other hardware or required software licenses are reduced or eliminated [8] [9].

Software as a Service (SaaS) cloud model operates on a virtualized and a pay-per-use billing model. In this model, software applications are rented to contracted customers by SaaS service providers. The applications are accessible from various client devices through a client interface, such as a web browser. The client does not operate nor control the fundamental cloud infrastructure including servers, storage, network, operating systems, nor individual application features, with the possible exception of application configuration settings, as specified by the service provider. Applications on providers cloud infrastructure are always up-to-date, pre-installed and configured. This reduces the time spent for installation and configuration of software on local machines. It also allows SMBs to use a software that would otherwise be too costly due to a high licensing fees. Maintenance costs are reduced as well, as the SaaS provider owns the environment and the cost is split among all customers using that solution [8] [9].

C. *Cloud deployment models*

A public cloud is a model that provides users access to the cloud over the Internet, using client interfaces, such as common web browsers. It is typically based on a pay-per-use model. This helps cloud clients to better match their IT expenditure at an operational level by decreasing its capital expenditure on IT infrastructure. Public clouds are less secure than other cloud models because they place an additional burden of ensuring all applications and data accessed on the public cloud are protected from malicious attacks [8]. Security is a very challenging problem in the public cloud model. Service providers are responsible not only for storing data, but also for securing it. Moreover, many organizations and private users are sharing the same resources [7], which means the provider must also implement isolation between different customers.

A private cloud is set up for organization's internal use. Service provider is no longer a third party associate, but the organization itself. There are several variations in a private cloud setup. Software components are always run privately. However, hardware modules may be owned by the organization, but also hosted elsewhere. A private cloud approach provides more control over the data ownership and privacy. However, it also implies the organization's own responsibility for setting up both the infrastructure and a software environment, as well as security measures implementation and, later on, the maintenance of each system component. Fig. 2

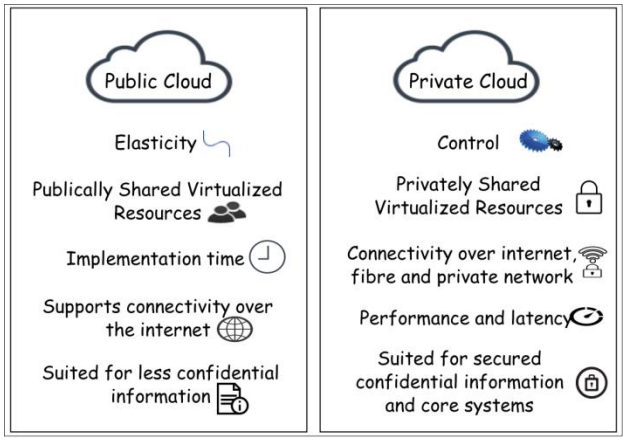


Figure 2. Differences between private and public cloud solutions

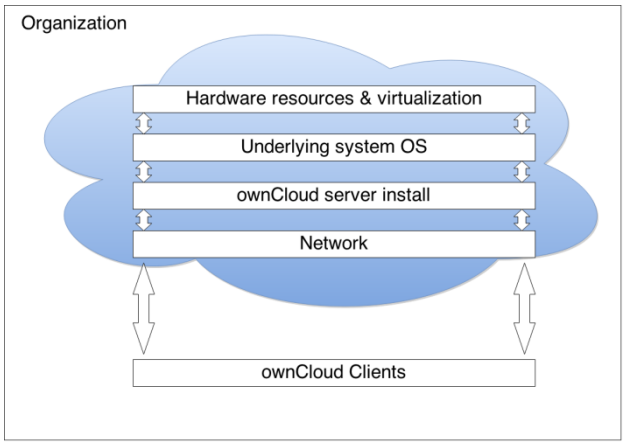


Figure 3. OwnCloud architecture components in an organization

summarizes advantages and disadvantages of private and public cloud solutions.

A hybrid cloud model combines benefits of both private and public cloud. It is composed of two or more different cloud infrastructures that are bound together [9].

III. PRIVATE CLOUD

A private cloud may be set up in-house or hosted with a third party service provider. Typically, a private cloud is owned by the organization itself. Organizations may benefit from a private cloud in many ways. When internal infrastructure is used, organization is the sole owner of its data and the only entity having access to it. This reduces or eliminates several potential public cloud service problems, such as sharing platform with other users, possible unauthorized access to the data, changes of providers privacy policy, unexpected service outages, data content restrictions, dependence on a third party organization, control over the equipment and flexibility in a form of cloud bursting.

On the other hand, disadvantages of the private cloud model also have to be taken into consideration. Since an organization is running the entire cloud platform, planning and preparation is essential. Initial investment in hardware and software sometimes proves to be a serious obstacle, although, using existing hardware resources with several equipment upgrades, implementation of virtualization techniques and open source software, may significantly reduce the initial cost. After the infrastructure is set up, an organization needs to deploy a private cloud system, implement security measures and enable seamless remote access. The cloud also needs to be maintained and monitored for security and performance. This requires skilled personnel whom not all organizations can afford.

Some authors suggest private clouds can obtain some of the public cloud benefits by merging into a hybrid cloud approach [10].

IV. IMPLEMENTATION, SECURITY AND FEATURES OF A PRIVATE CLOUD SOLUTION

A. *Demo environment*

In this section we explore setup and security aspects of a full-featured open source system for a private cloud implementation. Demo environment was built using virtualized servers on own infrastructure. As an underlying OS, a Linux server distribution was chosen. OwnCloud server was installed on top of the virtualized server. In a typical organization, the cloud would only be connected to the organization's internal network. To gain access to the cloud, remote clients would first need to access organization's LAN, usually by means of a VPN connection. Then, they would be able to access the cloud in the same manner as if they were physically present at the organization's premises. Fig. 3 illustrates components of a private cloud setup in an organization using ownCloud.

Lab setting allowed us to explore the system in a real everyday use-case scenario. Testing was done on a smaller scale environment, but interpretation of results is applicable to larger deployments. OwnCloud was chosen as a cloud solution, being the most popular open source solution for SMBs at the moment.

B. *OwnCloud implementation and security*

Currently, ownCloud supports Linux and Windows servers as a system base. Installation of server-side ownCloud is supported in different scenarios. In a lab environment, we used Linux packages as a preferred installation method. One of the most important system configuration steps is to implement basic security measures. Using the Secure sockets layer (SSL) and the Transport layer security (TLS) for protected access is highly recommended. SSL/TLS provide both data confidentiality and integrity for TCP/IP transmissions. Client-server communication using TLS protocol should be safe from many security attacks. End-to-end authentication and data confidentiality are provided by cryptographic algorithms in SSL/TLS. It provides a one way authentication, as the servers identity is already known to the client, whereas the client rests

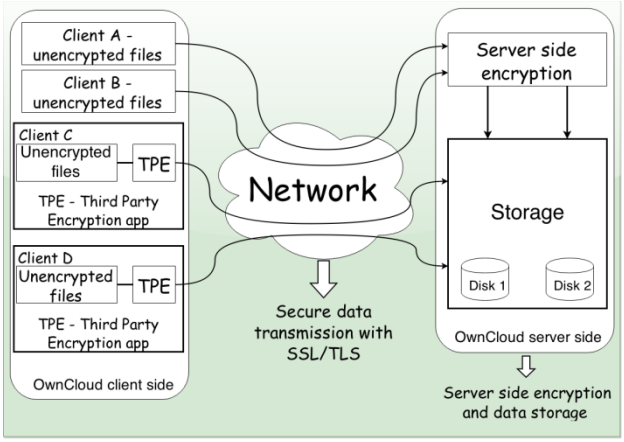


Figure 4. File encryption in a private cloud system

unauthenticated. Two endpoints using TLS connection can be certain that communication is indeed between those two devices, without the presence of a third party [11].

Another strongly advised feature is usage of the ownCloud provided Encryption app for the server-side encryption. In this case, after client data arrives over a secure SSL/TLS connection to the server side, the Encryption app executes automatic creation of a 4096-bit private and public key pair for each user. Private keys are encrypted using users login credentials with AES 256 algorithm [12]. Using the keys, the data is encrypted and protected from theft. In case of lost user credentials, administrator may access protected data to recover information. This encryption model is claimed in [12] to be highly secure and optimized to perform well with a larger amount of data, large file sizes and a great number of users. OwnCloud administrators have full control over the encryption keys. Using this encryption model, stored files should be protected from unauthorized access.

If absolute confidentiality is required, client-side encryption must be implemented. The client side encryption refers to encryption of data on a client's own device, before sending it through a secure connection towards the server. This way, the data never leaves a user's device unencrypted and not even an administrator can gain access to it. Although neither of ownCloud editions offer a client-side encryption, it can be implemented using a third party software, such as Credeon. Fig. 4 shows the aforementioned file encryption scenarios.

C. *OwnCloud client and features*

OwnCloud offers data access, synchronization and data sharing across devices, everything in a controlled environment. Its architecture is extensible and works with any storage [13]. Synchronizing ownCloud with various devices is reinforced with client applications. Desktop clients are supported for multiple operating systems: Windows, Mac, different Linux distributions and versions, as well as Android and iOS for mobile clients. When stored in the ownCloud, user's files, folders, contacts, calendars and images are accessible from PCs, laptops, smartphones and other mobile devices at any given time or location. From a user's point of view, only a network connection is needed for data synchronization among devices. Previously synchronized data on desktop devices are accessible without a network connection, since the data are also stored locally. OwnCloud web server acts as a main repository of users data. When a client connects to the server, a system automatically updates two lists of data: the list of data that has remained unchanged, and the list of data that has changed since the last synchronization. In this manner, a client is able to download and send only a list of files that have been changed. This speeds up the process of scanning and analyzing which files need to be updated. Storing and synchronizing is different with mobile devices, as their limited memory is not well-suited for keeping gigabytes of data. Consequently, mobile clients require network connection for data access.

D. *OwnCloud scalability*

Both ownCloud editions, the ownCloud Community and the ownCloud Enterprise, were designed to accommodate whole range of application areas [13], and for that they needed a scalable architecture. Scalability issues were analyzed in detail in [14]. SMBs that invest in a dedicated infrastructure for ownCloud should have no need to implement anything beyond the simple ownCloud setup. However, in case that ownCloud is accommodated on an existing infrastructure or an organization grows significantly, a performance or storage bottlenecks could occur.

A standard approach to building a scalable solution is to have separate servers for each of the private cloud service tiers, such as database, data storage or application tiers. As this setup is unnecessarily complex and too costly for SMBs, a more adequate solution would be to use a model with collapsed tiers. In a consolidated configuration, shown in Fig. 5, a single computing node can be used to host a web server, the ownCloud software, a storage server and a MySQL server. Having two or more such computing nodes behind a load balancer enables the clients' requests to be served by each of the nodes in a round-robin manner. OwnCloud allows scale-out of each component to be performed while running, thus providing continuous service - keeping the data accessible and applications uninterrupted.

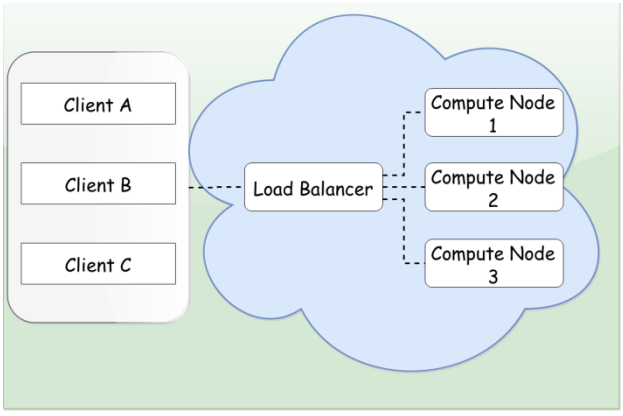


Figure 5. OwnCloud scalable deployment model

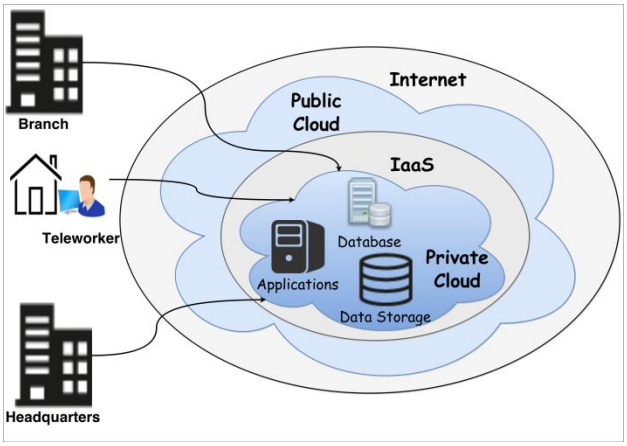


Figure 6. Hybrid approach to ownCloud deployment

E. OwnCloud in a cloud

By now, we only discussed deployment of ownCloud on a private infrastructure, i.e. within the organization. Another option could be a hybrid approach, in which the infrastructure is provided by an outer cloud service provider [1]. In this configuration, a private cloud system is hosted in a public cloud using leased infrastructure. The approach could be very interesting for businesses looking to minimize initial infrastructure investment and, later on, maintenance costs. Although the infrastructure is leased, with careful cloud implementation, data security should not be significantly impacted. Some organizations may be reluctant to invest in their own infrastructure and this approach might be a compromise which would allow them to have enough benefits of private clouds to alleviate concerns they might have with public cloud systems.

Another advantage of a hybrid cloud is flexibility in terms of changing capacity. With a pay-per-use model, an organization can momentarily scale-up or scale-down as necessary, something that is not always possible, especially in the case of SMBs, with own infrastructure. Therefore, a hybrid cloud in some cases may be a better option and a more cost efficient solution, easier and quicker to deploy than a private cloud, but also more secure than public cloud. Running ownCloud on an external infrastructure is shown in Fig. 6.

V. CONCLUSION

Public cloud services are very popular, however, organizations are still cautious with moving their businesses into the cloud. In case of trust or data security being a serious concern, we suggest using private cloud solutions. A private cloud approach eliminates the trust issue. It preserves data ownership and can improve both privacy and security. The solution we explored is scalable and is suitable not only for SMBs but also for larger deployments. For organizations which do not have enough resources to invest into their own infrastructure or to maintain it, we suggest a hybrid cloud solutions. In this case, a private cloud is hosted on a public cloud infrastructure, making a tradeoff between security and privacy on one side and cost on the other. It allows even SMBs with very limited resources to leverage benefits of the private cloud solutions.

ACKNOWLEDGMENT

The work presented in this paper is supported by the University of Rijeka research grant Grant 13.09.2.2.16.

REFERENCES

[1] K. Khan, Q. Malluhi, “Establishing trust in cloud computing,” in IT Pro, September/October 2010, pp. 20-26.

[2] F. Doelitzscher, A. Sulistio, C. Reich, H. Kuijs, D. Wolf, “Private cloud for collaboration and e-Learning services: from IaaS to SaaS,” Hochschule Furtwangen University, Germany

[3] J. Peng, X. Zhang, Z. Lei, B. Zhang, W. Zhang, Q. Li, “Comparison of several cloud computing platforms,” in Information Science and Engineering (ISISE), 2009 Second International Symposium, pp. 23-27.

[4] R. Giordanelli, C. Mastroianni, “The cloud computing paradigm: Characteristics, opportunities and research issues,” Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR), 2010.

[5] K. Atefi, Y. Saadiah, A. Atefi, “A survey on digital forensics investigation of Seafile as a cloud storage,” in International Journal of Engineering Research And Management (IJERM), vol. 01, October 2014.

[6] J. Mościcki, M. Lamanna, “Prototyping a file sharing and synchronization service with Owncloud,” in Journal of Physics: Conference Series, vol. 513, no.4, IOP Publishing, 2014.

[7] R. Grossman, “The Case for Cloud Computing,” in ITpro, pp. 23–27, March/April 2009.

[8] S. Ramgovind, M. M. Elof, E. Smith, “The management of security in cloud computing,” IEEE 978-1-4244-5495-2/10, University of South Africa, Pretoria, South Africa

[9] P. Mell, T. Grance, “The NIST definition of cloud computing,” NIST special publication 800-145, National institute of standards and technology, September 2011.

[10] M. Armbrust, A. Fox, et al, “A View of Cloud Computing,” Comunications of the ACM, vol. 53., April 2010, pp.50–58.

[11] L.Ertaul, S.Singhal, G.Saldamli, “Security challenges in cloud computing,” CSU East Bay, Hayward, CA, USA

[12] OwnCloud, “OwnCloud’s data encryption model"

[13] Official ownCloud web page

[14] “Scale out file sync and share: Deploying ownCloud and Red Hat Storage Server on HP ProLiant SL4540 Servers,” OwnCloud whitepaper

–

Testing in the Private Cloud

Tomislav Pavić, Ivan Vrbovčan, Marija Šoša Anić
Hrvatski Telekom d.d., Zagreb, Croatia
tomislav.pavic@t.ht.hr, ivan.vrbovcan@t.ht.hr, marija.sosaanic@t.ht.hr

Abstract - Due to a requirement for increasing the development speed of new telecom services and enhancing functionalities of existing ones, telecom providers are trying shorten and automate creation of testing environments. Using cloud infrastructure installed on its own premises for that purpose opens myriad of new possibilities and gives an opportunity to improve traditional development process.

In this paper we will provide an overview of our approach for creating testing environments in the private cloud using open source solutions that works on top of the OpenStack cloud infrastructure. The overview will include description of the approach related to the existing development process, high level architecture and requirements for successful implementation of the presented approach.

I. INTRODUCTION

Due to a requirement for increasing the development speed of new telecom services and enhancing functionalities of existing ones, telecom providers are trying shorten and automate creation of testing environments. Using cloud infrastructure installed on its own premises for that purpose opens myriad of new possibilities and gives an opportunity to improve traditional development process.

In this paper we will show the possibilities of improving the process of application development and deployment through the automation of creation and administering of testing environments which required a great amount of time and effort from our IT staff. In this approach we are using an open source private cloud O/S (OpenStack) enhanced with technologies for automation (Chef) and run-time control (Cloudify) – see Figure 1. These tools make it possible to automate the process of creating the environment for testing purposed and thus minimizing the need for manual work.

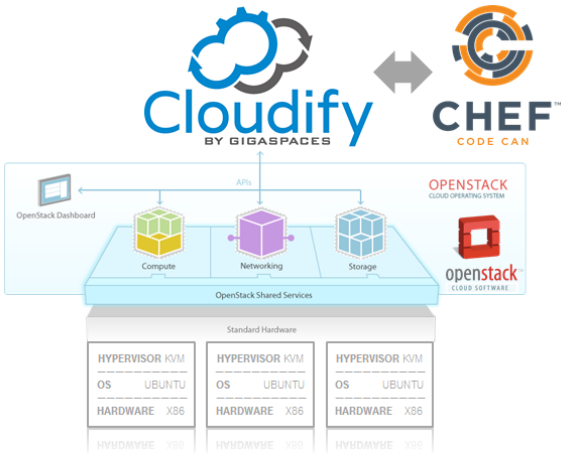


Figure 1: Components and how they Work Together [10]

Hrvatski Telekom (HT) started to evaluate open-source cloud computing solutions starting in 2011 [1]. As result of this effort OpenStack (http://www.openstack.org/) was chosen as the most viable solution for creation open-source cloud infrastructure commonly known as an Infrastructure as a Service (IaaS). OpenStack project has been started with high expectations to run large cloud installation. It is characterized by rapid development and fast growing level of community support. High level of modularity and robustness is suitable for vendor customizations and development of value-added features. Recent releases offer many new features such as distributed scheduler for deploying VMs, strong image management service, unified identity management system, IPv6 and VPN support. List of recent improvements include major improvements to OpenStack's networking, orchestration, and metering systems in addition to many features that support software development and big data analysis. Common cloud operations are performed via a web-based OpenStack Dashboard user interface. OpenStack has many important players on its side and became one of the leaders in cloud computing arena.

Having selected open-source solution for IaaS we have recognized that we need capable tool for application installation and configuration. Based on findings from joint venture project with FER we selected Chef for that purpose [4]. Chef (https://www.chef.io/) is a systems and cloud infrastructure automation framework which can be used for easy deployment servers and applications to any physical, virtual or cloud location with no matter the size

of the infrastructure. It relies on abstract definitions known as cookbooks and recipes. Cookbooks are used to describe how a specific part of infrastructure should be built and managed declaratively. Applying these definitions we are able to fully automate our infrastructure (not only cloud based). Rich repository (cookbooks, recipes, versions, attributes, resources, providers, libraries, and templates), roles, data bags, environments and configuration files simplified development of custom tailored recipes and can be used for complex production environments.

Third component in our tool-box is Cloudify (http://getcloudify.org/) which is an open-source cloud computing Platform as a Service (PaaS) solution by GigaSpaces Technologies. It can be integrated with Chef so the services lifecycle can be simplified by using Chef to take care of keeping service configurations up-to-date with current specifications as defined by roles and cookbooks. This approach isolates the application completely from the underlying cloud runtime. Cloudify recipes provide the execution plans for installing, starting, orchestrating and monitoring an application stack including actions to ensure self-healing and horizontal scalability.

II. SOLUTION TEST AND ACCEPTANCE PROCESS

Existing Solution Test and Acceptance Process (see Figure 2) as part of development process in HT is based on ITIL V3 [5-9].

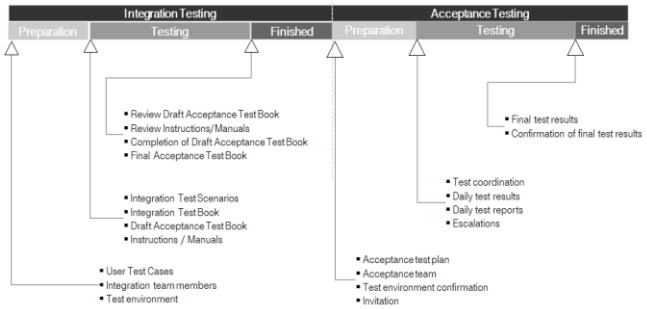


Figure 2: Solution test and acceptance process [3]

Solution Test and Acceptance Process is designed to ensure quality control of IT systems. The main objective is to ensure interoperability of systems through the activities of analyzing and documenting detected errors ensuring correct behavior of software systems prior to going live. The goal of the process is to improve resolution of problems raised from the testing process with the test time reduced as much as possible. This can be done by improving the allocation of tasks and improving utilization of used resources (both physical and human).

Test management as is defined in ITIL Service Transition stage [7] includes the planning, control and reporting of activities. Prior any testing activity detailed planning of required test resources should be done which includes prioritizing and scheduling what is to be tested and when. During the testing activities any incident, problem, errors, non-conformance, risk and any other issue should be properly managed so that incoming known errors and their documentation can be timely processed. Monitoring progress and collecting feedback from validation and test activities are essential to reduce errors that may end up into production. Other activities include capturing configuration baseline, collection of test metrics, analysis and reporting.

Test management should be able to identify activities that can impose delays. These issues needs to be proactively managed through the activities of mitigating risks and implementing changes identified during testing activities.

Clearly this process is huge time and resource consumer and HT internal analysis [3] prove that. Another finding stated that automation of Solution Test and Acceptance Process in HT is currently on very low level and for integration testing totally missing. That was the main driver for pilot implementation of testing in the private cloud.

III. PILOT IMPLEMENTATION

To create and administer pilot environment we used the following Open Source tools – see Figure 3:

- OpenStack Havana: for private cloud infrastructure installation based on four physical servers
- Chef Server 11.0.8: to automate O/S configuration (which is installed manually) and installation/configuration databases and other applications
- Cloudify 2.7.0: PaaS as a tool for run-time control and deployment

This pilot is based on real world applications that are used in T-HT:

- Donat: internally developed client/server CRM application based on Oracle and Microsoft technologies
- Citrus: web application based on Postgresql, Apache, Ruby on the Rails and Memcached used as application inventory

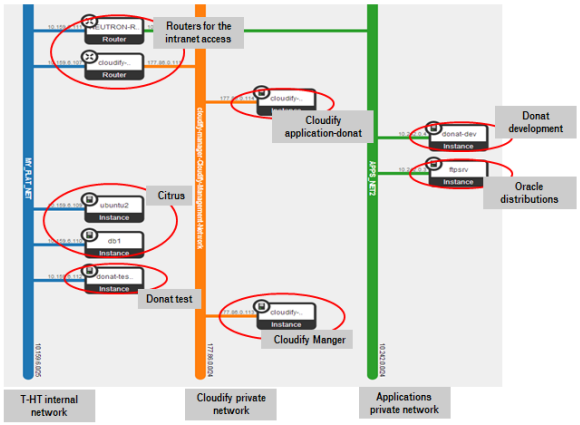


Figure 3: Pilot Environment

Deployment process is defined as follows – see Figure 4:

1. Connect to the OpenStack cloud from the Cloudify Management virtual machine (VM). This initial connection ensures OpenStack cloud is operable and accessible.
2. Create one or more application VMs on OpenStack. Number of VMs depends on initial requirements of application workload.
3. Get IPs addresses of application VMs. This IPs are used by Chef and Cloudify Management for further activities that need to be done on these VMs and for needs for maintain possible dependences of created VMs (e.g. for establishing communication between application and database servers).
4. Install Cloudify client on application VM which is responsible for communication with Cloudify Management and execution of local tasks
5. Install Chef client on application VM which is responsible for communication with Chef server.
6. Register Chef client to Chef server and ensure secure communication
7. Provide Chef recipes to the Chef client to be installed on local VM
8. Install and configure required software on application VM

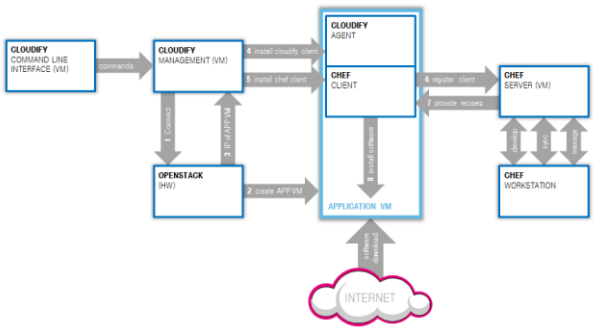


Figure 4: Deployment process [10]

Based on existing Chef recipe for the Oracle Database 11g Enterprise Edition it was easy to install Oracle database and later on to create Cloudify recipe based on that service:

```
service {
  extend "../../../services/oracle"
}
application {
  name="donat"
  service {
    name = "oracle"
  }
}
```

As can be seen from this example an application recipe describes the application name, the services it requires and their interdependencies – see Figure 5. A service is defined as set of service instances in form of cluster from which an application tier is created (e.g. web cluster of service instances). In this context it is important to distinguish between a service and a service instance. A service is defined as a cluster which is made from at least one service instance. Every service instance always resides on one host (typically this is an VM created exclusively for that service instance), whereas a service is the collection of all service instances and in case that thre is more than one service instances than obviously it would span multiple machines. It is important to understand that each service is installed, managed, and monitored by a service recipe.

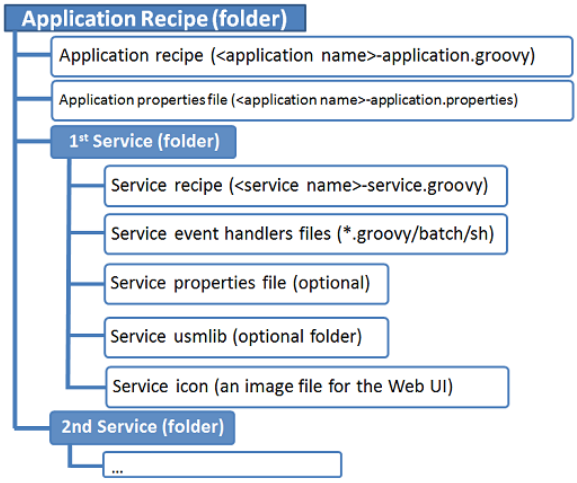


Figure 5: Application's recipe folder structure [11]

IV. CONCLUSION

In this work we have presented high level solution concept for testing in the private cloud. Based on this approach initial pilot implementation on simplified scenario was done to prove validity of proposed models and concepts. After successfully finished pilot we have identified following positive aspects of this approach:

- Easy to install OpenStack using appropriate Chef Cookbooks (in this Pilot we have used Rackspace Private Cloud cookbooks)

- Common cloud operations are performed via web-based OpenStack Dashboard user interface
- Easy to deploy servers and applications using Chef to any physical or cloud location
- Cloudify recipes provide the execution plans for installing, starting, orchestrating and monitoring an application stack
- Easy to use and modify existing recipes (Chef and/or Cloudify) - no matter the size of the infrastructure

Besides these positive aspects still careful planning is required because conversion of legacy systems can be difficult especially if there is no existing recipe (Chef and/or Cloudify) for the specified application.

Although, pilot was successfully finished more work on automation and orchestration has to be done, and it remains for the future phases.

LITERATURE

[1.] I. Voras, B. Mihaljević, M. Orlić, M. Pletikosa, M. Žagar, T. Pavić, K. Zimmer, I. Čavrak, V. Paunović, I. Bosnić, S. Tomić: Evaluating Open-Source Cloud Computing Solutions. 34th International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija: 2011.

[2.] I. Vrbovčan, T. Pavić, M. Šoša Anić: Connecting clouds. 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija: 2014.

[3.] Z. Kamčev: Reduce effort for IT Development testing. HT internal document, Zagreb: 2012.

[4.] M. Žagar, B. Mihaljević, M. Orlić, I. Voras: Analysis and Selection of the Automated Application Provisioning Solutions for the Cloud. HT internal document, Zagreb: 2013.

[5.] ITIL: Service Strategy. TSO, Edinburgh: 2007.

[6.] ITIL: Service Design. TSO, Edinburgh: 2007.

[7.] ITIL: Service Transition. TSO, Edinburgh: 2007.

[8.] ITIL: Service Operation. TSO, Edinburgh: 2007.

[9.] ITIL: Service Improvement. TSO, Edinburgh: 2007.

[10.] A. Sobania: Cloud Application Management. DT internal document, Bonn: 2014.

[11.] GigaSpaces: Cloudify 2.7 Documentation

NETWORKS & NETWORKS PERFORMANCE

The methodology for the optimal choice of access technologies

Zeljko Bozinovic, Nasuf Hadziahmetovic
BH Telecom d.d. Sarajevo/R&D Department, Sarajevo, BiH
zeljko.bozinovic@bhtelecom.ba, nasuf.hadziahmetovic@bhtelecom.ba

Abstract - There are many access technologies which it is possible to realize new and reconstruct the existing access networks. The goal is to find those technologies that will best meet the quality parameters, use of existing resources and provide support to current and future services. In considering this problem there are many criteria to be taken into account. In consideration we started from the hypothesis that technology can change the status of optimality when changing bandwidth and density of users. Possible optimization methods do not provide a universal approach to this problem. In this paper, the methodology used, which also takes into account the quality criteria and criterion optimal utilization of existing network resources. As a result, this paper offers a new algorithm and optimization methods suitable for the optimization of both fixed and mobile access networks. Also, the algorithm and optimization methods perform optimal technology choice with corresponding topology at the given conditions. The results confirm the initial hypothesis.

I. INTRODUCTION

Construction of new and reconstruction of existing access networks expressed the problem of choosing the optimal solution. The optimal solution depends on the parameters of quality, quantity of resources needed and the total cost. There are many different methods and software packages used for optimization and planning [1], [3] and [4]. Unfortunately, most of these methods of optimization are not universally usable and provide a solution taking into account only one criteria. Even the methods that are set as a universal do not include all segments that are necessary for quality optimization. The reason for this is the desire to be applied in the broader spectrum of possible optimization process. On the other hand, the most commonly treated problems of the new networks without taking into account the possibilities of exploiting the resources of existing access networks.

It was observed that the increase in bandwidth and user density in access networks, could cause the initially less optimal solution becomes the best solution and vice versa [2]. This hypothesis, which is partly proved in the paper [2], because we considered only a passive part of the access network, we have taken as a starting hypothesis in this paper.

To prove our initial hypothesis, we developed a new algorithm and optimization method for techno-economically optimal selection of access technologies. The algorithm takes into account the quality parameters that are eliminatory for selection of optimal solution. He also compares the needs of the considered solution for resources and available resources of the existing access network. The algorithm allows an optimization of wired, wireless and mobile access technologies and their combinations. With minor adjustments it can be used also for the other areas of optimization.

By applying the algorithm to the technologies analyzed in the paper [2], we received confirmation of the initial hypothesis. The reason for using technology from the work [2] the possibility of utilization of part of the previously obtained results. On the other hand, we have shown that the change of solutions optimality is influenced by the increase in bandwidth, user density, and the relationship between prices of active equipment and fiber optics.

II. DEVELOPMENT OF MATHEMATICAL MODEL AND ALGORITHM

A. *Initial assumptions*

For mathematical modeling of a passive part of the network we used the mathematical model and the initial assumptions elaborated in the work [2].

In addition to the initial assumptions given in [2], we introduced the following assumptions, no impairment of generality:

- We will observe only technologies that have been considered in the paper [2], for the use of the obtained results and the possibility of mutual validation.
- Mobile technologies currently is not be taken into consideration;
- We will ignore the parameters of quality, because the choice of parameters can vary from case to case;
- The impact of resource utilization of the existing access network is mathematically modeled, but it will not be considered due to variations from case to case;
- Instead of the actual price, we will use the normalized prices. They are obtained by dividing the required amount of resources with the lowest necessary amount of resources;
- Passive network components are included into the price of required active elements of technology under consideration. Exceptions are optical cables.

B. Mathematical model

By studying the optimization method, which is used to optimize the access networks, we noticed that they generally use one optimization criterion or parameter for ranking solutions. By studying the access networks, it is clear that the optimization problem is significantly more complex, and that cannot be reduced to a single parameter or criterion.

Optimization problem is reduced to finding a balanced relationship between quality, cost and constructed network resources. As a solution imposed by the combined methods of optimization of all these parameters simultaneously.

In our case, the length and number of optical fibers, and the amount of passive and active network elements directly dependent on bandwidth and users density [2]. Also, the amount of resources and their arrangement depend on the parameters of the quality required. While the quality parameters are stricter it is generally necessary needed higher amount of passive and active equipment, regardless of the applied solutions. If it is not possible to reach some of the set quality parameters it is necessary to improve observed solution. This means more passive and/or active elements, and therefore more expensive solution. The entire procedure must be repeated for all considered bandwidths.

Required passive and active resources, are necessary to compare with the available resources of the existing access network. If resources are sufficient then the additional cost solutions are zero. If the available resources are insufficient then the additional cost are given by equation (1).

$$C_{ij}^{add} = C_{ij}^{as} + C_{ij}^{ds} \tag{1}$$

where is:

C_{ij}^{add} – additional price of the i-th solution and the j-th bandwidth,

C_{ij}^{as} – additional price of the aggregation segment of the i-th solution and j-th bandwidth,

C_{ij}^{ds} – additional price of the distribution segment of the i-th solution and j-th bandwidth.

In equation (1) are:

$$C_{ij}^{as} = \left(L_{ij}^{as} - L_{av}^{as}\right) \cdot C^{\frac{as}{1m}} \tag{2}$$

$$C_{ij}^{as} = \left(L_{ij}^{ds} - L_{av}^{ds}\right) \cdot C^{\frac{ds}{1m}} + C_{ij}^{add\ bs} \tag{3}$$

where are:

L_{ij}^{as} – the required length of the aggregation segment of the i-th solution and j-th bandwidth,

L_{av}^{as} – available length of the aggregation segment of the existing network,

$C^{\frac{as}{1m}}$ – price of the aggregation segment per 1m length,

L_{ij}^{ds} – the required length of the distribution segment of the i-th solution and j-th bandwidth,

L_{av}^{ds} – Available length of the distribution segment of the existing network,

$C^{\frac{ds}{1m}}$ – price of the distribution segment per 1m length,

$C_{ij}^{add\ bs}$ – price missing resources for the base station of the i-th solution and j-th bandwidth, which do not include active equipment base station.

The total normalized price for i-th solution is given by equation (4).

$$C_i^{nor} = \sum_{j=1}^{m} \left(C_{ij}^{nor\ opt} + C_{ij}^{nor\ eq} + C_{ij}^{nor\ add} + C_{ij}^{nor\ bs}\right) \tag{4}$$

In equation (4) are:

$$C_{ij}^{nor\ opt} = \frac{L_{ij}^{opt}}{L_{j\,\min}^{opt}} = \frac{L_{ij}^{as} + L_{ij}^{ds}}{L_{j\,\min}^{opt}} \tag{5}$$

$$C_{ij}^{nor\ eq} = \frac{C_{ij}^{eq}}{C_{j\,\min}^{eq}} \tag{6}$$

$$C_{ij}^{nor\ add} = \frac{C_{ij}^{add}}{C_{j\,\min}^{add}} \tag{7}$$

$$C_{ij}^{nor\ bs} = \frac{C_{ij}^{bs}}{C_{j\,\min}^{bs}} \tag{8}$$

In equation (4), (5), (6), (7) and (8) are:

C_i^{nor} – normalized total price of the i-th solution for all considered bandwidths simultaneously,

$C_{ij}^{nor\,opt}$ – normalized price required optical fibers of the i-th solution and the j-th bandwidth,

L_{ij}^{opt} – the total length of optical fibers of the i-th solution and the j-th bandwidth,

$L_{j\,\min}^{opt}$ – the minimum required length of optical fibers for the j-th bandwidth,

$C_{ij}^{nor\,eq}$ – normalized price of the necessary active equipment of the i-th solution and the j-th bandwidth,

C_{ij}^{eq} – price of necessary active equipment of the i-th solution and the j-th bandwidth,

$C_{j\,\min}^{eq}$ – the minimum price of the necessary active equipment of the j-th bandwidth,

$C_{ij}^{nor\,add}$ – normalized additional price of the i-th solution and the j-th bandwidth,

$C_{j\,\min}^{add}$ – minimal additional price for the j-th bandwidth,

$C_{ij}^{nor\,bs}$ – normalized price of the active base stations equipment of the i-th solution and the j-th bandwidth,

C_{ij}^{bs} – the price of the active base stations equipment of the i-th solution and the j-th bandwidth,

$C_{j\,\min}^{bs}$ – the minimum price of the active base stations equipment for the j-th bandwidth,

m – number of the bandwidths under consideration.

When the algorithm does not evaluate mobile access technologies then $C_{ij}^{bs} = 0$, and $C_{ij}^{nor\,bs} = 0$, in (4).

Selection of techno-economic optimal solution is made from the (9).

$$C_{\min} = C_1^{nor} \;\; for\; i = 1\; and$$
$$C_{\min} = C_i^{nor} \;\; if \;\; C_i^{nor} < C_{\min} \;\; for\; 1 < i < n \qquad (9)$$

Where is:

$C_{\min}$ – the minimum price of the considered solutions for all considered bandwidths simultaneously.

Often it is not possible to compare the prices of network elements due to differences in technology, or in a case when we do not know the unit prices of these elements. This is most often case with active equipment. In this case for the purpose of price comparisons we introduce a_{kj} parameter in the following way:

$$L_{ij}^{opt} \cdot C^{opt} + N_{ij}^{eq} \cdot C_i^{eq} = L_{kj}^{opt} \cdot C^{opt} + a_{kj} \cdot N_{kj}^{eq} \cdot C_{\min}^{eq} \qquad (10)$$

If we adopt:

$$C_i^{eq} = C_{\min}^{eq} \qquad (11)$$

Then from equations (10) and (11) follows:

$$a_{kj} = \frac{C^{opt} \cdot \left(L_{ij}^{opt} - L_{kj}^{opt}\right)}{N_{kj}^{eq} \cdot C_{\min}^{eq}} + \frac{N_{ij}^{eq}}{N_{kj}^{eq}} \qquad (12)$$

In equations (10), (11) and (12) are:

C^{opt} – price of the optical fibers independent of technology,

N_{ij}^{eq} – the amount of active equipment of the i-th solution and j-th bandwidth,

C_i^{eq} - the price of the active equipment of the i-th solution,

$C_{\min}^{eq}$ - the minimum unit price of active equipment,

a_{kj} - regulator equipment costs of the k-th solution and the j-th bandwidth.

From equation (10) we notice that the parameter a_{kj} indicates how much the cost of equipment can be higher / lower than the minimum costs of active equipment, if equation (11) is satisfied.

If it is $a_{kj} > 0$, then it is possible to realize the price of the active equipment in real conditions. In that case total price of k-th solution can be more preferable than total price of i-th solution. In the case where $0 < a_{kj} < 1$ the price of the k-th solution should be a_{kj} times less than the minimum cost $C_{\min}^{eq}$. If it is $a_{kj} = 1$ than the price of k-th solutions active equpment is equal to the minimum price of active equpment. When is $a_{kj} > 1$ price of k-th solutions active equipment can be a_{kj} times larger than the minimum costs of active equipment.

In the case where $a_{kj} \le 0$ price is in undefined area. When $a_{kj} = 0$, equipment have not got a value. When $a_{kj} < 0$ the price of the equipment must have negative value. We conclude that for a given $C_{\min}^{eq}$, with $a_{kj} \le 0$ there is not exist the k-th solution which satisfy equation (10) in real conditions.

When we have to determine the parameter a_{kj} then it is determined for all solutions and all bandwidths. By adding parameters a_{kj} for all bandwidths, we get the summary parameter a_k for k-th solution according to the equation (13).

$$a_k = \sum_{j=1}^{m} a_{kj} \qquad (13)$$

In this case, the optimal solution is determined using the expression (14).

$$C_{\min} = a_1 \;\; for\; i = 1\; and$$
$$C_{\min} = a_k \;\; if \;\; a_k > C_{\min} \;\; for\; 1 < k < m \qquad (14)$$

C. *The optimization algorithm*

Based on the mathematical model and the above considerations in part B, we formed the optimization algorithm, which is shown in shorter form at Figure 1.

The algorithm begins by defining the available resources of existing access network, coverage area, user's density, bandwidths, quality parameters and characteristics of access technologies.

As a first step, we examine whether we consider radio technology. If we consider wire technology, then we determine the number and total length of required optical fibers and the number of active elements, as a function of bandwidth and users density. If we are considering a radio technology then we determine the number and type of

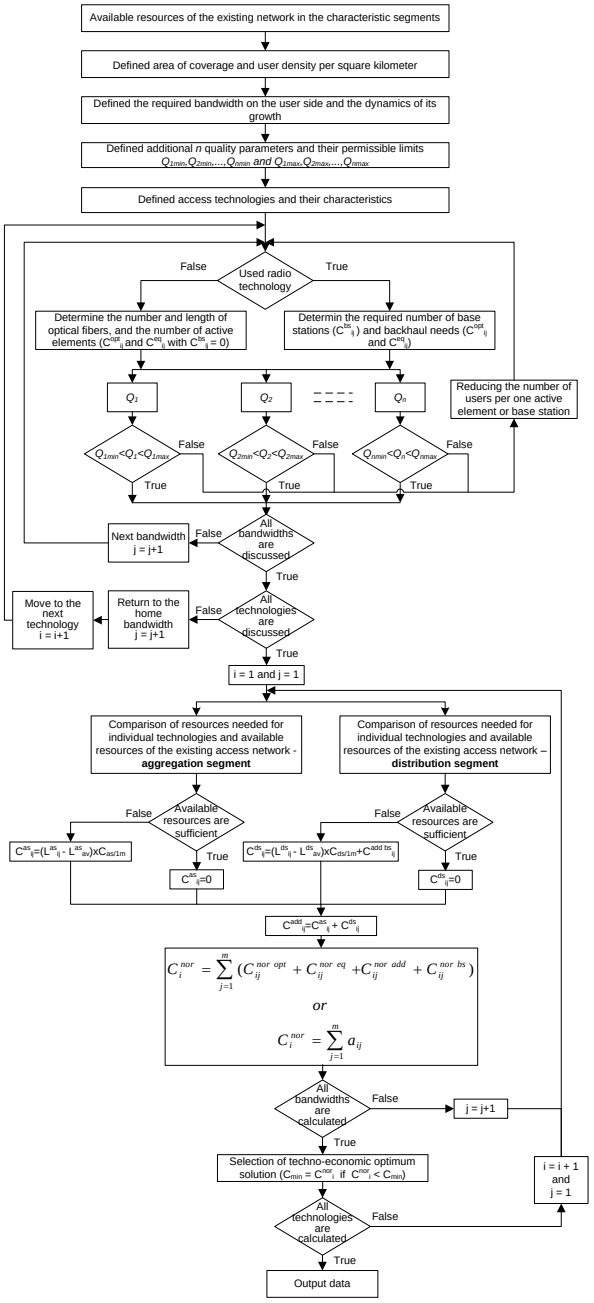


Figure 1. The optimization algorithm for wired, wireless and mobile access networks

base stations. Also, we determine the required amount of active equipment, and the number and length of optical fibers to connect base stations to the rest of the network. These calculations are made on the basis of mathematical model given in [2]. Based on the initial assumptions, mathematical model for radio technology was not observed in this paper.

After determining the necessary resources for the analyzed solution we perform its comparison with the quality parameters. If any of the parameters is not within the allowed limits we reduce the number of users per one active element or a base station. The process is repeated until the obtained solution does not satisfy the defined quality parameters within the permitted limits. This procedure is repeated for all considered bandwidths. If the preliminary proceedings conducted for all bandwidths, the same procedure is repeated for the next technology.

When the procedure is finalized for all technologies, the required resources of individual solutions are compared with the available resources of the existing access network. The aim of this comparison is the maximum utilization of existing resources while reducing the cost of a new product. When the available resources are sufficient then the extra cost are zero. Otherwise, they are calculated according to equations (1), (2) and (3).

All prices are normalized in equations (5), (6), (7) and (8). In this way, prices may not be expressed in monetary units, and they can be combined among themselves for the various elements of the network.

If the elements of technological solutions differ, and we do not know the real price, our solution to this problem is the introduction of a parameter a_{kj}. The parameter a_{kj} is determined by the equation (12). Normalized price for i-th solution we calculate according to equation (13). In this case the solution with the minimum cost is determined using the equation (14).

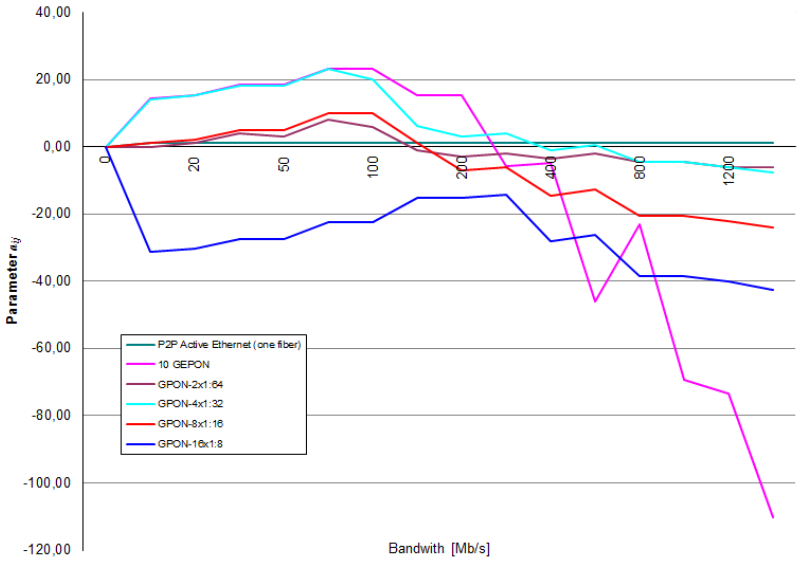
If the total normalized price is calculate according to equation (4), then the solution with the minimum cost determined using the equation (9).

One solution, or more of them, which possess a minimum price presents an optimal solution for the given initial conditions observed access network.

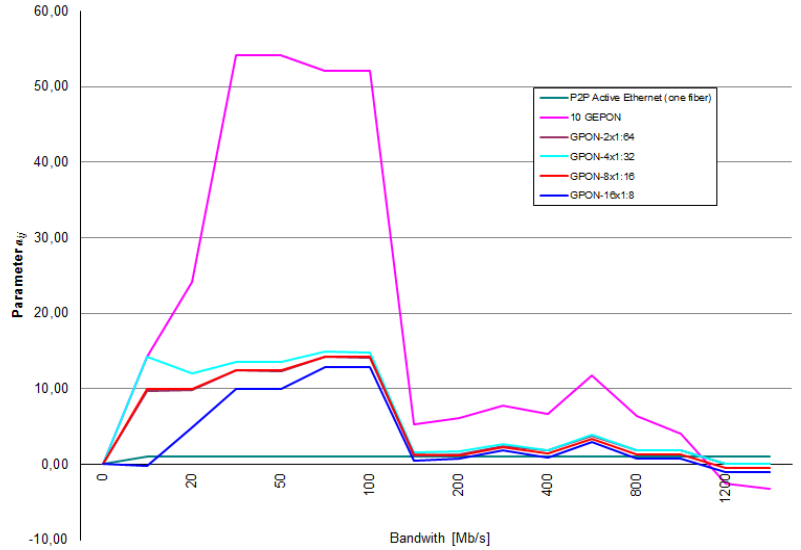
III. RESULTS

Taking into account the previously adopted assumptions and using the developed mathematical model and algorithm, we obtained results that confirm our initial hypothesis. They also verify the results that were previously obtained in the work [2]. At Figure 2 a) and b) are presents the results of behavioral parameters depending on the changes in bandwidth, the users density of 10 and 100 users / km2, respectively. The same calculations are made and the users density of 1000, 10 000 and 100 000 users / km2. For this users density are obtained almost identical parameter relates, such as shown in Figure 2. b). The difference is mainly related to the equalization parameters for GPON solutions, and its increase by 10 GEPON solution in relation to the other. As expected, the solution with active Ethernet has a constant value for all bandwidths, due to the assumption of a minimum price of its active equipment.

On Figure 3. Are presented the results of behavioral parameter a_k in dependence of users density, when $C^{opt}/C^{eq}_{\min}$ = 1/10. When $C^{opt}/C^{eq}_{\min}$ has got a values 1/100, 1/1000 and 1/10 000, behavior of the parameter a_k is similar to the behavior of parameter a_{kj}.



a)



b)

Figure 2. a) and b) The behavior of the parameter a_{ij}, with respect to the change of bandwith, for users density of a) 10 users/km^2 and b) 100 users/ km^2

Solution for active Ethernet in this case is constantly for all user densities.

IV. DISCUSSION

At lower user density results show that changes solutions benefits caused by changing bandwidth much more pronounced than at higher user densities. Also, in Figure 3. we can see that at different user's densities completely different solutions become the most appropriate for the application. Through that we have argued the initial hypothesis.

When users density reaches a certain value of the convenience use is stabilizing and the difference in the benefits of individual solutions becomes approximately constant. The lower limit users density at which stabilize the relationship between the benefits for the use of individual solutions is reduced to a lower value of relationships $C^{opt}/C^{eq}_{\min}$.

On Figure 3. It is visible that for the small user densities up to 100 users / km2, the most suitable solution is GPON 4x1: 32. For higher users density the most suitable solution is 10 GEPON. With the reduced relations $C^{opt}/C^{eq}_{\min}$ gradually on all user densities suitable solution becomes 10 GEPON.

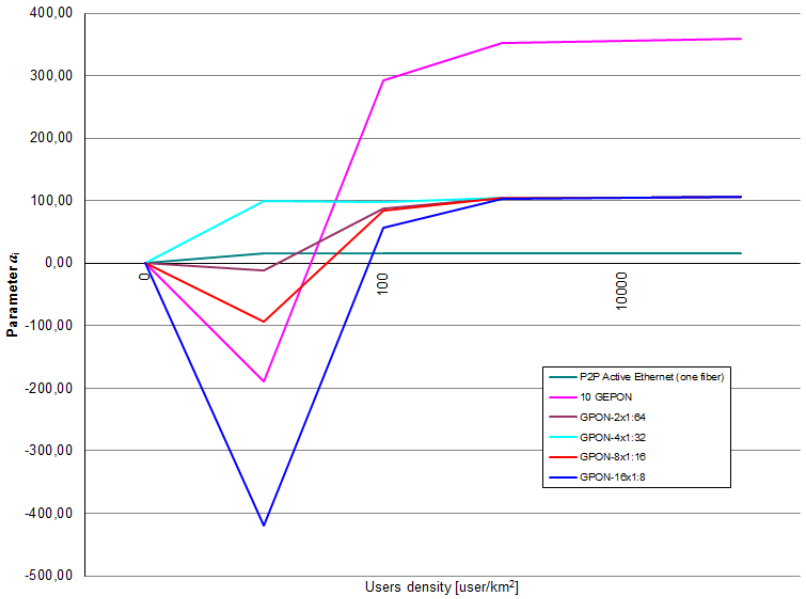


Figure 3. The behavior of the parameter a_i, with respect to the change of users density and $C^{opt} / C^{eq}_{\min} = 10$

Considering the behavior of the parameter a_k depending on user density and relations $C^{opt} / C^{eq}_{\min}$ we notice that he does not give specific values of prices of individual solutions. He gives limits in which individual solutions can be most favorable. If the actual cost of the active equipment meet this limit it becomes the most suitable solution. If the actual cost of the active equipment does not meet the limits given by the parameter a_k then solutions with this active equipment will not be the best solution.

In this way we do not get the final but conditionally optimal solution. The algorithm and implemented method finds the final optimal solution in the case of recognition of real price of active and passive equipment.

The subject of further research by this method would be the development of mathematical models for wireless and mobile access technologies. Continued research will entail adequate quality parameters. New research will also include other access technologies, which were not the subject of this study.

The above research will not affect the generality of the results obtained in this paper.

V. CONCLUSION

The initial hypothesis is confirmed, and at the same time are verified the results obtained in the paper [2]. Achieved contribution of this paper is a new method in the process of optimizing access networks, formed mathematical model, and a new algorithm that describes each step of the new method.

The method and algorithm are applicable for all access technologies. With minor adjustments they can be applied in other optimization processes. For a new needs it is necessary to adjust some parts of the mathematical model. The most part of the mathematical model is universally accepted and can be used for other optimization tasks.

LITERATURE

[1] V. Grimaila, N. Listopadskis, „Optimization of Telecommunication access Network," ISSN 1392-1215 Electronics and electrical engineering, 2006, No. 8(72), pp. 25-30.

[2] Ž. Božinović, H. Dizdarević and S. Dizdarević, „Optimal techno-economic selection of the optical access network topologies," IEEE Xplore Digital Library, 2012.

[3] M. P. Bendsoe, O. Sigmund, „Topology Optimization – Theory, Methods and Applications," Springer, 2004.

[4] Diwekar, Urmila, „Introduction to Applied Optimization," Springer, 2008.

An Open Source TCP/UDP-Based Network Probing Tool for Real-Time Packet Loss Estimation

Dr. Ala' Khalifeh
Assistant Prof., Communication Eng.
German Jordanian University
Amman, Jordan
ala.khalifeh@gju.edu.jo

Ahmed Abbad
Computer Science
German Jordanian University
Amman, Jordan
a.abbad@gju.edu.jo

Dr. Khalid Darabkeh
Associate Prof., Computer Eng. Jordan University
Amman, Jordan
k.darabkeh@ju.edu.jo

***Abstract*— This paper demonstrates the design, implementation and the evaluation of a new open source tool for estimating end-to-end packet loss between a source and a destination node over the lossy Internet. Different from the most of the proposed tools in the literature that uses either TCP or UDP, and may require long time to accurately predict the packet loss rate, the proposed tool utilizes both features of the TCP and UDP, thus making it able for accurately predicting the packet loss of a communication link within a short probing time, which makes it appropriate for real-time multimedia applications. Our performance evaluation shows the effectiveness of the proposed tool even in relatively short probing time.**

Keywords—Network probing, active probing, packet loss estimation and prediction, multimedia applications

I. INTRODUCTION

Internet packet loss is considered one major factor that greatly affects the quality of many Internet applications such real-time multimedia services including audio and video streaming, VoIP, video conferencing, etc. [1]. These applications are becoming very essential in our daily life and users are continuously demanding better performance. The main problem that the Internet is not originally designed to transport real-time Internet traffic. As such, different mechanisms have been used to improve the transmission quality of these applications [2]. Some examples of these techniques are the use of joint source and channel coding for determining the optimal source/channel encoding scheme for a given network packet loss, such that the expected received video or audio quality is maximized. Another approach is to dynamically change the multimedia encoding rates and qualities such that a low encoding rate with a low media quality is used when the network is congested, where high packet loss is encountered, while higher rate with higher quality is used for less congested networks [3]. However, all these proposed algorithms and techniques rely on estimating the network packet loss, so the appropriate action can be taken. Moreover, predicting the packet loss rate helps in network management and routing the Internet traffic to less congested links [4]. Furthermore, many of the proposed QoS routing algorithms rely on estimating the packet loss rate for different routes and communication links [5, 6].

Packet loss estimation can be performed using passive or active methods [7]. In passive methods, the behavior of the ongoing TCP sent packets and their re-transmitted packets are observed and an estimation for the end-to-end packet loss is obtained, these methods can be useful for monitoring the network and evaluating the QoS of different transmission links, which is very valuable for Internet Service Providers (ISPs) [8] who try to direct the traffic to the less congested links. These methods assume the existence of a TCP traffic between a source and a destination node, which may not be available in case these two nodes are exchanging real-time multimedia contents that normally use UDP traffic. The other probing technique samples the network by sending probing packets between the source and the destination nodes, the sent packets can be an ICMP, TCP or UDP packets. In the literature, there are many tools proposed to predict and measure the network packet loss. In what follows, a brief description of the most closely related work in the literature is provided. Authors in [9] designed a tool named *tulip* that utilizes the ICMP protocol in sending ICMP echo packets to a destination node, and by exploiting the ICMP features such as the timestamps and IP identifier counters, return remote timing and ordering information, the path packet loss can be estimated. However, ICMP-based tools have two main problems as reported in [10] that reduce their accuracy and practicality. Authors in [10] proposed a TCP-based networking measurement tool called *Sting*. Sting analyzes the TCP stream to predict packet loss in both directions between the receiver and the transmitter nodes. However, due to the TCP protocol operation nature and features such as: three way handshaking, packets re-transmission in case of packet loss, packets acknowledgment in case of packet reception, etc.; it will require significant time to estimate the packet loss. In addition, TCP packets have large packet header size, which increases the transmission overhead and delay. As such, TCP- based techniques may not be timely efficient for real-time multimedia applications that require a quick method to estimate the link packet loss.

The rest of the paper is organized as follows: Section II discusses the tool design, its features, and implementation. Section III describes the test-bed setup used in the experimental evaluation. Section IV provides performance evaluation of the tool. Finally, Section V concludes the paper and suggests future works.

II. TOOL DESIGN AND IMPLEMENTATION

In this section, a detailed description of the tool design implementation and architecture is provided. Further, the tool operation and how it can estimate network packet loss is demonstrated and supported by some examples.

A. NetProb design and operation probing modes

As shown in Figure 1, the tool is composed of two main components; the transmitter and the receiver. Suppose the network packet loss of the one way path between two computers over the Internet is to be estimated. Then, the transmitter component should run on the first computer while the receiver component should run on the second computer. Each one is composed of two different modules, TCP and UDP modules. The transmitter component operates as a TCP and UDP sender, while the other component operates as a TCP and UDP receiver. The TCP sender role is to establish the connection through the handshaking process shown in Figure 2. The handshaking packet will notify the receiver of the probing settings that will be used by the transmitter, mainly, the *Notification Period (NP), the Probing Window (PW), and the Probing Sleeping Period (PSP)* defined later, After the TCP receiver receives the handshaking packet, it will accept this connection by sending an acknowledgment signal to the transmitter, then the UDP sender at the transmitter starts probing UDP packets according to the probing settings configured at the transmitter. These settings include the NP, PW, PSP, and also the size of the probing packets. The NP defines the time period that the receiver will wait to receive the probed packets sent by the transmitter. At the end of this period, the receiver will estimate the packet loss and send it to the transmitter. In fact, the receiver is able to estimate the packet loss since the transmitter during the handshake stage sent how many UDP packets will be transmitted during this probing period. The number of packets sent during a probing period is defined as the PW. After the transmitter sends the UDP packets as defined in PW, it sleeps for a period of time defined as the Probing Sleeping Period. The PSP is simply the time duration in which the transmitter will sleep and stop from probing packets, after this time elapses, the probing period ends and the transmitter will record the reported packet loss sent by the receiver during the probing period, and then repeat the probing process again (i.e. sends a TCP handshaking packet and then start sending UDP packets). Notice that varying the PSP is inversely proportional with the probing rate, which is defined as the number of probed packets during the probing period. Consequently, increasing the PSP while fixing the PW will reduce the probing rate and vice versa. For example, if PW was set to 5, and the receiver only receives 4 packets during NP, then it will report 20% packet loss estimation for this probing period.

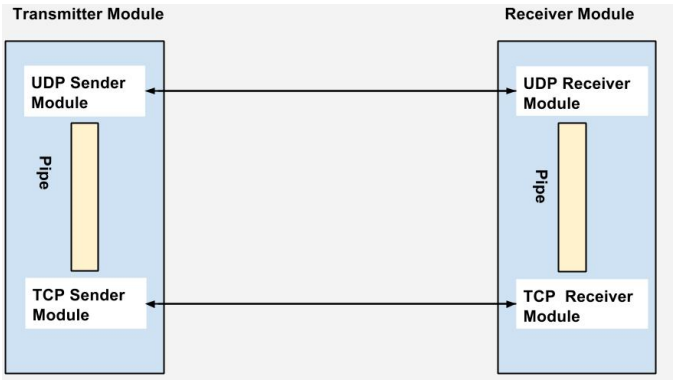


Figure 1: The NetProb software architecture

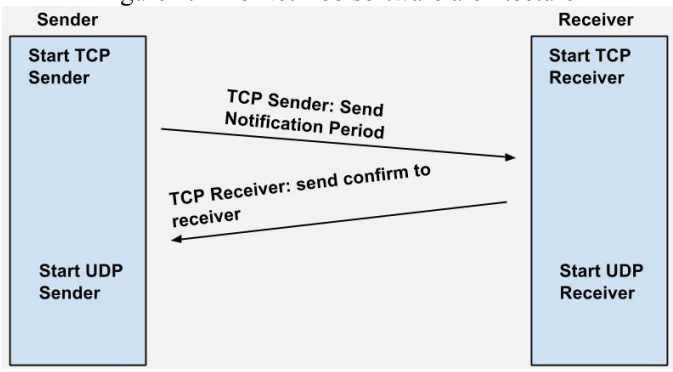


Figure 2: NetProb Handshaking steps

Figure 3 depicts the steps after the transmitter and receiver initialized the connection through the handshaking process, through which NP, and PW were both set to 5. After that, the transmitter starts sending the UDP probing packets to the receiver (Step 1), then it sleeps for 5 seconds (PSP = 5), while the receiver node waits for a period of NP seconds (5 in this example), it counts the number of received UDP packets, estimates the packet loss, and sends its estimation to the transmitter using a reliable TCP connection (Step 2). The transmitter then saves the received packet loss ratio, and also calculates the overall average loss ratio for the all probed packets (Step 3). Notice that this value will be the same as the one received by the receiver for the first probing period, but after that, it will be the average of the subsequent estimated packet loss for each probing period. Then the sender will start a new probing period after the PSP (5 seconds in this example) finishes (Step 4). Once this step is finished, the NetProb finishes one probing period, and restart another period by issuing a new handshaking packet to the receiver node. Notice that this combination of both TCP and UDP assures synchronization between the transmitter and the receiver.

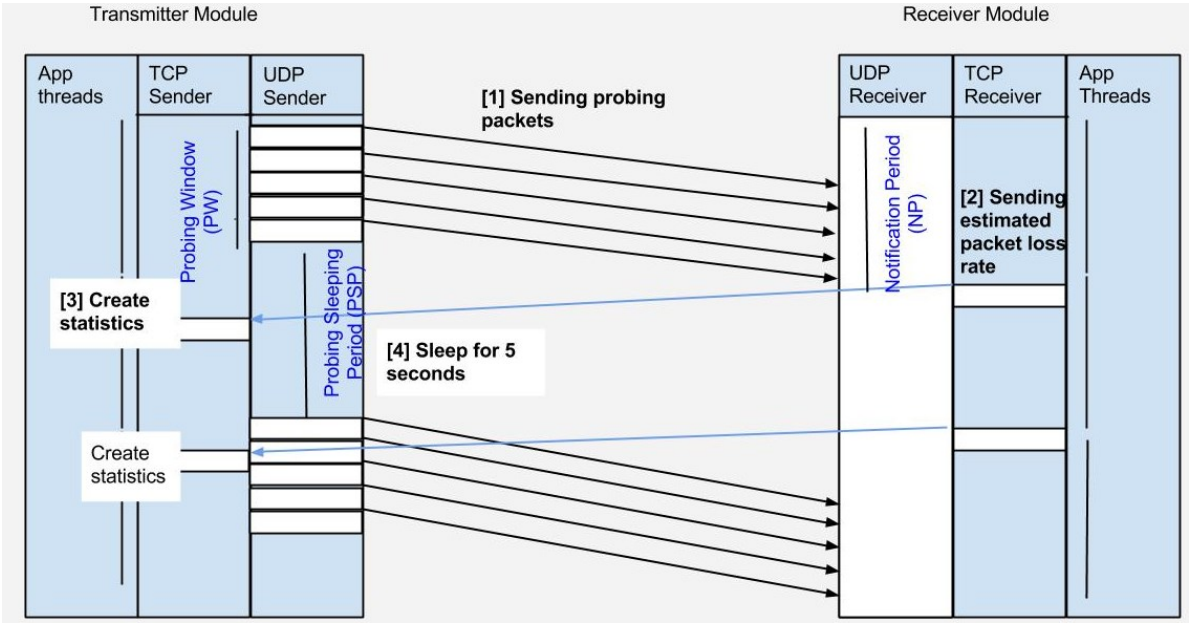


Figure 3: NetProb operation steps for one probing period

B. *Probing modes*

NetProb is designed to operate in two different modes; the first one is the fixed probing mode, in which the same probing settings are used for all probing periods. The other mode allows NetProb to probes packets non-uniformly; hence packet probing can follow certain statistical distribution. In this mode, NetProb changes its parameters on each probing period. So it can have different number of probing packets, probing packet size, and different sleeping periods between probing rounds. Figure 4 shows an example of non-uniform probing, as shown in the figure, the tool is configured to use a PW of 5 packets, and PSP of 3 seconds on the first probing period, while the values are changed in the second probing period to 4 packets, 6 seconds for the PW, and PSP, respectively.

C. *NetProb implementation*

NetProb is built using Python under Microsoft Windows Operating system. It provides the user with simple GUI, which was built using the Tkinter library [11]. The tool has three main threads, the main application thread, which initiates the tool and also starts the other two threads; the TCP and the UDP threads, these two threads open a TCP, UDP sockets, respectively between the sending and the receiving nodes. Notice also since the receiving node has to report the estimated packet loss for each probing period to the sending node, a reliable connection using TCP is used for this purpose, as such, the receiving nodes has to establish another TCP socket with the transmitting node. Further, since the program utilizes both TCP and UDP sockets, Inter-Process Communication (IPC) is performed by using pipes as shown in Figure 1.

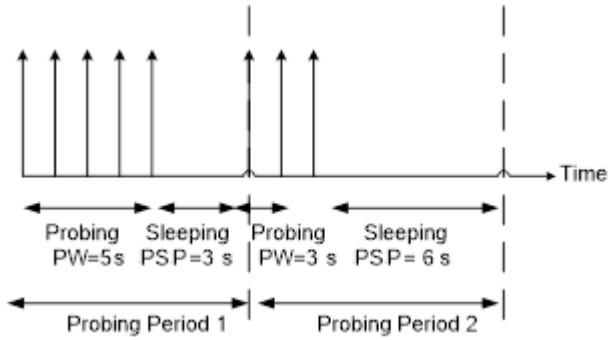


Figure 4: NetProb non-uniform probing mode

III. TEST-BED SETUP

In order to verify the tool performance and accuracy in predicting packet loss, a test-bed shown in Figure 5 is built. As depicted in the figure, the test-bed consists of two computers (laptop 1 and laptop 2), both running Microsoft Windows 7 operating system, the two computers are connected together through a direct Ethernet cable. However, in order to emulate the Internet behavior in a controlled way, the Microsoft Network Emulator Toolkit NEWT [12] is installed at one of the two computers; laptop 1 in this figure. The emulator is capable of emulating different characteristics of the Internet such as packet loss and transmission delay using different statistical models. More details about the emulator settings are provided in the following subsection.

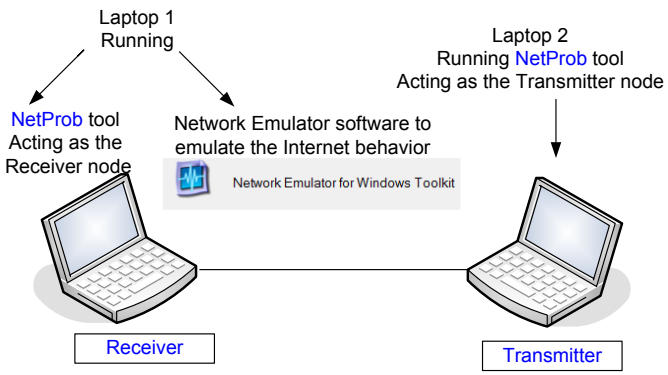


Figure 5: Test-bed setup

A. *Network emulator settings*

The emulator is used to emulate the Internet packet loss in a controlled way. In this paper, the Gilbert Eliot (GE) packet loss model is used to emulate packet loss [13]. As shown in Figure 6, this model uses two state Markov chains to describe the lossy characteristic of the Internet. It characterizes the network into two states: a Good state (**G**), which represents a less congested network situation with probability of packet loss equals to (1 - K), which has small values, while the other Bad state (**B**) represents a highly congested link with high probability of packet loss equals to (1 - H). Further, the network changes its status from Good to Bad and vice versa with a transition probabilities P and R, respectively. All these probabilistic parameters of the GE model determine the average packet loss rate.

Finally, the emulator can provide a background traffic emulation with certain transmission rate, we used the Pareto On/Off probabilistic model [14] with a shape factor, packet size, traffic rate, burst-time, idle- time equals to 1.5, 500 Bytes, 1Mbps, 1 second, 1 second , respectively.

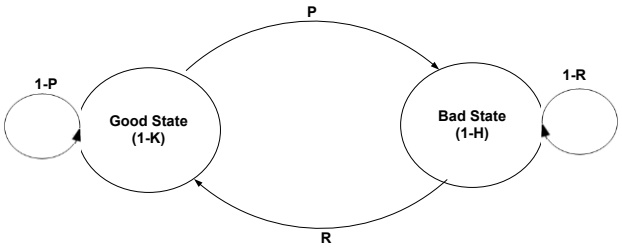


Figure 6: Gilbert Eliot packet loss model

B. *NetProb settings*

The tool provides the user with many various parameters that determine the probing behavior such as: probing packet size, number of probed packets in the probing window, idle probing time or the sleeping period, the notification period, and finally and the total probing time. In the following experiments, we tried different values of the above parameters. The Probing Accuracy (PA) is used in estimating the network packet loss accuracy which is calculated using Equation (1):

$$PA = (1- abs (L_{avg}- L_{est}) / L_{avg}) * 100\% \qquad (1),$$

where L_{avg} is the average packet loss set by the NEWT emulator, which can be controlled using the GE loss model parameters, and L_{est} is the estimated packet loss predicted by the NetProb tool.

IV. PEFORMANCE EVLAUATION

Several experiments have been conducted to verify the tool operation and it's capability in predicting the packet loss of the end-to-end path between a source and a destination node. In the conducted experiments, we varied the parameters of GE loss model to obtain various packet loss rates. Table I shows one of these experiments where the following parameters have been used to setup an average packet loss of 10% on the GE model (1 - K = 0.01, 1- H = 0.1, P = 0.99, and R = 0.01). The NetProb tool has been configured to use a probing packet size of 500 bytes, the NP, and the PSP are both set to 1 second, the PW was set to 5 packets. Notice that the optimal choice of these parameters is still one of our ongoing research works. As depicted in Tables I, II, III, and IV, the NetProb estimation accuracy was very high even for short probing times and for various average packet losses (5%, 10%, 15%, and 20%) which proves the effectiveness of the tool in packet loss estimation within a short operation time.

Table I: Probing accuracy when the average packet loss approximately equals to 5% under various probing durations. GE model settings are (1 - K = 0.01, 1 – H = 0.05, P = 0.99, and R = 0.01)

Probing time (s)	L_{avg}	L_{est}	PA (%)
100	4.85	5.14	94.0
120	5.20	5.78	88.8
140	5.24	5.76	90.1
160	4.95	5.34	92.1
180	4.95	5.19	95.19
200	4.96	5.07	97.7

Table II: Probing accuracy when the average packet loss approximately equals to 10% under various probing durations. GE model settings are (1-K =0.01, 1-H=0.1, P= 0.99, and R= 0.01)

Probing time (s)	L_{avg}	L_{est}	PA (%)
100	9.5	9.6	98.9
120	9.7	10.8	88.6
140	9.8	10.48	93.1
160	9.5	9.5	100
180	9.64	10.55	91.0
200	9.75	10.5	92.30

Table III: Probing accuracy when the average packet loss approximately equals to 15% under various probing durations. GE model settings are (1 - K = 0.01, 1- H = 0.15, P = 0.99 , and R = 0.01)

Probing time (s)	L_{avg}	L_{est}	PA (%)
100	14.9	13.6	91.3
120	15.0	16.16	92.3
140	14.7	15.5	94.5
160	14.6	15.5	93.8
180	14.5	13.66	93.8
200	14.79	14.77	99.9

Table IV: Probing accuracy when the average packet loss approximately equals to 20% under various probing durations. GE model settings are (1 - K = 0.01, 1 – H = 0.2, P = 0.99, and R = 0.01)

Probing time (s)	L_{avg}	L_{est}	PA (%)
100	18.8	19.0	98.9
120	19.95	20.0	99.74
140	20.1	22.5	88.1
160	20.05	22.5	87.5
180	19.97	21.56	92.0
200	19.78	20.9	94.3

V. CONCLUSION AND FUTURE WORK

In this paper, a network probing tool based on the TCP and UDP protocols has been proposed, where TCP is used only for connection control while UDP traffic is used for packet loss estimation. The tool showed accurate and fast estimation for the network packet loss, which makes it a good choice for estimating packet loss for real-time multimedia applications. As a future work, we are currently working on studying the optimal settings of the tool parameters such as the probing rate, the probing packet size, the probing period duration, etc., that give the best estimation accuracy with minimum probing overhead. We are also working on making the tool capable of estimating other network parameters such as the end-to-end delay and the bandwidth. Finally, we are looking forward receiving feedback from the research community after using the tool, which will be taken into consideration for further developments and improvements.

REFERENCES

[1] O. E. Adebomi and J. S. Mwela, "Impact of packet losses on the quality of video streaming, " M.Sc. Thesis, School of Science in Electrical Engineering, School of Computing at Blekinge Institute of Technology, 2010.

[2] A. Khalifeh and H. Youse'zadeh, "Optimal audio transmission over error-prone wireless links, " IEEE Transactions on Multimedia, 12(3):204{214, April 2010.

[3] F. Sabrina, and J.-M. Valin, "Priority based dynamic rate control for VoIP traffic, " in Proceedings of Globecom, 2009.

[4] A. Fonte, M. Curado and E. Monteiro, "A prediction criterion for selecting popular destinations," in Proceeding of GLOBECOM Workshops 2011.

[5] A. Khalifeh and A. El-Mousa,"QoS Routing of VoIP using a modified widest-shortest routing algorithm," IEEE/ACS International Conference on Computer Systems and Applications (AICCSA), 2007.

[6] A. Khalifeh and A. Elmousa, "Performance evaluation of VoIP using shortest-widest and modified widest-shortest QoS routing algorithms", International Conference of Computer Science and Engineering (ICCSE), 2007.

[7] H. Ningning and P. Steenkiste, "Evaluation and characterization of available bandwidth probing techniques," IEEE Journal on Selected Areas in Communications, vol.21, no.6, pp.879, 894, Aug. 2003.

[8] P. Benko and A. Veres, "A passive method for estimating end-to-end TCP packet loss," in Proceeding of IEEE Globecom 2002.

[9] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson. "User-level Internet path diagnosis," in Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP '03).

[10] S. Savage, "Sting: a TCP-based Network Measurement Tool, " in Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems (USITS) 1999.

[11] https://wiki.python.org/moin/TkInter [last accessed July, 1, 2014].

[12] Emulator Windows Toolkit, NEWT http://network-emulator-for-windows-toolkit.software.informer.com/ [last accessed July, 1, 2014].

[13] H. Gerhard and H. Oliver, "The Gilbert-Elliott model for packet loss in real time services on the Internet," in Proceedings of Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008.

[14] X. Yang, "Designing Traffic Profiles for bursty Internet traffic", in Proceedings of IEEE Global Internet, 2002.

Improvements in failure detection for emergency service centers in IMS network

Ivan Krevatin*, Željko Presečki*, Mate Gudelj**
* Services, Platforms and Core Network Development Department
**IMS Shared Service Center Hub Department
Hrvatski Telekom
Zagreb, Croatia
ivan.krevatin@t.ht.hr, zeljko.presecki@t.ht.hr, mate.gudelj@t.ht.hr

Abstract— **Migration of customers from PSTN to IMS network brings challenges in replacing the existing services with equal services on the IMS network. Emergency service centers are special types of customers that require high availability of the service and fast responses to potential failures. Differences in technology require different approaches for handling same requirements. This paper describes the method for detection of failures on emergency service centers in IMS network that we have implemented in our network. The method is based on charging events for the detection of alternative call routing cases in the situations of unavailability. In the paper we describe the motivation for such approach and analyze the results and types of failures we encountered. In addition, the paper provides wider aspects of the implementation of the emergency service centers in IMS network.**

Keywords: emergency call handling; IMS; failure detection; PSTN migration;

I. INTRODUCTION

Migration of customers from PSTN (Public Switched Telephone Network) to IMS (IP Multimedia Subsystem) network brings challenges in replacing the existing services with equal services on the IMS network. Emergency service and emergency service centers are one of the most delicate services and customers to migrate. Apart from the most important requirement of keeping the service availability and minimizing the risk of outages during migration, emergency service centers are treated as very important customers and often special solutions are designed for them. In this paper we are describing the solutions for migration of emergency service centers that we have developed in IMS network with the focus on the improvements in failure detection and the analysis of the results. IMS network standardization bodies have recognized the need for special handling of emergency sessions and have dedicated technical specifications [1] addressing this subject. Implementation of emergency call handling architecture on open-source project Open IMS is described in [2] with the focus on providing a test bed for verification of technical specifications. An overview of challenges in implementation of emergency call handing in production IMS network is given in [3]. As we can see from the papers, there are several topics that should be addressed when implementing and migrating emergency call centers from PSTN to IMS network. While localization of customers, selection of nearest PSAP (Public Service Answering Point) and recognition and prioritization of emergency calls are handled by the IMS network, mechanisms for alternative routing of calls in case of unavailability is unlike to PSTN not supposed to be handled by the IMS network thus it should be handled differently e.g. on application server (AS) using call forwarding service.

In this paper we are describing the configurations of two types of emergency call centers that are migrated to IMS: emergency centers with local PBX (Private Branch Exchange) connected over ISDN BRA and emergency centers with local analogue PBX where PBX functionality is provided by the network. On each type of access different causes of failure can happen resulting in service unavailability which triggers alternative call routing that is registered in charging events. These charging events can be used as triggers for alarming the network operations center which can then take necessary steps to determine and eliminate the cause of failure. Specific to emergency call centers is that they are used primarily to receive calls therefore it is much harder for them to detect the faulty line i.e. it is the caller who first encounters the faulty line. For this reason alternative call routing and network detected failure play very important role in emergency call handling.

The paper is organized as follows. Section II describes the principles of configurations of emergency call centers on IMS and the idea of the failure detection mechanisms. Section III describes the architecture of the failure detection mechanisms. Section IV analyzes the results and provides the insight in the types of encountered failures, conducted analysis and ideas for improvements. Section V summarizes the conclusions of this paper.

II. EMERGENCY CALL CENTER IN IMS

IMS network is a telecommunication network that is part of a layered architecture and unlike to PSTN doesn't have dedicated last mile connection but relies on IP connectivity which is expected to be provided by lower layers. Multi layered approach brings more elements in consideration when it comes to possible failure points and it also brings split responsibility for failure detection and remedy. In this chapter we are describing configurations of emergency centers in our network and possible types of failures.

A. Types of emergency call centers

Figure 1 shows ISDN PBX connected to PSTN network. This is a typical configuration of medium sized emergency call center in Croatia. Up to four ISDN links from operator's telephone exchange are installed at customer's premises. Each of the links is terminated in one ISDN network termination (NT) equipment. The emergency call center connects its own PBX to the NTs. Usually a recording device is also placed in between the PBX and the ISDN NT. Typically regular phones connected to the PBX are used for receiving and placing calls.

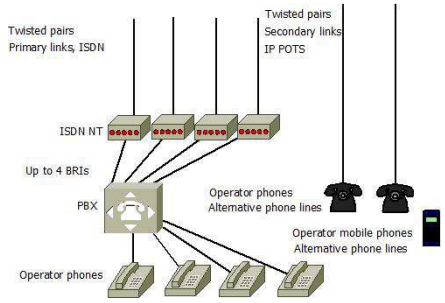


Figure 1 Emergency call center in PSTN with ISDN BRA PBX

On the right side of the figure alternative lines for receiving emergency calls are shown. In some emergency call centers alternative lines are placed at the same location as primary lines i.e. they serve only as backup lines when primary lines are not working. In other emergency call centers, alternative lines are actually primary lines of another emergency call center i.e. in case of failure another emergency center beside its own traffic handles also the traffic of the center which is in failure.

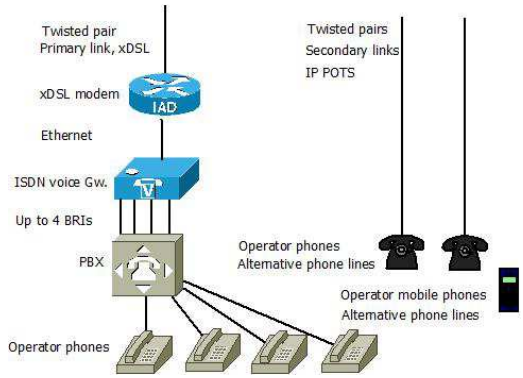


Figure 2 Emergency call center in IMS with ISDN BRA PBX

Figure 2 shows the same emergency call center migrated to IMS network. The difference on the customer premises is that the ISDN NTs are replaced with the xDSL modem and ISDN voice gateway (GW). The xDSL modem is connected to the operator's network over a single twisted pair cable meaning that all other lines used for ISDN connection became obsolete after the migration. The role of the xDSL modem is to provide the IP connection to the ISDN voice GW which then provides the ISDN BRI interfaces to the customer's PBX and the SIP user agent functionality for registration and communication with the IMS network. ISDN GW is the main element at customer's premises where operator's configuration and customization for proper functioning of the service is configured.

Another type of emergency service center is shown on Figure 3. Customer uses the analogue PBX which is connected over analogue lines (POTS lines) to operator's telephone exchange and has several phones connected to it. The operator's telephone exchange provides the functionality of serial distribution of incoming calls to analogue lines.

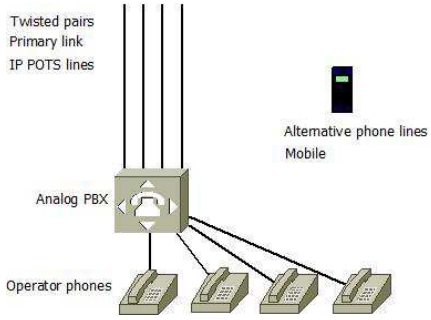


Figure 3 Emergency call center with analogue PBX

In this scenario when migrating the emergency call center form PSTN to IMS nothing changes on the last mile for the customer. Nevertheless, from the network point of view the differences are huge. The entire service logic has to be emulated by the IMS network. In IMS network the serial distribution of incoming calls and other call features are provided by the application server (AS). The emulation of POTS interface is provided by the multi-service access node (MSAN) which is a node located in operator's access network that acts as a SIP user agent towards the IMS network.

Besides the just described configurations, there are some special call features required from emergency call centers and that need to be developed especially for their purpose i.e. such features are not provided to regular customers. Example of such features is the type of call transfer which is used by the 112 center in Croatia where the transferred emergency call from 112 center to another emergency center is actually a new call which carries the identity presentation of the original caller. Requirements of this kind are handled by special configurations on the AS, ISDN GW and on session border controller (SBC). SBC as the first node in the access to the IMS network handles also the requirements for the prioritization of calls.

To be able to efficiently handle these specifics, in our implementation, we have introduced emergency subdomain in which we have separated all emergency call centers from regular IMS users. Using a subdomain as a differentiator of emergency call centers a special call handling and many dedicated features to emergency call centers can be applied. One of such features that we implemented is triggering alarms in case when a call arrives to an emergency call center which is for some reasons not reachable. The architecture of such

mechanisms and benefits of implementation are explained in next chapters.

B. *Failures and possibilities for failure detection*

High availability is crucial when designing a call handling solution for emergency call centers. Obviously, failures cannot be completely avoided therefore it is important to detect them as soon as possible. By failure we are referring to any circumstance that is preventing the emergency call center to receive or place calls. Usually it can be caused by faulty equipment, wrong configuration of equipment or connectivity problems.

In PSTN network failure detection on the line or in the communication with the PBX is much easier than it is in the IMS network. This is especially the case on the ISDN where such features are more evolved end it is easier to monitor the lines. If we compare the installation of the emergency call center with ISDN BRA PBX in IMS network to the installation in the PSTN network we see the higher risk of failure that is introduced with additional equipment and lower number of physical lines (twisted pairs) at customer premises. This doesn't refer only to faulty or misconfigured equipment, but also to human factor on the site which can result in accidental disconnections and/or wrong connections between elements. Additionally, instead of dedicated telephony equipment (ISDN NT), we are placing the multipurpose equipment (xDSL modem). The ISDN NT equipment follows more rigorous standards for stability, protocol adherence and quality of implemented maintenance functions. That, in turn, results with better fault detection system.

In the case of analogue PBX configuration compared to the ISDN BRA PBX configuration advantages in failure detection are in the fact that there is no new and additional equipment at customer premises and that SIP user agent is placed on the MSAN node which is located in central office. Since MSAN node is part of operator's network equipment, placed in operator premises and serves many customers it is integrated into operator's real-time monitoring system. This brings much higher probabilities for faster fault repair. Unfortunately, real-time monitoring of xDSL modems and ISDN GWs is not very practical because of their very large number in the network and the fact that a lot of customers switch them off on purpose, e.g. out of working hours for preventing electricity consumption.

In IMS network failures can be detected based on SIP response codes which are generated as responses to SIP requests for sessions, registration, re-registration etc. Failures can be detected on network side or by the SIP user agent. When the SIP user agent receives the error response code on registration or on session initiation attempt or it doesn't receive the response at all, it usually tries to reattempt to register over the alternative SIP proxy which it has been configured in it. On the other side the IMS network can only detect if the SIP user agent is registered and if the registered user agent properly responds to SIP requests. There's no way the network can force or ping an unregistered user agent to register. That is because SIP registration is the mechanism for associating the user agent with its location and unregistered user agents remain unavailable to the network. Since SIP registration is periodical activity, possible failure situation is when the network sees the registered user agent which is in fact unavailable due to network conditions and cannot respond to SIP requests. In such cases the user agent is not aware of session requests or any problems with the network until it attempts to re-register. Such and other unavailability cases have to be handled by the network in the way to provide the routing to an alternative emergency destination, since the primary is not reachable. In our case, we are doing the alternative routing on the IMS application server which triggers the call forwarding service on not reachable user agent. Such functionality should be only operator administered since is vital for proper functioning of the emergency service. It is important to note that such call forwarding for emergency call centers should be free of charge therefore special charging logic for such scenarios should be developed.

III. FAILURE DETECTION SYSTEM ARCHITECTURE

The architecture of the failure detection system is shown on Figure 4. The figure shows the IMS network with the application server (AS) which is providing the services to the emergency call centers. In our network we are using two types of AS which are handling the emergency call centers. The type of AS is determined by the type of access (MSAN or xDSL) to which the emergency call center is connected. Nevertheless, the alarming logic is the same for both types of AS. Each AS is connected to the charging mediation function which analyzes offline charging records and sends the alarm events to the network monitoring system.

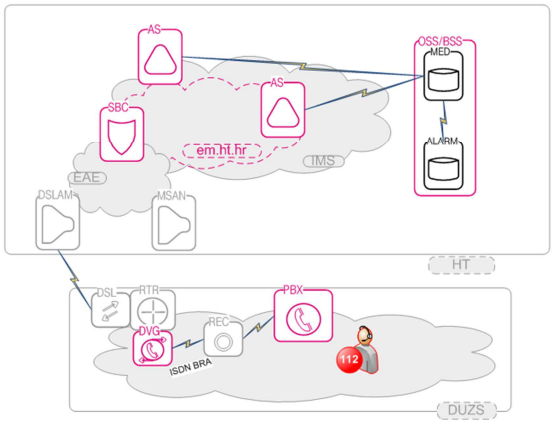


Figure 4 Architecture of the alarming system

The charging mediation logic is implemented in the way to detect a forwarded call of the IMS users which are in the emergency domain i.e. the emergency call centers. During each call the AS sends charging account records [4] to the charging system which is performing real-time analysis of such records. Such records, among other, contain information about the called domain, indication that the call has been forwarded, type of call forwarding service and the reason for forwarding the call. Alarm is triggered when the forwarded call from emergency call center is finished.

IMS as a network can only detect failures of this type based on SIP response error codes which are generated by the involved network elements. Other ways for detecting such failures would be to monitor SIP signaling and filter desired SIP response codes, or to trigger alarms directly from the AS. Since we are interested only in particular SIP response error codes and only for particular IMS users and call scenarios, we found it more practical to implement the alarm generation logic on the charging system. Generally, unavailability of user equipment will be more frequent when the IMS network will serve also mobile devices in VoLTE scenario, meaning that unlike with fixed terminals for mobile terminals is a normal behavior that it is switched off or located in the area with no network coverage.

IV. Analysis of results

After the implementation of the described alarming system we have performed initial analysis of the alarms occurred over the period of two months. The goals were to find issues which are causing the alarms and to establish a solid fault repair procedure which will be triggered when an alarm occurs. First results showed a lot more alarms that are occurring than expected. Such amount of alarms required to search for alarm occurrence patterns and to perform alarm filtering to eliminate alarms which were not caused by unavailability of endpoints.

Correlation between alarms and all calls placed towards specific emergency call center showed two types of alarm occurrence: (*i*) periodic, where alarms are occurring between successful calls and (*ii*) continuous, where alarms are occurring on each call attempt. Continuous are most problematic from the emergency call center perspective but are also most easy to detect and to find the problem and a solution it is just matter of reaction time and problem type. On the other hand, periodic failures are most difficult to detect, and often it is most difficult to find the cause of the problem. It is in those cases where our failure detection system brings more value. Detecting a periodic failure from emergency call center is difficult because its primary role is to receive calls. Our emergency call handling configuration provides alternative routing in cases of unavailability of one emergency center. That means that the user always reaches another emergency call center (or even the same emergency call center but over the alternative line) if the primary is not reachable so it is not experiencing service outage and doesn't notice that there is a reason to report a problem. This brings us to the conclusion that such types of failures should be detected by the network and with dedicated failure detection logic.

Simulation of alarms in test environment showed that the MSAN generates busy response code in some cases of failure on its POTS line e.g. short circuit, contact of the line with ground etc. For this reason we have configured the application logic to forward the calls from the analogue PBX on the alternative mobile device even in the busy case. This situation in the alarm analysis introduced a lot of confusion since there was no way to distinguish the failure from the regular busy case ad such alarms had to be discarded.

Analysis of the remaining alarms showed that most commonly following types of failures are:

- Wrong connections of PBX and ISDN GW. ISDN GW is equipped with four BRI interfaces may not all be configured. How many ports are configured depends on the installed capacity, usually it is just two. In such cases it can happen that the user for some reasons disconnects the PBX from the right port on the ISDN GW and connects it to the port that is not configured resulting in reduced capacity (two instead of four ISDN channels). This type of misconfiguration cannot be easily detected by the user himself.
- Wrong configuration on the ISDN GW resulting in reduced capacity.
- Unregistered SIP user agent(s) on the ISDN GW.
- Defects on the BRI link between the ISDN GW and the PBX resulting in instabilities of the link meaning that the link is irregularly up or down.

Some of these failures would have been difficult to detect without the implemented alarming system. Additionally, during the analysis of endpoint configurations some other misconfigurations have been detected like the one on the ISDN GW which resulted in the inability of making outgoing calls. Repairing detected failures resulted in lower number of alarms.

V. Conclusions

In this paper, we have described the configurations of emergency call centers in IMS network which are based on analogue PBX and ISDN BRA access points, the implementation of the failure detection system based on charging events and provided the analysis of alarms generated by this system. Implemented alarming system proved to be useful for detecting failures that would otherwise be very difficult to notice. Most of the detected failures showed that there was a problem on the end point which resulted in repairing action of the endpoint configuration meaning the ISDN GW or the customer PBX. Further work on the subject is to check each endpoint which is generating occasional alarms and to apply alarm filtering for the busy case. In this way we will reduce the number of alarms and be able to setup a real-time reaction procedure in the case when an alarm occurs. Additional improvements in the alarm logic could be to detect the occurrence of the subsequent successful call for alarm seizure or classification. ISDN PRA emergency center access points were not described in this paper because at the moment they are not implemented in our network, but the alarming logic also for these types of access points remains the same.

Analysis also showed that there is a need for alternative routing of emergency calls in cases of unavailability of endpoints and that it increases the general reliability of the emergency call handling.

REFERENCES

[1] ETSI, Universal Mobile Telecommunications System (UMTS);LTE;IP Multimedia Subsystem (IMS) emergency sessions, TS 123 167 V11.9.0, December, 2013

[2] A. A. Onofrei *et ali.*, „Emergency Services in IMS“, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Volume 46, pp 643-652, 2011

[3] I.Krevatin, I.Viskovic, “Emergency Communication in IMS network: challenges in implementation” MIPRO 2014.

[4] ETSI, IP Multimedia Subsystem (IMS) charging (3GPP TS 32.260 version 12.6.0 Release 12, December 2014.

BER Analysis in Relay-Based DF Cooperative Diversity Systems with Relay & Destination Interferers

Amer M. Magableh[1], Mamoun F. Al-Mistarihi[1], Rami Mohaisen[1], and Ashraf H. Sharaqa[2]

(1) Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
ammagableh@just.edu.jo, mistarihi@just.edu.jo, rami.fayez@hotmail.com

(2) WorleyParsons / Communication and Security Projects Division, Al-Khobar, Saudi Arabia
elengashraf@yahoo.com

Abstract - **Incremental relaying has been recently introduced to overcome the problem with the restricted resources utilization associated with fixed relaying schemes in a pre-determined manner. Only when the direct path is not usable, the incremental relaying exploits the feedback from the destination to inform the relay about the necessity to send the source signal. In this paper, the performance analysis in terms of the bit error rate (BER) of the decode-and-forward (DF) incremental relaying over mixed fading channels is provided. The effects of the Co-channel interference near the destination and the relay caused respectively by the multiple L and N interferers are presented. Numerical results are provided to prove the solidity of our assumptions and compare the results of the system with and without the effects of the interference at the relay.**

Keywords: ***Incremental relaying, cooperative diversity, decode-and-forward, interference, bit error rate (BER), signal to noise ratio (SNR), signal to interference ratio (SIR), maximal ratio combiner (MRC).***

I. INTRODUCTION

The deployment of the cooperative diversity has been considered due to its appealing attributes in term of extending the battery life of the communication devices, extending the coverage area and the spatial diversity gain at the destination [1-3]. Cooperative wireless networks are implemented by placing the intermediate nodes between the sending node (the source) and the receiving node (the destination) in an attempt to enhance the reception of the source signal at the destination.

Incremental relaying came into play to mitigate for the deficiencies of the fixed schemes by saving the channel resources in which the relay will get involved only when the direct path is not usable [4].

Performance analysis of relay systems with interference at the relay was presented in [5]. Closed-form expressions for outage probability and BER in relay-based cooperative diversity systems over different fading channels with the presence of multiple interferers near the destination were introduced in [6-7].

In this paper, we discuss the decode-and-forward incremental relaying cooperative diversity by introducing an intermediate single relay and multiple L and N interferers located near the destination and relay, respectively, which will affect the performance of the system in response to the Co-channel interference over mixed fading channels between the source-destination and the source-relay-destination paths.

This paper is further presented as follows; Section II provides the system model. Section III provides the performance analysis and derived tight closed form formula of bit error rate. Results and discussion are presented in section IV. Finally we concluded this paper in section V.

II. SYSTEM MODEL

In Fig. 1, the channel coefficient between the source and the destination is assumed to follow the Rayleigh distribution whereas the Channels coefficients between the source-relay and the relay-destination are assumed to follow the Nakagami-m distribution. The relay will participate with the source in transmitting its signal to the destination upon the reception of the destination's negative feedback that request the relay to send the source signal within the second time slot. The relay then will decode the source signal, re-encode it and then send it to the destination.

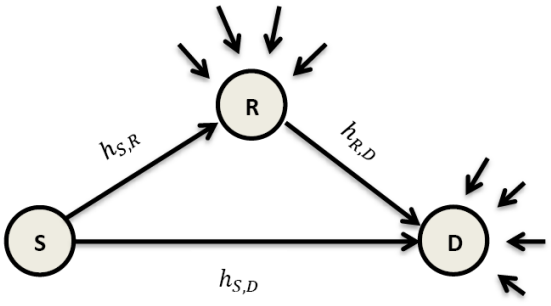


Fig. 1 Cooperative Diversity Network with Multiple Interferers L and N Located near the Destination and the Relay, respectively.

The noise at the destination is assumed to be Additive White Gaussian Noise with a unity variance N_o. Beside that, near the destination and the relay, the channels from the L and N interferers respectively that cause a Co-channel interference

are assumed to be a sum of identically, independently and exponentially distributed Rayleigh random variables which consequently will degrade the overall performance of the system. All the nodes in the system are equipped with one single antenna and Time Division Multiple Access (TDMA) is used as the multiple access technique, it is worth to mention that Frequency Division Multiple Access (FDMA) can be used too.

Within the first time slot, the destination will receive a signal directly from the source and based on the quality of this signal, the destination will decide whether or not the relay should participate in the transmission. From instance, if the direct transmission was successful and the destination can properly process the source signal from the direct transmission, then a positive feedback from the destination will be sent to both the relay and the source respectively informing them about the successfulness of the direct transmission and requesting the source to send a new signal within the second time slot. Otherwise, if the destination failed to receive the source signal or the direct path was in outage, then the relay will be requested to resend the source signal within the second time slot which by then will be combined with the source signal at the destination using the MRC.

The received signals at both the relay and the destination within the first time slot are:

$$y_{S,D}(t_1) = h_{S,D}\sqrt{E_s}\, x(t_1) + n_1(t_1) + I(t_1) \quad (1)$$

$$y_{S,R}(t_1) = h_{S,R}\sqrt{E_s}\, x(t_1) + n_2(t_1) + I(t_1) \quad (2)$$

where $x(t)$ is the symbol signal with a normalized energy. $n_1(t_1)$ and $n_2(t_1)$ are the AWGN noise terms at the destination and E_s is the signal energy. The interference effects at the destination and the relay within the first time slot are represented by $I(t_1)$.

If the relay participates in the transmission within the second time slot, then the received signal at the destination from the relay is denoted as:

$$y_{R,D}(t_2) = h_{R,D}\sqrt{E_s}x_r(t_2) + n_3(t_2) + I(t_2) \quad (3)$$

where $x_r(t_2)$ is the regenerate signal at the relay within the second time slot . $I(t_2)$ is the interference effects near the destination within the second time slot.

The destination will assess the usability of the direct link by comparing the signal to noise ratio of the direct path against $\gamma_o$which is the minimal threshold for which the destination can receive the source signal merely depending on the direct path without the help of the relay.

In the interference environment, in most of the communication systems, rather than working on the SNR at the destination a transgression is made to merely work on the signal to interference ratio SIR since the noise effects near the destination is not comparable to that of the Co-channel effects and since the interference limited environment is a common practice in the literature.

In the following section, we will derive the building blocks in deriving the performance metrics, namely the probability density function (PDF) and the cumulative distribution function (CDF) of both the SIR at the relay from the source link and the SIR at the destination from both the relay and source respectively.

III. PERFORMANCE ANALYSIS

To derive one of the performance metrics, namely the BER, of the incremental relaying scheme, both the PDF and the CDF of the SIR at the destination from the source (direct path) and the relay need to be derived. Incorporating the effect of the Co-channel interference near the destination, it is a common practice in the literature to work merely in an interference limited environment due to the fact that the noise effects at the destination and the relay are not as severe as the effects of the interferers.

A. Probability Density Function of SIR's $\boldsymbol{\eta_{S,D}}$, $\boldsymbol{\eta_{R,D}}$ and $\boldsymbol{\eta_{S,R}}$

Given that the SNR of the direct path $\gamma_{S,D}$ follows the exponential distribution and the fading channels from the interferers to the destination ($\gamma_b = \sum_1^L \gamma_k$) is assumed to be a sum of identically, independently and exponentially distributed random variables. The PDF of the SIR ($\boldsymbol{\eta_{S,D}}$) at the destination from the source path is given as follows [6]:

$$f_{\eta_{S,D}}(\eta_{S,D}) = \frac{\varsigma_{S,D}}{(\lambda_{S,D}+\eta_{S,D})^{L+1}} \quad (4)$$

where $\lambda_{S,D} = \frac{\overline{\gamma_{S,D}}}{\overline{\gamma_d}}$, $\varsigma_{S,D} = L(\lambda_{S,D})^L$, $\overline{\gamma_d} = \frac{\mathcal{E}}{N_o}E({\alpha_k}^2)$ is the average SNR of each independent interfering channel, and $\gamma_b = \frac{\mathcal{E}}{N_o}\sum_{k=1}^{L}{\alpha_k}^2 = \sum_{k=1}^{L}\gamma_k$ where γ_k is the instantaneous SNR from the $k-th$ interfering channel and L is the number of interfering channels near the destination. $\gamma_k's$ are mutually and statistically independent .

The PDF of the SIR $\boldsymbol{(\eta_{R,D})}$ at the destination from the relay path is given as follows**:**

$$f_{\eta_{R,D}}(\eta_{R,D}) = \frac{\zeta_{R,D}{\eta_{R,D}}^{m-1}}{(\eta_{R,D}+\lambda_{R,D})^{L+m}} \quad (5)$$

where $\lambda_{R,D} = \frac{\overline{\gamma_{R,D}}}{m\overline{\gamma_d}}$, $\zeta_{R,D} = \frac{(L+m-1)!}{m^L\Gamma(m)(L-1)!}(m\,\lambda_{R,D})^L$

Following the same approach used in [6] to derive (4) and (5), the PDF of the SIR $\boldsymbol{(\eta_{S,R})}$ at the relay from the source is given as follows**:**

$$f_{\eta_{S,R}}(\eta_{S,R}) = \frac{\zeta_{S,R}{\eta_{S,R}}^{m-1}}{(\eta_{S,R}+\lambda_{S,R})^{N+m}} \quad (6)$$

where $\lambda_{S,R} = \frac{\overline{\gamma_{S,R}}}{m\overline{\gamma_r}}$, $\zeta_{S,R} = \frac{(N+m-1)!}{m^N\Gamma(m)(N-1)!}(m\,\lambda_{S,R})^N$

B. The Cumulative Distribution Function of SIR's $\boldsymbol{\eta_{S,D}}$, $\boldsymbol{\eta_{R,D}}$, and $\boldsymbol{\eta_{S,R}}$

The CDF of the SIR $\boldsymbol{(\eta_{S,D})}$ at the destination from the source path is given as follows [6]:

$$F_{\eta_{S,D}}(\eta_{S,D}) = \left[\frac{(\lambda_{S,D}+\eta_{S,D})^L - {\lambda_{S,D}}^L}{(\lambda_{S,D}+\eta_{S,D})^L}\right] \quad (7)$$

The CDF of the SIR $(\boldsymbol{\eta_{R,D}})$ at the destination from the relay path is given as follows [6]:

$$F_{\eta_{R,D}}(\eta_{R,D}) = \frac{\zeta_{R,D}{\eta_{R,D}}^m}{m(\lambda_{R,D})^{L+m}}\ {}_2F_1\left(L+m, m; m+1; -\frac{\eta_{R,D}}{\lambda_{R,D}}\right) \quad (8)$$

where ${}_2F_1(a,b;c;d)$ is the Gauss Hyper-geometric function [10].

Following the same approach used in [6] to derive (7) and (8), The CDF of $\boldsymbol{\eta_{S,R}}$ at the relay from the source path is given as follows:

$$F_{\eta_{S,R}}(\eta_{S,R}) = \frac{\zeta_{S,R}{\eta_{S,R}}^m}{m(\lambda_{S,R})^{N+m}}\ {}_2F_1\left(N+m, m; m+1; -\frac{\eta_{S,R}}{\lambda_{S,R}}\right) \quad (9)$$

C. Bit Error Rate Analysis

The incremental relaying average unconditional bit error rate is expressed as:

$$p_e = \Pr(\eta_{S,D} \le \gamma_o) \times P_{div}(e) + \left(1 - \Pr(\eta_{S,D} \le \gamma_o)\right) \times P_{direct}(e) \quad (10)$$

where the first term characterizes the case when the relay helps the destination in the transmission and the second term is the case without the help of the relay and this is emphasized by the complement of $\Pr(\eta_{S,D} \le \gamma_o)$ in the second term. $P_{div}(e)$ is the average probability that an error occurs in the combined diversity transmission from the source and the relay to the destination by the MRC where as $P_{direct}(e)$ is the probability of error at the destination given that the destination decides that the relay should not forward the source signal, in such a case the destination depends only on the direct transmission.

The conditional probability of error takes the form $\alpha\, Q\left(\sqrt{\beta\eta_{S,D}}\right)$ with α and β represent the constellation parameters. For BPSK, α and β are 1 and 2 respectively, for M-PSK, $\alpha = 1$ and $\beta = 2\sin\left(\frac{\pi}{M}\right)^2$, for M-QAM, $\alpha = 4$ and $\beta = \frac{3}{(M-1)}$, $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{\left(-\frac{y^2}{2}\right)}dy$ is the Gaussian Q-function [8].

The approximated expression of $P_{direct}(e)$ was derived in [6] which is given as:

$$P_{direct}(e) \approx \alpha\rho \sum_{i=1}^{2} A_i \left\{ e^{-\gamma_o a_i \beta} \sum_{k=1}^{L} \frac{(k-1)!(-a_i\beta)^{L-k}}{(L)!(\gamma_o+\lambda_{S,D})^k} - \frac{(-a_i\beta)^L}{(L)!} e^{a_i\beta\lambda_{S,D}} E_i\left[-a_i\beta(\gamma_o+\lambda_{S,D})\right] \right\} \quad (11)$$

where $E_i[x]$ is the Exponential Integral function [10, (8.211.1)].

When the relay helps in relaying the source signal, the MRC at the destination will combine the two signals and the probability of error $P_{div}(e)$ is expressed as [6]:

$$P_{div}(e) = p_{SR}(e)p_x(e) + (1 - p_{SR}(e))p_{com}(e) \quad (12)$$

where $p_{SR}(e)$ characterizes the probability of error at the relay and $p_x(e)$ is the probability of error at the destination given that the relay decoded unsuccessfully and is bounded by 0.5 as stated in [11] and [12]. The term that represents the error probability at the destination given that the relay decode correctly is $p_{com}(e)$.

$p_{SR}(e)$ can be expressed as:

$$p_{SR}(e) = \alpha \int_0^\infty f_{\eta_{S,R}}(\eta_{S,R})\, Q\left(\sqrt{\beta\eta_{S,R}}\right) d\eta_{S,R} \quad (13)$$

using the Prony's approximation of the Q-function, $p_{SR}(e)$ can be expressed as:

$$p_{SR}(e) \approx \frac{\alpha\zeta_{S,R}}{\sqrt{2\pi}\Gamma(N+m)(\lambda_{S,R})^{N+m}}\left(\frac{2}{\beta}\right)^m \times G_{3,2}^{1,3}\left(\frac{2}{\beta\lambda_{S,R}}\middle|\begin{matrix}1-N-m, 1-m, \frac{1}{2}-m \\ 0, -m\end{matrix}\right) \quad (14)$$

where $G_{p,q}^{m,n}\left(x\middle|\begin{matrix}a_1,\ldots,a_n \\ b_1,\ldots,b_q\end{matrix}\right)$ is the Meijer's G-function [10,(9.3)].

For the second event, if the relay decoded the source signal correctly, then the spatial diversity will be obtained once the destination combines the received signals from both the source and the relay and this is emphasized by the complement of $p_{SR}(e)$ which is $(1 - p_{SR}(e))$ confirming the success of the decoding process at the relay, in this case, the error probability $p_{com}(e)$ will be expressed as [6]:

$$p_{com}(e) \approx \alpha \int_0^\infty f_{\eta_{R,D}}(\eta_{R,D}|\eta_{S,D} \le \gamma_O)\, Q\left(\sqrt{\beta\eta_{R,D}}\right) d\eta_{R,D} \approx \frac{\alpha\zeta_{R,D}}{\sqrt{2\pi}\Gamma(L+m)(\lambda_{R,D})^{L+m}}\left(\frac{2}{\beta}\right)^m \times G_{3,2}^{1,3}\left(\frac{2}{\beta\lambda_{R,D}}\middle|\begin{matrix}1-L-m, 1-m, \frac{1}{2}-m \\ 0, -m\end{matrix}\right) \quad (15)$$

To get the overall expression of the BER substitute (15), (14), (12) and (11) into (10).

IV. NUMERICAL RESULTS

In this section we analyze the performance of the decode-and-forward incremental relaying system over mixed fading channel encompassing the effects of the interference near the destination and the relay. The effect of changing Nakagami-m order and the threshold value γ_o on the system behavior is investigated too.

As can be seen from Fig. 2, the impact of the interference at the destination is much greater than that near the relay. For example, a system with (L=1, N=5) outperforms a system with (L=3, N=1) when the BER is 10^{-1} by about 3dB. and that explains the degradation on the system performance of the interference at the destination compared with the interference near the relay.

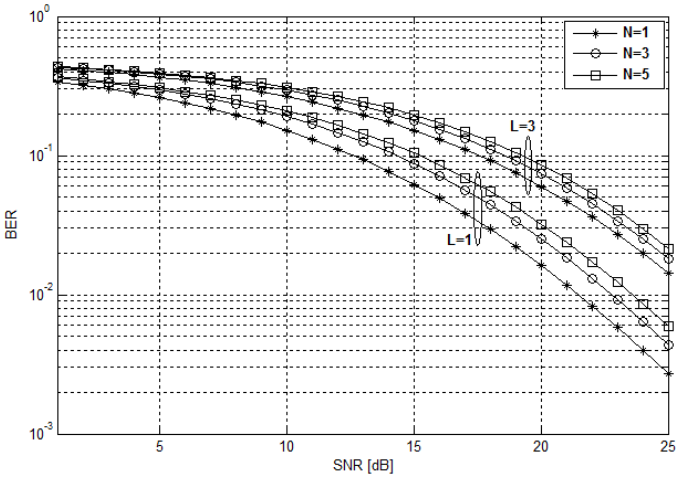


Fig. 2 Bit Error rate fir BPSK modulation with Different number of interferers at the destination (L=1,3), Different number of interferers at the relay (N=1,2,3), (m=1) and (γ_o= $\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB).

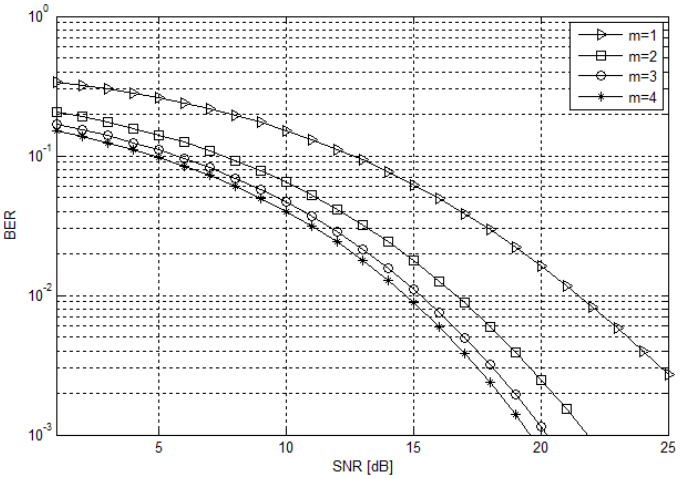


Fig. 3 Bit Error Rate for BPSK modulation with Different values of the nakagami order (m=1,2,3,5), (L=N=1) and (γ_o= $\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB).

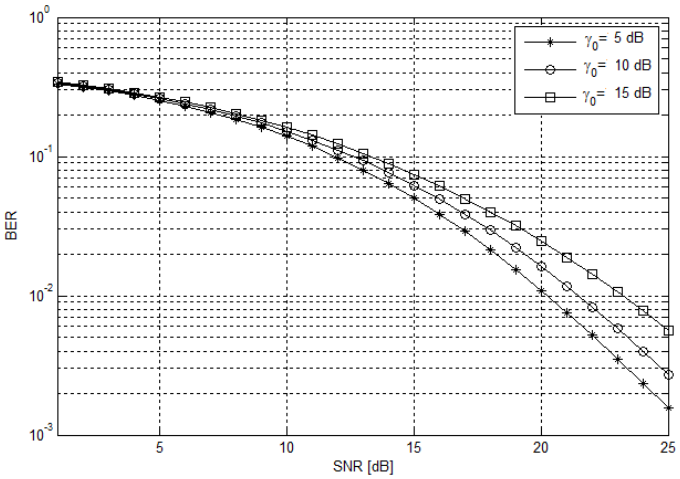


Fig. 4 Bit error rate for BPSK modulation for different values of (γ_o=5,10,15 dB), (L=N=1), (m=1) and ($\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB).

As depicted in Fg. 3, as the nakagami order increases, then system performance will be enhanced in response to that. For instance, at a BER of 10^{-2}, the system with (m=4) outperforms the system with (m=1) by about 7dB. which reflects the impact of increasing the Nakgami order m on the system performance.

The degree of degradation of the system performance when changing γ_o, $\overline{\gamma_r}$ and $\overline{\gamma_d}$ will vary. For instance, Fig. 4 represents the effects on the system when changing γ_o and it's apparent that when we increase γ_o, then the system will be degraded and as an example on that, given that the BER is 10^{-2}, then the system with γ_o=5 dB outperforms the system with γ_o=15 dB by 4 dB.

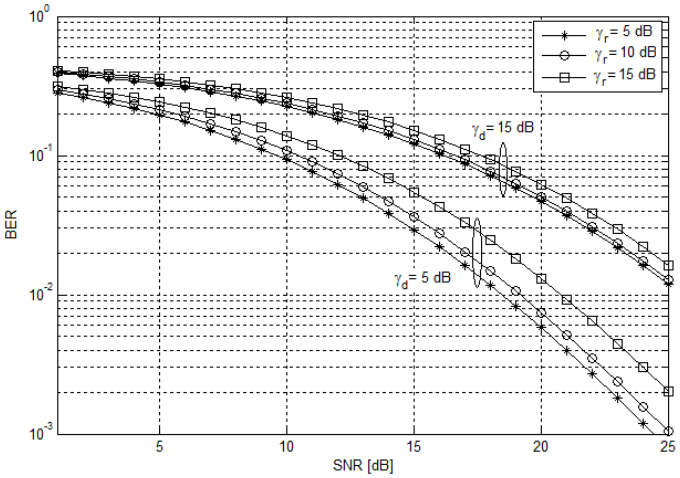


Fig. 5 Bit error rate for BPSK modulation for different values of (γ_o=5,10,15 dB), (L=N=1), (m=1) and ($\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB).

Likewise, if we increase $\overline{\gamma_d}$ then the system performance will be degraded consequently and this has more impact on the system than increase $\overline{\gamma_r}$. In other words, the increment of $\overline{\gamma_r}$ is incomparble to that of $\overline{\gamma_d}$ and this is obvious as depuicted in fig.5. For instance, given that the BER is 10^{-2}, then for a system with ($\overline{\gamma_d}$=15 dB, $\overline{\gamma_r}$ = 5 dB) to have the same performance as that with ($\overline{\gamma_d}$=5 dB, $\overline{\gamma_r}$ = 15 dB) then an increment of SNR of about 5 dB is required in order to acheive the same behavior.

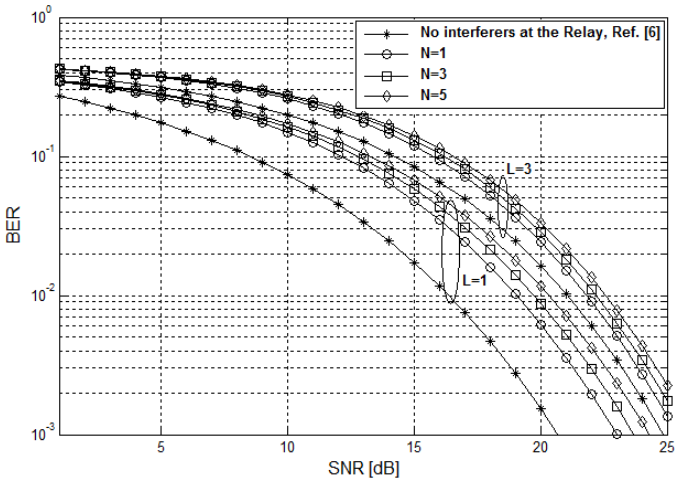


Fig. 6 BER comparison with and without the presence of the interference at the relay where (L=1,3), (N=1,3,5), (γ_o= $\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB) and (m=3).

In order to see the impact of the interference at the relay and measure the results on the system performance compared with the case where there is no interference at the relay, Fig. 6

depicts that comparison and shows that the impact of the presence of the interference at the relay is considerably has less degradation on the BER performance of the system compared to the presence of the interference at the relay. As a consolidation for that, the difference between a system that have (L=1, N=5) and that with (L=3) is only 1 dB when the BER is 10^{-2}. Another comparison is depicted in Fig. 7 that consolidates the results of Fig. 6 and shows the case if we increase the SNR from each interferer near the relay and compare it with the interference near the destination. Apparently, the required increment on the SNR for a system with (L=1, $\overline{\gamma_r}$=17 dB) to behave exactly the same as that with (L=4) is around 3 dB when the BER is 10^{-2} which is incomaparble to the required increment of the SNR if we compare the two systems with (L=1) and (L=4), which is around 8 dB.

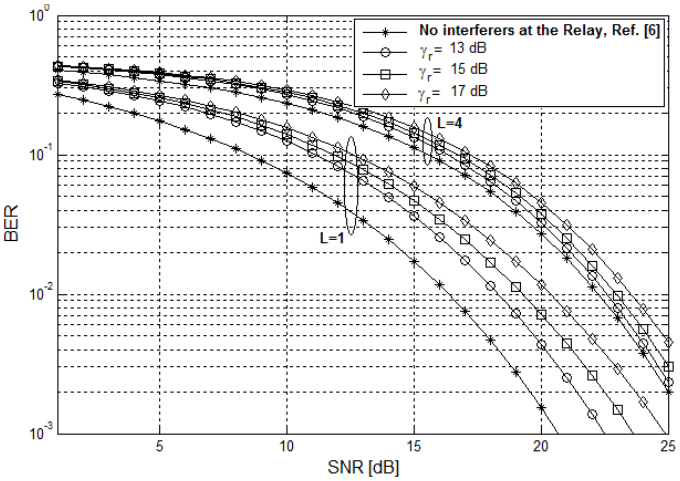


Fig. 7 BER comparison with and without the presence of the interferrence at the relay where (L=1,4), (N=1), (γ_o=$\overline{\gamma_d}$=10 dB) ($\overline{\gamma_r}$=13,15,17 dB) and (m=3).

V. CONCLUSION

This paper discusses and analyzes the performance of the decode-and-forward incremental relaying with the presence of multiple interferers near the destination and the relay. The incremental relaying systems achieve the spatial diversity at the destination and utilize the channel resources. Numerical results show that the improvement on the BER of a decode-and-forward incremental relaying over mixed fading channels can be achieved by increasing the Nakagami order m and will be drastically degraded if we increase the interferers at the destination L, γ_o and $\overline{\gamma_d}$ which have more impact on the system as if we increase the interferers at the relay N and $\overline{\gamma_r}$. Numerical results also compare the system behavior with and without the presence of the interference at the relay and shows that the interference at the relay have a relatively less impact on the system performance compared to that near the destination.

REFERENCES

[1] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, “Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, Dec. 2004.

[2] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, “Distributed spacetime-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 49, pp. 2415-2425, Oct. 2003.

[3] A. Sendonaris, E. Erkip, and B. Aazhang, “User cooperation diversity, part I: system description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927-1938, Nov. 2003.

[4] S. Ikki and M. H. Ahmed, ”Performance Analysis of Cooperative Diversity Wireless Networks over Nakagami-m Fading Channel,” *IEEECommun. letter*, vol. 11, pp. 334-336, April, 2007.

[5] H. A. Suraweera, H. K. Garg and A. Nallanathan, "Performance analysis of two hop amplify-and-forward systems with interference at the relay," IEEE Commun. Lett., vol. 14, pp. 692-694, Aug. 2010.

[6] Amer Magableh, Mamoun F. Al-Mistarihi, and Rami Mohaisen, “Closed-form Expression for Bit Error Rate in Relay-Based Cooperative Diversity Systems over Multipath Fading Channels with Interference,” *Proc. 9th International Wireless Communications and Mobile Computing, IWCMC 2013,* Sardinia, Italy, July 1-5, 2013.

[7] Mamoun F. Al-Mistarihi, Amer Magableh, and Rami Mohaisen, “Closed-form Expression for Outage Probability in Relay-Based Cooperative Diversity Systems over Multipath Fading Channels with Interference,” *Proc. IEEE 11th International Multi-Conference on Systems, Signals & Devices, SSD 2014,* Castelldefels-Barcelona, Spain, February 11 - 14, 2014.

[8] J. G. Proakis, *Digital Communications.* McGraw-Hill, Fourth ed., 2001.

[9] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York: McGraw-Hill, 2002.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.

[11] P. Herhold, E. Zimmermann, G. Fettweis, ”A Simple Cooperative Extension to Wireless Relaying,” *in Int. Zurich Seminar on Communications*, Zurich, Switzerland, Feb. 2004.

[12] A. Adinoyi and H. Yanikomeroglu, ”Cooperative relaying in multiantenna fixed relay networks,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 533-544, February 2007.

Outage Probability Analysis in Relay-Based DF Cooperative Diversity Systems with Relay & Destination Interferers

Rami Mohaisen[1], Mamoun F. Al-Mistarihi[1], Amer M. Magableh[1], and Ashraf H. Sharaqa[2]

(1) Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
rami.fayez@hotmail.com, mistarihi@just.edu.jo, ammagableh@just.edu.jo

(2) WorleyParsons / Communication and Security Projects Division, Al-Khobar, Saudi Arabia
elengashraf@yahoo.com

Abstract - **Incremental relaying was firstly introduced in order to accommodate for the restriction associated with the fixed relaying schemes by using the intermediate relays only when they are needed by using a feedback from the destination to determine whether or not to use the relay which depends mainly on the usability of the direct path between the source and the relay. When the relay receives this negative feedback from the destination then it will send the signal to the destination within the second time slot which by then will be combined with the source signal using MRC, and by this, the system resources will be utilized efficiently. This paper provide the performance analysis of a decode-and-forward incremental relaying cooperative relaying over Rayleigh and Nakagami-m mixed fading channels with the presence of multiple interferers L and N near the destination and the relay respectively, which in turn will degrade the overall system performance and also compares the system performance with and without the presence of the interferers near the relay. Numerical results are provided to validate our assumptions.**

Keywords: ***Incremental relaying, cooperative diversity, decode-and-forward, interference, outage probability (OP), signal to noise ratio (SNR), signal to interference ratio (SIR), maximal ratio combiner (MRC).***

I. INTRODUCTION

The appealing advantages of using cooperative diversity make it the promising solution for the deployment of the future wireless communication systems. Cooperative diversity is implemented by the introduction of the multiple relaying nodes between the source and the destination in order to prolong the battery life, extend the coverage area and achieve the spatial diversity at the destination [1]. In fixed relaying schemes, the destination will combine the two received signals from both the source and the relay by the MRC.

Aside from the advantages of implementing the fixed cooperative diversity schemes in terms of its implementation simplicity and the spatial diversity gain at the destination, they don't utilize the channel resources efficiently since the relay forwards the source signal regardless of the channel conditions. Since most of the time the direct path is usable which makes the utilization of the relay path a waste of resources.

Performance analysis in terms of symbol error probability (SEP) and outage probability (OP) for the amplify-and-forward (AF) and decode-and-forward (DF) relaying over Rayleigh and Nakagami-m fading channels have been thoroughly analyzed and discussed [2-3].

Incremental relaying was firstly introduced to overcome these restrictions by using the assistance of the relay node only when needed [4] by the virtue of using a feedback channel from the destination to both the source and the relay informing them about the quality assessment of the direct transmission which also will inform the relay to participate in forwarding the source signal in an opportunistic manner. The destination will combine the two received signals from both the source and the relay using MRC as stated in [5].

Performance analysis of relay systems with interference at the relay was presented in [6]. Closed-form expressions for outage probability and BER in relay-based cooperative diversity systems over different fading channels with the presence of multiple interferers near the destination were introduced in [7-8].

In this paper, we discuss the decode-and-forward incremental relaying cooperative diversity by introducing an intermediate single relay and multiple L and N interferers located near the destination and the relay, respectively, which will affect the performance of the system in response to the Co-channel interference over mixed fading channels between the source-destination and the source-relay-destination paths and compares the result and effect of the interference presence near the relay to that near the destination.

This paper is further presented as follows; Section II provides the system model. Section III provides the performance analysis and derived tight closed form formula of outage probability. Results and discussion are presented in section IV. We concluded this paper in section V.

II. SYSTEM MODEL

In Fig. 1, the channel coefficient between the source and the destination is assumed to follow the Rayleigh distribution whereas the Channels coefficients between the source-relay

and the relay-destination are assumed to follow the Nakagami-m distribution. The relay will participate with the source in transmitting its signal to the destination upon the reception of the destination's negative feedback that request the relay to send the source signal within the second time slot. The relay then will decode the source signal, re-encode it and then send it to the destination. The noise at the destination is assumed to be Additive White Gaussian Noise with a unity variance N_o. Beside that, near the destination and relay, the channels from the L and N interferers respectively that cause a Co-channel interference are assumed to be a sum of identically and independently distributed Rayleigh random variables which consequently will degrade the overall performance of the system. All the nodes in the system are equipped with one single antenna and Time Division Multiple Access (TDMA) is used as the multiple access technique, it is worth to mention that Frequency Division Multiple Access (FDMA) can be used too.

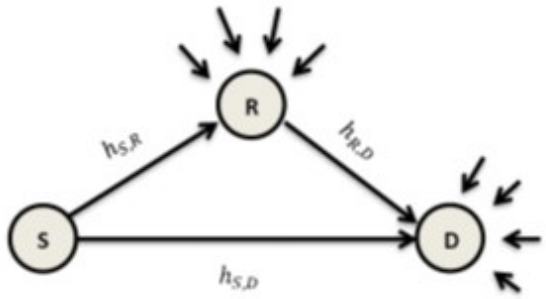


Fig. 1 Cooperative Diversity Network with Multiple Interferers L and N Located near the Destination and the Relay, respectively.

Within the first time slot, the destination will receive a signal directly from the source and based on the quality of this signal, the destination will decide whether or not the relay should participate in the transmission. From instance, if the direct transmission was successful and the destination can properly process the source signal from the direct transmission, then a positive feedback from the destination will be sent to both the relay and the source respectively informing them about the successfulness of the direct transmission and requesting the source to send a new signal within the second time slot. Otherwise, if the destination failed to receive the source signal or the direct path was in outage, then the relay will be requested to resend the source signal within the second time slot which by then will be combined with the source signal at the destination using the MRC.

The received signals at both the relay and the destination within the first time slot are:

$$y_{S,D}(t_1) = h_{S,D}\sqrt{E_s}\, x(t_1) + n_1(t_1) + I(t_1) \quad (1)$$

$$y_{S,R}(t_1) = h_{S,R}\sqrt{E_s}\, x(t_1) + n_2(t_1) + I(t_1) \quad (2)$$

where $x(t)$ is the symbol signal with a normalized energy. $n_1(t_1)$ and $n_2(t_1)$ are the AWGN noise terms at the destination and E_s is the signal energy. The interference effects at the destination and the relay within the first time slot are represented by $I(t_1)$.

If the relay participates in the transmission within the second time slot, then the received signal at the destination from the relay is denoted as:

$$y_{R,D}(t_2) = h_{R,D}\sqrt{E_s}x_r(t_2) + n_3(t_2) + I(t_2) \quad (3)$$

where $x_r(t_2)$ is the regenerate signal at the relay within the second time slot . $I(t_2)$ is the interference effects near the destination within the second time slot.

The destination will assess the usability of the direct link by comparing the signal to noise ratio of the direct path against $\gamma_o$which is the minimal threshold for which the destination can receive the source signal merely depending on the direct path without the help of the relay.

In the interference environment, in most of the communication systems, rather than working on the SNR at the destination a transgression is made to merely work on the signal to interference ratio SIR since the noise effects near the destination is not comparable to that of the Co-channel effects and since the interference limited environment is a common practice in the literature.

In the following section, we will derive the building blocks in deriving the performance metrics, namely the probability density function (PDF) and the cumulative distribution function (CDF) of both the SIR at the relay from the source link and the SIR at the destination from both the relay and source respectively.

III. PERFORMANCE ANALYSIS

The derivation of the outage probability necessitates is to derive the PDF and the CDF of the SIR at the destination. And since the interference is higher than the noise near the destination and its impact is more severe than the noise effects, based on this assumption, a transgression is made to merely work on the SIR at the destination.

A. Probability Density Function of SIR's $\boldsymbol{\eta_{S,D}}$, $\boldsymbol{\eta_{R,D}}$ *and* $\boldsymbol{\eta_{S,R}}$

Given that the SNR of the direct path $\gamma_{S,D}$ follows the exponential distribution and the fading channels from the interferers to the destination ($\gamma_b = \sum_1^L \gamma_k$) is assumed to be a sum of identically, independently and exponentially distributed random variables, the PDF of the SIR ($\boldsymbol{\eta_{S,D}}$) at the destination from the source path is given as follows [7]:

$$f_{\eta_{S,D}}(\eta_{S,D}) = \frac{\varsigma_{S,D}}{(\lambda_{S,D}+\eta_{S,D})^{L+1}} \quad (4)$$

where $\lambda_{S,D} = \frac{\overline{\gamma_{S,D}}}{\overline{\gamma_d}}$, $\varsigma_{S,D} = L(\lambda_{S,D})^L$, $\overline{\gamma_d} = \frac{\varepsilon}{N_o}E({\alpha_k}^2)$ is the average SNR of each independent interfering channel near the destination, and $\gamma_b = \frac{\varepsilon}{N_o}\sum_{k=1}^{L}{\alpha_k}^2 = \sum_{k=1}^{L}\gamma_k$ where γ_k is the instantaneous SNR from the $k-th$ interfering channel and L is the number of interfering channels near the destination. $\gamma_k's$ are mutually and statistically independent .

The PDF of the SIR $(\boldsymbol{\eta_{R,D}})$ at the destination from the relay path is given as follows [7]**:**

$$f_{\eta_{R,D}}(\eta_{R,D}) = \frac{\zeta_{R,D}\ {\eta_{R,D}}^{m-1}}{(\eta_{R,D}+\lambda_{R,D})^{L+m}} \quad (5)$$

where $\lambda_{R,D} = \frac{\overline{\gamma_{R,D}}}{m\overline{\gamma_d}}$, $\zeta_{R,D} = \frac{(L+m-1)!}{m^L\Gamma(m)(L-1)!}(m\,\lambda_{R,D})^L$

Following the same approach used in deriving (5), the PDF of the SIR $(\boldsymbol{\eta_{S,R}})$ at the relay from the source is given as follows [7]**:**

$$f_{\eta_{S,R}}(\eta_{S,R}) = \frac{\zeta_{S,R}\ {\eta_{S,R}}^{m-1}}{(\eta_{S,R}+\lambda_{S,R})^{L+m}} \quad (6)$$

where $\lambda_{S,R} = \frac{\overline{\gamma_{S,R}}}{m\overline{\gamma_r}}$, $\zeta_{S,R} = \frac{(N+m-1)!}{m^N\Gamma(m)(N-1)!}(m\,\lambda_{S,R})^N$, $\overline{\gamma_r} = \frac{\varepsilon}{N_o}E({\alpha_k}^2)$ is the average SNR of each independent interfering channel near the relay, and $\gamma_b = \frac{\varepsilon}{N_o}\sum_{k=1}^{L}{\alpha_k}^2 = \sum_{k=1}^{L}\gamma_k$ where γ_k is the instantaneous SNR from the $k-th$ interfering channel and N is the number of interfering channels near the destination. $\gamma_k's$ are mutually and statistically independent.

B. The Cumulative Distribution Function of SIR's $\boldsymbol{\eta_{S,D}}$, $\boldsymbol{\eta_{R,D}}$, and $\boldsymbol{\eta_{S,R}}$

The CDF of the SIR $(\boldsymbol{\eta_{S,D}})$ at the destination from the source path is given as follows [7]:

$$F_{\eta_{S,D}}(\eta_{S,D}) = \left[\frac{(\lambda_{S,D}+\eta_{S,D})^L-{\lambda_{S,D}}^L}{(\lambda_{S,D}+\eta_{S,D})^L}\right] \quad (7)$$

The CDF of the SIR $(\boldsymbol{\eta_{R,D}})$ at the destination from the relay path is given as follows [7]:

$$F_{\eta_{R,D}}(\eta_{R,D}) = \frac{\zeta_{R,D}{\eta_{R,D}}^m}{m(\lambda_{R,D})^{L+m}}\ {}_2F_1\left(L+m, m; m+1; -\frac{\eta_{R,D}}{\lambda_{R,D}}\right) \quad (8)$$

where ${}_2F_1(a,b;c;d)$ is the Gauss Hyper-geometric function [9].

The CDF of $\eta_{S,R}$ at the relay from the source path is given as follows:

$$F_{\eta_{S,R}}(\eta_{S,R}) = \frac{\zeta_{S,R}{\eta_{S,R}}^m}{m(\lambda_{S,R})^{N+m}}\ {}_2F_1\left(N+m, m; m+1; -\frac{\eta_{S,R}}{\lambda_{S,R}}\right) \quad (9)$$

Proof:

The CDF of $\eta_{S,R}$ is found by integrating the PDF of $\eta_{S,R}$ as follows [10]:

$$F_{\eta_{S,R}}(\eta_{S,R}) = \zeta_{S,R}\int_0^{\eta_{S,R}} f_{\eta_{S,R}}(y)dy \quad (10)$$

Solving the integral in (10) using [9,(3.381.1)], the compact form of the CDF of $\eta_{S,R}$ will be as in (9).

C. Outage Probability Analysis

In the incremental relaying system, if the direct path is in outage, then the destination will request the relay to assist in relaying the source signal in attempt to mitigate for the fading severity in the direct path. Still there is a probability that the overall combined signal by the MRC at the destination will be in outage. (i.e, the SNR of the combined signals from the source and the relay within the second time slot by the MRC will fall below a pre-specified threshold γ_o).

Formulating all possible scenarios of the system outage probability, p_{out} is given as [7]:

$$p_{out} = \Pr(\eta_{S,D} \le \gamma_o) \times P_r[\min(\gamma_{S,R}, \eta_{S,D} + \eta_{R,D} \le \gamma_O | \eta_{S,D} \le \gamma_o)]$$

$$p_{out} = P_r(\eta_{S,D} \le \gamma_o)\left[1 - P_r(\eta_{S,D} + \eta_{R,D} > \gamma_o | \eta_{S,D} \le \gamma_o)\right] \times P_r(\gamma_{S,R} > \gamma_o | \eta_{S,D} \le \gamma_o)$$

$$p_{out} = P_r(\eta_{S,D} \le \gamma_o) \times \left[1 - \frac{P_r(\eta_{S,D}\le\gamma_o) - P_r(\eta_{S,D}+\eta_{R,D}\le\gamma_o)}{P_r(\eta_{S,D}\le\gamma_o)} P_r(\gamma_{S,R} > \gamma_o)\right] \quad (11)$$

where the first argument within the min(•|•) operator in the above expression represents the outage at the relay and the second one represents the outage at the destination .

Further simplifications on the above expression yields [7]:

$$p_{out} = \Pr(\gamma_{S,R} \le \gamma_o)\Pr(\eta_{S,D} \le \gamma_o) + \Pr(\gamma_{S,R} > \gamma_o)\Pr(\eta_{S,D} + \eta_{R,D} \le \gamma_o) \quad (12)$$

When using the MRC, the value of $\eta_{R,D}$ will be much higher compared with $\eta_{S,D}$ since we used the relay path in the first place to compensate for the deficiency of the direct transmission. And hence, we can rely on the SIR at the destination from the relay path and totally ignore the SNR from the direct path to get the approximated outage probability expression as in (13).

IV. NUMERICAL RESULTS

Fig. 2 compares the impacts of increasing the number of interferes near the destination and the relay, namely the effects of L and N respectively, on the system performance, apparently as the number of interferers near the destination increases, the system performance will be degraded and affected drastically compared with increasing N near the relay. As depicted from the figure, in order for a system with two interferers near the destination to have the same performance as that with only one interferer, then an increment of around 4 dB on the SNR is required in order to achieve the same performance when the outage probability is 0.5. This case is incomparable with the number of interferers at the relay which is accommodated by a slight increment on the SNR.

$$p_{out} \approx \frac{\zeta_{S,R}{\gamma_o}^m}{m(\lambda_{S,R})^{N+m}}\left[\frac{(\lambda_{S,D}+\gamma_o)^L-\lambda_{S,D}{}^L}{(\lambda_{S,D}+\gamma_o)^L}\right] {}_2F_1\left(N+m,m;1+m;-\frac{\gamma_o}{\lambda_{S,R}}\right)+\frac{\zeta_{R,D}{\gamma_o}^m}{m(\lambda_{R,D})^{L+m}}\left[1-\frac{\zeta_{S,R}{\gamma_o}^m}{m(\lambda_{S,R})^{N+m}}\ {}_2F_1\left(N+m,m;1+m;-\frac{\gamma_o}{\lambda_{S,R}}\right)\right]\ {}_2F_1\left(L+m,m;1+m;-\frac{\gamma_o}{\lambda_{R,D}}\right) \quad (13)$$

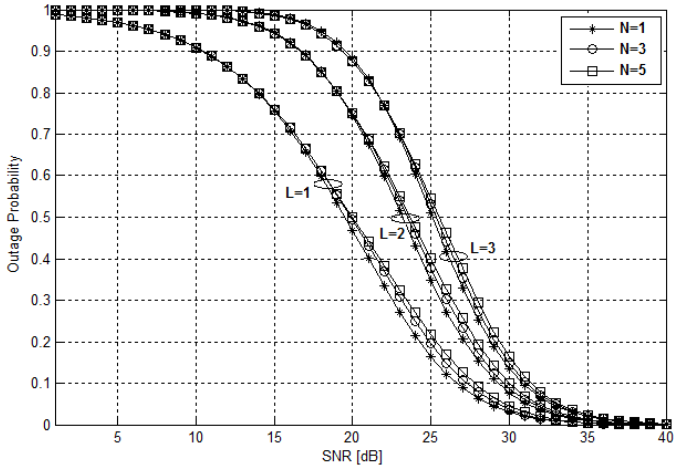


Fig. 2 Outage Probability for different number of interferers (L=1, 2,3), (N=1,3,5), (m=2) and (γ_o= $\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB).

Fig. 3 compares the effects of increasing γ_o along with the number of interferers near the relay. From the figure, it is obvious that increasing N near the relay is incomparable to the increment of γ_o and have a slight impact on the system. For instance, for a system with (γ_o=10 dB) to have the same performance as that with (γ_o=15 dB) then an increment of (SNR=5 dB) is required in order to achieve the same performance when the outage probability is 0.5 while only a fraction of an SNR is required for a system with (N=5) to have the same performance as that with (N=1).

Elaborating on the enhancement parameters on the system performance, Fig. 4 depicts the enhancement on the system when we increase the Nakagami order m. As can be seen from the figure, the impact of increasing m will take effect after SNR=22 dB. For example, a system with (m=1) will have the same performance as that with (m=5) after increasing the SNR of about 3 dB when the outage probability is 0.2. This increment will be greater if we have more interferers near the destination and less if we have only one interferer near the destination.

In comparing the effect of the interferers near the relay and the destination, Fig. 5 elaborates on this and depicts the results on the system performance if we increase $\overline{\gamma_d}$ and $\overline{\gamma_r}$ respectively. For example, a fraction of SNR is required for a system with ($\overline{\gamma_r}$=5 dB) to have the same performance of a system with ($\overline{\gamma_r}$=15 dB) if we have (L=N=1) which is incomparable with the required increment for a system with ($\overline{\gamma_d}$=5 dB) to have the same performance as that with ($\overline{\gamma_d}$=15 dB) which is apparently around 8 dB SNR when the outage probability is 0.4.

As a result, the degradation on the system performance is apparent if we increase γ_o, $\overline{\gamma_d}$, $\overline{\gamma_r}$, L and N. But at the same time the severity of increasing these parameters will vary depending on the parameters themselves. As explained in this section, increasing L, γ_o, $\overline{\gamma_d}$ will have more impact on the system performance unlike the case if we increase N and $\overline{\gamma_r}$. Finally, increasing m will improve the system performance and this enhancement will take effect after certain SNR, namely, when the SNR is beyond 22 dB.

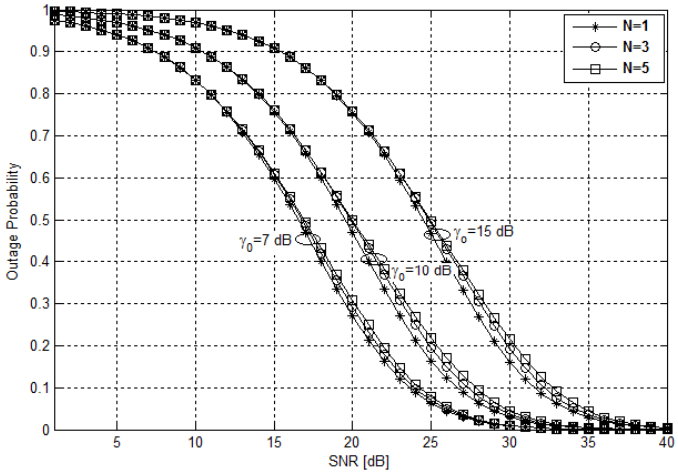


Fig. 3 Outage Probabilty for Different vlause of (N=1,3,5), (γ_o=7, 10,15 dB), (L=1), (m=2) and ($\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB).

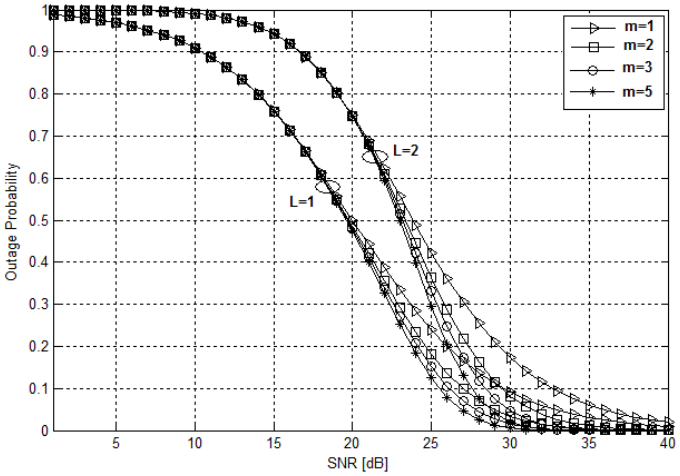


Fig. 4 Outage Probability for Different values of (m=1,2,3,5), (L=1,2), (N=2) and (γ_o= $\overline{\gamma_r}$=$\overline{\gamma_d}$=10 dB).

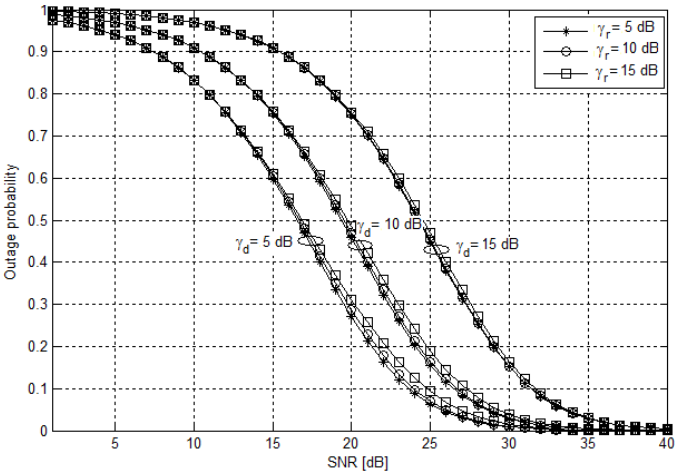


Fig. 5 Outage Probability for Different vlaues of ($\overline{\gamma_d}$=5, 10, 15 dB), ($\overline{\gamma_r}$=5, 10,15 dB), (L=N=1) and (γ_o=10 dB)

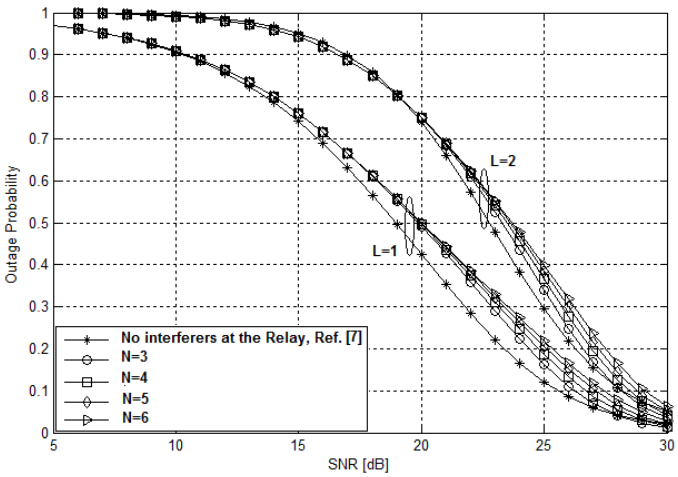


Fig. 6 Outage Probability Comparison representing the performance of two systems with and without the interference at the relay when ($\overline{\gamma_d} = \overline{\gamma_r} = \gamma_o$ =10 dB), (L=1, 2), (m=4) and different values of (N=3,4,5 and 6).

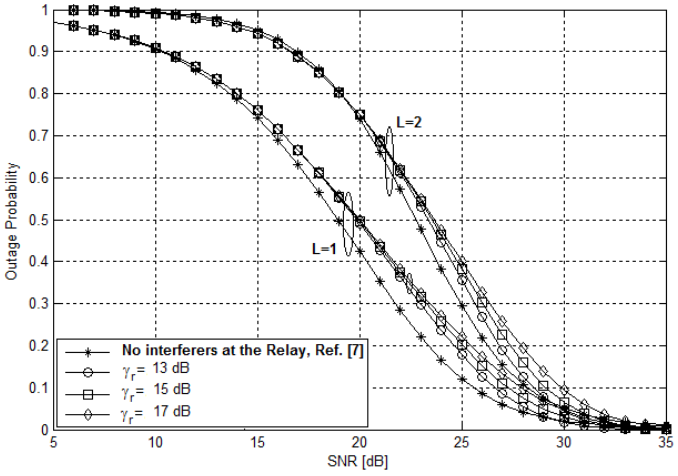


Fig. 7 Outage Probability Comparison representing the performance of two systems with and without the interference at the relay when ($\overline{\gamma_d} = \gamma_o$ =10 dB), ($\overline{\gamma_r}$= 13,15.17 dB,)(L=1, 2), (m=4) and (N=2).

As a comparison between the performance of the system with and without the interference at the relay, Figure 6 presents the impact of the interference at the relay on the outage probability which is apparently a minor degradation compared with the impact of the interferers at the destination. For example, when we consider an outage probability of 0.2, then of a system with five interferers near the relay (N=5) to have the same performance as that without interferes near the relay and (L=1) then a SNR of about 3 dB is needed which emphasizes the fact that the severity of the interference at the destination has much more impact on the system performance compared with that of the interference at the relay.

Elaborating on the comparison between the two cases with and without interference at the relay, Fig. 7 shows the effects on the system performance when we vary the SNR of each interfering signals and compare them with the case where we don't have interferers at the relay. For instance, impact of the presence of the interferers near the destination has more severity than increasing the individual SNR of the interferers near the relay. As an example on that, the required SNR for a system without interferers near the destination to have the same performance as that with 2 interferers at the relay and ($\overline{\gamma_r}$ =17 dB) is about 3 dB which is incomparable to that required by a system with two interferers at the destination and no interferes at the relay.

V. CONCLUSION

The adaptive cooperative diversity draws a considerable attention from the researchers since they exploit the channel resources efficiently as opposed to the fixed relaying schemes which do so inefficiently. This work discusses the performance namely, the Outage Probability of the Decode-and-Forward incremental relaying system with the presence of the multiple L and N interferers near the destination and relay respectively over Rayleigh and Nakagami-m mixed fading channels which indicates that as the Nakagami-m order increased and the number of the interferers and the value of γ_o decreased a considerable enhancement in the overall performance of the system will be achieved. Numerical results showed that the effects of the interference near destination are more severe than that near the relay.

REFERENCES

[1] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inform. Theory*, vol. 50, Dec. 2004.

[2] S. Ikki and M. H. Ahmed, "Performance Analysis of Cooperative Diversity Wireless Networks over Nakagami-m Fading Channel," *IEEE Commun. letter*, vol. 11, pp. 334-336, April, 2007.

[3] Mamoun F. Al-Mistarihi, and Rami Mohaisen, "Performance Analysis of Multiuser Diversity in Multiuser Two-Hop Decode-and-forward Cooperative Multi-Relay Wireless Networks," *Proc. International Conference on Wireless Communication and Mobile Computing, ICWCMC 2011*, Bangkok, Thailand, pp. 238-242, December 25-26, 2011.

[4] S. Ikki and M. H. Ahmed, "Performance of multiple relay cooperative diversity systems with best relay selection over Rayleigh fading channels," *EURASIP J. Advances Signal Process.*, 2008.

[5] S. Ikki and M. H. Ahmad, 'Performance Analysis of Decode-and-Forward Incremental Relaying Cooperative-Diversity Networks over Rayleigh Fading Channels,' *Proc. IEEE Technology Conference-Spring* (VTC-Spring'09), Barcelona , Spain, April 2009.

[6] H. A. Suraweera, H. K. Garg and A. Nallanathan, "Performance analysis of two hop amplify-and-forward systems with interference at the relay," IEEE Commun. Lett., vol. 14, pp. 692-694, Aug. 2010.

[7] Mamoun F. Al-Mistarihi, Amer Magableh, and Rami Mohaisen, "Closed-form Expression for Outage Probability in Relay-Based Cooperative Diversity Systems over Multipath Fading Channels with Interference," *Proc. IEEE 11th International Multi-Conference on Systems, Signals & Devices, SSD 2014,* Castelldefels-Barcelona, Spain, February 11 - 14, 2014.

[8] Amer Magableh, Mamoun F. Al-Mistarihi, and Rami Mohaisen, "Closed-form Expression for Bit Error Rate in Relay-Based Cooperative Diversity Systems over Multipath Fading Channels with Interference," *Proc. 9th International Wireless Communications and Mobile Computing, IWCMC 2013,* Sardinia, Italy, July 1-5, 2013.

[9] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.

[10] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York: McGraw-Hill, 2002.

Closed Form Expression of Outage Probability in DCSK Cooperative Communication Systems over Nakagami-m Fading Channels

Mamoun F. Al-Mistarihi*, Amer M. Magableh*, and Maiss M. Al-Khasawneh*
* Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
mistarihi@just.edu.jo, ammagableh@just.edu.jo, mays_m_87@hotmail.com

Abstract – **In this paper, two user Cooperation systems based on differential chaos shift keying (DCSK) with orthogonal sub-channel in the broadcast phase and cooperative phase using Walsh code sequence is proposed. Single relay of decode and forward (DF) type is investigated according to the conventional cooperation protocol. Outage probability is derived and verified by simulation. Independent three paths Rayleigh and Nakagami-m fading channels are considered. Numerical results show an enhancement of the system performance for the Nakagami-m fading channel over the Rayleigh fading channel. Monte Carlo simulation shows a high agreement with the numerical results.**

Keywords: ***Outage probability, differential chaos shift keying (DCSK), multipath Rayleigh fading channel, multipath Nakagami-m fading channel, cooperative communication system.***

I. INTRODUCTION

Due to the characteristic of the chaotic signal such as security in transmission, resistance of jamming, and low power consumption, its use in communication systems gabbed a great attention. A Number of chaos-based communication schemes have been proposed and analyzed. Such schemes included Chaos Shift keying (CSK) [1], Chaos-Based DS-CDMA , Differential Chaos Shift keying (DCSK) [2]-[4], and Frequency-Modulated DCSK (FM-DCSK) [5], [6].

In DCSK and frequency-modulated (FM)-DCSK, there are no need to have a knowledge of the channel estimation or the code acquisition and synchronization, so many ultra-wideband systems based on them have been proposed lately [7]-[9].

The performance of any system in a wireless network is affected by the fading and intersymbol interference (ISI), which appears in the multipath propagation environment. Diversity technique, by which many replicas of the same signal are transmitted over independent channels, can be used to mitigate the effect of signal fading.

Multiple-antenna diversity technique is one of the most attractive techniques that can be easily used and combined with other diversity techniques. However, due to the variation of the carrier, it has been difficult to use multiple antenna diversity on a DCSK based system. Moreover, in wireless sensor networks (WSN) multiple-antenna diversity cannot be used because of the size and the cost of the sensors. Thus, a new spatial diversity technique is needed [10], [11]. User cooperation diversity is a good substitute to the multiple antenna diversity [12].

In this paper, a cooperative system consisting of two-users based on DCSK over a Nakagami-m fading channel is proposed. The performance of the system has been studied in terms of the outage probability.

The rest of this paper is organized as follows. Section II presents the principle of the two-user cooperative system based on DCSK and the outage probability of the proposed system. Performance evaluation and simulation results are presented in Section III. Finally we conclude our paper in section IV.

II. SYSTEM MODEL

A. DSCK

In DCSK system, the carrier is a chaotic signal generated using one of the chaotic mapping methods like Cubic and Logistic mapping method, while the modulator is a differential shift keying.

In Binary DCSK systems, the symbol duration is divided into two halves; in the first one, a reference copy of the chaotic signal is transmitted, while in the second half, the digital transmitted information will be repeated if it was '1', or reserved if the digital information was '0'.

Two orthogonal basis functions, $g_1(t)$ and $g_2(t)$ are used to represent the modulated signal as follows:

$$s(t) = s_{m1}g_1(t) + s_{m2}g_2(t)$$
$$(s_{11}\ s_{12}) = \left(\sqrt{E_b}\ \ 0\right)$$
$$(s_{21}\ s_{22}) = \left(0\ \ \sqrt{E_b}\right) \quad (1)$$

where $s(t)$ is the modulated signal and E_b is the bit energy.

The orthogonal basis functions $g_1(t)$ and $g_2(t)$ are represented as:

$$g_1(t) = \begin{cases} +c(t), & 0 \le t < \frac{T}{2} \\ +c\left(t - \frac{T}{2}\right), & \frac{T}{2} \le t < T \end{cases} \quad (2)$$

$$g_2(t) = \begin{cases} +c(t), & 0 \le t < \frac{T}{2} \\ -c\left(t - \frac{T}{2}\right), & \frac{T}{2} \le t < T \end{cases} \quad (3)$$

where T is the bit duration and $c(t)$ is the chaotic signal.

At the receiver side, using suboptimum Autocorrelation receiver (AcR), and depending on the received signal, $r(t)$, an observation signal will be given by:

$$z = \int_{T/2}^{T} r(t) r\left(t - \frac{T}{2}\right) dt \quad (4)$$

B. Two-User Cooperation Model

Consider a cellular system with two mobiles communicating with a base station. The links between the users in such a system are assumed to be independent as well as the links between users and base station. The channel state stays the same during the cooperative period. In other words, frequency selective fading channel is used. The investigated system uses conventional cooperation protocol, single relay of the type DF, and half-duplex transmission.

Fig.1 below shows the transceiver model of the proposed system.

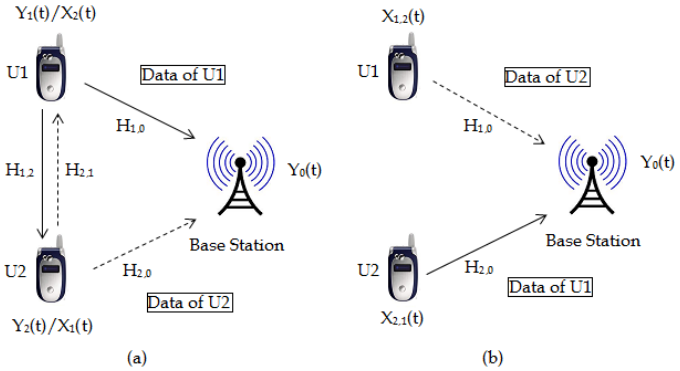


Fig.1 Cooperative transmission scheme (a) Odd period (b) Even period.

As can be seen from the figure, the received signal at the base station, user 1 and user 2 are, $Y_0(t)$, $Y_1(t)$ and $Y_2(t)$ respectively, which can be expressed as in (5) and (6).

where $X_i(t)$, $i = 1,2$ is the transmitted signal from users and $Z_i(t)$, $i = 1,2$ is the white Gaussian noise with zero mean and power spectral density $N_i/2$. $H_{i,j}(t)$ is the multipath channel linear time-invariant impulse response given by $H_{i,j}(t) = \sum_{l=1}^{L} \alpha_l \delta(t - \tau_l)$, where L is the number of multipath components, the attenuation and delay of the l^{th} path is α_l and τ_l, respectively. In the even period, user one will use $Y_1(t)$ to reconstruct the data of user two and transmit it to the base station. Using $Y_2(t)$, user two will do the same.

C. Outage Probability Performance

The outage probability is defined as the probability that the instantaneous SNR (γ_b) falls below a predetermined ratio (γ_{th}).

$$P_{out} = P_r(\gamma_b < \gamma_{th}) \quad (7)$$

Equivalently, it is the Cumulative Distribution Function evaluated at γ_{th}:

$$P_{out} = F_\gamma(\gamma_{th}) = \int_0^{\gamma_{th}} f_\gamma(\gamma) d\gamma \quad (8)$$

In this section, outage probability is derived for a DF cooperative relaying system over Rayleigh and Nakagami-m multipath fading channels.

An error free relay of the type Decode and Forward (DF) is used in order to derive the outage probability of the DCSK-CC, as well as the use of conventional cooperative protocol that is divided into two phases; the first half is called the 'broadcast phase' and the second is called 'cooperation phase'. Denote f as the global spread spectrum factor. During modulation, and since the proposed system has only two users, we force the length of each carrier segment, $\beta = f/4$ to make sure that the global spread-spectrum factor kept constant.

During the 'broadcast phase', each user transmits his own data, which is received and detected by the base station (BS) and the other user. For example the signal transmitted by user 1 is $X_1(t)$, user 2 will receive these data according to:

$$Y_2(t) = H_{1,2}(t) \otimes X_1(t) + Z_2(t) \quad (9)$$

which will be used by user 2 to make a hard decision to estimate $b^{(1)}$ of user 1.

1. Rayleigh Fading Channel:

Using the PDF for the Rayleigh multipath fading channel, the outage probability is given by (10).

where, $\beta_1 = \frac{1}{\overline{\gamma_1}(\overline{\gamma_1} - \overline{\gamma_3})}, \beta_2 = \frac{\overline{\gamma_3}}{(\overline{\gamma_1} - \overline{\gamma_3})^2}, \beta_3 = \frac{\overline{\gamma_1}\,\overline{\gamma_3}}{(\overline{\gamma_1} - \overline{\gamma_3})}$

Let $A_1 = \frac{{\beta_1}^2}{6}$, $A_2 = \beta_1\beta_2$, $A_3 = {\beta_2}^2 + 2\beta_1\beta_2\beta_3$, $A_4 = 2{\beta_2}^2\beta_3 + 2\beta_1\beta_2{\beta_3}^2$, $A_5 = {\beta_2}^2$, then P_{out} is given by (11).

2. Nakagami-m Fading Channel:

The PDF of γ_b is the convolution of the PDF's of the multipath channels. Hence,

$$f(\gamma_b) = f(\gamma_1) \otimes f(\gamma_2) \otimes \ldots \otimes f(\gamma_L) \quad (12)$$

where $\gamma_b = \left(\frac{E_b}{N_o}\right) \sum_{l=1}^{L} {\alpha_l}^2$, assuming α_l is an independent and identically-distributed (i.i.d) Nakagami-m random variable, and the PDF of γ_l, $f(\gamma_l)$, will have a gamma distribution that is given by:

$$f(\gamma_l) = \frac{{m_l}^{m_l} {\gamma_l}^{m_l - 1}}{\Gamma(m_l) \overline{\gamma_l}^{m_l}} \exp\left(-\frac{m_l \gamma_l}{\overline{\gamma_l}}\right)$$

where $\gamma_l = \frac{E_b}{N_o} {\alpha_l}^2$, $\overline{\gamma_l} = \left(\frac{E_b}{N_o}\right) E\{{\alpha_l}^2\}$.

For simplicity, assuming L=3, the PDF of γ_b is given by (13).

where $a_l = \frac{{m_l}^{m_l}}{\Gamma(m_l) \overline{\gamma_l}^{m_l}}$ and $b_l = \frac{m_l}{\overline{\gamma_l}}$.

On the other hand, at the base station, the total signal to noise ratio (SNR) is given by $\gamma_b^{BS} = {\gamma_b}^{odd} + {\gamma_b}^{even}$, therefore, $f(\gamma_b^{BS})$ is obtained using:

$$f(\gamma_b^{BS}) = f({\gamma_b}^{odd}) \otimes f({\gamma_b}^{even}) \tag{14}$$

The PDF of γ_b^{BS} is given by (15).

where $c = {a_1}^2 a_3 B(m_1, m_3) B(m_1 + m_3, m_1)$.

Substituting equation (15) in (8), the outage probability is given by (16).

Using the series representation of the confluent hypergeometric function ${}_1F_1(b, a, \beta\gamma)$

$${}_1F_1(b, a, \beta\gamma) = \sum_{n=0}^{\infty} \frac{b^{(n)}(\beta x)^n}{a^{(n)} n!} \tag{17}$$

where $a^{(n)}$ and $b^{(n)}$ is the rising factorial given by:

$$x^{(n)} = \frac{\Gamma(x+n)}{\Gamma(x)} \tag{18}$$

Let $c = \frac{\left(a_1{}^2 a_3 B(m_1,m_3) B(m_1+m_3,m_1)\right)^2 \Gamma(2m_1+m_3)^2}{\Gamma(4m_1+2m_3)}$, the outage probability is given by:

$$P_{out} = c \sum_{n=0}^{\infty} \frac{b^{(n)} \beta^n}{a^{(n)} n!} {b_1}^{-(n+a)} \gamma(n + a, b_1 \gamma_{th}) \tag{19}$$

where $\gamma(.,.)$ is the upper incomplete gamma function.

III. Results And Discussions

Fig.2 shows a significant match between the two carves of the system outage probability over Rayleigh and Nakagami-m fading channels with m=1. Fig. 3 shows A comparison between the exact expressions of the outage probability and the simulated results for the DCSK-CC system over Rayleigh fading channel for different values of γ_{th}. As demonstrated, the simulation results match with the exact results.

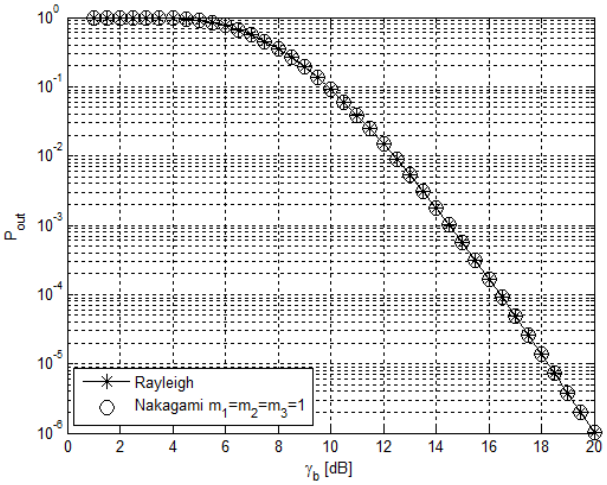


Fig .2 Outage probability for the DCSK-CC system over Rayleigh and Nakagami-m fading channel with m=1 .

$$odd\ period \begin{cases} Y_1(t) = H_{2,1}(t) \otimes X_2(t) + Z_1(t) \\ Y_2(t) = H_{1,2}(t) \otimes X_1(t) + Z_2(t) \\ {Y_0}^{odd}(t) = H_{1,0}(t) \otimes X_1(t) + H_{2,0}(t) \otimes X_2(t) + {Z_0}^{odd}(t) \end{cases} \tag{5}$$

$$even\ period \quad {Y_0}^{even}(t) = H_{1,0}(t) \otimes X_{1,2}(t) + H_{2,0}(t) \otimes X_{2,1}(t) + {Z_0}^{even}(t) \tag{6}$$

$$P_{out} = \int_0^{\gamma_{th}} \left(\left(\frac{1}{6} {\beta_1}^2 {\gamma_b}^3 - \beta_1 \beta_2 {\gamma_b}^2 + {\beta_2}^2 \gamma_b + 2\beta_1 \beta_2 \beta_3 \gamma_b - 2{\beta_2}^2 \beta_3 - 2\beta_1 \beta_2 {\beta_3}^2 \right) \exp\left(\frac{-\gamma_b}{\overline{\gamma_1}}\right) + \left({\beta_2}^2 \gamma_b + 2{\beta_2}^2 \beta_3 + 2\beta_1 \beta_2 {\beta_3}^2 \right) \exp\left(\frac{-\gamma_b}{\overline{\gamma_3}}\right) \right) d\gamma_b \tag{10}$$

$$P_{out} = \left\{ \left[\left(-A_1 \overline{\gamma_1} {\gamma_{th}}^3\right) + \left(\left(-3A_1 {\overline{\gamma_1}}^2 - A_2 \overline{\gamma_1}\right) {\gamma_{th}}^2 \right) + \left(\left(\left(-6A_1 {\overline{\gamma_1}}^3\right) + \left(-2A_2 {\overline{\gamma_1}}^2\right) + \left(A_3 \overline{\gamma_1}\right) \right) \gamma_{th} \right) \right] \left(\exp\left(\frac{-\gamma_{th}}{\overline{\gamma_1}}\right) \right) \right\} + \left\{ \left(-6A_1 {\overline{\gamma_1}}^4 - 2A_2 {\overline{\gamma_1}}^3 + A_3 {\overline{\gamma_1}}^2 - A_4 \overline{\gamma_1}\right) \left(exp\left(\frac{-\gamma_{th}}{\overline{\gamma_1}}\right) - 1 \right) \right\} + \left\{ -A_5 \overline{\gamma_3} \gamma_{th} exp\left(\frac{-\gamma_{th}}{\overline{\gamma_3}}\right) \right\} + \left\{ \left(-A_5 {\overline{\gamma_3}}^2 + A_4 \overline{\gamma_3}\right) \left(exp\left(\frac{-\gamma_{th}}{\overline{\gamma_3}}\right) - 1 \right) \right\} \tag{11}$$

$$f(\gamma_b) = {a_1}^2 a_3 B(m_1, m_3) B(m_1 + m_3, m_1) \exp(-b_1 \gamma_b) {\gamma_b}^{2m_1+m_3-1} \, {}_1F_1(m_3, 2m_1 + m_3, (b_2 - b_3)\gamma_b) \tag{13}$$

$$f({\gamma_b}^{BS}) = \frac{c^2 \Gamma^2(2m_1+m_3)}{\Gamma(4m_1+2m_3)} \exp(-b_1 \gamma_b) {\gamma_b}^{4m_1+2m_3-1} \, {}_1F_1(2m_3, 4m_1 + 2m_3, (b_2 - b_3)\gamma_b) \tag{15}$$

$$P_{out} = \int_0^{\gamma_{th}} \frac{\left(a_1{}^2 a_3 B(m_1,m_3) B(m_1+m_3,m_1)\right)^2 \Gamma(2m_1+m_3)^2}{\Gamma(4m_1+2m_3)} e^{(-b1\gamma)} \gamma^{a-1} \, {}_1F_1(b, a, \beta\gamma) dx \tag{16}$$

Fig. 4 shows a comparison between numerical and simulated results of the outage probability for different values of the fading parameter m. Obviously, there is an agreement between numerical and simulation results. It can be seen from the figure that the performance of the system is enhanced with the increase of value of m.

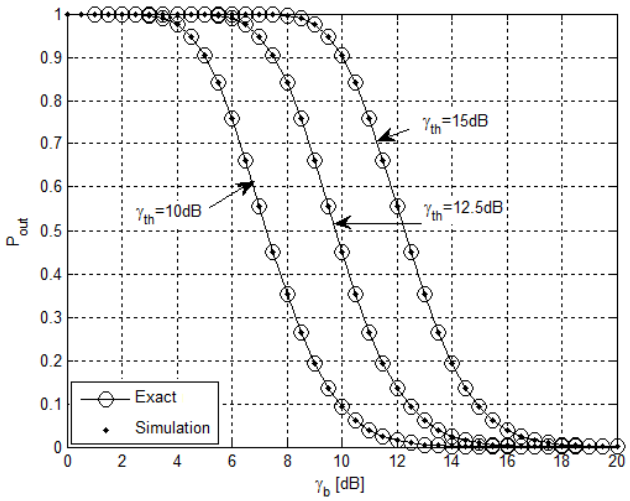


Fig.3 Outage probability of DCSK-CC system over Rayleigh fading channel at different values of γ_{th}.

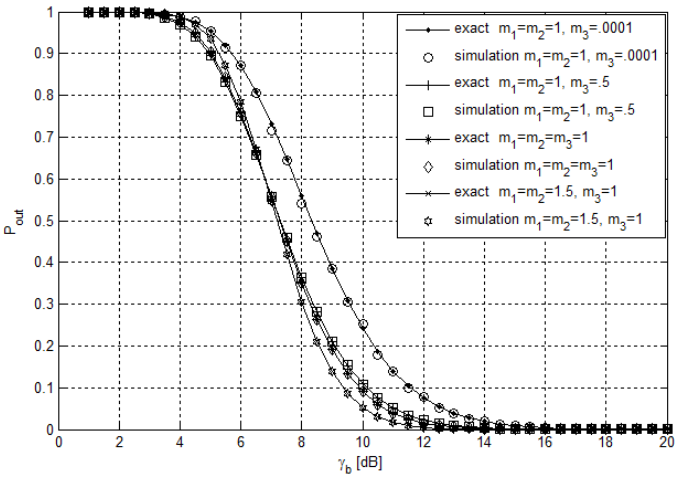


Fig.4 Outage probability of DCSK-CC system over Nakagami-m fading channel with different values of the fading parameter m , at fixed γ_{th}.

IV. CONCLUSION

In this paper, a two-user Cooperative system based on DCSK modulation over Rayleigh and Nakagami-m fading channels have been proposed. A discussion on the performance behavior of the system is provided by evaluating its performance in terms of outage probability. Closed form expression of this metric has been derived. The higher values of the fading parameter (m) provide better performance as expected. Simulation results using Monte Carlo method show a high agreement with the numerical results of outage probability.

REFERENCES

[1] W. M. Tam, F. C. M. Lau, and C. K. Tse, "Analysis of bit error rates for multiple access CSK and DCSK communication systems," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 50, pp. 702–707, May 2003.

[2] Georges Kaddoum and Daniel Roviras, "Performance analysis of differential chaos shift-keying over an m-distributed fading channel,", *Proc. of 2010 IEEE International Symposium on circuits and systems (ISCAS)*, pp. 2462-2465, Paris, May 30-June 2, 2010.

[3] Yongxiang Xia, Chi K. Tse and Francis C. M. Lau , "Performance of Differential Chaos-Shift-Keying Digital Communication Systems Over a Multipath Fading Channel With Delay Spread," *IEEE Transactions on Circuits and Systems—II: Express Briefs* , vpl. 51, no. 12, December 2004.

[4] Soumyajit MANDAL and Soumitro BANERJEE "Performance of Differential Chaos Shift Keying Communication over Multipath Fading Channels," IEICE Trans. , vol.E85–A, no.1, January 2002 .

[5] Yongxiang Xia, Chi K. Tse, Francis C. M. Lau and G´eza Kolumb´an "Performance of FM-DCSK Communication System Over a Multipath Fading Channel with Delay Spread," *Proc. of 2004 International Symposium on Nonlinear , Theory and its Applications (NOLTA2004)*, Fukuoka, Japan, Nov. 29 - Dec. 3, 2004.

[6] L. Wang, C. X. Zhang, and G. R. Chen, "Performance of an SIMO FM-DCSK communication system," IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 55, pp. 457–461, May 2008.

[7] L. A. De Michele, G. Cimatti, R. Rovatti, and G. Setti, "Joint design of a DS-UWB modulator and Chaos-based spreading sequences for sensor networks," *Proc. of IEEE ISCAS*, pp. 1441–1444, New Orleans, LA, May 2007.

[8] G. Mazzini, R. Rovatti, and G. Setti, "Chaos-based spreading in DS-UWB sensor networks increases available bit rate," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 6, pp. 1327–1339, Jun. 2007.

[9] C. C. Chong and S. K. Yong, "UWB direct chaotic communication technology for low-rate WPAN applications," *IEEE Trans. Veh. Technol.*, vol. 57, pp. 1527–1536, Mar. 2008.

[10] Weikai Xu, Lin Wang and Guanrong Chen "Performance of DCSK Cooperative Communication Systems Over Multipath Fading Channels" *IEEE Transactions on Circuits and Systems—I: Regular*, vol. 58, no. 1, January 2011.

[11] Chen Yiping, Shi Ying and Li Haijun, "Analysis of Performance for DCSK over Rayleigh Fading Channel," Proc. of 5th International Conference on Wireless Communications, Networking and Mobile Computing, WiCom '09, Beijing, September 24-26, 2009.

[12] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," IEEE Trans. Inform. Theory, vol. 50, Dec. 2004.

[13] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, 7th ed. Academic Press, 2007.

Outage Probability of Dual-Hop Amplify-and-Forward Cognitive Relay Networks Under Interference Power Constraints Over Nakagami-m Fading Channels

Mohammad Shurman[1], Mamoun F. Al-Mistarihi[2], and Moawiah Alhulayil[2]

(1) Jordan University of Science and Technology/ Network Engineering and Security Department, Irbid, Jordan
alshurman@just.edu.jo
(2) Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
mistarihi@just.edu.jo, muawiah-hlayil@live.com,

Abstract - **In this paper, the performance of the outage probability (OP) for cognitive dual-hop relay networks with amplify-and-forward (AF) relay under spectrum sharing constraints on primary user (PU) over independent and identically distributed (i.i.d.) Nakagami-m fading channels is investigated. An exact closed-form expression for the OP of the proposed system is derived under the peak interference power (I_p) at the primary user. The impact of the PU location on the OP performance is studied also.**

Keywords: **Nakagami-m fading, cognitive relay network, dual hop, cognitive radio, amplify-and-forward relaying, outage probability.**

I. INTRODUCTION

One of the most important topics that have been recently considered in the research areas is the Cooperative Diversity (CD), which makes a virtual antenna arrays among the cooperative nodes to combat the multipath effects and to extend the coverage area of the network [1]. In our wireless communication evolution the need for spectrum bands becomes more and more and it considers a big challenge in all applications either in Ad-hoc networks or in wireless local area networks (WLANs).

Cognitive Radio (CR) which was proposed in 1998 by Joseph Mitola is a promising technology to solve the spectrum shortage and to utilize the underutilized spectrum, where in CR we define two users, the first one is the primary user (PU) which have a licensed spectrum access and have the priority for transmission and the second one is the secondary user (SU) which have unlicensed spectrum access. The SU can access the channel if the PU does not use it, but suddenly if the PU decides to transmit its own data, the SU shall stop its transmission; this access strategy of spectrum sharing is called overlay strategy [2-4]. In the underlay strategy, the spectrum is accessible to the SU under an interference constraint at the PU receiver. Therefore, the SU is allowed to access the channel as long as its transmit power is below a certain threshold [5].

OP is one of the important performance metrics of wireless networks. For the outage probability definition; some authors defined it as the probability that the symbol error probability (SEP) exceeds a maximum tolerable level [6, 7], some authors defined it as the probability that the instantaneous signal-to-noise ratio (SNR) at the destination goes below a predefined threshold [8].

The OP for cognitive decode-and-forward (DF) relay networks has been derived in [9, 10]. In [11] the authors derived the OP of cognitive dual-hop relay networks with AF relay over Rayleigh fading channels under spectrum sharing constraints on a PU. Since the Nakagami-m distribution gives the best fit for land-mobile [12] and indoor mobile [13] multipath propagation, closed form expression for cognitive dual-hop relay networks with AF relay under spectrum sharing constraints on PU over i.i.d. Nakagami-m fading channels is derived in this paper. The improvement for the transmission using AF relaying over the conventional transmission (without relays) in cognitive radio networks (CRNs) is depicted, and the impact of the PU location on the OP performance is studied also.

This paper is further presented as follows; Section II provides the system and channel models, as well as the structure of the CRN components. Section III provides the performance analysis and derived expression of outage probability. Results and discussion are presented in section IV. We concluded this paper in section V.

II. SYSTEM AND CHANNEL MODELS

Consider a dual-hop cooperative spectrum sharing system as shown in Fig.1, where (S) is the SU source node, (D) is the SU destination node, (R) is the AF SU relay, and (P) is the PU receiver. Furthermore, we suppose that the direct link between nodes (S) and (D) is not available due to the severe shadowing and path loss. All nodes are provided with a single antenna and operate in a half-duplex mode.

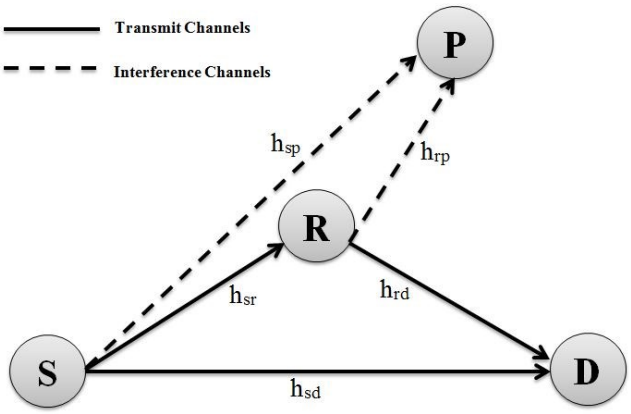


Fig.1: Dual-hop Cognitive relay network

Due to the interference power constraints, both the SU source and the SU relay should satisfy the two following relations respectively:

$$P_s \left| h_{sp} \right|^2 \le I_p \tag{1}$$

$$P_r \left| h_{rp} \right|^2 \le I_p \tag{2}$$

where P_s and P_r are the transmit power at the SU source and at the SU relay, respectively, I_p is the maximum allowable interference power on the PU receiver, and h_{sp} and h_{rp} are the channel coefficients links of S→P and R→P respectively.

In the 1st-hop transmission, the SU source transmits its signal to the relay under a transmit power constraint as mentioned before. Thus; the maximum transmit power at (S) can be expressed as:

$$P_s = \frac{I_p}{\left| h_{sp} \right|^2} \tag{3}$$

The received signal at (R) can be written also as:

$$y_r = \sqrt{P_s}\, x\, h_{sr} + n_r \tag{4}$$

where x is the transmitted information symbol, h_{sr} is the channel coefficient of the link S→R, and n_r is the additive white Gaussian noise (AWGN) at (R), having the power of N_o. In the second-hop transmission, the relay amplifies the received signal from (S) and forwards it to the destination in its new form given by:

$$\begin{aligned} y_D &= G\, h_{rd}\, y_r + n_D \\ &= \sqrt{P_s}\, x\, G\, h_{rd}\, h_{sr} + G h_{rd}\, n_r + n_D \end{aligned} \tag{5}$$

where h_{rd} is the channel coefficient of the link R→D, and n_D is the additive white Gaussian noise (AWGN) at (D), having the power of N_o, and G is the relay gain. Our choice of the relay gain will take into account the relay noise and assume channel state information (CSI)-assisted. So we can express the relay gain as:

$$G^2 = \frac{P_r}{P_s \left| h_{sr} \right|^2 + N_o} = \frac{\bar{\gamma} \left| h_{sp} \right|^2}{\left| h_{rp} \right|^2 \left(\bar{\gamma} \left| h_{sr} \right|^2 + \left| h_{sp} \right|^2 \right)} \tag{6}$$

where $P_r = I_p / \left| h_{rp} \right|^2$ is the transmit power at (R), and $\bar{\gamma} = I_p / N_o$.

We assume that the fading channels are i.i.d. flat Nakagami-m channels, and the channel power gains $X \in \{ \left| h_{sp} \right|^2, \left| h_{sr} \right|^2, \left| h_{rp} \right|^2, \left| h_{rd} \right|^2 \}$ have a gamma distribution, with means power Ω_{sp}, Ω_{sr}, Ω_{rp}, and Ω_{rd}, respectively. Thus the cumulative distribution function (CDF) of each one of these power gains can be expressed as:

$$F_X(x) = 1 - \frac{\Gamma(m_X, \lambda_X x)}{\Gamma(m_X)} \tag{7}$$

and the probability density function (PDF) as:

$$f_X(x) = \frac{{\lambda_X}^{m_X}}{\Gamma(m_X)} x^{m_X - 1} e^{-\lambda_X x} \tag{8}$$

where $m_X \in \{m_{sp}, m_{sr}, m_{rp}, m_{rd}\}$ represents the corresponding fading severity parameter m, $\lambda_X = m_X / \Omega_X$, and $\Gamma(.,.)$ is the incomplete gamma function given by [14, (8.350.2)].

III. PERFORMANCE ANALYSIS

To start our analysis for this cognitive relay network, we have to determine the instantaneous end-to-end signal-to-noise-ratios (SNRs) for each hop:

$$\gamma_1 = \frac{P_s \left| h_{sr} \right|^2}{N_o} = \frac{\bar{\gamma} \left| h_{sr} \right|^2}{\left| h_{sp} \right|^2} \tag{9}$$

$$\gamma_2 = \frac{P_r \left| h_{rd} \right|^2}{N_o} = \frac{\bar{\gamma} \left| h_{rd} \right|^2}{\left| h_{rp} \right|^2} \tag{10}$$

where γ_1 and γ_2 are the instantaneous end-to-end SNRs for the 1st-hop and the 2nd-hop, respectively. Thus; the instantaneous end-to-end SNR at (D) can be expressed as [15],

$$\gamma_{e-e} = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2 + 1} \tag{11}$$

The last equation may be expressed in a more mathematically tractable form. To achieve that; a tight upper-bound approximation was suggested in [16] using γ_D instead of γ_{e-e}, where γ_D is expressed as:

$$\gamma_D = \min(\gamma_1, \gamma_2) \tag{12}$$

Since γ_D is the minimum of two random variables, we need to find the CDF for the SNR of each hop and then apply the following equation to find the CDF for γ_D:

$$F_{\gamma_D}(\gamma) = 1 - \left(1 - F_{\gamma_1}(\gamma)\right)\left(1 - F_{\gamma_2}(\gamma)\right) \tag{13}$$

Theorem 1: The CDF of γ_1 is given by:

$$F_{\gamma_1}(\gamma) = 1 - \Upsilon_{sp,sr}(\gamma) \tag{14}$$

where $\Upsilon_{sp,sr}(\gamma)$ is given as in (15).

$$\Upsilon_{sp,sr}(\gamma) = \frac{(\lambda_{sp}\bar{\gamma})^{m_{sp}}(\lambda_{sr}\gamma)^{m_{sr}}}{m_{sp}\,(\lambda_{sp}\bar{\gamma}+\lambda_{sr}\gamma)^{m_{sp}+m_{sr}}} \cdot \frac{1}{B(m_{sp},m_{sr})} \cdot {}_2F_1\left(1, m_{sp}+m_{sr}; m_{sp}+1; \frac{\lambda_{sp}\bar{\gamma}}{\lambda_{sp}\bar{\gamma}+\lambda_{sr}\gamma}\right) \quad (15)$$

$$\Upsilon_{rp,rd}(\gamma) = \frac{(\lambda_{rp}\bar{\gamma})^{m_{rp}}(\lambda_{rd}\gamma)^{m_{rd}}}{m_{rp}\,(\lambda_{rp}\bar{\gamma}+\lambda_{rd}\gamma)^{m_{rp}+m_{rd}}} \cdot \frac{1}{B(m_{rp},m_{rd})} \cdot {}_2F_1\left(1, m_{rp}+m_{rd}; m_{rp}+1; \frac{\lambda_{rp}\bar{\gamma}}{\lambda_{rp}\bar{\gamma}+\lambda_{rd}\gamma}\right) \quad (17)$$

where $B(.,.)$ is the Beta function given by [14, (8.384.1)], and ${}_2F_1(.,.;.;.)$ is the Gauss hypergeometric function given by [14, (9.100)].

***Proof*:** The proof is given in Appendix A.

Theorem 2: The CDF of γ_2 is given by:

$$F_{\gamma_2}(\gamma) = 1 - \Upsilon_{rp,rd}(\gamma) \quad (16)$$

where $\Upsilon_{rp,rd}(\gamma)$ is given as in (17).

***Proof*:** The proof is given in Appendix B.

Substituting (14) and (16) into (13) the CDF of γ_D is:

$$F_{\gamma_D}(\gamma) = 1 - \Upsilon_{sp,sr}(\gamma)\,\Upsilon_{rp,rd}(\gamma) \quad (18)$$

Substitute γ_{th} into (18) the P_{out} will be as:

$$P_{out} = \Pr(\gamma_D \le \gamma_{th}) = 1 - \Upsilon_{sp,sr}(\gamma_{th})\,\Upsilon_{rp,rd}(\gamma_{th}) \quad (19)$$

IV. Results and Discussion

In this section, numerical results are given to validate our analysis, where the outage threshold is defined to be 1 dB. We consider the co-linear topology for all nodes in our system, where the distance between the SU source (S) and the SU destination (D) is normalized to be one. As a result the (S) and (D) are located at coordinates (0,0) and (1,0), respectively. The SU relay (R) location is designed to be at the middle distance between (S) and (D) with a coordinates (0.5,0).

The path loss of each channel follows an exponential decay model; which means that the channel mean power between any two nodes is modeled as $\Omega_X = d^{-\alpha}$, where α is the path loss exponent and it can be set to four in the free-space propagation with non-line-of-sight model.

Fig.2 depicts the OP performance of cognitive spectrum sharing system for two different transmission scenarios including the AF relaying (no direct link), and the direct transmission. Hence, the OP of the direct spectrum sharing scenario is given as:

$$P_{out}{}^{direct\ transmission} = 1 - \Upsilon_{sp,sd}(\gamma_{th}) \quad (20)$$

As can be seen from Fig.2, the first scenario of transmission using AF relaying outperforms that of direct transmission, and more than 5 dB gain can be obtained using the AF relaying transmission compared to the conventional one. One note to be mentioned too is that the PU is at the coordinate (0.6,0.6) and the severity parameter is two for channels.

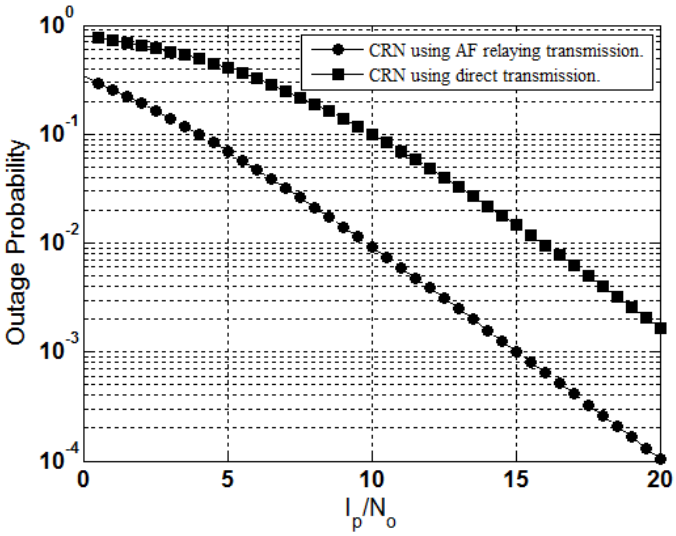


Fig. 2 OP of cognitive network using AF relaying transmission and direct transmission.

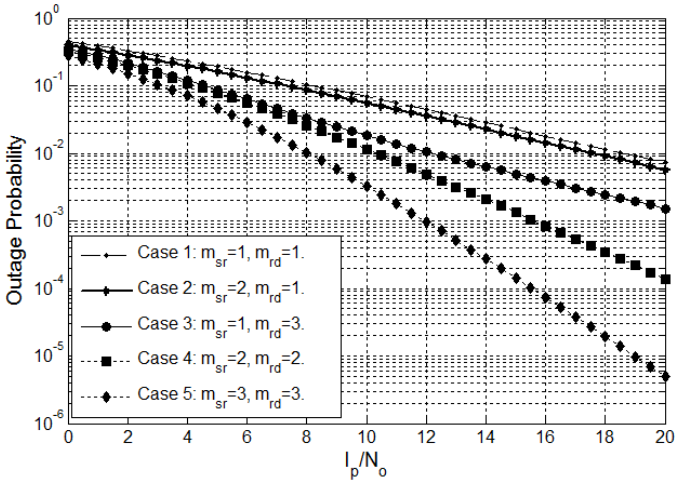


Fig. 3 OP of cognitive relay network (varying the severity parameters of the channels).

Fig.3 shows the impact of fading severity of the secondary network links on the outage performance where the PU location is at (0.6,0.6) coordinates and the fading severity parameters for the interference links are (m_{sp}=m_{rp}=1). The figure shows five cases with different fading severity parameters; Case 1: the two hops have the same severity parameters (m_{sr}=m_{rd}=1); which represents the Rayleigh case, Case 2: the first hop is less severe than the second hop (m_{sr}=2, m_{rd}=1), Case 3: the second hop is less severe than the first hop (m_{sr}=1, m_{rd}=3), Case 4: the two hops have the same severity

parameters (m_{sr}=m_{rd}=2), and Case 5: the two hops have the same severity parameters (m_{sr}=m_{rd}=3). The first thing to be noted here is that our results conform the results in [11] which describes the Rayleigh fading channels and it is considered the worst outage performance case in all cases. The results show that the outage performance depends on the minimum fading severity parameter of the two hops and the best outage performance case is case 5; since the two hop have the same severity parameters which equal three.

The outage performance for different locations of the PU is depicted in Fig.4. It is obvious that the best performance can be achieved for the case when the PU at coordinate (0.8,0.8), in which the minimum interference power of the PU is satisfied. From the figure, the outage probability will decrease as the PU moves away from the secondary network.

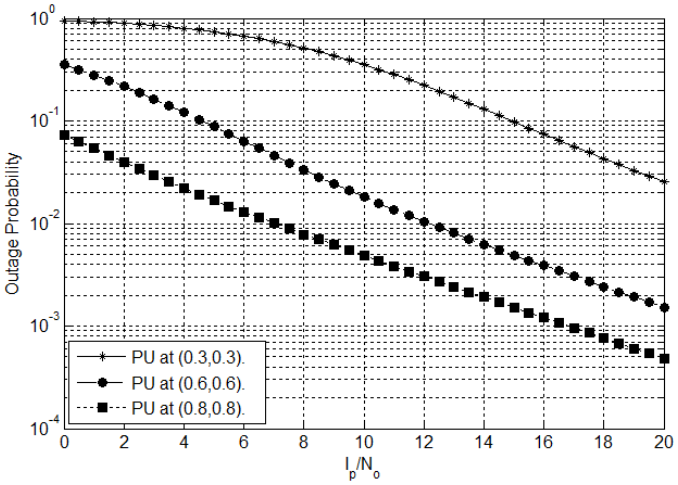


Fig. 4 OP of cognitive relay network (varying the PU location, and m_{sp}=m_{rp}= m_{sr}=1, m_{rd}=3).

V. CONCLUSION

In this paper, considering a spectrum sharing scenario, the outage probability of an underlay cognitive AF relay network over i.i.d. Nakagami-m fading channels under interference power constraiants on the PU has been investigated. From the results we see that the cognitive AF relaying transmission outperforms the conventional direct transmission, and the outage performance depends on the minimum fading severity among the two hops. Furthermore, the impact of the PU location has been studied and it is shown that the PU location plays a main role in the outage perfmorance; as the PU moves away from the secondary network, the outage probability decreases. As a result, these findings are very important for the wireless networks design.

APPENDIX A

For notational simplicity, let us denote the following terms: $X_1 = |h_{sr}|^2$ and $Y_1 = |h_{sp}|^2$.

In order to find the CDF of the SNR for the 1st-hop, $F_{\gamma_1}(\gamma)$, we need to rewrite γ_1 as:

$$\gamma_1 = \frac{\bar{\gamma}|h_{sr}|^2}{|h_{sp}|^2} = \bar{\gamma}\frac{X_1}{Y_1} \quad (A.1)$$

So we can express $F_{\gamma_1}(\gamma)$ as:

$$F_{\gamma_1}(\gamma) = \int_0^\infty F_{X_1}\left(\frac{\gamma y}{\bar{\gamma}}\right) f_{Y_1}(y)dy \quad (A.2)$$

Then,

$$F_{\gamma_1}(\gamma) = \int_0^\infty \left[1 - \frac{\Gamma\left(m_{sr}, \frac{\lambda_{sr}\gamma y}{\bar{\gamma}}\right)}{\Gamma(m_{sr})}\right]\left[\frac{\lambda_{sp}{}^{m_{sp}}}{\Gamma(m_{sp})} y^{m_{sp}-1} e^{-\lambda_{sp} y}\right] dy \quad (A.3)$$

and for simplictiy we can write the above integral as:

$$F_{\gamma_1}(\gamma) = I_1 - I_2 \quad (A.4)$$

where

$$I_1 = \int_0^\infty \left[\frac{\lambda_{sp}{}^{m_{sp}}}{\Gamma(m_{sp})} y^{m_{sp}-1} e^{-\lambda_{sp} y}\right] dy \quad (A.5)$$

and

$$I_2 = \int_0^\infty \left[\frac{\Gamma\left(m_{sr}, \frac{\lambda_{sr}\gamma y}{\bar{\gamma}}\right)}{\Gamma(m_{sr})}\right]\left[\frac{\lambda_{sp}{}^{m_{sp}}}{\Gamma(m_{sp})} y^{m_{sp}-1} e^{-\lambda_{sp} y}\right] dy \quad (A.6)$$

$I_1 = 1$ using (A.7) below,

$$\int_0^\infty y^{\mathrm{n}} e^{-\mathrm{a}y} dy = \frac{\Gamma(n+1)}{a^{n+1}} ; n > -1 , a > 0 \quad (A.7)$$

using [14, (6.455.1)] we can express I_2 as:

$$I_2 = \frac{(\lambda_{sp}\bar{\gamma})^{m_{sp}}(\lambda_{sr}\gamma)^{m_{sr}}}{m_{sp}\,(\lambda_{sp}\bar{\gamma}+\lambda_{sr}\gamma)^{m_{sp}+m_{sr}}} \cdot \frac{1}{B(m_{sp},m_{sr})} \cdot {}_2F_1\left(1, m_{sp}+m_{sr}; m_{sp}+1; \frac{\lambda_{sp}\bar{\gamma}}{\lambda_{sp}\bar{\gamma}+\lambda_{sr}\gamma}\right) \quad (A.8)$$

where $B(.,.)$ is the Beta function given by [14, (8.384.1)], and ${}_2F_1(.,.;.;.)$ is the Gauss hypergeometric function given by [14, (9.100)]. By substituting I_1 and I_2 in (A.4) we get $F_{\gamma_1}(\gamma)$ as:

$$F_{\gamma_1}(\gamma) = 1 - \Upsilon_{sp,sr}(\gamma) \quad (A.9)$$

where

$$\Upsilon_{sp,sr}(\gamma) = \frac{(\lambda_{sp}\bar{\gamma})^{m_{sp}}(\lambda_{sr}\gamma)^{m_{sr}}}{m_{sp}\,(\lambda_{sp}\bar{\gamma}+\lambda_{sr}\gamma)^{m_{sp}+m_{sr}}} \cdot \frac{1}{B(m_{sp},m_{sr})} \cdot {}_2F_1\left(1, m_{sp}+m_{sr}; m_{sp}+1; \frac{\lambda_{sp}\bar{\gamma}}{\lambda_{sp}\bar{\gamma}+\lambda_{sr}\gamma}\right) \quad (A.10)$$

APPENDIX B

For notational simplicity, let us denote the following terms: $X_2 = |h_{rd}|^2$ and $Y_2 = |h_{rp}|^2$.

In order to find the CDF of the SNR for the 2nd-hop, $F_{\gamma_2}(\gamma)$, we need to rewrite the γ_2 as:

$$\gamma_2 = \frac{\bar{\gamma}\,|h_{rd}|^2}{|h_{rp}|^2} = \bar{\gamma}\frac{X_2}{Y_2} \quad \text{(B.1)}$$

So; we can express $F_{\gamma_2}(\gamma)$ as:

$$F_{\gamma_2}(\gamma) = \int_0^{\infty} F_{X_2}\left(\frac{\gamma y}{\bar{\gamma}}\right) f_{Y_2}(y)\,dy \quad \text{(B.2)}$$

using the same procedure in Appendix A, we get the CDF of the SNR for the 2nd-hop as:

$$F_{\gamma_2}(\gamma) = 1 - \Upsilon_{rp,rd}(\gamma) \quad \text{(B.3)}$$

where

$$\Upsilon_{rp,rd}(\gamma) = \frac{(\lambda_{rp}\bar{\gamma})^{m_{rp}}(\lambda_{rd}\gamma)^{m_{rd}}}{m_{rp}\,(\lambda_{rp}\bar{\gamma}+\lambda_{rd}\gamma)^{m_{rp}+m_{rd}}} \cdot \frac{1}{B(m_{rp},m_{rd})} \cdot {}_2F_1\left(1, m_{rp}+m_{rd}; m_{rp}+1; \frac{\lambda_{rp}\bar{\gamma}}{\lambda_{rp}\bar{\gamma}+\lambda_{rd}\gamma}\right) \quad \text{(B.4)}$$

REFERENCES

[1] Laneman, J. N., Tse, D. N. C., and Wornell, G. W., "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, 2004.

[2] Huang, Senhua, Xin Liu, and Zhi Ding, "Opportunistic spectrum access in cognitive radio networks," *Proc. 27th IEEE Conference on Computer Communications, INFOCOM 2008*, Phoenix, AZ, April 13-18, 2008.

[3] Lee, Won-Yeol, and Ian F. Akyildiz, "Optimal spectrum sensing framework for cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, issue 10, pp. 3845-3857, 2008.

[4] Tan, Sheu-Sheu, James Zeidler, and Bhaskar Rao, "Opportunistic channel-aware spectrum access for cognitive radio networks with periodic sensing," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*, pp. 5333 – 5337, Vancouver, BC, May 26-31, 2013.

[5] Le, Long Bao, and Ekram Hossain, "Resource allocation for spectrum underlay in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7,issue 12, pp. 5306-5315, 2008.

[6] A. Conti, M. Z. Win, M. Chiani, and J. H. Winters, "Bit error outage for diversity reception in shadowing environment," *IEEE Commun. Lett*., vol. 7, issue 1, pp. 15-17, 2003.

[7] A. Conti, M. Z. Win, and M. Chiani, "On the inverse symbol-error probability for diversity reception," *IEEE Trans. Commun.*, vol. 51, issue 5, pp. 753-756, 2003.

[8] Bao, V., Duong, T., Da Costa, D., Alexandropoulos, G., and Nallanathan, A, "Cognitive amplify-and-forward relaying with best relay selection in non-identical Rayleigh fading," *IEEE communication letters*, vol. 17, issue 3, 2013.

[9] Luo, L., Zhang, P., Zhang, G., and Qin, J., "Outage performance for cognitive relay networks with underlay spectrum sharing," *IEEE Commun. Lett.* vol. 15, issue 7, pp. 710–712, 2011.

[10] Guo, Y., Kang, G., Zhang, N., Zhou, W., and Zhang, P., "Outage performance of relay-assisted cognitive-radio system under spectrum sharing constraints," *Electron. Lett.* vol. 46, issue 2, pp. 182–184, 2010.

[11] Duong, T. Q., Vo Nguyen Quoc Bao, and H-J. Zepernick, "Exact outage probability of cognitive AF relaying with underlay spectrum sharing," *Electronics letters.* vol. 47, issue 17, pp. 1001-1002, 2011.

[12] W. R. Braun and U. Dersch, "A physical mobile radio channel model," *IEEE Trans. Veh. Technol.* VT-40, pp. 472–482, 1991.

[13] A. U. Sheikh, M. Handforth, and M. Abdi, "Indoor mobile radio channel at 946 MHz: measurements and modeling," *Proc. IEEE Veh. Technol. Conf. VTC'93*, pp. 73-76, Secaucus, NJ, May 18-20, 1993.

[14] Jeffrey, Alan, and Daniel Zwillinger, eds, *Table of integrals, series, and products*, 7th ed. Academic Press, 2007.

[15] M. O. Hasna and M-S. Alouini, "Outage probability of multi-hop transmission over Nakagami fading channels," IEEE Commun. Lett., vol. 7, issue 5, pp. 216–218, 2003.

[16] Ikki S., Ahmed M.H., "Performance analysis of cooperative diversity wireless networks over Nakagami-m fading channel," *IEEE communication letters,* vol. 11, issue 4, pp. 334 – 336, 2007.

Outage Probability of Amplify-and-Forward Underlay Cognitive Relay Networks with Selection Diversity over Nakagami-m Fading Channels

Moawiah Alhulayil[1], Mamoun F. Al-Mistarihi[1], and Mohammad Shurman[2]

(1) Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
muawiah-hlayil@live.com, mistarihi@just.edu.jo
(2) Jordan University of Science and Technology/ Network Engineering and Security Department, Irbid, Jordan
alshurman@just.edu.jo

Abstract - **In this paper, the outage performance is investigated for cognitive dual-hop relay networks with amplify-and-forward (AF) relay and selection combining (SC) receiver at the destination under spectrum sharing constraints on primary user (PU) over independent and identically distributed (i.i.d.) Nakagami-m fading channels. We derive an exact closed-form expression for the outage probability (OP) of the considered system under the peak interference power I_p at the primary user, and the impact of the PU location on the OP performance is studied also. Moreover; the effects fading severity for the two hops on the OP is investigated also for this cognitive relay network (CRN).**

Keywords: **Nakagami-m fading, cognitive relay network, dual hop, cognitive radio, amplify-and-forward relaying, selection diversity, outage probability.**

I. INTRODUCTION

Recent measurements have shown that most of the allocated spectrum is under-utilized and the new wireless services need more spectrum bands. To improve inefficient use of allocated spectrum, the concept of Cognitive Radio (CR) has been introduced. CR is one of the important technologies for wireless applications that has considered as a spectrum opportunistic technology. The main goal of CR is to exploit licensed bands without a license, where the unlicensed user (Secondary User) can use the spectrum band of the licensed user (Primary User) when the primary user is idle [1-3].

One of the important issues in CR is the sensing accuracy which refers to the precision in detecting the presence of the primary user[4]. Therefore; there are many solutions are existed to deal with this issue such as cooperative spectrum sensing where multiple secondary users cooperate together to improve the sensing accuracy and it takes smaller time to check the presence of the primary user resulting both decreasing in the sensing time period and increasing in the transmission time period for secondary user[5,6].

The spectrum is accessible to the SU the underlay strategy under an interference constraint at the PU receiver. Therefore, the SU is allowed to access the channel as long as its transmit power is below a certain threshold [7]. For the outage probability definition, some authors defined it as the probability that the symbol error probability (SEP) exceeds a maximum tolerable level [8, 9], some authors defined it as the probability that the instantaneous signal-to-noise ratio (SNR) at the destination goes below a predefined threshold [10].

The OP for cognitive decode-and-forward (DF) relay networks has been derived in [11, 12]. In [13] the authors derived the OP of cognitive dual-hop relay networks with AF relay over Rayleigh fading channels under spectrum sharing constraints on a PU. In [14] an exact OP of cognitive dual-hop relay networks wth AF relay and selection combining receiver at the destination over Rayleigh fading channels has been derived. While in [15] closed-form lower-bounds and asymptotic expressions for the outage probability of dual hop cognitive relay network with a single AF relay has been derived using different strategies to determine the transmit power of the SU source using the power series for the lower incomplete gamma function.

Comparing AF with DF relay, AF requires much less delay as the relay node operates time slot by time slot and it requires much less computing power as no decoding operation, and since the Nakagami-m distribution gives the best fit for land-mobile and indoor-mobile [16] multipath propagation; exact closed form expression for the outage probabilitiy of cognitive relay network with AF relay and selection diversity under spectrum sharing constraints on PU over i.i.d. Nakagami-m fading channels is derived in this paper. The impact of the of the PU location on the outage probability will be studied by varying the location of the PU with respect to the SU network and the effects of the fading severity for the two hops on the outage probability will be investigated also for this CRN.

The main difference in opportunistic relay selection between CR networks and non-CR networks that is in CR networks due to the channel gain of the link from the secondary user source to the primary user the individual SNR

of each path contains the same random variable, which is not easy for analysis. So we will deal with a problem of multiple statistically dependent random variables.

This paper is further presented as follows; Section II provides the system and channel models, as well as the structure of the CRN components. Section III describes the analysis framework for evaluating the outage probability. Numerical results and discussion are presented in section IV. We concluded this paper in section V.

II. SYSTEM AND CHANNEL MODELS

Consider a dual-hop cooperative spectrum sharing system as shown in Fig.1, it is shown that the communication between the SU source (S) and the SU destination (D) may be satisfied using two paths; the first path is by using the direct conventional transmission and the second one is by using the AF relay link. For the direct link, the transmission takes palce in one phase where the SU source (S) transmits its signal to the SU destination (D) under a transmit power constraint I_p with a transmit power

$$P_s = \frac{I_p}{|h_{sp}|^2} \tag{1}$$

where h_{sp} is the channel coefficient of the link S→P. The transmission using the AF relay takes place in two phases. In phase I, the SU source transmits its signal to the relay under a transmit power I_p constraint with a transmit power $P_s = I_p / |h_{sp}|^2$. In phase II, the relay amplifies the received signal from (S) and forwards it to the destination, and the signal is transmitted with a transmit power of $P_r = I_p / |h_{rp}|^2$, where h_{rp} is the channel coefficient of the link R→P.

Both signals from the relaying path and from the direct path are combined together at the destination using a SC receiver to choose the best path for transmission. All nodes in our system are provided with a single antenna and operate in a half-duplex mode . Moreovere, we assume the channel state information (CSI)-assisted case.

We assume that the fading channels are i.i.d. flat Nakagami-m channels, and the channel power gains $X \in \{ |h_{sp}|^2, |h_{sr}|^2, |h_{rp}|^2, |h_{rd}|^2, |h_{sd}|^2 \}$ have a gamma distribution, with means power Ω_{sp}, Ω_{sr}, Ω_{rp}, Ω_{rd}, and Ω_{sd}, respectively. Thus the cumulative distribution function (CDF) of each one of these power gains can be expressed as:

$$F_X(x) = 1 - \frac{\Gamma(m_X, \lambda_X x)}{\Gamma(m_X)} \tag{2}$$

and the probability density function (PDF) as

$$f_X(x) = \frac{{\lambda_X}^{m_X}}{\Gamma(m_X)} x^{m_X - 1} e^{-\lambda_X x} \tag{3}$$

where $m_X \in \{m_{sp}, m_{sr}, m_{rp}, m_{rd}, m_{sd}\}$ represents the corresponding fading severity parameter m, $\lambda_X = m_X / \Omega_X$, and $\Gamma(.,.)$ is the incomplete gamma function given by [17, (8.350.2)].

III. PERFORMANCE ANALYSIS

In order to find the outage probability performance, we have to determine the instantaneous end-to-end signal-to-noise-ratios (SNRs) received at the destination (D) from the two paths. The isnataneous SNR at (D) reffering to the direct link transmission can be expressed as

$$\gamma_{direct} = \frac{P_s |h_{sd}|^2}{N_o} = \frac{\bar{\gamma} |h_{sd}|^2}{|h_{sp}|^2} \tag{4}$$

where $\bar{\gamma} = I_p / N_o$. The instantaneous SNR at (D) referring to

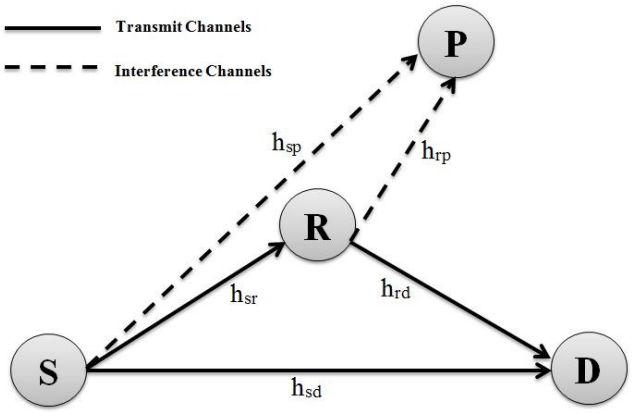


Fig.1: System Model for cognitive AF relay network with a selection combining receiver.

the AF relaying link transmission can be expressed as [18]:

$$\gamma_{relay} = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2 + 1} \tag{5}$$

where γ_1 and γ_2 are the instantaneous end-to-end SNRs for the 1st-hop and the 2nd-hop, respectively, given by:

$$\gamma_1 = \frac{P_s |h_{sr}|^2}{N_o} = \frac{\bar{\gamma} |h_{sr}|^2}{|h_{sp}|^2} \tag{6}$$

$$\gamma_2 = \frac{P_r |h_{rd}|^2}{N_o} = \frac{\bar{\gamma} |h_{rd}|^2}{|h_{rp}|^2} \tag{7}$$

Thus; we can rewrite γ_{relay} as follows:

$$\gamma_{relay} = \frac{\frac{\bar{\gamma} |h_{sr}|^2}{|h_{sp}|^2} \frac{\bar{\gamma} |h_{rd}|^2}{|h_{rp}|^2}}{\frac{\bar{\gamma} |h_{sr}|^2}{|h_{sp}|^2} + \frac{\bar{\gamma} |h_{rd}|^2}{|h_{rp}|^2} + 1} \tag{8}$$

It can be seen from (4) and (8) that γ_{direct} and γ_{relay} are statistically dependent due to the presence of the common random variable h_{sp} which makes the performance analysis not easy to find. (8) may be expressed in a more mathematically tractable form, to achieve that; a tight upper-

bound approximation was suggested in [16]. Thus; we will use γ_t isntead of γ_{relay}, where γ_t is expressed as:

$$\gamma_t = min\ (\gamma_1, \gamma_2) \tag{9}$$

Since we have the SC receiver at the destination that chooses the best path for transmission; we may express γ_{SC} as

$$\gamma_{SC} = max\ (\gamma_{direct}, \gamma_t) \tag{10}$$

where γ_{SC} is the instantaneous SNR at the destination using a SC receiver.

Due to the presence of the common random variable h_{sp}, we will apply the conditional statistics on the channel coefficient of the link S→P, and it is important to note that the remaining random variables, $\{|h_{sr}|^2, |h_{rp}|^2, |h_{rd}|^2, |h_{sd}|^2\}$, are independent. Since γ_{SC} is the maximum of two random variables, we need to find the CDF conditioned on h_{sp} for the SNR of each path and to apply the following equation

$$F_{\gamma_{SC}}(\gamma/Y) = F_{\gamma_{direct}}(\gamma/Y)\, F_{\gamma_t}(\gamma/Y) \tag{11}$$

where $Y = |h_{sp}|^2$.

Theorem 1: The CDF of the SNR for the direct path conditioned on *Y* is given by:

$$F_{\gamma_{direct}}(\gamma/Y) = 1 - \frac{\Gamma\left(m_{sd}, \frac{\lambda_{sd}\gamma}{\bar{\gamma}}y\right)}{\Gamma(m_{sd})} \tag{12}$$

Proof: The proof is given in Appendix A.

Theorem 2: The CDF of the SNR for the AF-relaying path conditioned on *Y* is given by

$$F_{\gamma_t}(\gamma/Y) = 1 - \Upsilon_{rp,rd}(\gamma)\frac{\Gamma\left(m_{sr}, \frac{\lambda_{sr}\gamma}{\bar{\gamma}}y\right)}{\Gamma(m_{sr})} \tag{13}$$

where $\Upsilon_{uv,ij}(\gamma)$ is a function of γ and it is given by:

$$\Upsilon_{uv,ij}(\gamma) = \frac{(\lambda_{uv}\bar{\gamma})^{m_{uv}}(\lambda_{ij}\gamma)^{m_{ij}}}{m_{uv}\left(\lambda_{uv}\bar{\gamma} + \lambda_{ij}\gamma\right)^{m_{uv}+m_{ij}}} \cdot \frac{1}{B(m_{uv}, m_{ij})} \cdot {}_2F_1\left(1, m_{uv} + m_{ij}; m_{uv}+1; \frac{\lambda_{uv}\bar{\gamma}}{\lambda_{uv}\bar{\gamma} + \lambda_{ij}\gamma}\right) \tag{14}$$

where $B(.,.)$ is the Beta function given by [17, (8.384.1)] and ${}_2F_1(.,.;.;.)$ is the Gauss hypergeometric function given by [14, (9.100)].

Proof: The proof is given in Appendix B.

Using theorems 1 and 2 and after some mathematical manipulations, $F_{\gamma_{SC}}(\gamma/Y)$ can be written as follows,

$$F_{\gamma_{SC}}(\gamma/Y) = 1 - \Upsilon_{rp,rd}(\gamma)\frac{\Gamma\left(m_{sr}, \frac{\lambda_{sr}\gamma}{\bar{\gamma}}y\right)}{\Gamma(m_{sr})} - \frac{\Gamma\left(m_{sd}, \frac{\lambda_{sd}\gamma}{\bar{\gamma}}y\right)}{\Gamma(m_{sd})} + \Upsilon_{rp,rd}(\gamma)\frac{\Gamma\left(m_{sr}, \frac{\lambda_{sr}\gamma}{\bar{\gamma}}y\right)}{\Gamma(m_{sr})}\frac{\Gamma\left(m_{sr}, \frac{\lambda_{sd}\gamma}{\bar{\gamma}}y\right)}{\Gamma(m_{sd})} \tag{15}$$

Since the OP is the probability that the instatneous SNR at (D) goes below a predefined threshold, γ_{th}, we can find the OP as :

$$\mathrm{OP} = \Pr(\gamma_{SC} \le \gamma_{th}) = \int_0^\infty F_{\gamma_{SC}}(\gamma_{th}/Y) f_Y(y)dy \tag{16}$$

Substituting (15) into (16) and using [17, (6.455.1)], [17, (8.352.4)], and the multinomial theorem we can express the OP of underlay CRN with a single AF relay plus direct link over i.i.d. Nakagami-m fading channels using a SC receiver as:

$$OP = 1 - \Upsilon_{rp,rd}(\gamma_{th})\Upsilon_{sp,sr}(\gamma_{th}) - \Upsilon_{sp,sd}(\gamma_{th}) + \frac{{\lambda_{sp}}^{m_{sp}}\ \Upsilon_{rp,rd}(\gamma_{th})}{\Gamma(m_{sp})} \sum_{l,w,m_{sd},\lambda_{sd}}^{*} \sum_{t,u,m_{sr},\lambda_{sr}}^{*} \left(\frac{\gamma_{th}}{\bar{\gamma}}\right)^{S_1+S_2} \frac{\Gamma(S_1+S_2+m_{sp})}{[\lambda_{sp} + \frac{\gamma_{th}}{\bar{\gamma}}(\lambda_{sd}+\lambda_{sr})]^{S_1+S_2+m_{sp}}} \tag{17}$$

where

$$\sum_{l,w,m_X,\lambda_X}^{*} = \sum_{l_1 \dots l_{m_X}}^{*} \frac{1}{l_1! \dots l_{m_X}!} \prod_{w=0}^{m_X-1} \left(\frac{{\lambda_X}^w}{w!}\right)^{l_{w+1}} \tag{18}$$

$$S_1 = \sum_{j=0}^{m_{sd}-1} j\ l_{j+1} \tag{19}$$

$$S_2 = \sum_{j=0}^{m_{sr}-1} j\ t_{j+1} \tag{20}$$

IV. Numerical Results and Discussion

In this section, we show the numerical results to validate our analysis. We consider a co-linear topology for All nodes in our cognitive relay network model, where the distance between the SU source (S) and the SU destination (D) is normalized to be one, and the SU relay (R) is located at the middle distance between (S) and (D). Therefore, the coordinates for (S), (D), and (R) are (0,0), (1,0), and (0.5,0), respectively.

The outage threshold is defined to be 1 dB and the path loss of each channel follows an exponential decay model with an exponent can be set to four in the free-space propagation with non-line-of-sight model. The OP performance for different transmission scenarios of cognitive spectrum sharing including the conventional transmission (without relays), transmission via AF relaying, and cooperative transmission with selection combining are compared in this section. The

impact of fading severity of the secondary network links on the OP is evaluated for a fixed location of the PU. Finally, we evaluate the impact of the PU location on the OP by taking different locations for the PU receiver.

Fig.2 compares the outage performance for three different transmission scenarios of cognitive spectrum sharing network including the direct transmission, AF relaying transmission, and cooperative transmission with SC receiver. As can be seen from the figure that the cooperative spectrum sharing transmission with SC outperforms the two other transmission scenarios since the diversity order increased to two. More than 5 dB gain can be obtained using the AF relaying transmission compared to the conventional one. The PU is at the coordinate (0.6,0.6).

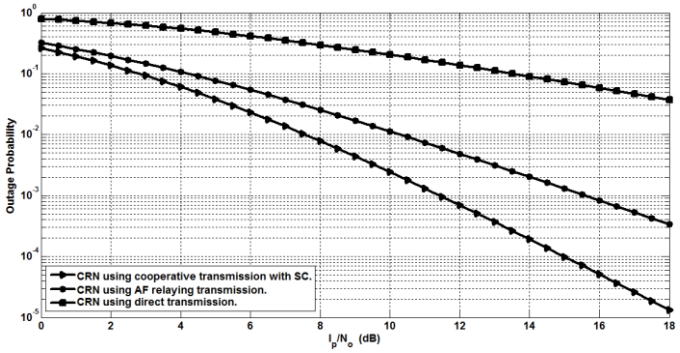


Fig.2: OP of cognitive network using direct transmission, AF relaying transmission, and cooperative transmission with SC (m_{sp}= m_{sr}= m_{rd}=2, m_{rp}= m_{sd}=1).

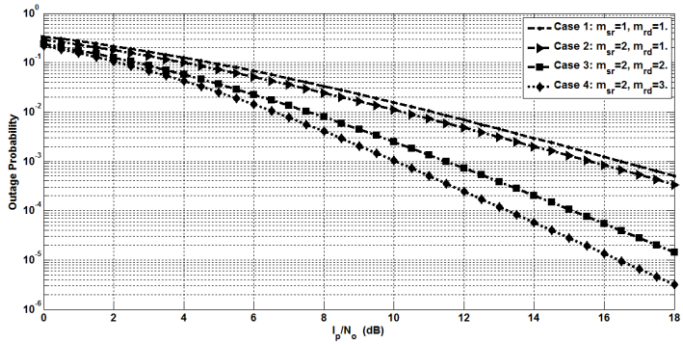


Fig.3: OP of cognitive relay network (varying the severity parameters of the channels).

Fig.3 shows the impact of fading severity of the secondary network links on the outage performance where the PU location is at (0.6,0.6) coordinates and the fading severity parameters for the interference links are (m_{sp}=m_{rp}=1) and the fading severity between (S) and (D) is (m_{sd}=1). The figure shows four cases with different fading severity parameters; Case 1: the two hops have the same severity parameters (m_{sr}=m_{rd}=1); which represents the Rayleigh case, Case 2: the first hop is less severe than the second hop (m_{sr}=2, m_{rd}=1), Case 3: the two hops have the same severity parameters (m_{sr}=m_{rd}=2), and Case 4: the second hop is less severe than the first hop (m_{sr}=2, m_{rd}=3). The first thing to be noted here that our results conform the results in [14] which describe the Rayleigh fading channels and it is considered the worst outage performance case in all cases. The results show that the outage performance depends on the minimum fading severity parameter of the two hops and the best outage performance case is case 4; since the two hops have the same severity parameters which equal two.

The outage probabilities for different locations of the PU is depicted in Fig.4. The best performance can be achieved for the case when the PU at coordinate (0.8,0.8), in which the minimum interference power of the PU is satisfied. So, the outage probability will decrease as the PU moves away from the secondary network.

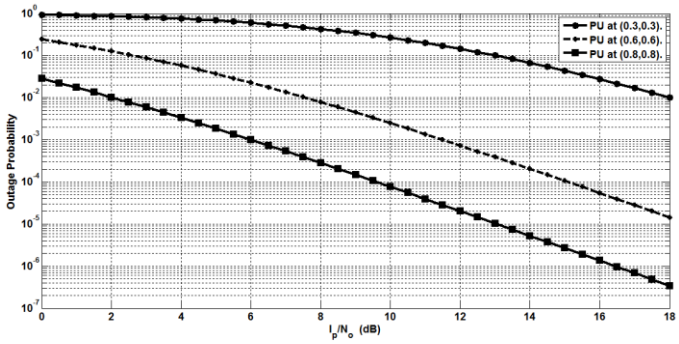


Fig.4: OP of cognitive relay network (varying the PU location, and m_{sp}=m_{rp}= m_{sd}=1, m_{sr}= m_{rd}=2).

V. CONCLUSION

In this paper, we have derived exact closed-form expression for the outage probability of cognitive AF relay network over i.i.d. Nakagami-m fading channels under interference power constraints on the PU. Many important issues for wireless network design have concluded. The first one; the cooperative spectrum sharing transmission outperforms the conventional direct transmission and the transmission using AF relaying since the the diversity order increased by two. Moreover, we have shown that the outage performance depends on the minimum fading severity among the two hops of the relaying link. Finally, the effect of the PU location on the network performance has been studied and it is shown that varying the location of the PU faraway from the SU network improves the outage performance.

APPENDIX A

For notational simplicity, let us denote the following terms: $X_0 = |h_{sd}|^2$ and $Y = |h_{sp}|^2$. In order to find the CDF of the instantaneous SNR for the direct path conditioned on h_{sp}, $F_{\gamma_{direct}}(\gamma/Y)$, we need to rewrite γ_{direct} as

$$\gamma_{direct} = \frac{\bar{\gamma}\,|h_{sd}|^2}{|h_{sp}|^2} = \bar{\gamma}\frac{X_0}{Y} \qquad (A.1)$$

Therefore; we can express $F_{\gamma_{direct}}(\gamma/Y)$ as follows,

$$F_{\gamma_{direct}}(\gamma/Y) = \Pr\left(\frac{X_0}{Y} \le \frac{\gamma}{\bar{\gamma}}\right) = F_{X_0}\left(\frac{\gamma}{\bar{\gamma}}y\right) \qquad (A.2)$$

After the substitution in the Nakgamai-m CDF , we can get

$$F_{\gamma_{direct}}(\gamma/Y) = 1 - \frac{\Gamma\left(m_{sd}, \frac{\lambda_{sd}\gamma}{\bar{\gamma}} y\right)}{\Gamma(m_{sd})} \quad (A.3)$$

APPENDIX B

In order to find the CDF of the instantaneous SNR for relaying path conditioned on h_{sp}, $F_{\gamma_t}(\gamma/Y)$, we need to find the CDFs for both hops conditioned on h_{sp} . For notational simplicity, let us denote the following terms: $X_1 = |h_{sr}|^2$ and $X_2 = \frac{\bar{\gamma}|h_{rd}|^2}{|h_{rp}|^2}$.

We can rewrite the instantaneous SNR for the 1st-hop as follows,

$$\gamma_1 = \frac{\bar{\gamma}|h_{sr}|^2}{|h_{sp}|^2} = \bar{\gamma}\frac{X_1}{Y} \quad (B.1)$$

Therefore; we can express the CDF for the 1st hop conditioned on h_{sp}, $F_{\gamma_1}(\gamma/Y)$, as follows,

$$F_{\gamma_1}(\gamma/Y) = F_{X_1}\left(\frac{\gamma}{\bar{\gamma}} y\right) = 1 - \frac{\Gamma\left(m_{sr}, \frac{\lambda_{sr}\gamma}{\bar{\gamma}} y\right)}{\Gamma(m_{sr})} \quad (B.2)$$

We will do the same analysis as in $F_{\gamma_1}(\gamma/Y)$ to find CDF of the SNR for the 2nd-hop conditioned on h_{sp} , $F_{\gamma_2}(\gamma/Y)$, then; the instantaneous SNR for the 2nd-hop, γ_2, does not depend on Y and using the notation that we supposed before, we can rewrite γ_2 as

$$\gamma_2 = \frac{\bar{\gamma}|h_{rd}|^2}{|h_{rp}|^2} = X_2 \quad (B.3)$$

Therefore;

$$F_{\gamma_2}(\gamma/Y) = F_{\gamma_2}(\gamma) = \int_0^{\infty}\int_0^{\frac{\gamma y}{\bar{\gamma}}} f_{|h_{rd}|^2}(x) f_{|h_{rp}|^2}(y) dy \quad (B.4)$$

After some manipulations and using [17, Eq.(6.455.1)], we can express $F_{\gamma_2}(\gamma/Y)$ as

$$F_{\gamma_2}(\gamma/Y) = 1 - \Upsilon_{rp,rd}(\gamma) \quad (B.5)$$

where $\Upsilon_{uv,ij}(\gamma)$ is a function of γ and it is given by:

$$\Upsilon_{uv,ij}(\gamma) = \frac{(\lambda_{uv}\bar{\gamma})^{m_{uv}}(\lambda_{ij}\gamma)^{m_{ij}}}{m_{uv}\left(\lambda_{uv}\bar{\gamma} + \lambda_{ij}\gamma\right)^{m_{uv}+m_{ij}}} \cdot \frac{1}{B(m_{uv}, m_{ij})} \cdot {}_2F_1\left(1, m_{uv} + m_{ij}; m_{uv}+1; \frac{\lambda_{uv}\bar{\gamma}}{\lambda_{uv}\bar{\gamma} + \lambda_{ij}\gamma}\right) \quad (B.6)$$

where $B(.,.)$ is the Beta function given by [17, (8.384.1)] and ${}_2F_1(.,.;.;.)$ is the Gauss hypergeometric function given by [17, (9.100)]. Then; $F_{\gamma_t}(\gamma/Y)$ can be expressed as

$$F_{\gamma_t}(\gamma/Y) = 1 - \Upsilon_{rp,rd}(\gamma)\frac{\Gamma\left(m_{sr}, \frac{\lambda_{sr}\gamma}{\bar{\gamma}} y\right)}{\Gamma(m_{sr})} \quad (B.7)$$

REFERENCES

[1] Huang, Senhua, Xin Liu, and Zhi Ding, "Opportunistic spectrum access in cognitive radio networks," *Proc. 27th IEEE Conference on Computer Communications, INFOCOM 2008*, Phoenix, AZ, April 13-18, 2008.

[2] Lee, Won-Yeol, and Ian F. Akyildiz, "Optimal spectrum sensing framework for cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, issue 10, pp. 3845-3857, 2008.

[3] Tan, Sheu-Sheu, James Zeidler, and Bhaskar Rao, "Opportunistic channel-aware spectrum access for cognitive radio networks with periodic sensing," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*, pp. 5333 – 5337, Vancouver, BC, May 26-31, 2013.

[4] Nagina Zarin, Imran Khan and Sadaqat Jan, "Relay Based Cooperative Spectrum Sensing in Cognitive Radio Networks over Nakagami Fading Channels," International Journal of Multidisciplinary Sciences and Engineering, vol. 3, No. 4, pp. 54-58, April 2012.

[5] Ganesan, Ghurumuruhan, and Ye Li, "Cooperative spectrum sensing in cognitive radio networks," *Proc. IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, DySPAN 2005*, pp. 137-143, Baltimore, MD, USA, Nov. 8-11, 2005.

[6] Ganesan, Ghurumuruhan, and Ye Li, "Agility improvement through cooperative diversity in cognitive radio," *Proc. IEEE Global Telecommunications Conference, GLOBECOM'05*, vol. 5, St. Louis, MO, Dec. 2005.

[7] Le, Long Bao, and Ekram Hossain, "Resource allocation for spectrum underlay in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7,issue 12, pp. 5306-5315, 2008.

[8] A. Conti, M. Z. Win, M. Chiani, and J. H. Winters, "Bit error outage for diversity reception in shadowing environment," *IEEE Commun. Lett*., vol. 7, issue 1, pp. 15-17, 2003.

[9] A. Conti, M. Z. Win, and M. Chiani, "On the inverse symbol-error probability for diversity reception," *IEEE Trans. Commun.*, vol. 51, issue 5, pp. 753-756, 2003.

[10] Bao, V., Duong, T., Da Costa, D., Alexandropoulos, G., and Nallanathan, A, "Cognitive amplify-and-forward relaying with best relay selection in non-identical Rayleigh fading," *IEEE communication letters*, vol. 17, issue 3, 2013.

[11] Luo, L., Zhang, P., Zhang, G., and Qin, J., "Outage performance for cognitive relay networks with underlay spectrum sharing," *IEEE Commun. Lett.* vol. 15, issue 7, pp. 710–712, 2011.

[12] Guo, Y., Kang, G., Zhang, N., Zhou, W., and Zhang, P., "Outage performance of relay-assisted cognitive-radio system under spectrum sharing constraints," *Electron. Lett.* vol. 46, issue 2, pp. 182–184, 2010.

[13] Duong, T. Q., Vo Nguyen Quoc Bao, and H-J. Zepernick, "Exact outage probability of cognitive AF relaying with underlay spectrum sharing," *Electronics letters.* vol. 47, issue 17, pp. 1001-1002, 2011.

[14] Duong, T. Q., et al, "Cooperative spectrum sharing networks with AF relay and selection diversity," *Electronics letters*, vol. 47, issue 20, pp. 1149-1151, 2011.

[15] Duong, Trung Q., et al. "Cognitive Amplify-and-Forward Relay Networks Over Nakagami-Fading." IEEE Transactions on Vehicular Technology, vol. 61, issue 5, pp. 2368-2374, 2012.

[16] A. U. Sheikh, M. Handforth, and M. Abdi, "Indoor mobile radio channel at 946 MHz: measurements and modeling," *Proc. IEEE Veh. Technol. Conf. VTC'93*, pp. 73-76, Secaucus, NJ, May 18-20, 1993.

[17] Jeffrey, Alan, and Daniel Zwillinger, eds, *Table of integrals, series, and products*, 7th ed. Academic Press, 2007.

[18] M. O. Hasna and M-S. Alouini, "Outage probability of multi-hop transmission over Nakagami fading channels," IEEE Commun. Lett., vol. 7, issue 5, pp. 216–218, 2003.

[19] Ikki S., Ahmed M.H., "Performance analysis of cooperative diversity wireless networks over Nakagami-m fading channel," *IEEE communication letters,* vol. 11, issue 4, pp. 334 – 336, 2007.

ENVIRONMENTAL PROTECTION

Simulation of Energy Efficient Telecommunication Network

G. Dimić, J. Seljan, and D. Šimunić
Faculty of Electrical Engineering and Computing, Department of Wireless Communications, Zagreb, Croatia
goran.dimic@fer.hr

Abstract - Energy efficiency in these days is as it should be, a part of every scientific research, especially in electronic communications in which it holds a notably significant role. When it comes to electronic communications, what matters is that the network is efficient not only for environmental reasons but for economical, as well. In telecommunication network, control of all the network aspects is crucial, because an inadequate control could easily end up with energy's over-usage. This paper presents a simulation of basic energy network components, which is a part of the project "Information and Communication Technology for Generic and Energy-efficient Communication Solutions with Application in e-/m- Health" (ICTGEN). The growing importance of ICT in everyday life such as the one of communication devices surrounding us, has lead us to bigger impact in using ICT equipment globally.

I. INTRODUCTION

Energy and its role in today every part of living is indisputable. This is the reason why energy efficiency is very important in all aspects of life. It is especially applied in research and development of new technologies with the main purpose being energy efficiency. Developing technologies influence global awareness of climatic changes. The ICT (Information and Communication Technology) is not exception. In the past, development of telecommunication networks has two main components: performance of system and profitability. As energy costs contribute significantly to the operational expenditures of telecommunication network operators, energy consumption has become a vital economic interest of network operators today. Thus, increasing performance and increasing profitability of system are inseparable with energy efficiency.

The traffic volume is predicted to grow at tremendous rates during the next few years. In addition, data centers are becoming increasingly important for the provisioning of telecommunication services, and there is a huge and increasing demand for using mobile services and applications. Thus, the very important question is what are the methods of optimization energy resources in telecommunication networks?

I. METHODOLOGY

Importance of ICT in daily living and number of communication devices that surround us is increasing. Consequently, this led to a significant increase of energy consumption of ICT equipment in total share of global energy consumption, which then led to important question of increasing energy efficiency of communication equipment and networks, mainly as increasing number of Internet users led to increasing number of smart devices based on wireless technology (such as smartphones, tablets, entertainment devices, medical devices, wireless sensors in smart houses and similar network equipment and devices). These days, in data centers energy consumption is one of main expenses; this is the reason for telecommunication companies to invest in research of energy efficient elements of network infrastructure. Until now, optimization of energy efficiency was focused to a dynamic control on hardware level (voltage controlling, clock speed and cooling methods). New methods are focused towards optimization of algorithm for controlling load of communication network. Here, energy saving is achieved in a way that we temporarily turn off network units which are not in use at a moment. In that way, it is possible to save energy in numbers that we cannot reach optimizing devices on hardware level. Furthermore, reduction of total energy from electricity power network are not only advantage of optimizing energy efficiency of end networked devices and sensor networks, but also increasing autonomy of battery powered devices. This has a positive influence on reliability, durability and long-term applicability for this kind of devices. In networks for environment parameter tracking that are used in smart home environment energy efficiency and optimization have a special focus.

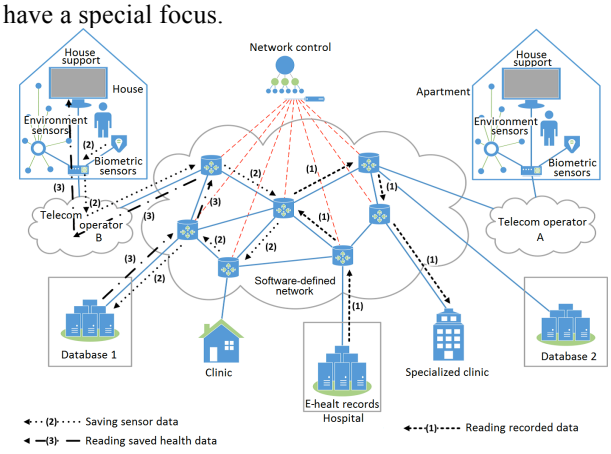


Figure 1 ICT network for application in e-/m- Health

Subject of active research in that field covers architecture development of Wireless Sensor Network (WSN) with very low power consumption specialized for environment parameter tracking in smart environment, which has to

allow maximal autonomy for network elements without constant power supply, also it has to allow high level of interoperability with other devices in communication network. Challenge for designing architecture and implementing algorithm for controlling energy efficiency is to synchronize multiple tasks, like a need for very low power consumption of network elements with need of a few years autonomy, interoperability within communication standards (like ZigBee, Wi-Fi and Bluetooth), implementation of complex algorithms for controlling energy efficiency of computer components. This is the reason why there is no standard for architecture of wireless area network for that purpose.
On second figure is shown a part of network and example of routing data. On figure 2 we can see that transfer of audio data and video data is achieved in two ways, because audio data must have the fastest route and video data must have sufficient bandwidth.

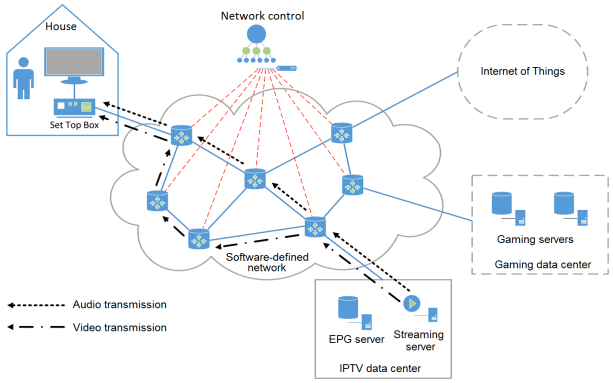


Figure 2 Routing of data

II. SIMULATION

The goal of ICTGEN project is to develop telecommunication network with end purpose in e-/m-health. Data which will be transferred in this network will be different in a purpose and in a requirements for network such as, medical records, video, home environment data collected via wireless sensor network etc. This information has a wide amount of data and can be very small from a few bites such as sensor data, to a few giga bites such as Magnetic Resonance Imaging (MRI). As requirements on network are big, energy efficiency can be successful by using elements of network where their usability is at maximum. As example, this work presents a simulation of wireless sensor network based on devices with ZigBee protocol (Fig. 3). With this, we ensure that network will have bigger autonomy of sensors and end devices with battery power supply. On the other hand, if video data or speech have to be exchanged with wireless network, WLAN devices based on 802.11 protocol will be used. These two networks are simulated and presented.

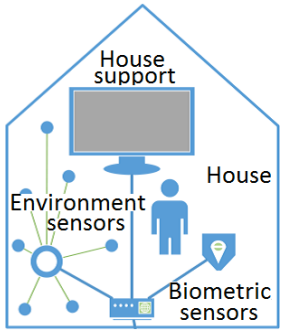


Figure 3 Wireless network in home environment

A. *WIRELESS SENSOR NETWORK BASED ON ZigBee DEVICES*

The OPNET ZigBee model uses four process models:

- ZigBee MAC model implements a model of the IEEE 802.15.4 MAC protocol.
- ZigBee Application model which represents a low fidelity version of the ZigBee Application Layer as specified in the ZigBee Specification. The process model initiates network joins and formations, generates and receives traffic and generates different simulation reports.
- ZigBee Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) model which implements the media access protocol of the MAC layer.
- ZigBee Network model which implements the ZigBee Network Layer as specified in the ZigBee specification. This model is responsible for routing traffic, process network join, formation requests and generating beacons.

Global MAC throughput is the total data traffic in bits/sec successfully received and forwarded to the higher layer by the 802.15.4 MAC in all the nodes of the WSN. Fig. 5 shows the global throughput for our ZigBee network. Network consists of one coordinator, four routers and eight end devices (Fig. 4)

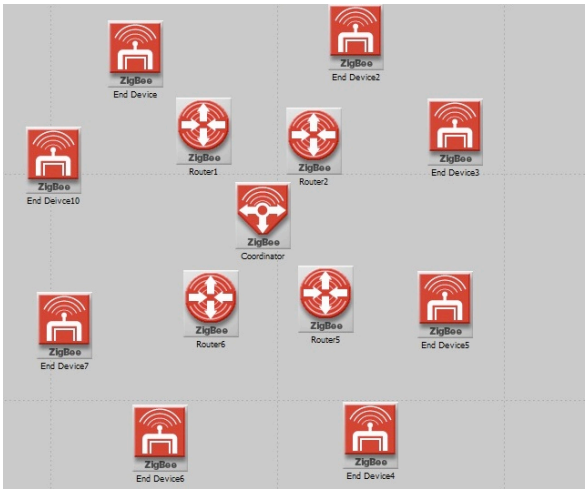


Figure 4 Topology of ZigBee network

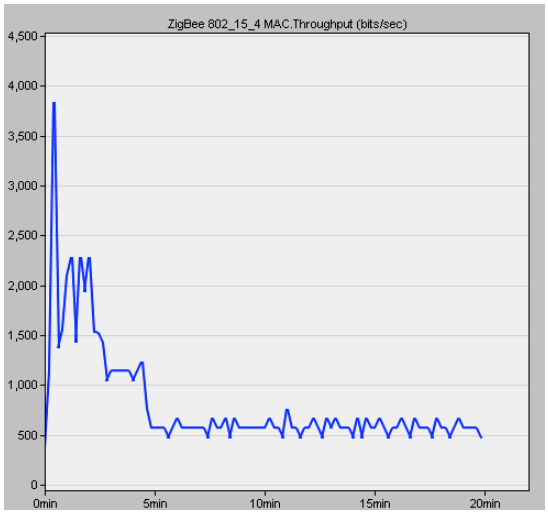


Figure 5 Global throughput for ZigBee network

B. *WLAN*

The field of wireless local area networks (WLANs) is being widely studied and used in various emerging research domains such as mobile and pervasive computing, where WLANs provide high-speed wireless connection and support accessing information from anywhere and anytime. WLANs support a wide range of applications, which may include simple applications such as web browsing etc. and the other ones, for instance, video streaming and video conferencing. IEEE 802.11 is a vital standard for wireless LAN, which adopts the standard 802 logical link control (LLC) protocol that is further divided into two sub layers: physical layer (PHY) and medium access control (MAC) layer. This configuration provides optimized functionality for wireless communication. The physical characteristics of 802.11a and 802.11g are identical – both are based on OFDM and support data rate of 54 Mbps. However they differ in operating frequency spectrum– 802.11a operates on 5 GHz band, while 802.11g on 2.4 GHz. 802.11bis based on DSSS and operates at 2.4 GHz band with transmission rate from 1 to 11 Mbps. 802.11a has significant advantage due to the wide range spectrum of 5 GHz, having more number of independent channels. Both 802.11b and 802.11g are compatible with each other as both operates on 2.4 GHz spectrum, but this may cause degradation in system performance as 2.4 GHz is a small band spectrum with a lesser number of independent channels.

In this simulation we are using 802.11g protocol for video streaming purpose.

On fig. 7 is shown received and sent traffic from one router (node_3) to another (node_2) and on fig. 8 is shown global throughput in WLAN.

Our network will be used in different ways, such are video, medical findings and data transmission, even the ones tracking surroundings of patients inside of their homes, or in the hospital. Energy efficiency comes from wise choosing of the right technology for the right use. In prior pictures (Fig. 5 and 8) we can see simulation results, such as the ones that come from differences in throughput when using ZigBee and WLAN. ZigBee is used for transmitting data from the sensor, and WLAN for video transmission. However, dual use of both together in some moments is unnecessary and uneconomical too, since while doing that we waste electricity and reduce the autonomy of our battery powered devices. Suggested solution is to shut down machines whenever it is necessary, depending upon their purpose. It can be done with software-defined networks and devices. By doing that a significant amount of energy might be saved.

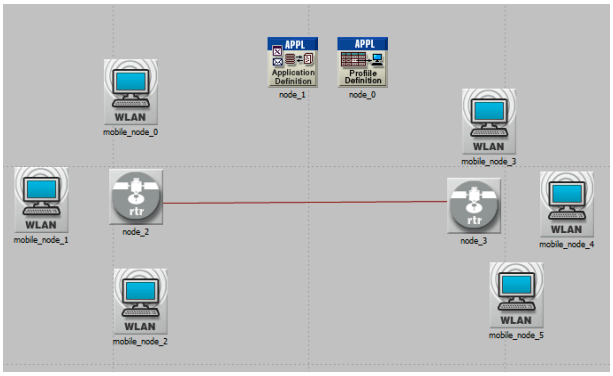


Figure 6 WLAN network

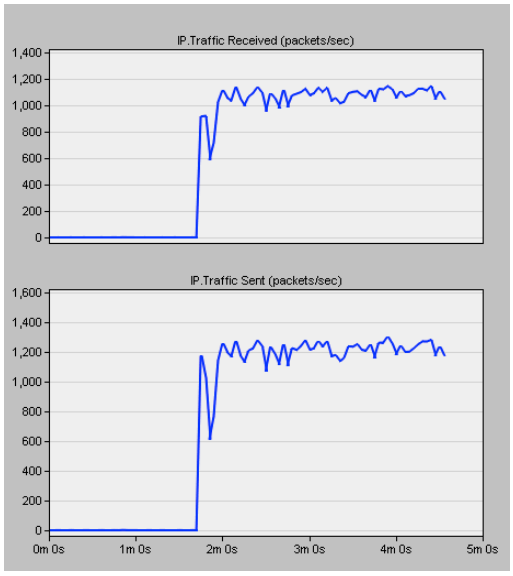


Figure 7 Traffic sent

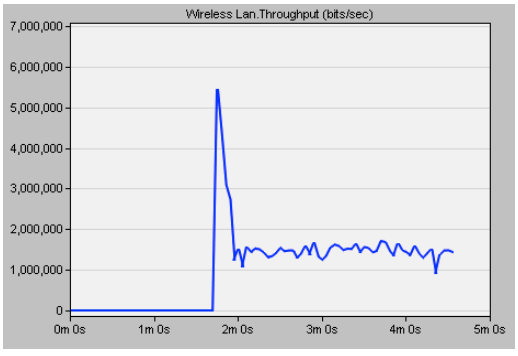


Figure 8. Global throughput of WLAN

III. FURTHER WORK

In order to properly estimate energy efficiency, the simulation of the full network is needed. Also, all protocols have to be performed in one simulation.
Also, for wireless part it would be very useful to implement some wireless devices which are battery powered, in which way we will have entirely system requirement on battery supply for that kind of devices.

IV. CONLUSION

In this paper the wireless part of telecommunication network with application in e-/m. health (optimized to be energy-efficient) is presented. Two parts of wireless network are presented and simulated, one based on ZigBee devices and other on WLAN. This network has multiple kinds of devices and protocols, because it has to transfer different kinds of information. Requirements for transferring information depend of information itself. MRI for example has few gigabytes so if we need to send or receive that kind of information wirelessly we will use WLAN. ZigBee devices, on the other hand, will be used for transferring information from sensors.

V. ACKNOWLEDGEMENT

The research leading to these results has received funding from the project "Information and communication technology for generic and energy-efficient communication solutions with application in e-/m-health (ICTGEN)" co-financed by the European Union from the European Regional Development Fund.

REFERENCES

[1] E. Husni, A. Narendra, “Energy Efficient Network Method Using Delay Tolerant Network”, IEEE International Conference in Green and Ubiquitous Technology, 2012, pp. 118-120.

[2] C. Lange, D. Kosiankowski, A. Betker, H. Simon, N. Bayer, D. von Hugo, H. Lehmann, and A. Gladisch, “Energy Efficiency of Load-Adaptively Operated Telecommunication Networks” IEEE Journal of lightwave technology, vol. 32, no. 4, February 15, 2014.

[3] C. Lange, D. Kosiankowski, R. Weidmann, and A. Gladisch “Energy Consumption of Telecommunication Networks and Related Improvement Options” IEEE Journal of selected topics in quantum electronics, vol. 17, no. 2, March/April 2011, pp. 285-295.

[4] I. S. Hammoodi, B. G. Stewart, A. Kocian, and S. G. McMeekin, “A Comprehensive Performance Study of OPNET Modeler For ZigBee Wireless Sensor Networks”, 2009 Third International Conference on Next Generation Mobile Applications, Services and Technologies IEEE, pp. 357-362.

[5] S. I. Jassim, S. W. Nourildean “IEEE 802.15.4 ZigBee-Based Wireless Sensor Network in Medical Application”, Iraqi Journal of Science, December 2012, Vol. 53, No. 4, pp. 1055-1066.

[6] W. Backman, K. Kordybach, and T. Sutinen, “Implementation of WLAN part of MERCoNe system in OPNET”, IEEE Telecommunications, 2008. ICT 2008. International Conference on 2008, pp. 1-6.

[7] M. F. Bari, R. Boutaba, R. Esteves, L. Z. Granville, M. Podlesny, M. G. Rabbani, Q. Zhang, and M. F. Zhani, “Data Center Network Virtualization: A Survey” IEEE Communications Surveys & Tutorials, vol. 15, no. 2, Second Quarter 2013, pp. 909-928.

Environmental Sustainability of Information and Communication Technologies for Advanced Metering and Home Area Networks

Vedad Mujan and Slavisa Aleksic
Vienna University of Technology/Institute of Telecommunications, Vienna, Austria
slavisa.aleksic@tuwien.ac.at

Abstract - The challenges imposed on the current electric power grid have led to the need of an improved and smarter electric power grid, i.e., the smart grid. The smart grid will enable not only a bidirectional transfer of electricity, but also a two-way data exchange between various systems essential for a correct operation of the advanced metering infrastructure (AMI) and home area network (HAN) applications. The realization of the smart grid will require the deployment of additional information and communication technology (ICT) equipment in various domains. This will unavoidably lead to an increase in electricity consumption. In order to obtain a meaningful environmental sustainability analysis, additional to the operation phase, various other ICT equipment life cycle stages, i.e., raw material extraction and processing, manufacturing and assembly, recycling and disposal, as well as transportation, have to be included in the assessment as well. In this paper, we present an exergy-based life cycle assessment (E-LCA) of ICT equipment for smart grids involved in AMI and HAN applications. Based on a model developed for the city of Vienna, it is shown that the cumulative embodied and operational exergy consumption distribution and their development significantly differ for the AMI and the AMI/HAN scenario.

I. INTRODUCTION

The current electric power grid does not fulfill the requirements of the 21st century [1]. An improved and smarter electric power grid, i.e., the smart grid, promises a reduction of greenhouse gas (GHG) emissions, introduction of huge amounts of renewable energy sources, and improvement of methods associated with current electric power production, delivery, and consumption [2]. Such benefits and features of the smart grid will heavily rely on a bidirectional information exchange between various systems, accompanied by a massive deployment of information and communication technologies (ICTs) in its different domains. The overall smart grid can be divided into seven domains according to the National Institute of Standards and Technology (NIST) smart grid conceptual reference model, namely into: 1. bulk power generation, 2. transmission, 3. distribution, 4. customer, 5. operations, 6. markets, and 7. service provider [2]. The customer and distribution domains of the smart grid are those that are expected to be equipped with the largest number of additional ICT equipment [3], such as smart meters, power line communication (PLC) modems, and data concentrators. This ICT equipment is required for various smart grid applications, making the customer and distribution domain more pronounced than the others. Additionally, user devices (UDs) like smartphones, tablets, notebooks, digital subscriber line (DSL) modems, and home energy management systems (HEMSs) are seen as an important part of the customer domain [2,4]. These devices will provide users with real-time information on their actual electricity consumption, and are essential for various management and control applications. Based on these considerations, we focus our research on the advanced metering infrastructure (AMI) and home area network (HAN) applications for smart grids. AMI includes the entire network infrastructure as well as the ICT equipment between customer smart meters and the meter data management system (MDMS) located in the data and control center (DCC) [4]. Some examples of AMI equipment are smart meters, PLC modems, data concentrators, base stations, switches, and routers. The HAN defines the network required to link various consumer appliances and systems with the HEMS, in order to achieve efficient management and control functions [4].

In this paper, we apply a holistic modeling method to assess the environmental sustainability of ICT equipment required for the AMI and HAN smart grid applications. While many studies analyze only the energy consumption of ICT equipment during its use (i.e., operational energy consumption), many other life cycle stages are not considered, such as raw material extraction and processing, manufacturing and assembly, recycling and disposal, as well as the transportation of various materials and components. To obtain a more meaningful environmental sustainability analysis of ICT equipment, the evaluation of these other life cycle stages has to be included in the assessment as well [5]. The *exergy-based life cycle assessment* (E-LCA) provides means for such an analysis, and takes the entire lifetime of ICT equipment into consideration, i.e., from cradle to grave. It represents a sound approach for the environmental sustainability analysis of ICT equipment, based on its life cycle exergy consumption. The obtained total exergy consumption of a component or a system serves as a measure for the attained environmental sustainability. The main benefit of an exergy analysis in comparison to an energy analysis is the fact that different forms of energy can be directly

compared with each other, which is related to its utilization of the second law of thermodynamics, not considered in an energy analysis. Moreover, the application of E-LCA for the environmental sustainability analysis of ICT equipment does not take as much time as a LCA, and leads to more meaningful and unambiguous results [6]. The E-LCA is therefore used here as the environmental sustainability indicator of choice for the assessment and evaluation of ICT equipment required for a proper operation of the AMI and HAN applications. The assessment based on the E-LCA framework confirms the assumption that the customer and distribution domains are indeed the most exergy consuming ones (considering the AMI and HAN smart grid applications), i.e., those domains are closely linked to environmental sustainability issues.

The paper is structured as follows. Section II provides basic insights into exergy and the E-LCA approach for ICTs. In Section III, the description of the overall model developed for the environmental sustainability analysis of the AMI and HAN applications is outlined. Moreover, basic assumptions are provided. Section IV presents and discusses the obtained results. Finally, a summary and conclusions are provided in Section V.

II. EXERGY-BASED LIFE CYCLE ASSESSMENT (E-LCA) OF ICT EQUIPMENT

E-LCA provides a very useful framework for the environmental sustainability assessment of ICTs. This has been recognized by many researchers, and there are already a few studies which utilize the exergy concept for the evaluation of the environmental sustainability of ICTs [5-9]. The quantity *exergy* itself relies on the first and second law of thermodynamics. It is defined as the maximum amount of useful work that can be attained from a system when brought into thermodynamic equilibrium with its reference environment [10,11]. Put more simply, it can be understood as the quantity of energy that can be converted into useful work, i.e., the portion of energy available to be consumed [12]. The major difference which distinguishes exergy from energy is the fact that it is never conserved due to irreversibilities (i.e., internal inefficiencies) attributed to most real processes [13]. The energy balance of a process states that the difference between input and output energy equals zero, i.e., $En_{in} - En_{out} = 0$ [13]. This means that the total amount of energy is conserved. According to the exergy balance, on the other hand, the difference between input and output exergy is greater than zero, i.e., $Ex_{in} - Ex_{out} > 0$, indicating a decrease in the quality of energy (i.e., exergy) of a process [11,13]. Based on that, more insights into the efficiency of diverse processes are provided. The exergy balance is, moreover, proportional to entropy generation ΔS, weighted by the temperature T_0 of the exergy reference environment:

$$Ex_{loss} = Ex_{in} - Ex_{out} = \Delta S T_0. \quad (1)$$

Relation (1) is also known as the law of exergy loss [13], and states that the difference between the exergy at the input and the output of a system corresponds to entropy creation, i.e., irreversibilities. The total exergy loss of a system is obtained by estimating and summing up the exergy losses of its subsystems or components, i.e., $Ex_{loss,system} = \sum Ex_{loss,component}$. The E-LCA considers the same life cycle phases just like a life cycle assessment (LCA) framework, but leads to a simple and unambiguous outcome (i.e., an exergy consumption value) that can be used for easy and more meaningful ICT equipment environmental sustainability assessment purposes. Moreover, it is not that time consuming and difficult to derive in contrast to a LCA [6]. Figure 1 depicts graphically the different exergy consumption life cycle stages of ICT equipment that are considered [5]. The exergy consumption of ICT equipment in each of these life cycle stages are estimated and analyzed using the E-LCA framework. As exergy cannot be conserved, it is assumed that the overall exergy losses accumulated during the ICT equipment lifetime correspond to their cumulative exergy consumption [13]. Therefore, the evaluation of various ICT equipment involved in AMI and HAN applications is based on their exergy consumption, which, on the other hand, serves as a measure for the attained environmental sustainability. Additionally, a distinction is made between *embodied exergy consumption* (EEC), i.e., the exergy consumed during raw material extraction and processing, manufacturing and assembly, recycling and disposal, as well as transportation, and *operational exergy consumption* (OEC), i.e., the exergy consumed during operation (including cooling if present). This will allow us to compare these two different exergy consumption categories with each other, and to indicate the most exergy consumption related one. Based on that, more information about category groups of ICT equipment with significant improvement potential is provided, i.e., the ICT equipment most strongly related to environmental sustainability issues.

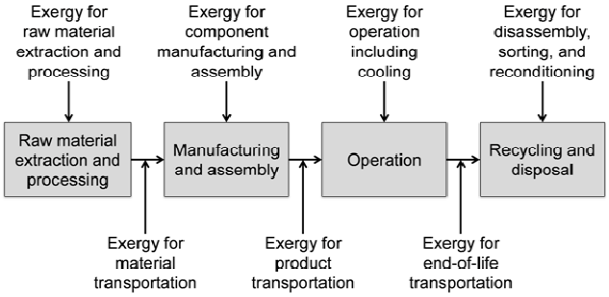


Figure 1. Life cycle exergy consumption stages of ICT equipment

III. DESCRIPTION OF THE MODEL

In this section, we present the overall smart grid model developed for the assessment of the AMI and HAN applications. Additionally, main assumptions required for the E-LCA of ICT equipment are provided. The considered overall smart grid model is composed of submodels for the home area network/building area network (HAN/BAN), neighborhood area network (NAN), radio access network (RAN), core network (CN), and the data and control center (DCC). A case study for the city of Vienna has been developed, with the aim to assess the EEC and OEC over a time period from 2020 to 2040 (i.e., operational duration of 20 years). Our assumption is that by the year 2020, the city of Vienna

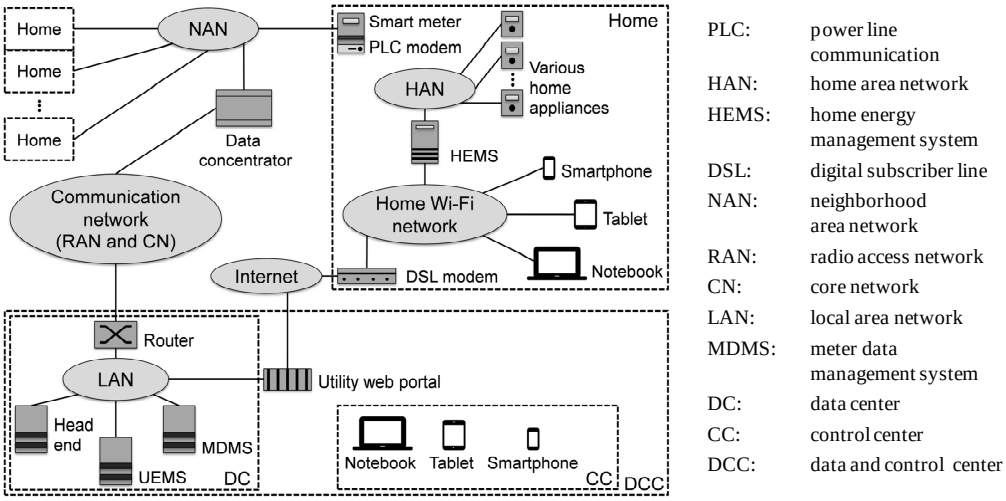


Figure 2. Overall smart grid model for the advanced metering infrastructure (AMI) and home area network (HAN) applications

will be equipped with appropriate smart metering, data processing and forwarding equipment (i.e., smart meters, PLC modems, and data concentrators), required for a correct operation of the AMI application [14]. The equipment needed for an appropriate functioning of the HAN application (e.g., smartphones, tablets, notebooks, DSL modems, and personal computer (PC) towers) is already in existence and widely utilized for various other purposes in almost every (if not every) household. These devices may as well be utilized for a multitude of different home energy management applications in connection with the smart grid concept. The PC tower, for example, may provide the functionality of a HEMS. For that purpose, adequate software programs would be required. Smartphones, tablets, and notebooks could be equipped with suitable mobile applications and software programs as well, enabling users to visualize their electricity consumption. This would, moreover, lead to an increased energy consumption awareness, and enable a greater involvement of consumers in the smart grid concept. The approach of dividing the overall smart grid model in different submodels will give us the opportunity to point out the most exergy consumption related ICT equipment, and with that the most dominant smart grid domains. Based on that, the domains that are closely related to environmental sustainability issues can be indicated.

A. *Overall Model*

The overall smart grid model is schematically depicted in Figure 2. The HAN/BAN equipment includes smart meters and PLC modems, as well as user devices (UDs) like smartphones, tablets, notebooks, DSL modems, and HEMSs required for a proper utilization of the HAN application. NAN equipment includes additionally to those of the HAN/BAN also data concentrators, required for data collection, processing, and forwarding purposes, and represents a very essential component of the smart grid. RAN equipment includes Global System for Mobile Communications (GSM) and Universal Mobile Telecommunications System (UMTS) equipment, i.e., base transceiver station (BTS), base station controller (BSC), Node B, and radio network controller (RNC) racks. The CN comprises serving General Packet Radio Service (GPRS) support node (SGSN) and gateway GPRS support node (GGSN) racks, as well as core switches. Copper and optical fiber cables, required to link RAN and CN equipment, are included in the model as well. Such a holistic approach allows us to obtain more meaningful conclusions on the environmental sustainability of ICTs associated with the AMI and HAN applications.

B. *Model for AMI and HANs*

The AMI represents the basic infrastructure of the future smart grid. It includes, e.g., smart meters, PLC modems, data concentrators, communication network equipment, and DCC equipment (see Figure 2). It is assumed that the data measured by smart meters are delivered to the data concentrator by means of the PLC technology. Moreover, the average number of smart meters (i.e., households) that can be supported by a data concentrator is assumed to be equal to 150. The forwarding of data for further evaluation and processing purposes from the data concentrator to the DCC is accomplished by means of cellular mobile communication systems, i.e., GSM and UMTS,. AMI is expected to bring a huge number of advantages with it, like increased reliability and energy efficiency, as well as a thorough insight into the condition of the entire smart grid. This will provide the staff at the DCC with advanced management and monitoring opportunities, and enable important remote control functions essential in the course of unusual or unexpected events [4].

The HAN considered here can be seen as an enhancement of the AMI. It extends the smart grid idea into the home, and enables important home energy management functions. It is used to link various consumer appliances with the HEMS by means of PLC or, e.g., the low-rate wireless personal area network (LR-WPAN) communication technology ZigBee. User devices such as smartphones, tablets, notebooks, DSL modems, and HEMSs are interconnected by means of the home Wi-Fi network and provide users with real-time electricity consumption information. This enables the users to see where and when their electricity consumption is at its peak, providing a way to cut down electricity bills. The

link between the HEMS and the utility energy management system (UEMS) is achieved by means of the Internet and the utility web portal [4], as depicted in Figure 2).

C. Assumptions and Scenarios

Information on the number of households in Vienna, their expected development, as well as the average number of persons per household is obtained from Statistics Austria. An operational duration of 20 years is assumed, i.e., a time period from 2020 to 2040, with the aim to gain more insight into the development of the lifetime exergy consumption of ICTs associated with the AMI and HAN smart grid applications. Based on that, the number of households in Vienna is expected to increase from 927,905 in 2020 to 1,027,846 in 2040 [15]. This corresponds to a yearly average household increase of 4,997.05 households. The average number of persons per household during this time period is assumed to be equal to 2. The total traveled distance of extracted and processed raw materials to their manufacturing and assembly location in Shenzhen, Guangdong in China is assumed to be equal to 5,000 km. From there, the final products are transported over Shanghai, China and Hamburg, Germany to the location where they will be deployed, namely to Vienna, Austria. The total traveled distance of these products was estimated to be equal to 22,403 km. For the end-of-life transportation, a recycling plant in Berlin, Germany is assumed. The total traveled distance of damaged or used up ICT equipment to this location was estimated to be equal to 675 km. The provided distances between the different ICT equipment life cycle stages (i.e., raw material extraction and processing, manufacturing and assembly, operation, recycling and disposal) were estimated by means of the Google Maps route planner. Moreover, various transportation modes (i.e., truck, rail, and ship) between these different locations are considered. Table I provides additionally some further ICT equipment assumptions required for the E-LCA case study, e.g., lifetime assumptions, EEC, and OEC of various devices. The OEC values listed in Table I were calculated based on the respective lifetime of the considered device. Material/component composition of smart meters, PLC modems, data concentrators, and HEMSs, required for the E-LCA, was in part based on analytical analyses, as well as information obtained from [16], [17], [18], and [5], respectively. Exergy consumption data of RAN and CN equipment (including copper and optical fiber cables), as well as that of routers and switches (i.e., rack/edge, aggregation, and core switches), was obtained from [7-9]. The lifetime of BTS and BSC racks is assumed to be equal to 7 and 8 years, respectively; that of Node B and RNC racks to 8 and 9 years, in that order. The lifetime of both the SGSN and GGSN racks is assumed to be equal to 10 years. The lifetime of routers, switches, and DSL modems is assumed to be 3 years. Data on technology and ICT equipment penetration was based on the information obtained from Statistics Austria and on the forecasts provided in [9].

It should be mentioned that the EEC of ICT equipment is, moreover, weighted by a usage factor (UF), accounting for the time it is used for the AMI and HAN applications. Smartphones, tablets, and notebooks will be used for other purposes most of the time (e.g., telephony, various internet services, mobile and video games). They will be used for home energy management applications for only about one hour daily (i.e., uptime of 4.17%). For that reason, their EEC is much smaller than that of smart meters, PLC modems, and data concentrators (see Table I). The same holds for the RAN and CN equipment, i.e., this equipment will be used for only a few minutes daily for the AMI application. The uptime of HEMSs and DSL modems, on the other hand, is assumed to be equal to 95% (i.e., 22.8 hours), however, with an average load of 50%. Based on these assumptions, the cumulative EEC and OEC of ICT equipment is investigated over the defined time period of 20 years for the AMI application alone (AMI scenario), and then in combination with the HAN application (AMI/HAN scenario). Moreover, the most exergy consuming ICT equipment in the case of these two scenarios is indicated. Based on that, the most dominant exergy consumption related (i.e., environmentally pronounced) smart grid domains are provided.

TABLE I. MAIN ASSUMPTIONS FOR THE E-LCA CASE STUDY

Device	***Lifetime [years]***	***EEC [MJ]***	***OEC [MJ]***
Smart meter	15	1,511.31	709.56
PLC modem	10	1,528.21	378.43
Data concentrator	7	2,849.66	529.80
HEMS	4	6,664.83	15,878.37
Smartphone	2	59.9	0.54
Tablet	2	62.13	1.17
Notebook	3	85.53	9.89

IV. RESULTS AND DISCUSSION

This section presents the results obtained from the E-LCA applied to ICTs involved in AMI and HAN applications. In the AMI scenario, only ICT equipment required for a correct operation of the AMI application is included in the assessment of the exergy consumption (i.e., environmental sustainability). The AMI/HAN scenario considers additionally UDs, i.e., smartphones, tablets, notebooks, DSL modems, and HEMSs, required for various home energy management applications. These devices are essential for a proper utilization of the HAN application.

Figure 3 depicts the distribution and the development of the cumulative EEC and OEC of ICTs for the AMI scenario over the defined period of 20 years. It is evident that the cumulative EEC dominates over the entire operational duration. The cumulative EEC and OEC values at the end of the operational duration of 20 years were estimated to be equal to 6.14 petajoules (PJ) and 2.127 PJ, respectively. This corresponds to 74.27% and 25.73% of the overall cumulative exergy consumption. Smart meters, PLC modems, and data concentrators, termed advanced metering, processing, and forwarding (AMPF) equipment in the scope of this paper, represent the ICT equipment category with the strongest contribution to the depicted cumulative EEC development. The large increase of the cumulative EEC between 2030

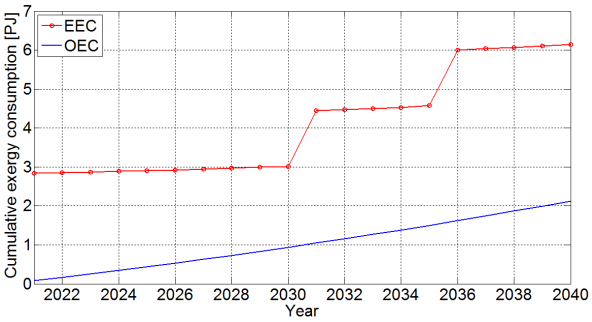


Figure 3. AMI scenario: distribution and development of the cumulative embodied exergy consumption (EEC) and the cumulative operational exergy consumption (OEC) of the overall smart grid model considering the AMI smart grid application

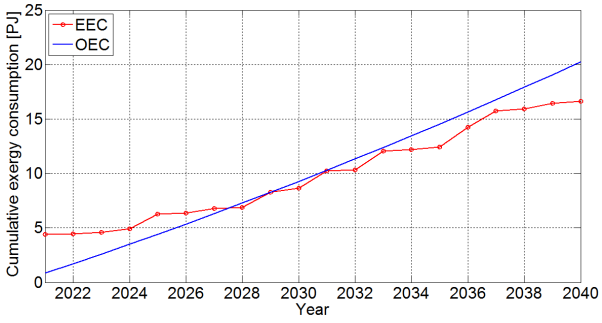


Figure 4. AMI/HAN scenario: distribution and development of the cumulative embodied exergy consumption (EEC) and the cumulative operational exergy consumption (OEC) of the overall smart grid model considering the AMI and HAN smart grid applications

TABLE II. AMI SCENARIO: CUMULATIVE EEC AND OEC DISTRIBUTION AMONG ICT EQUIPMENT CATEGORIES

ICT equipment category	***Cumulative EEC [%]***	***Cumulative OEC [%]***
AMPF	99.10	78.54
RAN	0.04	0.87
CN	0.01	0.41
DCC	0.86	20.19

and 2031, as well as between 2035 and 2036 arises due to the replacement of a huge number of PLC modems and smart meters, respectively, i.e., those that have been installed at customer locations in the year 2020. Table II provides additionally the breakdown of the cumulative EEC and OEC among the different ICT equipment categories of the overall smart grid model for the AMI scenario, obtained at the end of the operational duration of 20 years. As can be seen, the AMPF equipment contributes the most to the overall cumulative exergy consumption (i.e., EEC and OEC). Even though the number of ICT equipment deployed in the DCC is very much less than that of the AMPF equipment, their contribution to the cumulative OEC of about 20% is comparatively high. This is because the power consumption of the AMPF equipment is much lower than that of the ICT equipment utilized at the DCC. The share of the DCC equipment to the cumulative EEC is less than 1%, and therefore almost negligible. The share of the other ICT equipment categories (i.e., RAN and CN) to the cumulative EEC and OEC is very low compared to the AMPF and DCC equipment. Such a low contribution of RAN and CN equipment to the overall cumulative exergy consumption comes from the fact that it is utilized for the AMI application for only few minutes daily. These results clearly indicate the large importance of the customer and distribution domains of the smart grid (i.e., domains encompassing a large number of smart meters, PLC modems, and data concentrators), as well as the DCC because of their relatively high OEC. The cumulative EEC of ICT equipment was determined to be the largest contributor to the overall cumulative exergy consumption, during the entire operational duration, in the case of the AMI application. The AMPF equipment was ascertained to be the most dominant ICT equipment. The cumulative OEC is comparatively low, considering the large amount of AMPF equipment deployed in the customer and distribution domains, whereas the cumulative OEC of the DCC equipment contributes considerably to the overall cumulative OEC.

Figure 4 visualizes the distribution and the development of the cumulative EEC and OEC of ICTs for the AMI/HAN scenario. All assumptions made for the AMI scenario were left equal. However, the AMI/HAN scenario includes in the assessment of the cumulative exergy consumption besides the ICT equipment of the AMI scenario also UDs, namely smartphones, tablets, notebooks, DSL modems, and HEMSs. The quite different cumulative EEC and OEC distribution and development for the AMI/HAN scenario compared to the AMI scenario is clearly evident from Figure 4. It can be seen that the cumulative OEC becomes the most dominant exergy consumption category after approximately 7.5 years (i.e., between the year 2027 and 2028). The cumulative EEC and OEC for the AMI/HAN scenario were estimated to be equal to 16.628 PJ and 20.249 PJ at the end of the operational duration of 20 years, respectively. This corresponds to 45.09% and 54.91% of the overall cumulative exergy consumption. In Table III, the distribution of the cumulative EEC and OEC consumption among the different ICT equipment categories of the overall smart grid model for the AMI/HAN scenario, obtained at the end of the operational duration of 20 years, is provided. It is evident that the cumulative OEC becomes the major contributor to the overall exergy consumption, with the UDs being the most exergy consumption related ICT equipment category. The most dominant (i.e., exergy consumption related) component of the UDs was determined to be the HEMS. This is because the uptime of most other UDs (i.e., smartphones, tablets,

TABLE III. AMI/HAN SCENARIO: CUMULATIVE EEC AND OEC DISTRIBUTION AMONG ICT EQUIPMENT CATEGORIES

ICT equipment category	***Cumulative EEC [%]***	***Cumulative OEC [%]***
UDs	63.07	89.49
AMPF	36.59	8.25
RAN	0.01	0.09
CN	0.00	0.04
DCC	0.32	2.12

and notebooks) was assumed to be only about one hour daily. The uptime of the HEMS, however, was assumed to be equal to 95% (i.e., 22.8 hours). This fact makes the HEMS the most important (i.e., environmentally relevant) UD. Furthermore, it represents with 95.86% the strongest contributor to the cumulative OEC of all other UDs.

V. Summary and Conclusions

In this paper, we analyzed the environmental sustainability of ICT equipment involved in advanced metering (AMI) and home area network (HAN) smart grid applications. For that purpose, we applied an exergy-based life cycle assessment (E-LCA) framework. The assessment was based on a smart grid model developed for the city of Vienna. The cumulative embodied exergy consumption (EEC) and operational exergy consumption (OEC) of ICT equipment deployed in the various submodels of the overall smart grid model have been estimated and analyzed. The obtained exergy consumption of ICT equipment served as a measure for the attained environmental sustainability. We defined two scenarios, namely the AMI and the AMI/HAN scenario, and estimated the cumulative EEC and OEC for both of them. We proved that the customer and distribution domains are indeed the most exergy consumption related domains of the smart grid, and consequently, the most environmentally important domains considering the AMI and HAN smart grid applications.

The cumulative EEC was determined to be the most dominant exergy consumption related category over the entire operational duration of 20 years in the case of the AMI scenario. This result justifies and clarifies the importance of a holistic assessment approach, as it takes the entire ICT equipment life cycle exergy consumption into account and not just the exergy consumption during operation. The advanced metering, processing, and forwarding (AMPF) equipment, namely smart meters, PLC modems, and data concentrators was, moreover, ascertained to be the ICT equipment category with the strongest contribution to the overall cumulative exergy consumption (i.e., EEC and OEC). The cumulative OEC was determined to be the less exergy consumption related category over the defined operational duration in the case of the AMI scenario, and with that the category associated with lower environmental effects. The cumulative OEC of the DCC equipment was estimated to contribute by about 20% to the overall cumulative OEC, which is a comparatively high percentage considering the moderate number of ICT equipment deployed in the DCC. It can be concluded that in the case of the AMI scenario, the cumulative EEC of AMPF equipment strongly relates to environmental sustainability issues, i.e., it has the strongest potential to negatively impact the environment.

In the case of the AMI/HAN scenario, the cumulative OEC was determined to be the leading exergy consumption related category after about 7.5 years of operation. The strongest contribution to the overall exergy consumption is attributed to home energy management systems (HEMSs). Such a result arises due to a large number of deployed HEMSs and their respective high uptime of 95%, leading to a cumulative OEC share of 95.86% compared with all other UDs. The fact that HEMSs contribute the most to the overall cumulative OEC is, however, relaxed by the fact that their main purpose is the energy efficient management of consumers' electricity consumption. The energy consumption savings related to the use of HEMSs are believed to be larger than their own power consumption, leading to an indirect improvement of the environmental sustainability.

References

[1] V.C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, G.P. Hancke "Smart grid technologies: communication technologies and standards," IEEE Transactions on Industrial Informatics, vol. 7, no. 4, November 2011, pp. 529–539.

[2] National Institute of Standards and Technology, "NIST framework and roadmap for smart grid interoperability standards, release 2.0," NIST Special Publication 1108R2, U.S. Department of Commerce, February 2012.

[3] International Telecommunication Union, "Boosting energy efficiency through smart grids," 2012.

[4] K.C. Budka, J.G. Deshpande, M. Thottan, "Communication networks for smart grids," 2014.

[5] C. R. Hanneman, V.P. Carey, A.J. Shah, C. Patel "Lifetime exergy consumption as a sustainability metric for enterprise servers," Proceedings of ASME 2nd International Conference on Energy Sustainability, Jacksonville, Florida, USA, August 2008, pp. 1–8.

[6] D.J. Lettieri, C.R. Hannemann, V.P. Carey, A.J. Shah, "Lifetime exergy consumption as a sustainability metric for information technologies," Proceedings of the 2009 IEEE International Symposium on Sustainable Systems and Technology, vol. 00, IEEE Computer Society, 2009, pp. 1–6.

[7] S. Aleksic, M. Safaei, „Exergy based analysis of radio access networks" (invited), Proceedings of ICEAA - IEEE APWC – EMS 2013, 2013, pp. 1–8.

[8] S. Aleksic, M. Safaei, "Exergy consumption of cloud computing: a case study" (invited), NOC 2014, Milan, Italy, June 2014, pp. 1–6.

[9] M. Safaei, "Exergy-based life cycle assessment (E-LCA) of cloud computing," Master Thesis, Vienna University of Technology, Vienna, Austria, September 2013.

[10] G. Tsatsaronis, F. Cziesla, "Exergy and thermodynamic analysis," Exergy, Energy System Analysis and Optimization, vol. 1.

[11] M. A. Rosen, I. Dincer, M. Kanoglu, "Role of exergy in increasing efficiency and sustainability and reducing environmental impact," Energy policy, 36. Jg., Nr. 1, 2008, pp. 128–137.

[12] S. Aleksic, "Energy, entropy and exergy in communication networks," , Entropy, Vol. 15, No. 10, 2013, pp. 4484–4503.

[13] S. Herms, "Exergy flows in product life cycles – analyzing thermodynamic improvement potential of cardboard life cycles", MSc Thesis of Industrial Ecology, Delft University of Technology, Leiden University, August 2010.

[14] Wiener Netze Netzwervices, Smart Metering, available online: https://service.wienernetze.at/nc/ep/tab.do?pageTypeId=62006 (accessed on February 2, 2015).

[15] Statistics Austria, private households forecastes, Dezember 2014, available online: http://www.statistik.at/web_de/statistiken/bevoelkerung/demographische_prognosen/haushalts_und_familienprognosen/023529.html (accessed on February2, 2015).

[16] EDN Network, What's inside a smart meter? iFixit tears it down, August 2011, available online: http://www.edn.com/design/power-management/4368353/What-s-inside-a-smart-meter-iFixit-tears-it-down (accessed on February 2, 2015).

[17] Texas Instruments, Power Line Communication, Power Line Communication Modem Solutions from Texas Instruments, available online: http://www.ti.com/solution/power_line_communication_modem (accessed on February 2, 2015).

[18] Texas Instruments, Data Concentrators, Data Concentrator Solutions from Texas Instruments, available online: http://www.ti.com/solution/data_concentrator, (accessed on February 2, 2015).

Aspects of remote monitoring and recording system of non-ionizing electromagnetic radiation

D. S. Šuka*, M. I. Simić**, P. V. Pejović**
*Faculty of Electrical Engineering, University of East Sarajevo, BiH
**Faculty of Electrical Engineering, University of Belgrade, Serbia
darko.suka@etf.unssa.rs.ba
mira@etf.rs; peja@etf.rs

Abstract - The paper discusses the possibilities and aspects of continuous, remote monitoring system of non-ionizing electromagnetic radiation using certain networks for measuring and storage of measurement results. First are stated the motives for the usage and installation of such networks and preliminary activities necessary for the proper design of systems, as well. Further, the system architecture and methodology of such measurements are described, where defined requirements are placed in front of the remote monitoring station. Control of measurements is achieved from a remote distance via GSM, GPRS or other telecommunication services. Such systems allow full control of remote monitoring stations, as well as continuous recording of all measurement results. Measurement results from all monitoring stations are collected in control centre, where a comprehensive analysis of all measurement data can be done and published on certain website afterwards.

I. INTRODUCTION

Mobile telephony is nowadays a commonplace worldwide spread communication technology, which still increasingly spreads by new base stations continuous installations. Over 1.4 million of GSM/UMTS base stations exist in the world nowadays [1] and a new generation of 4G systems of these devices is already introduced and installed in some countries. Installation and operation of several wireless network bands in the last decade such as GSM, GPRS, UMTS, TETRA and Wi-Fi systems, increased public concern about exposure to radiation emitted from these sources, [1, 2]. Together with the traditional, well-established broadband systems (FM and UHF-VHF television, radar systems, etc.) this requires continuous care to inform the general public about levels of their exposure to non-ionizing radiation. More or less all human population is exposed to electromagnetic radiation. One of several reasons for the general public's fear is media announced and unconfirmed "scientific" studies, which cause feelings of uncertainty and perception of unknown and undiscovered possibility damage. Other factors are subtle care and a sense of control or expertise lack in the process of determining the location of new base stations or transmitters of any kind [1]. The so-called "ad hoc" measurements are the first step to estimate the level of electromagnetic field radiation. However, such measurements are carried out in a specific, fixed-time interval (1-2 hours) at certain date and under the certain circumstances, [2]. More appropriate and more acceptable way of estimating electromagnetic radiation levels to the general public represents 24 hours continuous measurements of electric field strength. This allows recording of the current situation on the field and continuous comparison of measured values with the reference levels of radiation prescribed by the relevant organizations, such as ICNIRP (International Commission for Non-Ionizing Radiation Protection) or the local authority organizations. Therefore, the network for continuous monitoring and recording of electromagnetic radiation levels in non-ionizing radiation band may represent the best solution to this problem. Another important reason for the installation of such networks is a possible way of measurement results presentation on certain "website" domain, which could be available to general public [2]. Monitoring networks are useful tool and help reducing the public's concern about potential hazard effects and impacts of non-ionizing radiation on human health. For this purpose, such networks across Europe and Africa (Italy, Greece, Spain, Portugal, UK, Malta and Egypt) are set and implemented, so far. These networks require overcoming of many challenges in terms of hardware certification, network control, daily maintenance, etc. [1, 2]. Therefore, there is a need for development of such networks in Balkan countries and that need is a result of widespread public fears about hazard effects and unknown consequences on human health caused by exposure to electromagnetic radiation. Such networks would have multiple targets and aims [1]:

- to inform the general public about current scientific researches;
- to inform the general public about radiation exposure levels and percentage of measured levels compared to them;
- to activate the procedure for additional measurements if they are higher than reference levels;
- to activate the procedure for reducing the reference levels, if needed;
- to reduce social tensions generated by the fear of radio communication networks installation.

Having all this on mind, we can freely conclude about necessity of development and implementation of such systems in Balkan countries in the nearer future.

II. PRELIMINARY ACTIVITIES

Before designing and installing such a complex network, it is necessary to carry out preliminary activities that will provide optimal selection of measurement points within the meaning of the most objective presentation of the current situation in the monitored area, from the electromagnetic radiation point of view. In order the quality selection of measuring points to be achievable, it is necessary to have all the information in the form of a database about: a) the schedule and location of all electromagnetic radiation sources at observed area with their full technical characteristics; b) terrain configuration with all associated buildings and structures, etc. Thanks to different modern radio-waves prediction propagation software, it is possible to predict the signal levels for each point in the observed area. More information is given in [3]. Many of these software are modular and open for adaptation to specific problems that can serve certain types of radio systems. The quality of all modern methods for radio-planning system largely depends on the quality of the information database available to the designer. Considering that, a great attention to the databases must be devoted. Therefore, it is necessary to provide or make several type of information for quality databases such as:

- A number of observed area scanned maps the (i.e. in the scale of 1:25,000, 1:50,000, 1:100,000, the territory of the major cities' wider surrounding in the scale of 1:150,000, 1:50,000 and 1:5000, etc.);
- A number of base station locations and technical characteristics of all radio transmitters;
- The types of used antenna systems;
- Digital base of terrain heights with 50×50m horizontal resolution;
- Digital database of vegetation on the earth's surface for the entire territory, if possible;
- Vector base of streets for major cities;
- Database of digitized satellite pictures.

Comprehensive analysis is carried out to eliminate a huge number of potential measurement points and in accordance to certain criteria (i.e. the intersection method) real shortlist for future measurement points is made. Then, these measurement points are verified by means of mobile measuring systems (spectrum analyzers). Afterwards, a remote-controlled measuring stations (RMS) in each verified measurement point is installed.

III. ARHITECTURE OF MONITORING SYSTEM AND METHODOLOGY OF WORK

The concept of the system should be based on the principle of ensuring the independence of performed measurements, so the general public can be sure that the results presented on a certain domain are not under the influence of various interest groups such as: Telecommunication operators, political parties, Government authorities, etc., [1, 4-7]. Therefore, the monitoring network should be funded with public money in total, without any sponsorship by the operator or other non-institutional subjects. Also, it would be decentralized so the local administration could have absolute control in their administrative areas. The network would be composed of several remote measurement stations (RMS), several local control units (CU) and one central unit, [1]. Each RMS must have communication with the local CU via public wireless communication networks such as GSM, GPRS or UMTS. This connection allows the installation of RMS in every area covered by the mobile operators' signals. CU control RMS units and save the measured data to be hosted on the "website", afterwards. Any interested party can access to stored results via Internet, [1, 2, 7]. Data from all local CU are collected in the central unit. Similar implementations are presented on Figure 1. [1] and Figure 2. [2].

A. Remote measuring stations

Measuring equipment or remote measurement stations (RMS), must be a completely autonomous units able to, [2, 7]:

- operate continuously, 24 hours a day;
- measure electromagnetic radiation in the frequency range that covers all or most of the sources that contribute to the overall radiation exposure;
- save the measurement results in internal memory and transmit them to control unit via wireless interface;
- be powered from the low voltage network of 220V or more often from the preferred solar panels;
- be completely remote-controlled, with no need for human presence;
- solve any kind of abnormal or unusual operations (such as self-diagnosis, overheating, loss of power, a supply failure, etc.) and report that to the control units (CU).

In order to meet all of these requirements in full, RMS should include broadband isotropic three-axis electric field sensor, the module with the necessary electronics, GSM/GPRS/UMTS modem which is used for communication with the relevant local control centre (the downlink is used for remote configuration, programming and polling, while the uplink is used for reporting the measurements and other information such as alarms) as well as temperature and humidity sensors, [2, 7]. All these should be calibrated prior to installation in the network and every 6-12 months thereafter. In order to obtain more relevant data, measurements should be performed within a wide and a narrow range of frequencies. Broadband RMS would record electric field strength and/or power density in the frequency range 100 kHz - 3GHz, in which the majority of the radiation sources would be included.

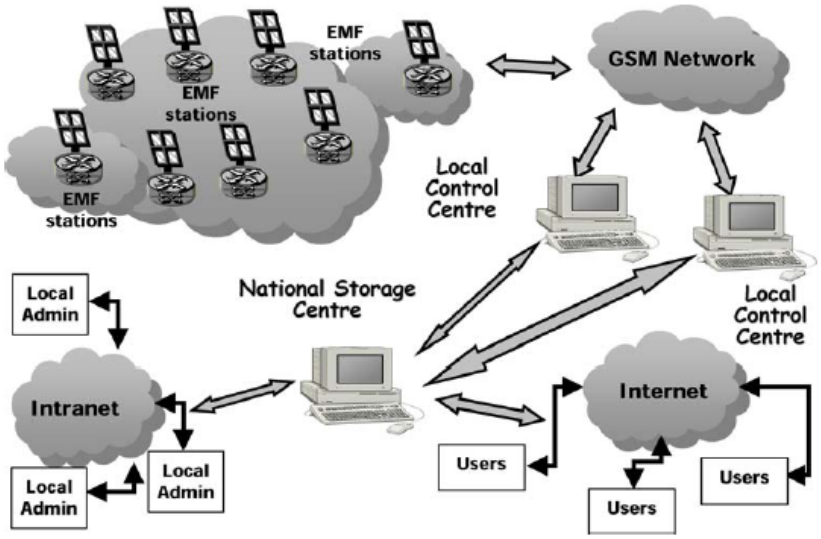


Figure 1. Architecture of monitoring system, National Storage Centre implemented

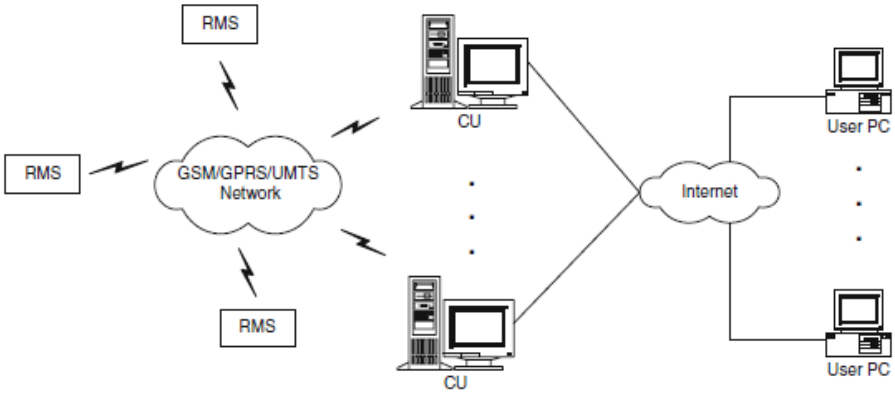


Figure 2. Architecture of monitoring system, local implementation

Figure 3. Examples of monitoring stations: EE4070-WB RMS station at Koskinou, Rhodes (left) and MCE-410 RMS station at Heraklion, Crete (right)

Other observed frequency ranges would be i.e. 880-960 MHz (GSM/UMTS 900 band), 1880-2100 MHz (DCS 1800 plus UMTS band), 2600 MHz (LTE - Long-term Evolution or 4G band) [8], etc. For frequencies above 100 kHz certain requirements must be met in order to avoid exceeding of electric field strength reference levels.

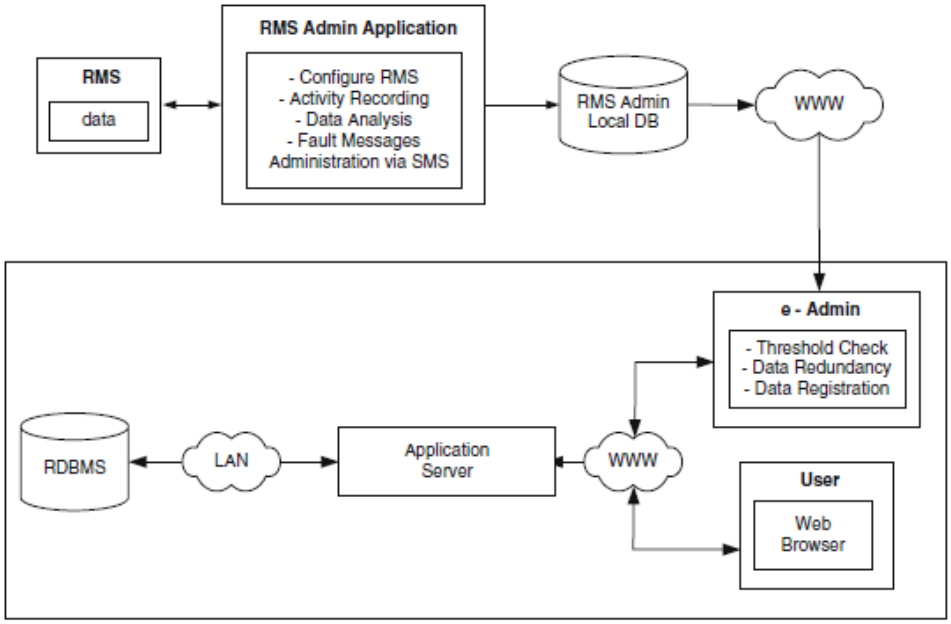


Figure 4. Software architecture

According to [9], the coefficient of total exposure to electric fields from multiple sources can be obtained by the following formula:

$$\sum_{i>1MHz}^{300GHz}\left(\frac{E_i}{E_{L,i}}\right)^2 \le 1 \qquad (1)$$

where E_i is measured intensity of electric field at frequency i, $E_{L,i}$ is the reference level of electric field given in Table I, according to ICNIRP standard. The relation (1) must be satisfied for all values measured by remote measuring stations.

TABLE I. ELECTRIC FIELD STRENGTH REFERENCE LEVELS FOR GENERAL POPULATION EXPOSURE, ICNIRP

Frequency f (MHz)	Electric Field (V/m)
0.15–1.0	87
1.0–10.0	$87/f^{1/2}$
10–400	28
400–2000	$1.375\,f^{1/2}$
2000–300000	61

Nowadays, there are wide range of measuring devices from different manufacturers, which could be used as a RMS [6, 7]. These devices are completely autonomous and can be remotely controlled via GSM/GPRS protocol, [2]. Examples of remote monitoring stations are presented on Figure 3. However, there are some restrictions on this type of equipment. The broadband measurement cannot determine the contribution of inndividual frequency bands or equivalent radiation sources. Additionally, measuring in only one point actually represents the situation of the measuring the whole band and that would require a very rigorous criteria for the selection of that point. Therefore, there are multiple criteria for the selection of the measurement sites (points). High population concentration is among the most important criteria. The special features of each location must be chosen to cover places of special interests, to include the presence of buildings of particular interest for the local community (e.g., schools, homes, hospitals and public places including zones like gardens, parks, shops, etc.) and to include the existence of antennas in the close vicinity of the measuring sites, as well [2, 3, 7]. One of proposals for measuring site selection criteria can include public opinions where the representatives of the population can decide where the measuring stations will be placed in their town or part of the city [1, 2]. To overcome these limitations, it would be necessary, in every area, to perform additional "ad hoc" measurements periodically, with different mobile equipment, every 6 to 12 months. These measurements would not be replaced by continuous measurement but would only amend them. The effect of that action would be twofold, [2]. First, by using a spectrum analyzer or a highly selective radiation meter a separation between the radiation sources would be achieved. Above all, a verification of continuous measurement accuracy would be possible by simple results comparing of these two methodologies. The time interval between these measurements can vary and it is limited by the need of human presence in the observed area.

B. Control units

The control unit would be a computer server equipped with a wireless modem and software applications which can interact with different types of monitoring stations. This software is necessary for, [2]:

- control and configuration of the RMS;

- collection of data from RMS via wireless network;
- saving and storing data on a hard drive;
- data processing and presentation of it via certain website, Figure 2.

The software architecture that could allow all of the above is shown in Figure 4. Usually, manufacturers of RMS provide software application for control units and data collection. With this application, a system administrator would have the ability to change the measurement points and measuring station communication parameters. The software application of CU would collect the measured value from measuring stations and save them in the local database (such as MS SQL Server or MySQL), [2]. In this way, a complete history of measurements for a certain administrative area would be stored in the local control unit database. In addition, there would be a web server (such as MS IIS or Apache) that could host the website in order to print the measured values and export it to text format, [1, 2]. For collection, consolidation and data presentation from all control units a web-based application should be developed and that application would be hosted on the central server. This application should collect data from all local databases, process them, filter and display them to the final user in an appropriate form. Examples of such applications are RDBMS which is is a Relational Data Base Main Server (built on MS SQL server) and "e - Admin" which is a web-based network administrative tool (developed in Java), [2].

C. *Central unit*

It would also be a computer equipped with a wireless modem which would be located in the national storage centre. It would have a software application for:

- collecting the processed data at the regional level by CU via wireless network
- storing data on the hard disk
- processing of data collected at the national level and presentation of it via the corresponding website, Figure 1.

D. *A website presentation*

The website could contain both static and dynamic pages and aim at presenting the measurement results for all the monitored sites. It could also provide information concerning the non-ionizing electromagnetic radiation exposure limits. Through the website, the visitors could view graphs of the measured values for custom date/time interval selection. Furthermore, a comparison of the measured values with reference levels could be presented. A feedback form would be also available to the public for them to express questions or comments to the system administrators.

IV. CONCLUSION

A proposed and described system for remote monitoring and recording of non-ionizing radiation is based on the usage of modern technological solutions and available systems. Automation of this system is achieved using computer-based units and their networks and remote communication is achieved using available mobile services. Web-based application allows permanent access to all measured data from any place. Automated measurement and monitoring of the whole system or network can be carried out remotely using GSM, GPRS or other telecommunication connections. The system provides the full ability to program or set and define the parameters for each monitoring station and to monitor and save all the measurement results in 3D form in an automated, remote-controlled and continuous way. The control centre collects measurement results from all monitoring stations and performs a complete analysis of all measured data and stores them for a long period of time. Such a system for monitoring and recording of non-ionizing radiation parameters would have a very useful social role. Direct communication with a large number of populations would provide full information to the general public, in terms of exposure to non-ionizing radiation and build a long-term confidence in such monitoring systems. Periodic analysis of monitoring system measurement results by additional measurement with different spectrum measuring equipment, could establish the trust even more. The network could easily be expanded in order to gain the full coverage of the observed territory. That way, the entire population, government, mobile operators and other interested parties would have a direct, continuous information about the levels of non-ionizing radiation, which would be one of the contributors preventing a possible harmful radiation on human health side-effects. At the end, the installation of the monitoring network would have, as final objective, the creation of a "national map" describing the real situation and highlighting the possible "over-radiated" territories.

LITERATURE

[1] F. Troisi, M. Boumis, P. Grazioso, „The Italian national electromagnetic field monitoring network" , Ann. Telecommun. 63:97-108, DOI 10.1007/s12243-007-0011-4, 2008.

[2] A. Gotsis, N. Papanikolaou, D. Komnakos, A. Yalofas, P. Constantiniou, „Non-ionizing electromagnetic radiation monitoring in Greece" , Ann. Telecommun. 63:109-123, DOI 10.1007/s12243-007-0006-1, 2008.

[3] Development Sector, "Radio Propagation Prediction Software, User's Manual Version 2.7", Remcom, USA, 2014.

[4] Narda Area Monitor System 2600, "Narda Safety Test Solutions",http://www.narda-sts.com/en/produkte/2600.htm

[5] http://www.pmm.it/main/safetyproducts.asp, PMM 8057 Electric and Magnetic Field Monitoring Station, "PMM Safety Products"

[6] http://www.eitsrl.it/iMCE.htm EE4070, EE4070S, MCE Monitoring System, "E.I.T s.r.l",

[7] Aaronia AG, "Manual for Spectran HF-4040 device" Aaronia AG, www.aaronia.de, Germany, 2005.

[8] W. Joseph, L. Verloock, F. Goeminne, G. Vermeeren and L. Martens, "In situ LTE Exposure of the General Public: Characterization and Extrapolation", DOI 10.1002/bem.21707 Wiley Online Library, Bioelectromagnetics 9999:1-10 (2012)

[9] Rulebook 37/2008 on the limitation of EM radiation emissions, Communications Regulatory Agency of Bosnia and Herzegovina.

Energetic Self-regulation of eWALL System

Ž. Martinović, D. Budić, D. Šimunić
Faculty of Electrical Engineering and Computing / Department of Wireless Communications
Unska 3, 10000 Zagreb, Croatia
E-mail: zeljko.martinovic@fer.hr, dora.budic@fer.hr, dina.simunic@fer.hr

In the last few years, there has been intensive research in the field of human motion and environment energy harvesting. That kind of energy harvesting used for power supply self-regulation in eWALL system (e WALL for Active Long Living) would contribute to increased energy efficiency of the system. The paper will represent possible methods and ways of energy harvesting in eWALL system. Statistical evaluation of the harvested energy will be given. The paper will also consider other, possible undeveloped ways of energy harvesting in terms of the practical implementation of the system. These methods and statistical estimates may relate to eWALL system, but also to other systems such as e.g. smart homes. Energy accumulation can be achieved from user motion, differences in ambient temperature, vibrations, radio frequency radiation and other technologies that are available to eWALL users.

I. INTRODUCTION

eWALL (*e Wall for Active Long Living*) is a system of intelligent services that will be possible to set up in any corner of our home or health institution, which aims to facilitate and extend active life expectancy of people with special needs. eWALL system is shown in Fig.1. Independent living is becoming a major challenge in terms of the population of the third age, which directly affects:

1. Family and friends;
2. An easier life with diseases;
3. Care providers;
4. EU health system;
5. Insurance companies

eWALL prototype is a result of partnership work on EU FP7 project after first year of research and development. The prototype is a result of integration and optimization work package led by the Faculty of Electrical Engineering and Computing, University of Zagreb. In Fig. 2 is shown the eWALL prototype integration environment

The prototype consist of a central *Gigabyte Brix* unit with the Ubuntu operation system and ELO "42" touch screen. Sensor network, which aims to harvest and send information about user in defined *Couch* database, consists of:

1. Environmental sensors (gases, light, humidity, temperature)
2. Portable sensors, based on *Arduino Lilypad* development tools, with built-in accelerometer.
3. Pressure sensor designed for the control of the user in bed.
4. Sensors for visual monitoring of the user's face based on *Kinect* solutions.
5. Sensors for motion control

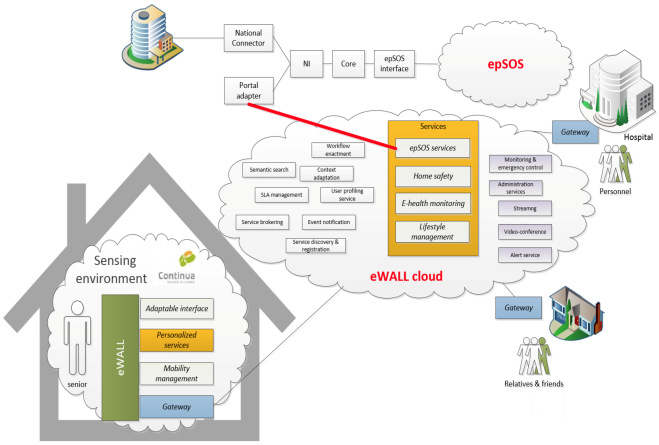


Figure 1. eWALL System

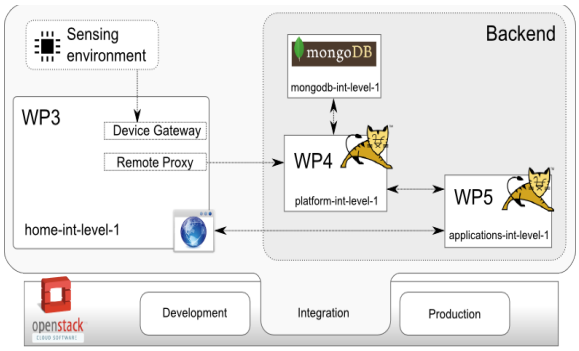


Figure 2. M12 Integration eWALL prototype environment

Besides the mentioned sensor network, prototype also consists of platform, application solutions for monitoring users and their activities. In the next period, the development of advanced versions of prototype will lead to an increase and diversity of sensors in order to give better user supervision and ensure timely response to negative scenarios. In accordance with development of prototype, one of the issues is system

power and energy self-regulation. Increase in energy efficiency of the system would greatly facilitate the use of the system, especially as seen from the user side. Regarding to this, we give an overview of possible user self-regulatory methods and statistical evaluation of possible energy harvesting methods.

II. OVERVIEW OF METHODS FOR ENERGY HARVESTING

Energy harvesting from user environment or user movement has, in recent years, become one of the most promising scientific research areas. The main goal of energy harvesting is to provide power to devices that consume small amounts of energy (mW). Energy harvesting models applicable to eWALL can be divided by energy source, represented by table 1. and Fig. 3.

TABLE 1. ENERGY SOURCES IN eWALL ENVIRONMENT

1. Kinetic energy from user motion;
2. Temperature difference between user and environment;
3. Environment vibrations (sound);
4. Radio frequency energy;
5. External source (sun, rain, snow, wind);

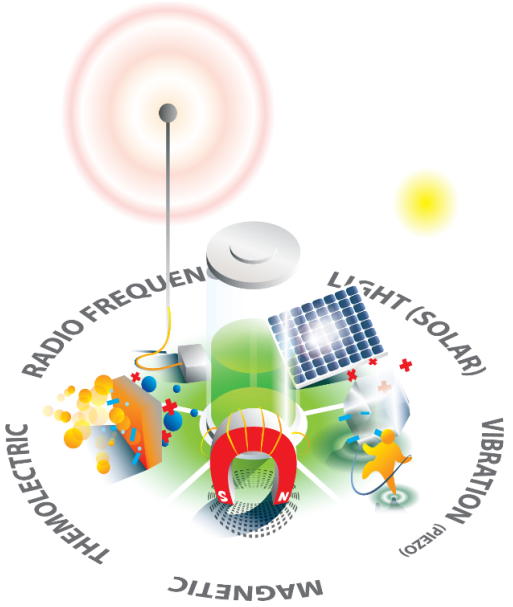


Figure 3. Energy sources in environment

There are two basic groups of sensitive materials that are used for the design and construction of micro mechanic sensor films (MEMS). These films are made of ZnO (*zinc - oxide*) and PZT (*lead zirconium titanate*) materials. The sensitivity to detect changes in the charge of ZnO films is much lower than in PZT. Because of that, ZnO is more commonly used then PZT materials for MEMS film. The sensitivity of such accelerometers was initially researched from year 2003 [6] when the defined sensitivity was from 0.1 to 7.6 pC/g.

Harvesting of kinetic energy from the movement of users is one of the initial motivations for scientists in this field. Statistic evaluation of energy harvesting from hand movement [4] provides 1.21 μW at one motion. The accelerometer is usually attached to the wrist. All movements of the arms, which have the force acting on the mass below 0.015 g are ignored. In further research, energy is harvested by placing accelerometers on the ankle, elbow and hip. The values obtained are in the range μW. Besides the systems with one degree of freedom, systems [5] that were taking into account the orientation of the force were also used. Results have not shown significantly increased amount of harvested energy.

The harvested energy, as we have previously mentioned, may be substituted for battery power in portable devices. One of these systems is the one [7], which provides charging a mobile device via walking. The problems are setting the device properly to harvest energy in running shoes or proper placement of sensitive polymer materials. For energy storage we used lithium batteries. During one hour, 4 steps per second generated 0.05 J energy, that is power about 13 μW. Therefore, to ensure the mobile phone power supply of 2 W, it is necessary to make millions of steps. The resulting power is controlled with DC-DC circuit whose efficiency is about 65 %.

Studies [11] have shown that it is possible to harvest 20 μW of power on the basis of difference in body temperature and the room temperature at 22 degrees. The required size of the film in this case is 9 cm^2. It was also found that film with 0.5 cm^2 can during run at 11.3 km/h harvest 3.7 μW. That provides 2.2 μW per cm^2 or 7.4 μW per cm^3 of film. The energy generated by thermal difference is harvested by the thermoelectric generator, which is set to a specific part of the body.

eWALL prototype consists of few hard disks that are used for processing and storage of data. It is known that in starting phase, hard disk consumes an average of 1.9 W, in idle phase 6.2 W and in operating phase 8 W of power. Energy can be harvested from the static stress and pressure during the operating phase of a disc. For the purpose of energy harvesting, energy harvester is designed [8]. Designed harvester consists of thin pressure and vibration sensitive film that is added to the flex cable of the hard disk. Research has shown that from the rotation of the hard disk, about 0.55 W of power can be harvested, 58% of it, or 0.32 W comes from hard disk vibrations. The remaining amount is harvested from the kinetic motion of the central engine torque.

To harvest energy, vibrations can combine with solar panels [3]. Solar panels have been set up along the road. Energy gain was based on the use of a double solar technology. The use of piezoelectric system has been introduced which uses car vibrations to power each solar collector. The vibrations were giving a sine wave with amplitude of 10 V. With voltage amplifier, voltage would go up to 20 V. That voltage would then

be converted by bridge to DC voltage from 17 V that can charge the battery.

One of the sources of energy is the radio frequency radiation. Initial studies were based on RFID technology. With designed RF - DC converter at distances of less than 10 m with the signals less than 50mV, the conversion efficiency reached 60%. Maximum theoretical standardized RF power harvested [17] on the frequency 900 MHz is 7 mW, and at 2.4 GHz is 1 mW for a distance of 40 m without obstacles. Research [18] shows that depending on input power of RF radiation (70 W/cm2 – 30 W/cm2) it is possible to harvest between 420 µW and 8 µW.

III. EWALL HUBS FOR ENERGY HARVESTING

Figure 4. eWALL energy hubs

Energy harvesting hubs represented with Fig. 4 and Table 2. would be set in eWALL environment to supply energy to low consumption sensors that have limited time use.

TABLE 2. EWALL ENERGY HARVESTING HUBS

1. Slippers (walking around the house)
2. Sneakers/Shoes (walking out of the house, running)
3. Termogenerator on the apron (kitchen)
4. Kitchen floor
5. Kitchen scale
6. Window hubs
7. Exercise bike
8. eWALL system hard disk
9. Pressure sensors in the bed
10. Modem
11. Washing machine
12. Dishwasher
13. Electric/wall plug

Those hubs can be expended by any element in the users' home that has power source, which can be utilized. That energy can power sensors with low power consumption such as: pressure sensors (bed), sensors to detect a gas leak, motion sensors, sensors for measuring temperature, sensors for measuring medical data (oxygen saturation, blood pressure, etc.). Energy harvesting is, according to many studies, particularly suitable for e-health sensors that have low energy consumption and are used in limited period of time. eWALL system is based precisely on the e-health sensors.

IV. ESTIMATION OF HARVESTED ENERGY

One of the most common and most easily applicable methods for energy harvesting is the installation of a piezo electric film in slippers or sneakers/shoes. According to the available measurements, 60 steps charge capacitor to the level of 21 V. According to equation (1) 92.5 µJ per one step is harvested.

$$E = \frac{1}{2} x V^2 x C \quad (1)$$

Users of the eWALL system will be stimulated with series of applications and reminders to physical activity. If we assume that the user will make between 5000 m and 10000 steps a day, we get the possible daily amount of harvested energy in the range between 7.7 mW and 0.016 W. On a monthly basis we are talking about the range between 0.23 W to 0.46 W.

If the user runs or workouts, the number of steps per second is on average 3 or 4. Assuming that the user runs in the countryside or on a bicycle for exercise 2 km a day, then we are talking about harvesting 0.65 µW daily, on a monthly basis, about 13 µW. One of the central devices in eWALL system will be *Xbox Kinect* device. It will be used, among other things, for exercise through interactive content. According to above-mentioned study, from one movement 1 µW can be harvested.

If user, every week, makes 100 movements in front of Kinetic device, on a monthly basis we are talking about the harvested 0.4 mW. eWALL should enable users to monitor their sleep quality. It can be done by monitoring users' motion in bed. That same motion can provide energy harvesting.

If we, according to Fig. 4., place 5 pressure sensors with DC-DC circuits in bed stand, we can use force to achieve power self-regulation in bed. For work purposes, bed sensor is needed, in the worst case, to ensure 1.25 mW of power. Assuming that the voltage generated by the force stored in the capacitor is 0.3 V per pressure, it is possible to provide a power of 0.75 mW and save 60% energy.

One of the more common sources for energy harvesting is vibration. Vibrations in the user's home can be caused by the washing machine, clothes dryer, hard drive, wind hitting the windows of the observed object, etc. Within the eWALL system, at least one hard disk will be used for data processing, and also one more for user data storage. Using energy harvesters, we can save up to 9 % of the energy in the operating/quiet mode.

Regarding to that, there is an extremely large number of devices for energy harvesting from vibrations, which are based on an electromagnetic, electrostatic, piezoelectric, and magnetostrictive mode. The electrostatic capacitive energy harvester [2] is a mechanical micro inverter with high sensitivity to vibration. The design of the inverter is reduced to designing four-layered capacity. Testing confirmed that the vibrations of 67 Hz, with large enough load resistance, could generate 175 W of power. For comparison, an asphalt leveling roller has the vibration of 67 Hz. Source of vibration was a vibrator. A device used in the assessment of vibration can simulate various types of intensity and vibrational forces. Power estimation is given by (2).

$$P = \frac{Q^2}{2\,x\,Cmin}\,x\,f \qquad (2)$$

Expression (2) shows us the way of calculating the resulting power. Q is the charge on the sensitive film, Cmin is minimum defined capacity of the unit, and f is vibration frequency. If we define the vibration as displacement of the object relative to the equilibrium axis, it is possible [12] to harvest 30 μW per millimeter displacement. Electrostatic energy harvested from the vibrations that is based on the setting up of static marble lines along the sensitive film, can harvest, from the displacement of 1.2 mm, around 0.18μW [13]. From the vibrations in limited periods of time, few hundred μW can be harvested.

One of the indirect ways of self-regulation is the optimization of sensor measurement methods, for example, energy regulation of CO2 [1] sensor by optimizing sampling. Battery consumption is proportional to sample size in any space. Reducing power consumption by reducing the size of samples, and at the same time preserving the quality of output sensor data is the way they can be self-regulated. In Table 3 summary of energy harvesting estimation is given.

TABLE 3. ESTIMATION OF HARVESTED ENERGY

Method	Energy harvested
Walking (per step)	92.5 μJ
Running (per km)	0,32 μW
Hand move (per movement)	1 μW
Body move (per pressure)	0,75mW
Vibration (per 1 mm displacement)	30 μW
RF (per bream)	2 - 4 μW

Possibility of energy harvesting by small samples exists also in radio frequency (RF) radiation. Previous studies have been conducted on different RF technologies (*WI-FI, Bluetooth, RFID*, etc.). Key characteristics in terms of RF energy harvesting are receiving antenna sensitivity, RF-DC converter and the generated output voltage. The utilization amount of reflected RF beams mainly depends on conversion factor of RF-DC converter. Research has shown [19] that it is possible to achieve an output voltage of 1.5V at the strength of the reflected beam from 1mW (-10dBm). The generated power would be the equivalent of 2-4 μW. The relatively low conversion factor of RF-DC converter, researchers are trying to annul by different methods. The last of them is designing RC-DC converter using metamaterials. For conversion, array of 7 antennas has been used and resulting efficiency was close to solar panel efficiency, which is 37%. If the mentioned converter becomes scientifically established (which would harvest 0.37mW on 1mW signal strength), it would be possible to power devices that consume significantly more energy.

V. CONCLUSION

The paper describes some of the existing methods of energy harvesting and possibility to apply them for the purpose of eWALL system power self-regulation. Statistics for possible energy harvesting has been shown. In developing energy harvesters, biggest shift is made with piezo electric harvesting methods. The research focus is on RF harvesting methods but also on external energy sources. Latest harvesting methods such as rainfall energy over the roof of the house built by a sensitive film shows us that awareness of recycling energy is growing unstoppably. Mentioned methods could harvest energy and directly power sensors or could be stored in batteries. This work will serve to further research and designing energy harvesters for the purposes of eWALL system and also as an overview of existing methods on a practical system such as eWALL.

LITERATURE

[1] J. Yan, L. Pei, M. Wang, C. Liu and S. Gao, "Local Adaptive Sampling for an Energy", Advanced Infocom Technology 2011 (ICAIT 2011), International Conference, pp. 1-4.

[2] I. V. Knyazev, V. Y. Dorzhiev, A. V. Gluhov and A. A. Panova, "Development and Manufacture of Electrostatic Capacitive Energy Harvesting Device", Micro/Nanotechnologies and Electron Devices (EDM), 2014 15th International Conference of Young Specialists, pp. 51-53.

[3] R. R. Chowdhury and M. S. Kabir, "Electrification of Streets of Dhaka City using Solar and Piezoelelectric Energy", Informatics, Electronics & Vision (ICIEV), 2014 International Conference, pp. 1-5.

[4] N. Yarkony, K. S. Pour and A. Possalo, "Statistical Modeling of Harvestable Kinetic Energy for Wearable Medical Sensors, World of Wireless Mobile and Multimedia Net-works" (WoWMoM), 2010 IEEE International Symposium, pp. 1-5.

[5] G. V. Merrett, H. Huang and N. M. White, "Modeling the Effect of Orientation on Human-Powered Inertial Energy Harvesters", IEEE Seniors Journal, 2014, pp. 434 – 441.

[6] X. Li, M. Guo and S. Dong, "A Flex-Compressive-Mode Piezoelectric Transducer for Mechanical Vibration/Strain Energy Harvesting, Ultrasonic Ferroelectrics and Frequency Control", IEEE Transactions 2011, pp. 698-703.

[7] M. Jain, U. Tiwari and M. Gupta, “Mobile charger Via Walk, Multimedia Signal Processing and Communication Technologies” (IMPACT), 2011 International Conference, pp. 149 – 152.

[8] J. Y. Chang, “Modeling and Analysis of Piezo Elastica Energy Harvester in computer hard disk drives”, Magnetics, IEEE Transactions 2011, pp. 1862 - 1867.

[9] H.T. Luong, C.M.T. Tien and N.S. Goo, “Performance of a Piezo-Composite Generating Element, Ubiquitous Robots and Ambient Intelligence” (URAI), 2011 8th International Conference, pp. 203-204.

[10] R. Vaddi and S. Dasgupta, “Enhanced Bias flip Restifier with Ultra low power for Piezo Electric Energy Harvester in the Microwatt Application Scenario”, Microelectronics and Electronics (PrimeAsia), 2012 Asia Pacific Conference, pp. 132-137.

[11] M. Wahbah, M. Alhawari, B. Mohamed, H. Saleh and M. Ismail, “Characterization of Human Body – based thermal and vibration energy harvesting for wearable devices”, Emerging and Selected Topics in Circuits and Systems, IEEE Journal 2014, pp. 354-363.

[12] S. Naifar, S. Bradai, T. Keutel, O. Kanoun, “Design of a Vibration Energy Harvester by Twin Lateral Magnetoelectric Transducers”, Instrumentation and Measurement Technology Conference Proceedings, 2014 IEEE International, pp. 12 -15

[13] Z. Yang, E. Halvorsen, and T. Dong, “Electrostatic Energy Harvester Employing Conductive Doplet and Thin Film Electret”, Microelectro-mechanical Systems, Journal 2013, pp. 315-323

[14] L. Deng, Z. Wen, X. Zhao, C. G. Luo and J. Mo, “High voltage Output MEMS Vibration Energy Harvester in d31 Mode With PZT Thin Film”, Microelectromechanical Systems, Journal 2014, pp. 855-861.

[15] Y. Li, M. Misra and S. Gregori, “Model and Prototype of a Green Electrostatic Harvester of Vibration Energy, Circuits and Systems (MWSCAS)”, 2014 IEEE 57th International Midwest Symposium, pp. 53-56.

[16] T. Le, K. Mayaram and T. Friez, Efficient Far-Field Radio Frequency Energy Harvesting for Passively Powered Sensors

[17] H. Jabbar, S. Song, and T. T. Jeong, “RF Energy Harvesting System and Circuits for Charging of Mobile Devices”, Consumer Electronics, IEEE Transactions 2010, pp. 247-253.

[18] T. Paing, J.Shin, R.Zane, and Z. Popovic, Resistor Emulation Approuch to Low-Power RF Energy Harvesting, Power Electronics Specialists Conference 2006, pp. 1-7.

[19] E. A. Kadir, A. P. Hu, and M. B. Abhari, “Indoor WIFI energy harvester with multiple antenna for lower power wireless applications”, Industrial Electronics (ISIE), 2014 IEEE 23rd International Symposium, pp. 526-530.

Investigation of electromagnetic compatibility of high frequency polymer-welding machines

Lyuben Iliev, Ivaylo Stoyanov, Teodor Iliev, Elena Ivanova
University of Ruse, Department of Electrical and Electronic Engineering, Ruse, Bulgaria
lailiev@uni-ruse.bg

Abstract - The article presents the results of a measurement of the electromagnetic disturbances of a High Frequency (HF) plastic welding machine and their analysis. These types of machines are used in food processing, medical field, plastic or other dielectric material welding or heating. Because they operate with high power high frequency electromagnetic fields, they emit electromagnetic disturbances with high amplitudes up to several GHz. For these reasons, these machines are strictly regulated by standards for electromagnetic compatibility.

I. INTRODUCTION

High frequency (HF) welding is the joining of material by supplying HF energy in the form of an electromagnetic field (27.12 MHz) and pressure to the material surfaces to be joined. A high-power high frequency generator produces the energy which is guided through a transmission line to the electrodes of the machine. The machine basically acts as a capacitor with high dielectric losses in the dielectric material used (Fig. 1). The electrical energy causes the molecules within the material to start moving, which generates heat that causes the material to soften and thereby fuse together. No external heat is applied, it is instead generated within the material. After cooling the welded surface under maintained pressure, the material is fused and a weld has been created. The weld seam can be at least as strong as the surrounding material – or even stronger.

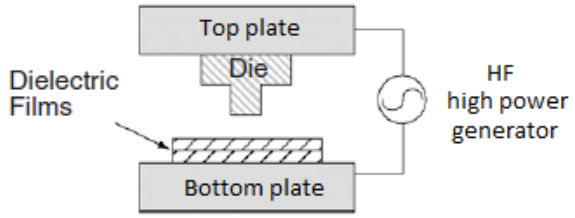


Figure 1 Diagram of the welding machine

Four important factors that affect the final weld result are pressure, welding effect, welding time and cooling time. These parameters can be adjusted and combined in different ways to achieve the optimal welding result for a specific material.

The material most commonly used with HF welding is PVC (polyvinyl chloride), sometimes called vinyl and PU (Polyurethane). The material can be thick or thin, reinforced or coated. It can also be plain, colored or patterned [1]. A photo of the tested welding machines can be seen on (Fig. 2).

Figure 2 Tested Welding machine series E

A. *EMC issues with HF welding machines*

The power of the HF welding machines can vary from less than 1 kW up to 20 kW or more. The frequency of the magnetic field is around 27.12 MHz. This means that high-power high frequency electromagnetic disturbances are produced during the working cycle of the machine. The harmonics of the generated frequency can reach up to several GHz with very high amplitudes. These disturbances might create many problems with electrical equipment operating near the machines, disturbances in the communication devices and can even have harmful effects on humans. For these reasons, such machines are subject to regulation according to a standard for electromagnetic compatibility - EN 55011:2010 [2]. The machines are intended to operate in industrial sites, that is why the standard has higher limits for the disturbances than in the domestic or urban areas.

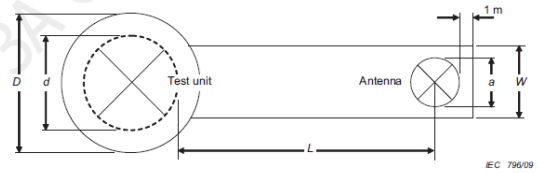
Electromagnetic compatibility ensures that the electrical devices can operate in a given surroundings without disturbing other devices or being disturbed.

The HF dielectric welding machines are usually heavy (several hundred kilograms or up to several tons). For this reasons they are not usually tested in laboratory conditions, but on the operation sites in normal conditions. The machine contains electronic components, it is a finished appliance intended for end user. According to the Guide for the EMC Directive 2004/108/EC it is classified as an apparatus [3]. It is intended for

incorporation into given fixed industrial installation. The directive provides set of requirements for these machines. According to the document the state of the art engineering practices should be applied for the machine manufacturing and also they should be documented and presented to the user and the authorities when required. Also if the machine creates disturbances which lead to improper operation or failure of other devices, then the manufacturer must take actions in order to fix the problems.

B. Standard EN 55011:2010

The title of the standard EN 55011:2010 is "Industrial, scientific and medical equipment - Radio-frequency disturbance characteristics - Limits and methods of measurement." In the Annex A of the standard it can be seen that the HF welding machine is classified as Group 2 "ISM RF equipment in which radio-frequency energy in the frequency range 9 kHz to 400 GHz." In the detailed description, plastic welding machines can be found.
The standard also defines the requirements about the test site an how the measurement must be held. Fig. 3 shows the distance requirements.



$D = (d + 2)$ m, where d is the maximum test unit dimension
$W = (a + 1)$ m, where a is the maximum test unit dimension
$L = 10$ m

Figure 3. Minimum size of the test area

The distance between the antenna and the test unit is defined to be 10m. There must be no other objects between the antenna and the measured machine. Also all other objects that might reflect the electromagnetic field must be removed from the room where the measurement takes place.
Table I shows the limits for the measurement results which are defined for 30 m distance from the test device and 10 m.

TABLE I LIMIT LINES FOR RADIATION MEASUREMENTS

Frequency range MHz	Limits for a measuring distance of D in m			
	On a test site D = 30 m from the equipment		On a test site D = 10 m from the equipment	
	Electric field	Magnetic field	Electric field	Magnetic field
	Quasi-peak dB(µV/m)	Quasi-peak dB(µA/m)	Quasi-peak dB(µV/m)	Quasi-peak dB(µA/m)
0,15 – 0,49	-	33,5	-	57,5
0,49 – 1,705	-	23,5	-	47,5
1,705 – 2,194	-	28,5	-	52,5
2,194 – 3,95	-	23,5	-	43,5
3,95 – 20	-	8,5	-	18,5
20 – 30	-	-1,5	-	8,5
30 – 47	58	-	68	-
47 – 53,91	40	-	50	-
53,91 – 54,56	40	-	50	-
54,56 – 68	40	-	50	-
68 – 80,872	53	-	63	-
80,872 – 81,848	68	-	78	-
81,848 – 87	53	-	63	-
87 – 134,786	50	-	60	-
134,786 – 136,414	60	-	70	-
136,414 – 156	50	-	60	-
156 – 174	64	-	74	-
174 – 188,7	40	-	50	-
188,7 – 190,979	50	-	60	-
190,979 – 230	40	-	50	-
230 – 400	50	-	60	-
400 – 470	53	-	63	-
470 – 1 000	50	-	60	-

At the transition frequency, the more stringent limit shall apply.
NOTE On a test site, class A equipment can be measured at a distance of 10 m or 30 m.

A quasi-peak detector should be used for the measurement in order to identify not only the peaks, but also the duration of the disturbances.

II. MEASUREMENT

A. Antenna

The antenna which is used for the measurement is log-periodic type with the following specifications:

- Freq.Range: 30-1000MHz
- Gain: -10~6.5dBi Typ..
- VSWR: 2.5:1 typ. (10.0:1 typ. @ 30MHz)
- Power Handling: 300W
- Connector: N-Female
- Material: Al
- Size: 1720mm(L) x 1450mm(H)

The antenna is shown on Fig. 4

Figure 4 Log-periodic antenna 30-1000 MHz

B. Spectrum analyzer

The spectrum analyzer used is a Hameg HM5014-2 with the following specifications (Fig 5):

- Frequency Range : 0.15 MHz…1.050GHz
- Frequency Resolution: 1kHz (6½ digit in readout)
- Resolution Bandwidths (RBW) @ 6dB: 1 MHz, 120 kHz and 9 kHz
- Video Bandwidth (VBW): 4kHz
- Sweep Time (automatic selection): 40 ms, 320 ms,1 s*)
- Measurement Range: -100dBm…+10dBm []

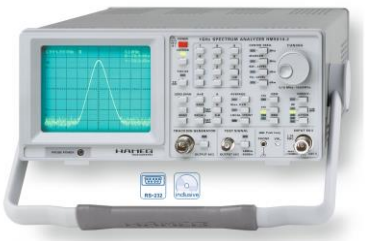


Figure 5 Hameg HM5014-2 spectrum analyzer

C. Software

The software used to record the measurement results is AS100E and is provided by Hameg. All the measurement components with their specifications can be set up in the software and all the values are calculated automatically and then the measurement is displayed graphically.

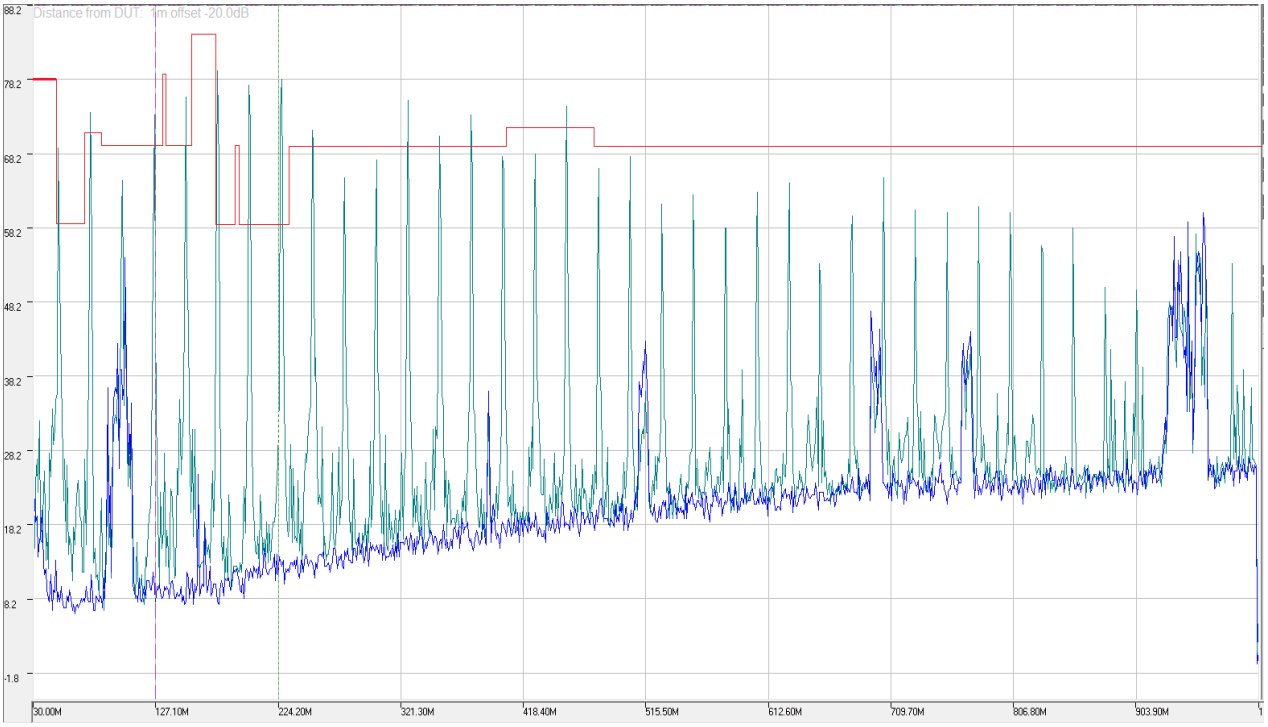


Figure 7 Measurement results for welding machine E200

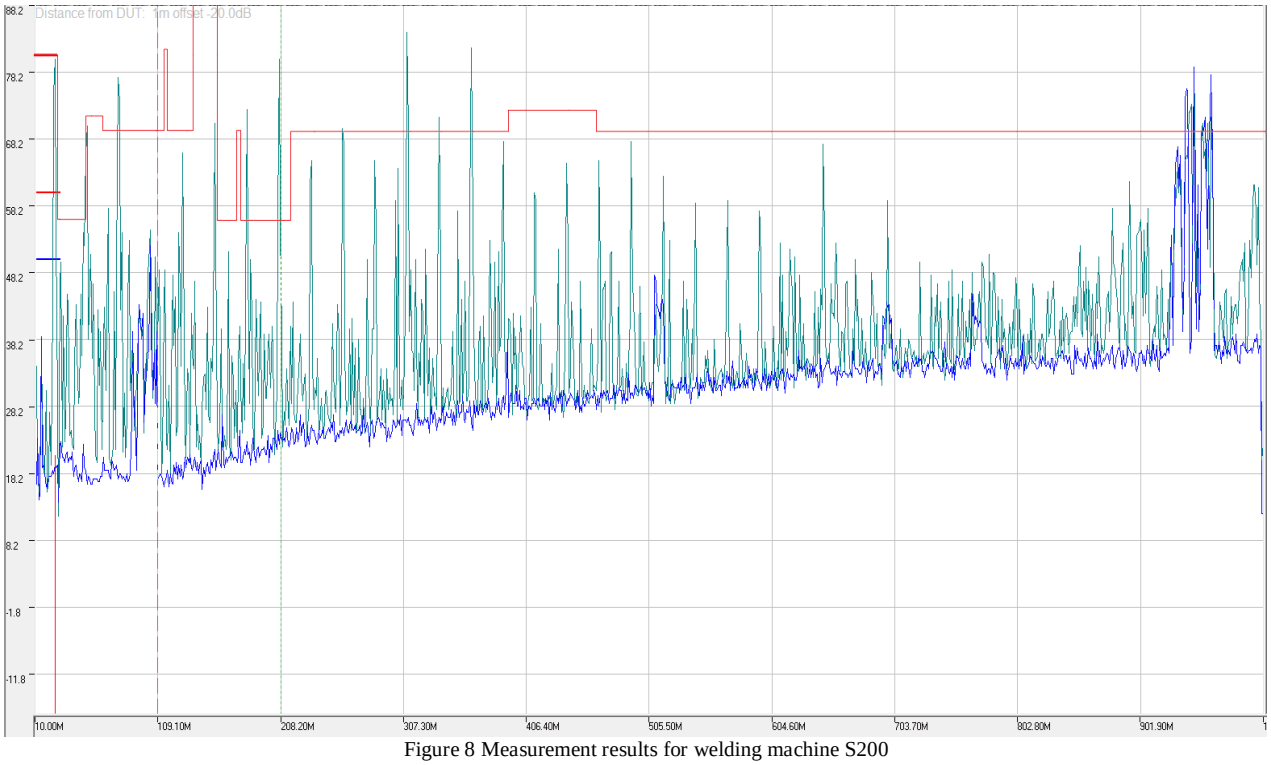


Figure 8 Measurement results for welding machine S200

D. Measurement set up

The measurement set up can be seen on fig. 6

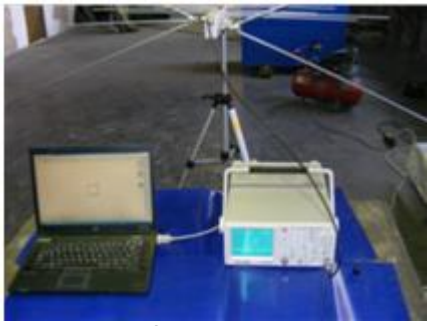


Figure 6 Measurement set up

Because of the smaller space of the testing site, the measurement was made at 3 m distance between the antenna and the machine. 10 dB are subtracted from the results in order to be comparable to 10 m distance, as in the standard. The measurements from 10 m are more correct than the ones held from 3 m, because of the near filed interactions in lower frequencies. The measurement data was recorded on a laptop PC with the special software installed. A 20 dB internal attenuator was used in order to protect the spectrum analyzer from the high energy electromagnetic pulses.

One welding cycle of the machine has a duration of about 15 seconds and then the machine stops. A 15 seconds sweep measurement from 10 MHz to 1 GHz was set up in the software in order to measure the whole welding cycle.

E. Measurement results

Two different machines were tested. Both must operate at 27,12 MHz. The first one, model E200, has a output power of 6kW and the other, model S200, has 9 kW output power.

Two measurements are made for every machine. First with machine turned off, in order to record the ambient electromagnetic field and the second the measurement of the machine in operational cycle. The actual value is obtained as a difference between the ambient and the operational electromagnetic field.

Fig. 7 Shows the measurement results for E200 machine.

The results show that the first harmonic is at 25 MHz and its value is 69 dBµV/m. The other harmonics can clearly be seen and they are repeated every 25 MHz.

From the measurements can be concluded that the disturbances from the machine E200 are at the edge of the compliance limit lines according to the standard EN 55011:2010. The machine disturbs nearby FM receivers, but no other harmful effects on other devices is detected.

Fig. 8 shows the measurement results for welding machine S200.

The results show that the first harmonic is at 26 MHz and its value is 60 dBµV/m. The harmonics repeating every 26 MHz can also be clearly seen on the graphic. The machine S200 is also very close to the limit lines but still under them. It also disturbs FM receivers nearby.

III. CONLUSION

The results of the measurements show that the disturbances from both plastic welding machines are very close to the limit lines specified by the standard for industrial equipment EN 55011:2010. A proper redesign of the machines will provide better results . Some of the suggestions are better geometric construction of the transmission line, shielding of the components and cables, better filtering of the supply voltage and communication wires.

ACKNOWLEDGMENT

"The present document has been produced with the financial assistance of the European Social Fund under Operational Programme “Human Resources Development”. The contents of this document are the sole responsibility of the “Angel Kanchev” University of Ruse and can under no circumstances be regarded as reflecting the position of the European Union or the Ministry of Education and Science of Republic of Bulgaria."

Project № BG051PO001-3.3.06-0008 “Supporting Academic Development of Scientific Personnel in Engineering and Information Science and Technologies

”REFERENCES

[1] Microwave/RF Applicators and Probes for Material Heating, Sensing, and plasma generation.

[2] Standard EN 550011:2010 Industrial, scientific and medical equipment - Radio-frequency disturbance characteristics

[3] DIRECTIVE 2004/108/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 15 December 2004

Energy consumption analysis of real metro-optical network

Josip Lorincz[1], Eldis Mujaric[2] Dinko Begusic[3]
[1, 3]FESB, University of Split, Croatia, e-mail: josip.lerinc@fesb.hr, dinko.begusic@fesb.hr
[2]Croatian Academic and Research Network – CARNet, Croatia, e-mail: eldis.mujaric@carnet.hr

***Abstract*— Significant part of the total energy consumption in distribution and core optical networks is contributed by the energy consumption of Internet Protocol (IP) routers. In this paper, an energy efficiency analysis has been performed for the case of real metro-optical network of a medium size city. Two real metro-optical network configurations, mores specifically dark fiber and the dense wavelength division multiplexing (DWDM) have been analyzed in terms of the energy efficiency. Obtained results provide insight into daily variations of routers' energy efficiency expressed in watts (W) per Mb/s of transferred traffic. In addition, cons and pros for selection of analyzed network configurations in real metro-optical implementations have been presented from the point of energy efficiency.**

Keywords*— optical, router, metro, network, energy-efficiency, dark fiber, WDM, green

I. INTRODUCTION

In this paper, energy-efficiency analyses for the real metro-optical network that is a part of e-Split project network infrastructure will be presented. The e-Split is a joint project of Croatian Academic and Research Network (CARNet) and the City of Split, Croatia. In the frame of this project, optical infrastructure based on a dark fiber has been developed. The dark fiber is a privately operated optical fiber infrastructure that is run directly by its owner, rather than by purchasing bandwidth or leasing line capacity. It is called "dark" because it is not in use until owner activates network resources [1].

Metro optical infrastructure connects CARNet member institutions in Split to the CARNet core network. Current implementation of given network infrastructure is based on dark fiber without implementation of WDM technology. WDM is a technology which multiplexes several optical carrier signals into a single fiber by using different wavelengths of light. This enables bidirectional communications over one fiber, as well as multiplication of capacity. Two most used WDM technologies are CWDM (Coarse Wavelength-Division Multiplexing) and DWDM (Dense Wavelength-Division Multiplexing). Basic CWDM characteristics are: used at distances below 50 km, without need to use optical amplifiers, up to 18 wavelengths in single fiber, 45 Gbps per single fiber, 2.5 Gbps per wavelength, less expensive than the DWDM systems.

Essential DWDM characteristics are: up to 240 wavelengths in single fiber, 32 C-band wavelengths are used in practice, 2.4 Tbps per single fiber, 10 Gbps per wavelength, at greater distance optical amplifiers are needed, more expensive then the CWDM. DWDM system (Figure 1) sends signals from several sources over a single fiber, based on signals multiplexing at the sending end and demultiplexing at the receiving end. Multiplexing is done by a multiplexer through taking optical wavelengths from multiple fibers and combining them into one light beam. At the receiving end demultiplexer has to separate out the components of the light so that they can be detected by separating the received beam into its wavelength components and coupling them to individual fibers.

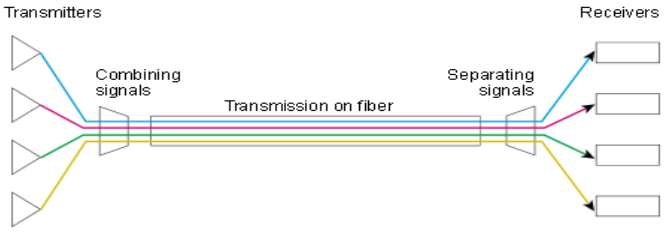


Figure 1: DWDM functional schema

In a unidirectional system, there is a multiplexer at the sending end and a demultiplexer at the receiving end. Two systems would be required at each end for bidirectional communication, and two separate fibers are needed. In case of bidirectional DWDM, multiplexer/demultiplexer must be at each end of the link because of the fact that single fiber is used for communication. Different wavelengths have to be used for full-duplex communication [2].

In the implemented real metro-optical network, ends of the optical fibers are closed with small form-factor pluggable (SFP) interfaces. The SFPs are popular industry format interface jointly developed and supported by many network component vendors. In the analyzed network, implemented SFP technology supports single wavelength transmission over single mode optical fiber.

Although expandable slots of the analyzed network equipment support WDM line cards, this technology is not used due to current user's needs which don't require WDM. However, constant increase of the network traffic transferred over implemented and analyzed metro-optical network, might demand upgrade to some of the WDM technologies in a near future. Besides operator increase in capital expenditures (CAPEX) due to implementation of additional optical elements such as multiplexers and demultiplexers, amplifiers and transceivers, upgrade to the WDM technology will also raise the operational expenditures (OPEX). This is because additional optical elements with router WDM line cards increase overall network energy consumption. This in turn increases overall operator OPEX. For that reason, differences in the energy efficiency among those two metro-optical network implementations based on single wavelength SFP and WDM technology are analyzed in the paper.

Research activates on this subject are supported by the EU project: Establish Pan-European Information Space to Enhance Security of Citizens - FP7 Grant Agreement No: 607078.

The rest of the paper is organized as follows: related work focused on improving energy efficiency of metro-optical networks is presented in Section 2. Physical and logical topology including architecture and active equipment of the analyzed network are introduced in Section 3. Section 4 shows results of real traffic patterns detected during seven days period. Power efficiency of network architecture based on dark fiber and DWDM technology has been discussed in Sections 5 and 6, respectively. Comparison of power efficiency among the dark fiber and DWDM technologies implemented in analyzed network are given in Section 7. Finally, some concluding remarks are given in Section 8.

II. RELATED WORK

Considering continuous increase for bandwidth in the core networks and overall Internet, satisfying power demands will be a fundamental challenge in the next generation networks (NGN). Authors in [3] present methods of potential savings achievable through power-aware network design and routing. Energy consumption model of large-scale optical network is discussed in [4]. It is shown that telecoms can improve energy efficiency of the network by changing hardware characteristics of active network equipment.

In addition, it is known that routers chassis and line cards consume significantly higher amount of energy in core networks. Different fill levels of the chassis results with different energy consumption. Higher fill level will result with more energy-efficient network. Even an empty chassis with no line cards consumes a large amount of energy. Therefore, a chassis with higher fill level has lower energy consumption per transferred bit [5].

Backbone networks transport large aggregated amount of traffic. Three different network architectures based on the WDM technology are compared in terms of power consumption in [6]. Methods for minimizing energy consumption in the WDM based transport architectures are presented in [7]. Particular realizations of static optical core networks are compared with the dynamic packet switched architecture based on the WDM in [8]. Survey of the most relevant research efforts considering backbone networks power consumption reduction is given in [9].

Considering all these facts, in this paper we perform energy consumption analyses of two network architectures, which, according to our knowledge haven't been previously analyzed. First network is static and fully operational real optical network maintained by the CARNet, and second is assumed dynamic optical network based on the xWDM technology. This network architecture might be used as future metro-optical network of e-Split project, and such network is widely implemented in large-scale interconnections, like interconnections inside large cities.

III. NETWORK ARCHITECTURE

A. Backbone optical ring

At the physical layer of analyzed network (Figure 2), there is the dark fiber cable with 192 fibers forming backbone optical ring which connects five network nodes (A-E)

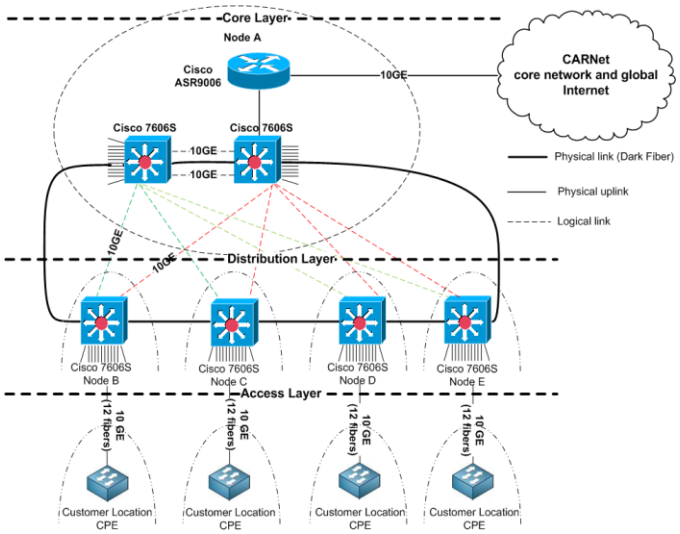


Figure 2: Analyzed network topology with dark fiber and SFPs

allocated inside one medium size city. Figure 2 illustrates logical and physical topology of the analyzed network based on the optical ring. The 24 fibers in the dark fiber cable are used for ensuring redundant logical connection to a main CARNet node (Node A). The dark fiber as point-to-point network architecture is chosen at the physical layer in order to satisfy requirements such as: simplicity, scalability and robustness.

Every customer location is connected in the same way (Figure 2): one optical cable with 12-fibers connects the end site of the costumer to the closest shaft. From that shaft, two fibers are used to connect end site to nearest node through the backbone ring. Two fibers are patched in the shaft as redundancy in case that fiber cut occurs. The remaining eight fibers are reserved for future use in the case when more connections will be needed. Each end site is connected to the geographically closest node (Figure 2).

Implemented architecture allows simple bandwidth increase through activation of new interfaces (line cards) at the end of the dark fiber optical links. This can be performed without coordination of carrier service provider, while full monitoring of all network devices by the owner of the network will remain. Passive network equipment at every end site includes: rack, optical patch panel, power socket extension, cable ducts, shelf for active network equipment.

B. Active network equipment

Main network node for connecting all other CARNet's nodes of the e-Split network to the CARNet backbone network is Node A (Figure 2). Main Node A is composed of two Cisco 7606 routers working in failover mode. All other nodes of the e-Split network have one Cisco 7606 router logical connection with the main Node A via redundant 10 Gbps links (Figure 2). Each node aggregates geographically closest end sites and each end site has access (L3) switch (such as Cisco 3560).

Cisco ASR9000 router is used in node A for interconnection with distribution layer as well as providing interconnection with main node in Zagreb (Figures 2 and 4). Node A is connected with a main node of the CARNet

TABLE 1: TECHNICAL CHARACTERISTICS OF THE DEVICES

Device (location)	CPU	RAM [Kb]
CISCO7606-S (core/distribution layer)	MPC8548_E, Version: 2.0	851,968
ASR9006 (core layer)	Intel 686 F6M14S4 processor at 2132 MHz, Revision 2.174	6,291,456

TABLE 2: CISCO ASR9006 ROUTER COMPONENTS

Card type	Model
Cisco ASR 9006 fan	ASR-9006-FAN
Route Switch Processor	A9K-RSP-4G
24-Port 10GE Packet Transport Optimized Line Card, Requires SFP+ optics	A9K-24X10GE-TR
Mod80 Modular Line Card, Packet Transport Optimized	A9K-MOD80-TR
4-Port 10GE Low Queue Line Card	A9K-4T-L

TABLE 3: CISCO 7606-S ROUTER COMPONENTS

Card Type	Model
CEF720 8 port 10GE with DFC	WS-X6708-10GE-3C
CEF720 48 port 1000Mb SFP	WS-X6748-SFP
Route Switch Processor 720	RSP720-3C-GE
Firewall service module	WS-SVC-FWM-1
4-port line card with Distributed Forwarding Card 3CXL	7600-ES+4TG3CXL
High Speed Fan Module for CISCO7606-S	FAN-MOD-6SHS

network in Zagreb city with the 10 Gbps uplink. Node in Zagreb provides access to the global Internet for all CARNet's users. Technical characteristics of the core and distribution network equipment in terms of central processing unit (CPU) types and random access memory (RAM) are shown in Table 1.

Cisco ASR9000 router components are shown in Table 2. Main part of Cisco ASR9000 router is A9K-RSP-4G Route Switch Processor comprising the following components: bidirectional nonblocking fabric, 40-GB hard drive, switch fabric, High-performance dual-core CPU, 4GB RAM and 1GB flash memory. It supports high 1/10/100-Gbps port throughputs. The A9K-MOD80-TR is modular line card, that is optimized for fast packet transport and requires modular port adapters to customize each slot. The A9K-24X10GE-TR is 24-Port 10 Gigabit Ethernet Line Card designed for various Cisco ASR platforms. It provides up to 10 Gbps of throughput per single port. Finally, the A9K-4T-L is modular 10-Gigabit Small Form-Factor Pluggable (XFP)-based line card. Short reach (SR), intermediate reach (IR), long reach (LR), CWDM, DWDM, and 10/100/1000BASE-T interfaces are supported.
Distribution layer includes Cisco 7606-S router with components shown in Table 3. The WS-X6708-10GE is 8-port Ethernet 10GBase-X module with 1 GB of RAM designed for various Cisco platforms. The WS-X6748-SFP is model of gigabit Ethernet SFP interface module. Router has two of these modules and one Route Switch Processor 720. Firewall module is stateful inspection firewall with application and protocol inspection engines. It provides up to 5.5 Gbps of throughput, 100,000 new connections per second, one million concurrent connections or 256,000 NATs (network address translations) and up to 80,000 ACL (access control list) entries. Up to four firewall services modules (FWSM) can be installed in a single chassis, providing scalability of up to 20 Gbps per chassis.

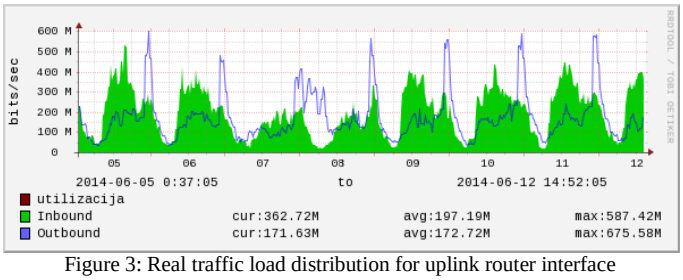


Figure 3: Real traffic load distribution for uplink router interface

IV. REAL TRAFFIC PATTERNS

For analyzed network, real inbound and outbound traffic patterns between end sites and the global Internet will be presented. One of the Cisco 7606-S routers aggregates all end sites (Figure 2) and provides connectivity to the CARNet backbone through Cisco ASR9000 router. Traffic data were obtained on uplink of this router which is link to the Cisco ASR9000 router (Figure 2). The other Cisco 7606-S router serves as backup (works in failover mode). Distribution of the real traffic load detected on the router for uplink interface is presented in Figure 3. Traffic loads presented in Figure 3 are the snapshot obtained from network traffic monitoring system. Data were taken in the interval from June 5 to June 12, 2014. Monitoring system is based on open-source Zenoss software that collects constantly (24/7) real-time data of network traffic.

It can be noticed that traffic pattern has typical day/night pattern, which is consequence of different user's activity during day and night. End locations reaches around 1.4 Gbps at maximum in 7 days period. Around 590 Mbps is maximal inbound and around 680 Mbps is maximal outbound traffic.

V. POWER EFFICIENCY OF CORE AND DISTRIBUTION LAYERS BASED ON DARK FIBER

In order to estimate power consumption of active network equipment in the distribution and core layer, online Cisco tool known as "Cisco Power Calculator" [10] were used. Application enables users to calculate the power supply requirements for a specific configuration of network element. Firstly, power consumption of distribution and core layer of the active network equipment presented in Figure 2, which uses dark fiber and SFPs (without WDM technology) are estimated. Then, power efficiency for network presented in Figure 4, with interface modules that supports DWDM technology is simulated. The data presented in Table 4 shows: output current/power and system's heat dissipation for the Cisco 7606-S router with elements supporting dark fiber (without DWDM line cards), having configuration presented in Table 2. The same parameters for CiscoASR9000 having configuration presented in Table 3, with elements supporting dark fiber (without DWDM links) are shown in Table 5.

In order to calculate power efficiency, traffic loads from distribution and core layers of the e-Split network infrastructure were selected. In addition, typical power consumption values characteristic for network operation at standard operating temperature (20° C) for each router configuration presented in Tables 4 and 5 were used. Traffic loads data were taken in typical periods of one working day (Wed, June 11 2014.) and one weekend day (Sun, June 8

TABLE 4: POWER CONSUMPTION AND HEAT DISSIPATION FOR CISCO7606-S IN THE DARK FIBER ARCHITECTURE

Power Consumption/Heat Dissipation Summary/Configuration Details

Slot	Line Card	Output Current (A)	Output Power (W)	Heat Dissipation (BTU/Hr)
FAN2	FAN-MOD-6SHS	7.40	311.00	1240.50
1	WS-X6708-10G-3C	10.58	444.36	1785.28
2	WS-X6748-SFP	6.07	254.94	1024.26
3	WS-X6748-SFP	6.07	254.94	1024.26
4	Empty	0	0	0
5	RSP720-3C-10GE	8.45	354.90	1425.86
6	RSP720-3C-10GE	8.45	354.90	1425.86
Total		**47.02**	**1975.04**	**7926.03**
Total Output Current	**47.02 Amps**			
Total Output Power	**1975.04 Watts**			
Total Heat Dissipation	**7926.03 BTU/Hr**			

TABLE 5: POWER CONSUMPTION AND HEAT DISSIPATION FOR CISCOASR9000 IN THE DARK FIBER ARCHITECTURE

Power Consumption/Heat Dissipation Summary/Configuration Details

Slot	Line Card	Output Current (A)	Typical Output Power (W)	Power Used at 40C(W)	Heat Dissipation At 40C(BTU/Hr)	Power Used At MAX(50C/55C) (BTU/Hr)	Heat Dissipation At MAX(50C/55C) (BTU/Hr)
SYSTEM-FAN	ASR-9006-FAN	0.00	125	275.00	937.75	375.00	1278.75
SYSTEM-FAN	ASR-9006-FAN	0.00	125	275.00	937.75	375.00	1278.75
RSP0	A9K-RSP-4G	0.00	175	205.00	699.05	235.00	801.35
RSP1	Empty	0	0	0	0	0	0
LC0	A9K-MOD80-TR	0.00	350	400.00	1364.00	420.00	1432.20
LC1	A9K-24X10GE-TR	0.00	775	850.00	2898.50	895.00	3051.95
LC2	Empty	0	0	0	0	0	0
LC3	Empty	0	0	0	0	0	0
Total		**10.13**	**1550.00**	**2005.00**	**6837.05**	**2300.00**	**7843.00**
Total Output Current	**10.13 Amps**						
Total Power Used At 40C	**2005.00 Watts**						
Total Heat Dissipation At 40C	**6837.05 BTU/Hr**						
Total Typical Output Power	**1550.00 Watts**						
Total Power Used At MAX(50/55C)	**2300.00 Watts**						
Total Heat Dissipation At MAX(50/55C)	**7843.00 BTU/Hr**						

2014.). Inbound and outbound traffic loads were taken at the uplink interfaces of distribution router CISCO7606-S in node A, which aggregates all customer end sites, and core router ASR9006 in Node A, which provides interconnection with main node in Zagreb. Power efficiency was calculated for each traffic direction (inbound and outbound). Traffic load and power efficiency for distribution layer devices (CISCO7606-S router) at working day and weekend are shown in Table 6.

Typical traffic load and power efficiency for core layer devices (ASR9006 router) at working day and weekend are presented in Table 7. Both tables show distinctions in the inbound and outbound traffic loads during different periods of the day and week, what influences on device power efficiency. From Tables 6 and 7 can be noticed that power efficiency for inbound traffic is always larger then efficiency for outbound traffic, due to higher traffic in inbound direction caused by larger user exploitation of services form Internet.

VI. POWER EFFICIENCY OF CORE AND DISTRIBUTION LAYERS BASED ON DWDM TECHNOLOGY

In addition to the power consumption analyses of network based on dark fiber technology with the SFPs, power consumption and heat dissipation analyses for network which enables data transfer over DWDM technology has been

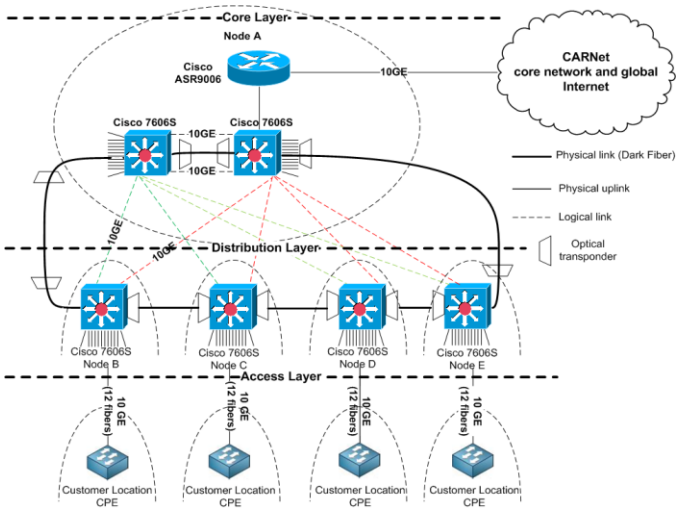


Figure 4: The proposed architecture based on DWDM

TABLE 6: TRAFFIC LOAD AND POWER EFFICIENCY FOR CISCO7606-S ROUTER ON WORK DAY AND WEEKEND

	Working day		Weekend	
Time (Wed, June 11)	**Traffic load (Inbound / Outbound) [Mbps]**	**Power efficiency (inbound/ outbound) [W/Mbps]**	**Traffic load (Inbound / Outbound) [Mbps]**	**Power efficiency (inbound/ outbound) [W/Mbps]**
00:00	254.3/411.39	7.76/4.8	221.91/413.53	8.9/4.77
08:00	222.95/81.99	8.86/24.09	37.03/178.14	53.34/11.09
12:00	396/127	4.99/15.55	133.73/113.91	14.77/17.34
16:00	284.97/192.4	6.93/10.27	181/151.82	10.91//13
20:00	316.64/228.87	6.24/8.63	126.81/119.18	15.57/16.57

TABLE 7: TRAFFIC LOAD AND POWER EFFICIENCY FOR CISCO ASR9006 ROUTER ON WORK DAY AND WEEKEND

	Work day		Weekend	
Time (Wed, June 11)	**Traffic load (Inbound / Outbound) [Mbps]**	**Power efficiency (inbound/ outbound) [W/Mbps]**	**Traffic load (Inbound / Outbound) [Mbps]**	**Power efficiency (inbound/ outbound) [W/Mbps]]**
00:00	263.5/419	5.88/3.7	226.5/417.54	6.84/3.71
08:00	257.39/118.33	6.02/13.1	40.3/277.73	38.46/5.58
12:00	472/135.63	3.28/11.43	137.6/119.95	11.26/12.92
16:00	316/196,62	4.91/7.88	197.69/162.34	7.84/9.55
20:00	326.42/232.19	4.75/6.68	134.08/127.44	11.56/12.16

performed. The proposed architecture for given network based on DWDM technology is shown in Figure 4. Results obtained from Cisco power calculator in terms of heat dissipation and power consumption are presented in Tables 8 and 9. Table 8 presents obtained results for the Cisco 7606-S router. Furthermore, power consumption and heat dissipation data for Cisco ASR9000 assuming hardware configuration which includes interfaces supporting the DWDM technology are presented in Table 9.

Also, for Cisco 7606-S and Cisco ASR9000 router, both having DWDM hardware components, power efficiency comparison with a real traffic load is given in Tables 10 and 11, respectively. Tables 10 and 11 show power efficiency for bidirectional traffic, which is represented as sum of inbound and outbound traffic. Due to performed analyses on the same network, it is possible to perform direct comparison among obtained results.

TABLE 8: POWER CONSUMPTION AND HEAT DISSIPATION FOR CISCO7606-S IN THE DWDM NETWORK

Power Consumption/Heat Dissipation Summary/Configuration Details

NOTE:

Chassis reserves power for Redundant Supervisor Engine if redundant Supervisor Engine slot is empty.

Slot	Line Card	Output Current (A)	Output Power (W)	Heat Dissipation (BTU/Hr)	Optional DFC	Power over Ethernet Capabilities
FAN2	FAN-MOD-6SHS	7.40	311.00	1240.50		
1	WS-X6708-10G-3C	10.58	444.36	1785.28	--	---
2	WS-X6748-SFP	6.07	254.94	1024.26	--	---
3	7600-ES+4TG3CXL	9.50	399.00	1603.04	WS-DFC-3CX	---
4	WS-SVC-FWM-1	4.09	171.78	690.15	--	---
5	RSP720-3C-GE	7.38	309.96	1245.31	--	--
6	-- Reserved Power --	7.38	309.96	--	--	--
Total		**52.40**	**2201.00**	**7588.54**		
Total Output Current	**52.40 Amps**					
Total Output Power	**2201.00 Watts**					
Total Heat Dissipation	**7588.54 BTU/Hr**					

TABLE 9: POWER CONSUMPTION AND HEAT DISSIPATION FOR CISCOASR9000 IN THE DWDM NETWORK

Power Consumption/Heat Dissipation Summary/Configuration Details

Slot	Line Card	Output Current (A)	Typical Output Power (W)	Power Used at 40C(W)	Heat Dissipation At 40C(BTU/Hr)	Power Used At MAX(50C/55C) (BTU/Hr)	Heat Dissipation At MAX(50C/55C) (BTU/Hr)
SYSTEM-FAN	ASR-9006-FAN	0.00	125	275.00	937.75	375.00	1278.75
SYSTEM-FAN	ASR-9006-FAN	0.00	125	275.00	937.75	375.00	1278.75
RSP0	A9K-RSP-4G	0.00	175	205.00	699.05	235.00	801.35
RSP1	Empty	0	0	0	0	0	0
LC0	A9K-24X10GE-TR	0.00	775	850.00	2898.50	895.00	3051.95
LC1	A9K-MOD80-TR	0.00	350	400.00	1364.00	420.00	1432.20
LC2	A9K-4T-L	0.00	310	320.00	1091.20	350.00	1193.50
LC3	Empty	0	0	0	0	0	0
Total		**11.74**	**1860.00**	**2325.00**	**7928.25**	**2650.00**	**9036.50**
Total Output Current	**11.74 Amps**						
Total Power Used At 40C	**2325.00 Watts**						
Total Heat Dissipation At 40C	**7928.25 BTU/Hr**						
Total Typical Output Power	**1860.00 Watts**						
Total Power Used At MAX(50/55C)	**2650.00 Watts**						
Total Heat Dissipation At MAX(50/55C)	**9036.50 BTU/Hr**						

VII. RESULTS COMPARISON

In the analyses, traffic load data in characteristic periods of a day (Table 6, 7 and 10, 11) are selected. It may be seen in Figure 3 that the maximum traffic loads occur around 00:00, and the minimum traffic loads occur around 08:00 AM. Bidirectional traffic load in specific timestamps and estimated power consumption are used for calculation of the power efficiency values [W/Mbps] for each network configuration (dark fiber and DWDM).

It shown in [11] that significant daily variation of traffic load has impact on instantaneous power consumption of network devices. However, our initial assumption is that typical power consumption of network device is constant in the regular network operation. This is the case when device is not loaded with a large amount of traffic, and when the value of memory and CPU utilization does not vary significantly in time. This is exact situation in our test case, since analyzed network is over capacitated for the purpose of accepting future traffic loads increase during period of next 10 years.

It can be clearly seen from Tables 6, 7 and 10, 11 that power efficiency is inversely proportional to the traffic load.

TABLE 10: CISCO7606-S POWER EFFICIENCY COMPARISON

Time (Wednesday, June 11)	Power efficiency (bidirectional dark fiber SFP interface) [W/Mbps]	Power efficiency (bidirectional DWDM interface) [W/Mbps]
00:00	1975/665.69 (2.97)	2201/665.69 (3,31)
08:00	1975/304.94 (6.48)	2201/304.94 (7.22)
12:00	1975/523.00 (3.78)	2201/523.00 (4.21)
16:00	1975/477.37 (4.14)	2201/477.37 (4.61)
20:00	1975/545.51 (3.62)	2201/545.51 (4.03)

Table 11: CISCOASR9000 POWER EFFICIENCY COMPARISON

Time (Wednesday, June 11)	Power efficiency (bidirectional dark fiber SFP interface) [W/Mbps]	Power efficiency (bidirectional DWDM interface) [W/Mbps]
00:00	1550/682.5 (2.27)	1860/682.5 (2.73)
08:00	1550/375.72 (4.13)	1860/375.72 (4.95)
12:00	1550/607.63 (2.55)	1860/607.63 (3.06)
16:00	1550/508.62 (3.05)	1860/508.62 (3.66)
20:00	1550/558.61 (2.77)	1860/558.61 (3.33)

For the higher values of the traffic load, power efficiency in W/Mbps is lower. This means that for the same consumed power, larger amount of data can be transferred.

In addition, values of power efficiency in core and distribution network based on dark fiber and DWDM technology are compared. Firstly, Cisco 7606 router in the distribution layer is analyzed. Its uplink towards the core layer has maximum traffic load around midnight. Its estimated total output power consumption with SFP line cards is 1,975 W, and the estimated total output power consumption with line card that supports DWDM is 2,201 W. Figure 5 presents data power efficiency curve interpolated for whole day.

Furthermore, Cisco ASR9006 router in the core layer was analyzed. Traffic transferred over its uplink towards the main node in Zagreb has the same shape as traffic pattern of the router in the distribution layer (maximum traffic load around midnight). It's estimated total power consumption when dark fiber and DWDM is implemented equals to 1,550 W and 1,860 W, respectively. Considering this, data power efficiency in characteristic times of the day is calculated and interpolated for whole day. Obtained results are shown in Figure 6.

Finally, we give power efficiency comparison for each device on working day and weekend (Sunday), in case of optical network based on dark fiber technology. Power efficiency comparison for Cisco 7606 and ASR 9006 on working day and weekend is shown in Figure 7 and 8, respectively. Larger differences between power efficiency on working and weekend day is consequence of significant difference in amount of traffic during this days Also, in comparison with DWDM configuration, Figures 5 and 6 shows that SFP implementation contributes to the reduction of power consumption. This is due to lower power consumption of SFP hardware components. Hence, SFP implementation is more favorable for networks with low CPU utilization and lower amounts of traffic. Based on results presented in Figures 7 and 8, it can be noticed that independently of implemented technology (SFP or DWDM), power efficiency of devices is better for working day, meaning that higher amount of traffic can be transferred for the same amount of consumed power.

Additionally, calculation of power consumption for a given network infrastructure based on dark fiber with SFPs (without

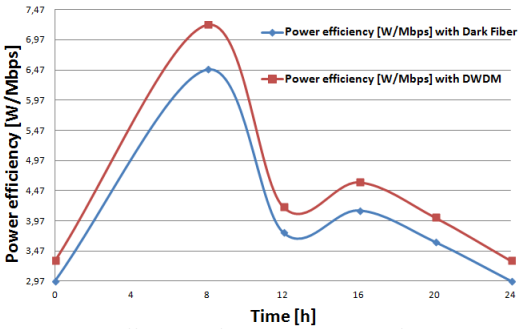


Figure 5: Power efficiency of Cisco 7606 in dark fiber and DWDM Optical networks

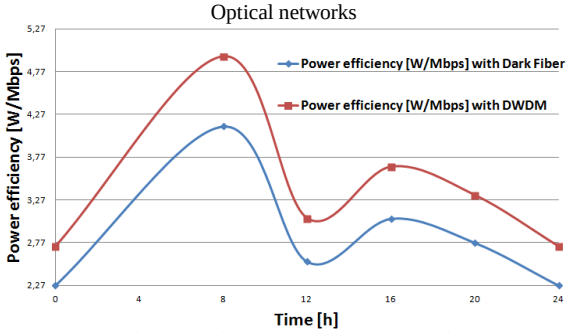


Figure 6: Power efficiency of Cisco ASR 9006 in dark fiber and DWDM optical networks

DWDM) is performed. As presented in Figure 2, network has five nodes with six Cisco 7606 routers and one ASR 9006. Each 7606 router consumes power of 1,975 W, while ASR 9006 typically consumes 1,550 W, what gives total instantaneous power consumption equal to the 13,400 W for routers in core and distribution layer of analyzed metro-optical network.

In case of the same metro-optical network with five nodes working with DWDM hardware, optical network architecture must contain ten additional 10 Gbps transponders (Figure 4) that typically consume 68.5 W each [12]. Total estimated power consumption in this case would be 14,085 W. This represents increase in power consumption of network core and distribution layer for more than 60 MWh yearly, what further gives favor on practical implementation of dark fiber.

VIII. CONCLUSION

In this paper, energy consumption analyses of the real metro-optical network based on prominent technologies such as the dark fiber and DWDM are presented. Analyses is performed for real traffic patterns, according to which, power efficiency of distribution and core level network equipment was estimated. It is shown that daily variation of traffic pattern influence on power efficiency (W/Mbps) of network devices, since higher amount of traffic can be transferred for the same amount of power consumption. Additionally, obtained results show that metro optical-networks of a small and medium size cities, based on dark fiber consumes less overall energy when compared with equal network based on the DWDM technology. Hence, dark fiber in terms of energy-efficiency is optimal solution for metro-optical networks of small and medium cities having low amount of data traffic and low

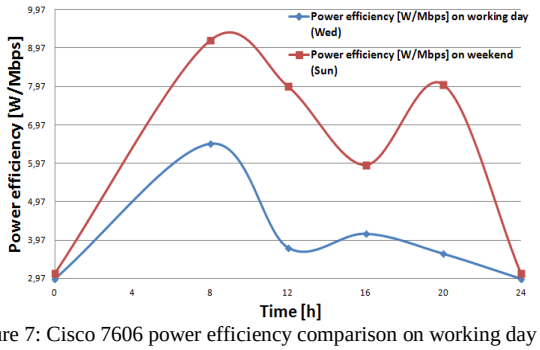


Figure 7: Cisco 7606 power efficiency comparison on working day and weekend

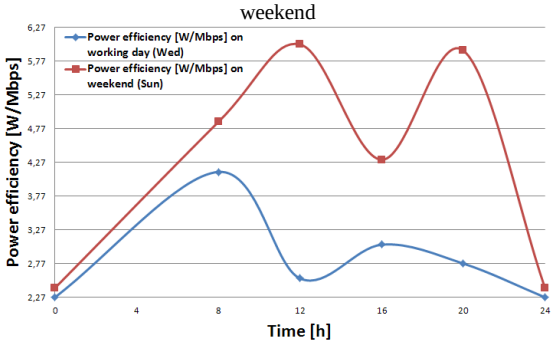


Figure 8: Cisco ASR 9006 power efficiency comparison on work day and weekend

network equipment CPU utilization. However, in case of large metro-optical networks where huge bandwidth demands must be satisfied, DWDM as technology has to be considered, regardless of lower energy-efficiency. Our future research activities will be focused on energy-efficiency analyses of optical amplifiers in metro-optical networks.

REFERENCES

[1] J. Bohman, T. Blomdahl, Dark fibre- market and state of competition, Report number: PTS-ER-2008:9National Post and Telecom Agency (PTS), Sweeden, 2008

[2] Introduction to DWDM for Metropolitan Networks, Cisco Syst., 2000.

[3] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, S. Wright, Power Awareness in Network Design and Routing, INFOCOM 2008.

[4] J. Baliga, R. Ayre, K. Hinton, W. V. Sorin and R. S. Tucker, Energy Consumption in Optical IP Networks, Journal of Lightwave Technology, Vol. 27, No. 13, July 1, 2009.

[5] Y. Zhang, P. Chowdhury, M. Tornatore, and B. Mukherjee, Energy Efficiency in Telecom Optical Networks, IEEE Comm. Surveys & Tutorials, Vol. 12, No. 4, Fourth Quarter 2010.

[6] F. Vismara, V. Grkovic, F. Musumeci, M. Tornatore, S. Bregni, On the Energy Efficiency of IP-over-WDM Networks, LATINCOM, 2010.

[7] G. Shen and R. S. Tucker, Energy-Minimized Design for IP Over WDM Networks, J. Opt. Commun. Netw./Vol. 1, No. 1/June 2009.

[8] S. Aleksic, W. Van Heddeghem and M. Pickavet, Scalability and Power Consumption of Static Optical Core Networks, GLOBECOM, 2012.

[9] M. N. Dharmaweera, R. Parthiban, Y. A. Sekercioglu, Towards a Power-Efficient Backbone Network: The State of Research, Comm. Surveys & Tutorials, IEEE (Volume:PP , Issue: 99), 2014.

[10] Cisco Power Calculator Application - http://tools.cisco.com/cpc/, Cisco Systems, Inc, 2014.

[11] J. Lorincz, T. Garma and G. Petrovic, "Measurements and Modelling of Base Station Power Consumption under Real Traffic Loads", *Sensors Journal*, Volume 12, Issue 04, pp. 4281-4310, 2012.

[12] Ward Van Heddeghem, Filip Idzikowski, Equipment power consumption in optical multilayer networks – source data, Photonic Network Comm., 2012.

Doprinos informacijsko-komunikacijskih tehnologija smanjenju emisije CO_2

Mirza Berković i Amel Kosovac

Fakultet za saobraćaj i komunikacije, Sarajevo, BiH
mirza.berkovic@gmail.com; amelkosovac@gmail.com

Sažetak – Cilj rada je analiza modela procjene doprinosa ICT-a na smanjenje CO_2 emisija i sistema procjene. Problem istraživanja je baziran na činjenici da ne postoji generički model mjerenja doprinosa ICT-a na smanjenje CO_2 emisija. U radu se definiše model strukture sistema procjene doprinosa te analiziraju funkcije sistema procjene uključenih komponenti ICT-a u skladu s opisanom svrhom. Pri čemu na funkcionisanje modela procjene utječu brojni faktori okoline, koji se dijele na unutrašnje i vanjske faktore okoline. Svrha definisanja modela i upravljanja doprinosom je reducirati vjerovatnoću pojave negativnih efekata i kontrolirati njihove posljedice kako bi se ostvarili pozitivni ciljevi projekta. Sistem procjene smanjenja CO_2 emisija primjenom ICT-a je dinamički sistem, što znači da tokom vremena dolazi do promjena čiji su uzroci procesi koji se odvijaju u elementima sistema, a njihovo djelovanje prenosi internom strukturom sistema. Ocjena CO_2 efikasnosti ICT sistemi i procjena mjera za smanjenje CO_2 otiska kritično zavise od prikupljenih podataka obzirom na veliki elemenata koji su duboko povezani sa socijalno-ekonomskim razvojem i akvizicijom podataka, pri čemu je vrlo teško izvršiti interpolaciju budućih scenarija. U radu je također korištena AHP metoda za višekriterijsko odlučivanje pri donošenju odluke o selekciji i izboru ICT servisa na postavljene kriterije za rješavanje problema smanjenja CO_2 emisija.

I UVOD

Generalna hipoteza istraživanja glasi da procjena ključnih doprinosi informaciono-komunikacionih tehnologija (ICT-a) na smanjenje CO_2 emisija predstavlja mjerljiv postupak u okviru napora uspješne kvantifikacije i ocjene utjecaja ICT-a, a u funkciji smanjenja karbonski emisija. Radi se o resursu velikog potencijala za povećanje energetske efikasnosti i ublažavanje prisutnih klimatskih promjena. Načini na koji ICT tehnologija utječe na smanjenje emisije CO_2 mogu biti iznenađujući, a rezultati značajni.

Da bi mogućnosti ICT-a, na nekom području, bile optimalno iskorištene neophodno je da postoji jasan matematički model procjene doprinosa istih na smanjenja karbonski emisija. Kompleksnost procesa definisanja matematičkog modela proizilazi iz činjenice nedovoljne rasvjetljenosti prirode utjecaja ICT-a na smanjenje emisija, odnosno činjenica da veze između ICT i životne sredine mogu biti direktne i indirektne (→ , ↳), pozitivne i negativne (+,-), jednosmjerne i dvosmjerne (⇄), itd. Upravo zbog toga je i glavni problem i izazov istraživanja da se definiše konkretan model procjene doprinosa ICT-a na smanjenje CO_2 emisija koji će omogućiti da se ovi utjecaji kvantifikuju, prate i kontrolišu te da se na osnovu istih efikasno i svrsishodno projektuju novi sistemi. Činjenica je da do sada nije provedeno istraživanje koje bi dalo prije svega jasnu empirijsku vezu između primjene ICT-a i smanjenja CO_2 emisija, prepoznalo najrizičnije faktore te dalo odgovor na pitanje kako iste uspješno mjeriti. Drugim riječima, razvoj generičkog modela za procjenu efekata postaje nužan uslov, za dalju dinamiku ulaganja u ICT tehnologije sa svrhom ublažavanja klimatskih promjena. Osnovu istraživanja u radu, predstavlja identifikacija, analiza i procjena ključnih matematičkih varijabli utjecaja ICT-a na smanjenje CO_2 emisija te analiza, odnosno analiza iz ugla ICT proizvoda, projekata i organizacija a u funkciji dizajniranja modela procjene. Podršku procesu dalju i različita tijela koja su udružili napore u pronalaženju adekvatne, ali prije svega standardizovane procedure za procjenu (kvantifikaciju, praćenje i izvještavanje) emisija štetnih polutanata za koje je odgovoran sektor ICT-a [1]. Proces matematičke formalizacije doprinosa ICT-a u osnovi se temelji na vrednovanju emisija pokretanja bita (informacija) umjesto pokretanja atoma (materije) referentnog sistema, što je bit procesa dematerijalizacije. Telekomunikacijski sistem, kao sistem za upravljivu distribuciju bita, nudi ključnu alternativu i pametnu nadgradnju različitih sistema sa značajno manjim nivoima emisije. Osnovni izazov leži o dualnoj prirodi ovog odnosa jer ICT-a sa jedne strane doprinosi smanjenju negativnih ekoloških utjecaja (kroz povećanje efikasnosti sistema, supstituciju prijevoza, itd), ali sa druge strane i sam sektor ICT-a je generator negativnih ekoloških utjecaja, prije svega emisija CO_2 [1].

Procjena ICT-a na smanjenja CO_2 emisija veoma je kompleksan proces i zahtjeva definisanje velikog broja varijabli. Sa jedne strane ovdje se radi o istraživanju dva fenomena, jednog tehničkog karaktera, koji ima jasno definisane i mjerljive varijable, a sa druge strane imamo prirodne fenome utjecaja koji imaju i određeni broj nezavisnih varijabli. Nezavisne varijable koje je veoma teško mjeriti, uključuju i subjektivni utjecaj koji se ne smije zanemariti, ali se ne može ni strogo uzročno determinirati i opisati. Istražuje se utjecaj upotrebe ICT-a kroz životni ciklus na smanjenje emisija CO_2 izvora i pri tome se definira kao pozitivan ako je dobivena vrijednost veća od vrijednosti utjecaja ICT-a, troškovi su manji ili jednaki.

U cilju definisanja modela mjerenja utjecaja, neophodno je izvršiti sistematizaciju utjecaja ICT-a na CO_2 emisije, odnosno sistematizovati pravce primjene ICT-a i sve relevantne faktore koji utječu na efekte primjene ICT-a.

II ZNAČAJ INFORMACIONO - KOMUNIKACIONIH TEHNOLOGIJA ZA SMANJENJE KARBONSKIH EMISIJA

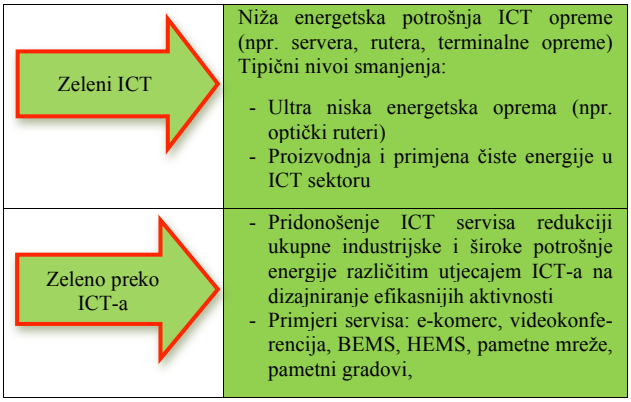
Donošenjem ispravnih odluka o načinima proizvodnje, kupnje i primjene proizvoda ICT-a imamo sposobnost osigurati goleme ekološke koristi. To može obuhvatati štednju energije kroz nadziranje i upravljanje njenom upotrebom,

stvaranje efikasnijih transportnih sistema te smanjivanje potreba za putovanjima primjenom video servisa i drugih oblika elektronskih transakcija. U tom kontekstu ICT tehnologije treba posmatrati s tri aspekta: društvenog i ekološkog utjecaja proizvodnje ICT proizvoda i kako se može smanjiti negativan utjecaj; utjecaja korištenja ICT-a proizvoda na održivi razvoj, posebno njegovu ulogu u proizvodnji energije i klimatskim promjenama i načina na koji se ICT proizvodi mogu primijeniti za poboljšanje životnih uslova. Prema [2], ICT se sastoji od hardvera, softvera, mreža i medija za prikupljanje, pohranjivanje, procesiranje, prenošenje i prezentaciju informacija (glasovnih, podatkovnih, tekstualnih i slikovnih), kao i od srodnih usluga. ICT se, prema istom izvoru, može podijeliti na informacijsku i komunikacijsku infrastrukturu (ICI) i na informacjisku tehnologiju (IT). ICI se odnosi na hardver i softver za prikupljanje, pohranjivanje, procesuiranje, prenošenje i prezentaciju informacija, a IT na telekomunikacijske sisteme i mreže, te na servise koji se njima koriste. Prema [6] definirana je djelatnost ICT sektora: ICT sektor obuhvata proizvodnju opreme i usluga za prikupljanje, prijenos te prikaz podataka i informacija elektronskim putem. ICT sektor su aktivnosti koje spadaju u uniju skupova aktivnosti informacijske tehnologije i telekomunikacija. Aktivnosti koje se nalaze u sektoru informiranja, a koje nisu u presjeku skupova s aktivnostima IT sektora i telekomunikacija, ne ubrajaju se u ICT sektor. Prema [2] postoje dvije ICT sektorske kategorije unutar kojih imamo specifične oblasti (TABLICA 1).

TABLICA 1. KATEGORIZACIJA ICT SEKTORSKE INDUSTRIJE

ICT industrija	OECD broj kategorije	Kategorije
Industrija ICT proizvodnje i trgovine	2610	Proizvodnja elektronskih komponenti i ploča
	2620	Proizvodnja računara i periferne oprem.
	2630	Proizvodnja komunikacijske opreme
	2640	Proizvodnja potrošačke elektronike
	2680	Proizvodnja magnetnih i optičkih medij
	4651	Veleprodaja kompjutera, periferne opreme i softvera
	4652	Veleprodaja elektronske i tk. opreme
Industrija ICT servisa	5820	Objavljivanje softvera
	6110	Bežične telekomunikacijske aktivnosti
	6120	Žičane telekomunikacijske aktivnosti
	6130	Satelitske telekomunikacijske aktivnosti
	6190	Ostale telekomunikacijske aktivnosti
	6201	Aktivnosti računarskog programiranja
	6202	Računarsko savjetovanje i upravljanje računarskom opremom
	6209	Aktivnosti IT-a i računarskih servisa
	6311	Obrada podataka, hosting i vezani servisi
	6312	Web portal
	9511	Popravak računarske i periferne opreme
	9512	Popravak komunikacijske opreme

Pri korištenju ICT-a, postaje sve značajnije razmatranje i procjenjivanje, ne samo mjerljivih i direktnih ekoloških utjecaja ICT sektora, nego i redukcijskih ekoloških koristi koje primjenom ICT-a možemo imati u smanjenju CO_2 emisija. U tom kontekstu, saradnja i optimizacija putem više ekonomskih sektora je značajna prilika. Termin „zeleno“ se danas često koristi za sistematično objašnjenje efekata primjene ICT-a. „Zeleno preko ICT-a“ se odnosi na dekarbonizaciju drugih sektora implementacijom ICT rješenja, dok „Zeleni ICT“ ukazuje na smanjenje karbonskog intenziteta samog ICT sektora (slika 1.). Efekti smanjenja CO_2 emisija putem „Zeleno preko ICT-a“ prema istraživanjima iz [4] mogu biti ekvivalentni sniženju od pet i više puta karbonskih emisija samog ICT sektora, uz pretpostavku da su ICT potencijali aktivno korišteni.



Sl. 1. Konceptualne razlike između “Zeleni ICT” i “Zeleno preko ICT-a”

III IDENTIFIKACIJA POZITIVNIH I NEGATIVNIH EFEKATA PRIMJENE ICT-a

ICT-a ima i pozitivne i negativne mjerljive efekte na okoliš (Sl. 2.) Negativan efekat je moguće izračunati pomoću mjerljivog ekološkog opterećenja uzrokovanog ICT uređajima i opremom, kao na primjer vrijednost potrošnje energije, prirodnih resursa i količina stvorenog ICT otpada.

Procesno stanje pozitivnog utjecaja, ICT usluge mogu napraviti kroz energetski efikasniji stil života i način poslovanja, jer npr. omogućavaju dematerijalizaciju, odnosno prijenos informacije umjesto fizičkih objekata, reduciraju kretanja ljudi i efikasniji transporta roba. U cilju balansiranog razvoja sa ekološkog stanovišta, neophodno je pristupiti i minimiziranju negativnih faktore ICT te maksimizirati pozitivne efekata. Iz tog razloga, i pozitivni i negativni ekološki efekti ICT trebaju biti kvantificirani, te ako je moguće, izvršiti procjenu u širem kontekstu.

Početni način kvantifikacije ova dva efekta je predložen od strane brojnih organizacija. Tako npr. prema [5] metodologija za procjenu ekoloških utjecaja ICT-a, razmatra utjecaje:

- ICT proizvoda, mreža, i servisa,
- ICT projekta,
- primjene ICT-a u poslovnim organizacijama,
- primjene ICT-a u gradovima (uključuje i druga administrativna područja koja mogu biti pokrivena sa procjenom ekoloških utjecaja u gradovima),
- primjene ICT-a na razini država.

Vrlo je važno na osnovu prethodnog da procjena utjecaja ICT-a nužno uključuje definiciju granica procjene, pri čemu srednje vrijednosti rezultata procjena ne bi smjeli značajno varirati ovisno o izboru granica. U svim slučajevima, definisane granice procjene trebaju biti adekvatno dokumentirane.

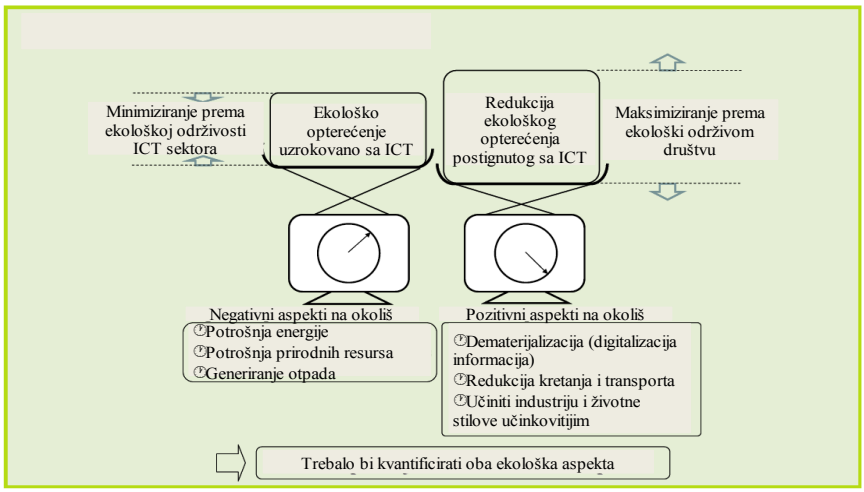
Dakle sistemski posmatrano ICT sektor ima dvostruku prirodu s obzirom na CO_2 emisije, odnosno postoje dva reda efekata s obzirom na ekološke utjecaje ICT-a:

- Efekti prvog reda (ili ekološko opterećenje ICT-a): utjecaji kreirani putem fizičkog postojanja ICT-a i uključenih procesa npr. CO_2 emisije, e-otpad, upotreba opasnih i rijetkih materija, ne-obnovljivih resursa, koji zajedno predstavljaju ekološko opterećenje vezano za ICT proizvode, mreže i usluge u toku njihovog životnog ciklusa. Dakle direktni utjecaji se odnose na utjecaje koje proizvođači i krajnji proizvodi industrije ICT-

a zajedno sa korisnicima tih proizvoda imaju na životnu sredinu. Glavni nosilac ovih utjecaja je ICT tehnologija.

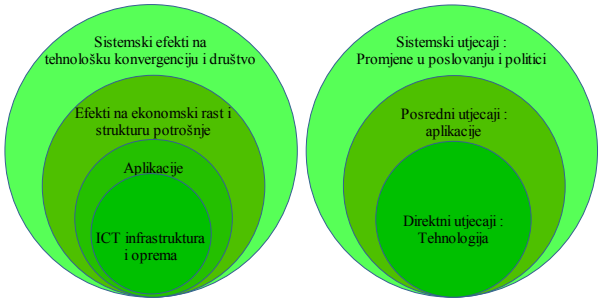
- Efekti drugog reda (ili redukcija ekološkog opterećenja): utjecaji i mogućnosti stvorene primjenom ICT-a. Ovo uključuje redukciju efekata ekološkog opterećenja koji mogu biti ili stvarni ili potencijalni, npr. substitucija putovanja, optimizacija saobraćaja, promjena radnog okruženja, upotreba ekoloških kontrolnih sistema, upotrebe e-poslovanja, e-vlada, itd.. Važno je spomenuti "Indirektne efekte" koji predstavlja mogućnost redukcije emisije stakleničkih plinova i to u pet velikih područja: dematerijalizacija, pametni motorni sistemi, logistika i transport, zgrade i pametne mreže. Indirektni utjecaji odnose se na rješenja ICT-a koje se koriste u ekonomskim i društvenim aktivnostima, a koje smanjuju utjecaj tih aktivnosti na životnu sredinu. U osnovi ovog utjecaja je činjenica da ICT učestvuju u proizvodnji i potrošnji većine proizvoda. Iako je ovaj utjecaj u osnovi pozitivan, jer omogućava racionalnije korištenje resursa, njegovu procjenu treba sprovoditi oprezno. Razlog je što često sistemi bazirani na ICT podrazumijevaju veću potrošnju energije (iako manju potrošnju radne snage) u odnosu na konven-cionalne (neautomatizovane) sisteme.

Pozitivni i negativni utjecaji ICT-a shematski opisuju sadržaj ekoloških utjecaja ICT-a na okoliš (Sl. 2.). Efekti prvog reda su direktno vezani za negativne efekte ICT-a na okoliš, a efekti drugog reda su poistovjećeni s pozitivnim efektom ICT-a. Za sve ove efekte, ključni indikatori ekoloških performansi trebaju biti korišteni.

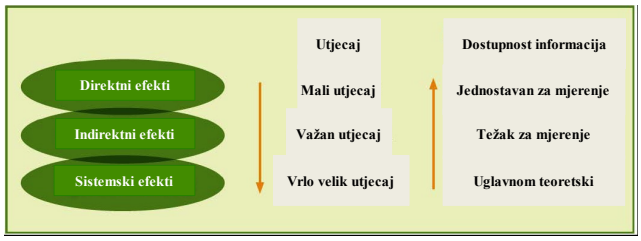


Slika 2. Pozitivni i negativni utjecaji ICT-a

Pored gornja dva tipa utjecaja ICT na životnu sredinu primjetno je postojanje i sistemskih utjecaja" (Sl. 3). Sistemski utjecaji ili efekti trećeg reda se odnose na faktore koji nisu tehnološki već organizacioni i upravljački. Glavni utjecaj ICT u ovom kontekstu je u širenju informacija koji će uticati da proizvođači i korisnici svoje ponašanje promjene u pravcu boljeg odnosa prema životnoj sredini. Međutim, i u ovom domenu treba biti oprezan pri procjeni utjecaja. Promjene na mikro nivou često ne rezultuju proporcionalnim koristima na makro nivou. Veća energetska efikasnost nekog proizvoda (npr. automobila) može smanjiti potrošnju energije za pojedinačnog korisnika, ali istovremeno zbog dostupnosti za širi krug potrošača, do-vesti do veće ukupne potrošnje na nivou regiona ili države [5].



Definisana su tri pravca djelovanja ICT-a na ekonomske procese: optimizacija (primjena različitih rješenja za povećanje energetske efikasnosti, na primjer u automobilskoj industriji); dematerijalizacija i supstitucija (za-mjena procesa i proizvoda digitalnim formama - telekonferencija umjesto sastanaka uživo); podsticaj (ICT mogu povećati potražnju za određenim proizvodima) i degrada-cija (ICT rješenja mogu otežati procese upravljanja otpadom).



Slika 3. Tri tipa utjecaja ICT na životnu sredinu

Nekoliko kategorija procjena ekoloških utjecaja može se izvršiti na polju ICT (tablica 2.). Kategorije su vezane za široko ciljani okvir procjene utjecaja. Spomenute kategorije mogu biti poboljšane i dalje razvijane.

TABLICA 2. KATEGORIJE PROCJENE EKOLOŠKIH UTJECAJA

1.	potencijal otopljavanja ili emisija stakleničkih plinova	6.	stvaranje fotohemijskog ozona,
2.	potrošnja energije	7.	acidifikacije zraka
3.	crpljenje resursa	8.	eutrophikacija vode
4.	iscrpljenje ozona	9.	emisije otrovne pare
5.	emisije otrovnih supstanci	10.	stvaranje opasnog otpada

Većina analiziranih metodologija se fokusira na energetsku potrošnju i emisije stakleničkih plinova. Selekcija je izvršena u kontekstu dva ključa područja: maksimalna proizvodnja nafte i klimatske promjene usljed ljudskih aktivnosti.

Korištenje ICT-a rješenja može pridonijeti također, potrebi nešto više materijala, više komponenti, manje energije i manje otpada. Redukcija ekološkog opterećenja može potjecati od utjecaja ICT na dematerijalizaciju temeljenu na zamjeni visoko karbonskih proizvoda i aktivnosti sa ICT mogućnostima nižih karbonskim alternativama (TABLICA 3), ovo uključuje:

TABLICA 3. ICT ALTERNATIVE

1.	online mediji	4.	e – trgovinu
2.	e – ticketing	5.	e – papir
3.	videokonferencije	6.	telerad ili drugi servisi

Efikasnost dematerijalizacije u smanjenju CO_2 emisija zavisi o behaviorističkim promjenama i dostupnosti infrastrukture. U tom kontekstu, moguće je imati efekat povratnog djelovanja (povećanom upotrebom opreme ICT-a). Treba uočiti da nagomilavanje efekata velikog broja ljudi koji koriste ICT, pravi složeniju procjenu ICT utjecaja u srednjeročnom roku. Primjetno je da analizirane mogućnosti redukcije emisije, u drugim sektorima putem upotrebe ICT-a, ima značajne povratne efekte. Upotreba ICT-a može povećati efikasnost potrošnje energije, može povećati učinkovitost i smanjenu proizvodnju i potrošnju resursa, te smanjiti kretanje ljudi i roba. Tablica 4. prikazuje listu kategorija pri čemu metrike redukcije ekološkog opterećenja se razlikuju po kategorijama ekoloških utjecaja. Spomenute vrste kategorizacija utjecaja mogu biti procijenjene na različitim poljima ICT-a:

- utjecaj ICT proizvoda,
- utjecaj ICT mreža,
- utjecaj ICT usluga.

Za proizvode, mreže i usluge funkcionalne jedinice mogu biti drugačije analizirane: na primjer, vrijeme tokom kojeg jedan korisnik telefonira, broj poslanih e-mail-ova, količina podataka prenesena preko mreže, vrijeme tokom kojeg jedan korisnik koristi videokonferencijski servis itd. Procjena utjecaja ICT proizvoda, mreža, i usluga zahtijeva definiciju granica, pri čemu granice trebaju biti dokumentirane. Procjena može biti ograničena na jedan dio životnog ciklusa komponente ili servisa ili pokriti ukupan životni ciklus, uključujući, naprimjer, ekstrakciju sirovina, proizvodnju komponenti, skladištenje, transport i distribucija, upotrebu, recikliranje i upravljanje otpadom. Za utjecaje proizvoda, mreža i usluga, odgovornost za utjecaje je raširena preko cijelog nabavnog lanca. Vremensko ograničenje je sintetičko, u smislu da postoji nagomilavanja procesa procjenjivanja utjecaja tokom dužeg vremena i prostora. Procjenjivanje može varirati tokom vremena ovisno o variranju samih procesa.

Za utjecaje proizvoda, mreža i usluga, također je potrebno izvršiti procjene životnog ciklusa (LCA) npr. Primjenom ISO 14040 standarda koji opisuje principe i radni okvir za LCA i ISO 14044 koji specificira zahtjeve i pruža smjernice za LCA uključujući: definiciju cilja i opsega LCA, fazu analize životnog ciklusa inventara (LCI), fazu procjene utjecaja životnog ciklusa (LCIA), fazu interpretacije životnog ciklusa, izvještavanje i kritičan pregled LCA, ograničenja LCA, odnos između LCA faza, i uslova za upotrebu izabranih vrijednosti i alternativnih elemenata. ICT proizvodi, mreže i usluge svojim razvojem pružaju mogućnosti za:

- poboljšanje energetske efikasnosti,
- redukciju u potrošnji energije,
- redukciju u emisijama kroz cjeli životni ciklus proizvoda i opreme,
- održivost materijala i eliminiranje opasnih materijala (uključujući način pakiranja),
- spajanje nižih utjecaja pojedinih komponenti (multifunkcionalni proizvodi).

ICT proizvodi, mreže i usluge mogle bi imati značajne mogućnosti ublažavanja CO_2 emisija korištenjem istih u projektima smanjenja velikih GHG emisija drugih sektora:

- industrija: motori i industrijska procesna automatizacija,
- uslužnu djelatnost: veća efikasnost,
- transport: logistika, privatni prijevoz, efikasna vozila, saobraćajni tok, pametna rasvjeta,
- zgrade: pametne zgrade, redukcija prostora i energije,
- energija i toplina: pametna mreža, efikasna proizvodnja struje, nadgledanje i kontrola potrošnje energije, kombinirano nadgledanje i upravljanje potrošnjom topline i energije.

Na polju ICT-a, postoji naročito zanimanje za uspoređivanje situacije s ICT servisom sa njegovim temeljnim baznim scenarijem. Uspoređivanje ekološkog utjecaja dvo-satnog sastanka putujući zrakoplovom između dva grada s ekološkim utjcajem koje ima dvo-satni sastanak ostvaren putem videokonferencijskog servisa. Usporedba utjecaja između dvije situacije implicira procjenu ekoloških utjecaja u drugim sektorima i potrebu integriranja rezultata.

Procjene utjecaja na organizacijskom nivou zahtijevaju definisanje granica tretirane organizacije, pri čemu granice trebaju biti dokumentirane u različitim vremenskim okvirima razmatranja, osim ako je drugačije naznačeno, pri čemu je vremenski okvir od jedne godine obično optimalno rješenje. Ovaj tip procjene može biti izvršen retroaktivno nakon proteklog izabranog vremenskog okvira. Izabrani podaci korišteni za izvođenje procjene trebaju biti dokumentirani i raspoloživi za verifikaciju, pri čemu trebaju biti korišteni oni podaci koji imaju najbolju raspoloživu kvalitetu. Ovaj tip procjene može biti verificiran od treće strane nakon završetka procjene.

Procjena utjecaja ICT na projektnom nivou se sastoji od procjene jednog ili nekoliko tipova ekoloških utjecaja povezanih s lansiranjem projekta ciljajući na reduciranje ovih utjecaja, npr. emisije stakleničkih plinova. Potvrđivanje procjena utjecaja se sastoji od procjene utjecaja projekta prije samog početka do datuma kada je procjena izvršena. Jedan od ciljeva ove vrste procjene je predvidjeti potencijalne utjecaje projekta, identificirati aktivnosti unutar projekta odakle dolaze glavni utjecaji i odlučivanju kako smanjiti neželjene utjecaje a povećati željene utjecaje. Potvrdna procjene utjecaja leži u procjeni utjecaja projekta, ili dijela projekta, koji je završio na datum kada je procjena izvršena. To može izvršiti treća strana. Selektovani podaci korišteni za izvođenje procjene trebaju biti dokumentirani i raspoloživi za ovjeru, pri čemu podaci najbolje raspoložive kvalitete trebaju biti izabrani.

Utjecaji ICT-a na nivou grada ili drugom administrativnom nivou mogu imati važnu ulogu u razvoju zelenog prostora. Naprimjer, pametne zgrade i pametne mreže mogu stvoriti veću energetsku efikasnost. Inteligentni transportni sistemi uključujući električna vozila, luke i aerodromsku infrastrukturu čine transport više energetski efikasnim. Multimedijski konferencijski sistemi i

telerad mogu smanjiti obim putovanja i njegov karbonski otisak. ICT može također imati važnu ulogu u dizajniranju energetski efikasne javne infrastrukture. Procjena utjecaja ICT-a u gradu ili bilo kojoj drugoj administrativnom nivou zahtijeva definiciju granice grada ili neke druge administrativne razina pri čemu granice trebaju biti dokumentirane. Procjena utjecaja ICT-a na nivou pojedinačne zemlje ili grupe zemalja, ili regija unutar zemlje, ili nekom drugom administrativnom nivou, zahtjeva definiciju regionalne, državne, međunarodne ili neke druge razine administrativne granice pri čemu granice trebaju biti dokumentirane. Izabrani podaci korišteni za izvođenje procjene trebaju biti dokumentirani i raspoloživi za provjeru. Podaci najbolje raspoložive kvalitete trebaju biti izabrani. Različiti vremenski okviri mogu biti razmotreni, okvir od jedne godine je optimalno rješenje.

TABLICA 4. EFEKTI ICT-a

Kategorija	Efekti
Potrošnja materijala	Kod smanjenja materijalne potrošnje ekološka opterećenja vezana za proizvodnju i upotrebu kao i otpad mogu se značajno reducirati.
Potrošnja energije /upotreba energije	Uz povećanje učinkovitosti snage i utroška energije uz smanjenje potrošnje, ekološka opterećenja vezana za proizvodnju i prijenos energije, itd., mogu biti redukovana.
Kretanje ljudi	Kod reduciranja kretanja ljudi, ekološko opterećenje transporta može se reducirati.
Transport materijala	Kod reduciranja transporta materijala, ekološko opterećenje transporta može se reducirati.
Poboljšanje efikasnosti ureda	Kod efikasnog korištenja uredskog prostora, potrošnja struje za rasvjetu, klimatizaciju, itd., može biti redukovana, a time i okolinsko opterećenje.
Skladištenje robe	Kod smanjenja robnog skladišnog prostora, potrošnja struje za rasvjetu, klimatizaciju, itd. može biti redukovana, a time i ekološko opterećenje.
Bolja radna efikasnost	Kod povećanja efikasnosti rada, ekološko opterećenje može se redukovati
Otpad	Kod smanjenja otpadnih emisija, ekološka opterećenost zahtijevana za ekološko očuvanje kao i za odlagalište smeća može se reducirati.

TABLICA 5. OČEKIVANE EMISIONE REDUKCIJE ILI RAST PUTEM ICT-A U IZABRANIM PODRUČJIMA

Područja redukcije CO_2	Predviđeni utjecaji (primjeri)
Zgrade	Manje prostore za skladišta, npr. trgovine na malo zamijenjene sa online kupovinom. Veća energetska efikasnost sa naprednim rashladnim i toplotnim sistemima.
Industrijska proizvodnja	Proizvodnja nekoliko proizvoda i roba putem dematerijalizacije. Niža energetska potrošnja po jedinici proizvoda kroz korištenja ICT-opremeljenih mašina kao što su pametni motori.
Otpad	Manje GHG emisije putem zamjenjivanja kompakt - diskova s muzike putem Interneta.
Sistemi distribucije energije	Smanjene gubitaka snage putem preciziranog zahtjeva i procjenjivanja opskrbe primjenom ICT-a. Manja upotreba fosilnih goriva i povećan udio obnovljive energije omogućeno upotrebom pametnih mreža.
Transport	Manje korištenje vozila zbog teleprisustva i videokonferenciskih sistema Veća efikasnost goriva vozila kroz ubrzanu vožnju predpostavljenu smanjenjem prometnih gužvi. Veća efikasnost goriva korištenjem ICT-a pametnih transportnih sistema, na primjer, autobuski informacijski sistemi i realna vremenska navigacija.
Ostalo	Više slobodnog vremena za ljude, koje može biti potrošeno za povećanje GHG emisija drugih aktivnosti (efekat povratne sprege).

III PROBLEMI PROCJENE UTJECAJA ICT-A NA SMANJENJE KARBONSKIH EMISIJA

Većina analizirane literatura objavljene nakon 2000. godine o utjecaju ICT-a na okolinu i klimatska promjene fokusira se na podizanje svijesti o različitim ICT uslugama sa pozitivnim kratkoročnim ekološkim efektima, ali bez mogućnosti da se potpuno uhvati u koštac sa multi-slojevitim utjecajima koje ICT rješenja imaju na okoliš. Razumijevanje interakcije između emisija i ICT usluga je ograničeno sa nekoliko značajnih poteškoćama.

Pojedina ICT rješenja su raspoređena kao dio višestrukih i među-povezanih procesa. Naprimjer, korisnici pametnih telefona koriste nekoliko usluga istovremeno (skidanja muzike, web-pretraživanje i navigacijske usluge). Ovaj aspekt višestruke upotrebe djeluje kao ograničenje obračunavanju specifične i direktne emisije ICT servisa. Korištenje ICT ima širok i dubok utjecaj na ekonomiju i društvo. Kao utjecaj ICT usluga ne samo da uključuje ekološke već i društvene probleme, pri čemu su karakteristike ICT efekata obično prevladavajuće i transformirajuće. One ne uključuju samo efekte prvog reda koji mogu biti zahvaćeni u kraćem vremenskom razdoblju, već također i efekte drugog reda za duže razdoblje.

ICT usluge uključuju upotrebu više scenarija. Ukupan utjecaj ICT usluga je praćen sa aspekta korisničkog ponašanja, a promjene mogu biti iznenadne odnosno teške za predviđanje i procjenjivanje. Početna istraživanja su pokušala pokriti složenu dinamiku procesa u pokušaju procjene potencijala ICT usluga u redukciji CO_2. Međutim, konkretniji i napredni pristupi su razvijeni samo za pojedinačni ICT servis s minimalnim scenarijem inputa, kao npr. servis mašina-mašina i aplikacije mjerenja.

Ne samo da postoje kompleksnosti procjene ICT usluga ograničenih okvira i već također brojne manjkavosti i nisko povjerenje u ICT-vezanu statistiku. ICT statistika je obično prikupljana za šira geografska područja, kao što je nivo državne ili kontinentalni nivo, fokusirajući se na proizvodnju ICT opreme ili korištenje ICT servisa brojnih korisnika. Zbog toga, ponekad je vrlo izazovno razvrstati ICT utjecaje na raspoloživu ICT statistiku. Iz sadašnjeg ugla posmatranja, najviše dosadašnjih analiza i izvještaja budućih projekcija smanjenja CO_2 emisija je uzeto sa pretpostavkom procjenjenog povećanja penetracije servisa te u smislu poboljšane efikasnosti postignute sa ICT tehnologijama. Obrnuto, historijske analize, koje bi mogle sugerirati koliko su se CO_2 emisije reducirale sa primjenom ICT servisima, još uvijek nisu sprovedene. Većina izvještaja kombinuje top down i bottom up pristup u razvijanju modela i budućih scenarija kako bi savladali ograničenje raspoložive statistike. Osim toga, proces kvantifikacije unosi rizik preklapanja utjecaja kod primjene ICT servisa u procjeni potencijala akumulirane CO_2 redukcije unutar određene geografske granice.

IV MATEMATIČKA FORMALIZACIJA EFEKATA UTJECAJA ICT NA SMANJENJE KARBONSKIH EMISIJA

Na osnovu teorije vidimo da ICT može imati i pozitivan (+) i negativan (-) utjecaj na životnu sredinu, a ključni izazov je da se sagleda i ocjeni krajnji bilans [x+(-y)=z] utjecaja.

Redukcija energetskepotrošnje	=	Efekat redukcije energetske potrošnje primjenom ICT-a	-	Energetska potrošnja kroz primjenu ICT-a

Rezultat komparativne analize između referentnog proizvodnog sistema i ICT sistema mogu se dobiti izračunom razlike u ekološkog utjecaja između referentnog proizvodnog sistema i sistema ICT -a. Razliku možemo nazvati sekundarni efekat, a obrazac glasi:

$$EI_{razlika,i} = EI_{referentni,i} - EI_{ICTservis,i}$$

gdje su:

EI = okolišni utjecaj,
i = i-ta komparativna kategorija,
$EI_{razlika,i}$ *= i-ti sekundarni efekat,*
$EI_{razlika,i}$ *= i-ti EI referentnog proizvodnog sistema,*
$EI_{ICTservis,i}$ *= i-ti EI sistema ICT-a.*

Sumiranjem ukupne $EI_{razlike,i}$ dobivamo $EI_{razliku}$ ili sekundarni efekat sistema ICT-a pokrivanog referentnog sistema. Jednačina ispod prikazuje formulu za izračun sekundarnog efekta.

$$Ukupno\ EI_{razlika} = \sum EI_{razlike,i}$$

Pozitivni rezultat (ukupno $EI_{razlika}$ je pozitivno) indicira pozitivan utjecaj na okoliš a negativna vrijednost (ukupno $EI_{razlika}$ je negativno) predstavlja negativan utjecaj. Pozitivni sekundarni efekti indiciraju redukciju okolišnog opterećenja uslijed uvođenja ICT sistema. Negativni sekundarni efekti indiciraju suprotno. Broj elemenata usporedbi ovisi o diskreciji praktičara a struktura podataka smije varirati između LCA. Podjela na direktne, posredne i systemske utjecaje predstavlja potreban, ali ne i dovoljan okvir za djelovanje u pravcu ostvarenja koristi od ICT i smanjenja negativnih utjecaja tehnologija.

V PREDLOŽENI RADNI OKVIR PROCJENE CO_2 EMISIJA

U nastavku su opisani neki od mogućih ciljeva i kriterija razvijenog višekriterijskog modela odlučivanja o procjeni doprinosa ICT alternativa (servisa i opreme). Analizirajući ICT servise i faktore koji utječu na realizaciju usluga sljedeći ciljevi i kriteriji su prepoznati kao najznačajniji na osnovnoj razini procjene. Radni okvir procjene obuhvata ocjenjivane ciljeve i kriterije procjene.

A) Ciljevi procjene

Prema prethodnoj analizi i tački 2.3.1. identificirana su tri cilja procjene: proizvod, organizacija i projekat. Međutim prema tački 2.3.5. definisana su još dva cilja procjene: gradovi i države.

a) Proizvod

Proizvod se definiše kao bilo koja roba ili servis (ISO 14040). Međutim, prema ITU-T, sa značajnom stručnošću u oblasti telekomunikacijama, "mreže" i "usluge" su identificirane kao dodatne ciljeve procjena. Prema ITU-T definiciji, "proizvod" obuhvata sve robe, mreže i usluge.

b) Organizacija

Organizacija (REF ISO 14064-1) se definira kao kompanija, korporacija, preduzeće, tijelo ili institucija, ili njihove kombinacije, javne ili privatne, koje imaju svoju funkcije ili uprave.

c) Projekat

Projekat je referenciran na GHG projekat koji je definiran ISO 14064-2 standardom kao aktivnost ili aktivnosti sposobne da mijenjaju uslove utvrđene u baznom scenariju koji uzrokuje smanjenje emisija ili poboljšanja uklanjanja GHG-a. Na temelju ove definicije, ITU-T je definirao "ICT projekat" kao GHG projekat korištenja ICT proizvoda, mreža i usluga, u cilju smanjenja CO_2 emisija ili poboljšanja uklanjanje GHG i / ili stakleнički projekat korištenja ICT proizvoda, mreže i usluga, s ciljem uštede energije i energetske efikasnosti.

d) Grad i država

Prema (REF ITU-T L.1440 i L.1450), neophodno je osigurati relevantne metodologije za procjenu utjecaja korištenja ICT proizvoda, mreža i usluga u gradovima ili zemljama, istražujući načine na koji ICT može pomoći u uštedi potrošnje energije i energetske efikasnosti.

B) Kriteriji procjenjivanja

Prethodnom analizom standarda i smjernica, različitih razvojnih organizacija za standardizaciju, identificirani su sljedeći kriterije procjene:

- Inventar: vrsta inventara razvijena za različite GHG izvore i kanale u cilju proračuna ukupne emisije.
- Karbonski otisak: definiše se kao ukupna suma GHG emisija i GHG uklanjanja tokom faza životnog ciklusa proizvoda (izražava se kao neto utjecaj globalnog zatopljenja u CO_2). U definiciji, karbonski otisak se temelji na procjeni životnog ciklusa.
- Procjena životnog vijeka (LCA): Metodologija procjene ulaza, izlaza i potencijalnih utjecaj na okoliš sistema proizvoda tokom cijelog njegovog životnog ciklusa. Procjena životnog ciklusa primjenjena na određeni proizvod obuhvata robu, mreže ili usluge. Ona se bavi svim fazama života proizvoda od ekstrakcije sirovina, obrade materijale, prijevoza u svim fazama, proizvodnje, korištenja, popravake i održavanja te odlaganja ili recikliranja. Prema životnom ciklusu, inventar je sastavljen od svih relevantnih energetskih i materijalnih inputa i izlaza.
- Proračun ICT mogućnosti: proračun za pozitivne utjecaje omogućene primjenom ICT-a i negativne utjecaje uzrokovane korištenjem ICT-a.
- Eko-dizajn proizvoda: zahtjevi i postupci za integraciju ekoloških aspekata u procese dizajna i razvoja proizvoda, uključujući kombinaciju proizvoda, materijala i komponenti.
- Označavanje prema evaluaciji: ovo pokriva sve zahtjeve koji ukazuju na ekološke aspekte proizvoda. Znak zaštite okoliša može poprimiti oblik izjave, simbola, grafičke oznake na proizvodu.
- Provjera i verifikacija: proces s kojim se GHG tvrdnje može ocijenjivati prema usvojenim valjanostima i provjerenim kriterijima.

Svrha upravljanja inventarom procjenom smanjenja emisija je smanjiti vjerovatnoću pojave rizika netačnih procjena i kontrolirati njihove posljedice kako bi se ostvarili ciljevi projekta. Ciljevi projekta su izgradnja sistema procjene i efikasnog ulaganja u ICT tehologije. Funkcija sistema upravljanja inventarom procjenom definirana je opisanom svrhom. Na funkcioniranje sistema procjene utječu razni faktori okoline, koji se dijele na:

- **faktore unutrašnje okoline**: ICT inventar, proces proizvodnje elemenata inventara,
- **faktore vanjske okoline**: legislativni, makroekonomski i sociološki, kvaliteta tima za procjenu, dokumentiranost prethodnih projekata, procedure i upute sektora, i sl.

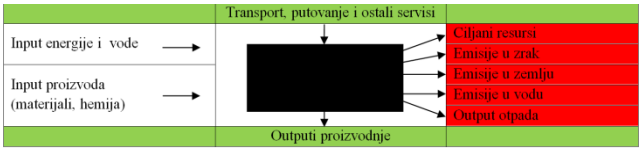
Elementi sistema upravljanja procjenom su:

1. identifikacija utjecaja sektora,
2. identifikacija granice i statističkih podataka,
3. analiza i razvoj dva scenarija za komparaciju,
4. procjena smanjenja CO_2 emisija i rizika,
5. projektovanje smanjenja CO_2 emisije do ciljne godine (kontrola procjene),
6. određivanje ukupnog neto efekta.

Preporuka je da podaci, bez obzira na korišteni model, treba da budu tačni, konzistentni, uporedljivi, jasni, sveobuhvatni, učestalo obnavljani, precizni i raspoloživi.

VI MODELIRANJE PONAŠANJA SISTEMA PROCJENE KARBONSKIH EMISIJA

Sistem upravljanja procjenom smanjenja CO_2 emisija primjenom ICT-a je dinamički sistem, što znači da kod njega tokom vremena dolazi do promjena čiji su uzroci procesi koji se odvijaju u elementima sistema, a njihovo se djelovanje prenosi internom strukturom sistema. Upravljati procjenom sa svrhom ostvarenja ciljeva projekta podrazumijeva provođenje procesa i aktivnosti tokom čitavog životnog ciklusa projekta. Jedinični proces tipično predstavlja opremu gdje je ICT proizvod proizveden, kancelariju ili trgovinu – ili izravnu aktivnost elementa gdje je servis proizveden. Jedinični proces može također biti vozilo ili " pokretna oprema " koja transportira ICT proizvode. Bilo koji jedinični proces može biti modelovan. Generički jedinični procesni model uključuje nekoliko ulaza i izlaza i može biti upućen na LCI modela opreme ili kraće model oprema. U mnogim slučajevima, rukovanje opremom nije samo ciljano na sistem proizvodnje sa LCA već također na druge proizvodne sisteme. Svi drugi ulazi i izlazi su definirani kao tokovi proizvodnje. Svaki input goriva i proizvoda, kao i svaki output otpatka, može uključiti transport koji je zatim dio input/output podataka povezanih na jedinicu procesa.



Slika 4. Inputi i outputi utjecaja ICT-a na životnu sredinu

Za razumijevanje ponašanja ili funkcioniranja sistema upravljanja procjenom, projekat procjene doprinosa ICT-a smanjenju CO_2 emisija moguće je analizirati kroz elemente dijagrama toka: priprema i planiranje projekta procjene u funkciji smanjenja CO_2 emisija. Na slici ispod je za fazu životnog ciklusa projekta procjene prikazan dijagram toka procesa upravljanja procjenom emisija. Pojedini proces ukratko je i verbalno opisan kako bi bilo jasno o kojem elementu sistema upravljanja procjenom se radi te koji se proces odvija unutar pojedinog elementa, u određenoj fazi životnog ciklusa projekta. Za grafički prikaz funkcioniranja sistema uobičajen je sistemski dijagram pod nazivom funkciogram, jer on omogućava izradu dinamičkog modela terminskog ponašanja nekog sistema.

VII MODELI ODLUČIVANJA U SISTEMU PROCJENE DOPRINOSA - MODEL DONOŠENJA ODLUKE O SELEKCIJI I IZBORU ICT SERVISA

U nastavku slijedi model za donošenje odluke o selekciji i izboru ICT servisa na postavljene kriterije za rješavanje problema smanjenja CO_2 emisija. Korišten je program Expert Chioce za AHP višekriterijsko odlučivanje.

A) Opis problema odlučivanja

Donesena je odluka o smanjenju CO_2 emisija primjenom ICT-a na određenom području. Na raspolaganju stoje brojni ICT servisi čije pozitivne i negativne utjecaje na nivo CO_2 emisija nemamo. U literaturi se nalaze brojni kriteriji ali je njihov odnos vrlo kompleksan i dvosmislen, zbog toga se javlja dilema koji ICT servis izabrati za najefikasnije ostvarenje postavljenog cilja.

B) Kriteriji za odabir

Za procjenu smanjenja CO_2 emisija izabrani su servisi koji su najpovoljniji u odnosu na slijedećih šest kriterija:

K1. - mogućnost identifikacije inventara emisija,
K2. - veličina ukupnog karbonskog otiska servisa,
K3. - procjena životnog ciklusa opreme,
K4. - mogućnosti procjene doprinosa ICT-a
K5. - nivo eko-dizajna proizvoda i ICT opreme,
K6. - mogućnosti validacije i verifikacije.
i dva pod kriterija:
K31. -nivo tehnološke zrelosti servisa,
K32. -% prihvačenost servisa od strane korisnika,

Dekompozicija problema - Za rješavanje konkretnog problema definisana je struktura prikazana u tabeli ispod. Glavni cilj je odabrati ICT servis sa visokim pozitivnim efektima na smanjenje CO_2 emisija i sa niskim negativnim efektima. Na nižoj hijerarhijskoj razini odlučivanja nalazi se šest kriterija odabira i dvapodkriterija, a na najnižoj razini su alternative – 19 raspoloživih servisa.

Nakon provedenog istraživanja, odabira relevantnih kriterija i definiranja modela višekriterijskog odlučivanja, podaci su obrađeni u programskom paketu Expert Choise 2000, a cilj je bio definisanje hijararhijskog stabla te odabir alternativa, koje bi prema istraživanjima bile najbolje opcije za smanjenje CO_2 emisija. Vrijednost kriterija i pod-kriterija nije prikazana zbog nekonzistentnih podataka dobivenih provedenom anketom. Nakon definisanja raspona potencijala utjecaja neophodno je izvršiti rangiranje potencijala utjecaja pojedinih servisa ICT-a na smanjenje CO_2 emisija.

TABLICA 6. RASPONI POTENCIJALA UTJECAJA

Opseg utjecaja (u skladu sa Indeksom spremnosti e-okruženja – EERI	**Raspon: -5 do + 5**
Visoko štetni negativni efekti	= -5
Srednje visoko štetni negativni efekti	= -4
Srednji štetni negativni efekti	= -3
Srednje niski štetni negativni efekti	= -2
Niski štetni negativni efekti	= -1
Ograničeni ili nepostojeći efekti	= 0
Niski pozitivni efekti	= 1
Srednje niski pozitivni efekti	= 2
Srednji pozitivni efekti	= 3
Srednje visoki pozitivni efekti	= 4
Visoki pozitivni efekti	= 5

TABLICA 7. SELEKTOVANI ICT SERVISI

Alternative	ICT servisi (po kriteriju smanjenja CO_2 emisija)
A1.	Navigacija u realnom vremenu
A2.	Autobuski informacioni sistem
A3.	E-logistika
A4.	Teleprisustvo i videokonferencije
A5.	Pametni rad
A6.	Sistem upravljanje energijom
A7.	Pametne mreže
A8.	E-zdravstvo
A9.	E-učenje
A10.	E-trgovina
A11.	E-vlada
A12.	E-civilni servisi
A13.	Digitalni sadržaj
A14.	Pametni motori u industrijskim procesima

Od odabranih 19 servisa, rangiranjem potencijala u odnosu na efekte drugog reda u BAU sistemima, odnosno na temelju vrijednosti kriterija smanjenja emisija, izdvojeno je 14 ICT servisa prikazanih u Tablici 7.

VIII Model donošenja odluke o području primjene alternativa ICT servisa

Područja primjene alternativa ICT servisa u funkciji cilja su vrednovana na osnovu dostupnih podataka sa visokom frekvencijom učestalosti servisa u području i kriterijem smanjenja emisija BAU sistema. Ključni indikatori koji su vrednovani prikazani su u tablici 8.

TABLICA 8. OBRASCI PRORAČUNA SMANJENJA ENERGETSKE POTROŠNJE

Ciljna područja ICT servisa	Kriteriji energetske redukcije
Potrošnja sredstava (papir, CD, DVD, itd)	Potrošnja energije za proizvodnju jedne jedinice proizvoda × umanjena veličina
Potrošnja energije/energetska potrošnja (gorivo, struja itd.)	Potrošnja energije po količini goriva × smanjena potrošnja goriva
Kretanje ljudi (automobilom, autobusima, itd.)	Potrošnja energije po količini goriva × smanjena potrošnja goriva
Kretanje roba (kargo itd)	Potrošnja energije po količini goriva × smanjena potrošnja goriva
Poboljšana efikasnost uredskog prostora	Potrošnja energije po površini ureda × smanjen iznos
Skladištenje roba (skladišni prostor, itd.)	Potrošnja energije po skladišnom prostoru × smanjen iznos
Poboljšana radna efikasnost	Potrošnja energije po površini (m^2) x površina korištena po osobi (m^2) x poboljšana radna efikasnost
Otpad (papir, industrijski otpad)	Potrošnja energije po količini otpada × umanjena količina

Rezultati softverske analize prikazani su su Tablici 9.

TABLICA 9. REZULTATI PROCJENE ALTERNATIVA PO SEKTORIMA

Područja smanjenja emisija	ICT alter	Potencijalno zahvaćeni sektori ICT alternativama		
		Transport	Gradnja	Industrija
Transport	A1.	Redukovana potrošnja goriva	Nije primjenjiv	Nije primjenjiv
Transport	A2.	Redukovana potrošnja goriva	Nije primjenjiv	Nije primjenjiv
Transport	A3.	Redukovana potrošnja goriva	Nije primjenjiv	Nije primjenjiv
Transport	A4.	Redukovana potrošnja goriva	Nije primjenjiv	Nije primjenjiv
Transport	A5.	Redukovana potrošnja goriva	Redukcija energetske potrošnje	Nije primjenjiv
Pametna gradnja	A6.	Nije primjenjiv	Redukcija potrošnje ener.	Nije primjenjiv
Pametna gradnja	A7.	Nije primjenjiv	Redukcija potrošnje ener.	Redukcija potrošnje ener.
Trgovina/ usluge	A8.	Redukcija potrošnje goriva	Redukcija potrošnje ener.	Nije primjenjiv
Trgovina/ usluge	A9.	Redukcija potrošnje goriva	Redukcija potrošnje ener.	Nije primjenjiv
Trgovina/ usluge	A10.	Redukcija potrošnje goriva	Redukcija potrošnje ener.	Redukcija potrošnje ener.
Trgovina/ usluge	A11.	Redukcija potrošnje goriva	Redukcija potrošnje ener.	Nije primjenjiv
Trgovina/ usluge	A12.	Redukcija potrošnje goriva	Nije primjenjiv	Nije primjenjiv
Trgovina/ usluge	A13.	Nije primjenjiv	Nije primjenjiv	Redukcija potrošnje ener.
Proizvodnja	A14.	Nije primjenjiv	Nije primjenjiv	Redukcija potrošnje ener.
Razine utjecaja		Visoki efekti	Negativni efekti	Ograničeni efekti

VI Zaključak

Rezultati istraživanja procjene doprinosa ICT-a na smanjenje CO_2 emisije pokazuju da dosadašnja metodologija procjene, kao skup parametara ili vrijednosti izvedenih na osnovu pozitivnih efekata, djelomično opisuje procesna stanje, te ne uključuje sistemske utjecaje koji prevazilaze domet same vrijednosti parametara karbonskog utjecaja. Analiza je pokazala da osim nepostojanja skupa relevantnih indikatora, dodatni problem kreira nedostatak modela prikupljanja podataka. Indikatori predloženi u literaturi mogu biti samo inicijalni indikatori procjene utjecaja. Za obuhvatniji i detaljniji pristup ovoj problematici neophodna su šira istraživanja i dostupnost terenskih podataka. Složenost procedura procjene i zahtijevani nivo informacione osnove čine da je posvećenost redovnoj ocjeni ekoloških utjecaja još uvijek samo dio pojedinačnih projekata. Kao i u svakom naučnom radu, u okvirima vremenskih i prostornih ograničenja, i ovdje se otvorilo niz dodatnih pitanja koje bi trebalo dalje obrađivati, kao što je vrednovanje i rangiranje efekata, odabir metode procjene efekata te definiranje modela prikupljanja podataka. Razvoj modela procjene doprinosa ICT-a na smanjenje CO_2 emisija nedvojbeno predstavlja ključni faktor uspješnog mjerenja pozitivnih doprinosa tehnologije. Bitno je naglasiti da kao što sistem nije numerički zbir elemenata i podsistema već njihov funkcionalni zbir [f(a,b,c)> >f(a)+f(b)+f(c)], tako i ovaj model nije samo numerički zbir parametara ICT komponenti i informacija nego uključuje i kompleksne relacije sa ostalim kompo-nentama i okolinom.

LITERATURA

[1] Berković, M., Model procjene doprinosa ICT-a na smanjenje CO_2 emisija sa osvrtom na područje BiH, Magistarski rad, Fakultet za saobraćaj i komunikacije, 2005.

[2] GeSI, Evaluating the carbon-reducing impacts of ICT – An assessment methodology, Global e-Sustainability Initiative Report, February 2011.

[3] Guidelines for the development of sustainability indicators, Michael Ditor, August 2001.,

[4] P. Nelson, P, P. Safirova, and M. Wells (2007). "Telecommuting and environmental policy: Lessons from the ecommute program. Transportation Research D 12, pp. 195-207.

[5] Bojkovic, N. i M. Petrovic, „Mogućnosti evaluacije doprinosa Informaciono – komunikacionih tehnologija smanjenju karbonskih emisija," Zbornik radova „Postel 2011", Beograd, pp. 78-86, 2011.

[6] REACT: Izveštaj o razvoju opšteg skupa indikatora za karbonski uticaj, Završni izvod iz Izveštaja o razvoju opšteg skupa indikatora za karbonski uticaj, D.3.3 iz EU FP7 REACT projekta

TELECOM INFRASTRUCTURE

A Low Energy Computer Infrastructure for Radio VOIP Supported Communication and SDR APRS in Education and Disaster Relief Situations

Kemal Hajdarevic*, Samim Konjicija**
*,** Faculty of Electrical Engineering, University of Sarajevo,
Sarajevo, Bosnia and Herzegovina
*khajdarevic@etf.unsa.ba, **skonjicija@etf.unsa.ba

Abstract – In long duration emergency and large scale disaster situations, information on asset position, information of current and information history of weather conditions, and communication are crucial for every rescue operation. Picture and voice are crucial so that affected population can make important decisions and take actions. Available technology offers affordable (low cost) and robust infrastructure solutions in disaster relief situations supported by radio amateur communication. In this paper we presented three types of services for which same hardware in essence can be used. First service is for GPS location mapping, second one is for VOIP communication, and third one is to provide real time video capturing. GPS location mapping was performed by using APRS-IS SDR, VOIP communication is performed via radio-svxlink-Internet using Echolink (svxlink which is linux compatible version) software, and for video capturing was used using Motion software. All above stated services is possible to install on low energy microcomputer such as Raspberry Pi which can be powered with 5 V battery or with mobile phone power adapter.

I. INTRODUCTION

In any emergency situation information on asset position, field and weather conditions, and communication are crucial for every rescue operation. We are all aware that these information are available to narrow specialized units such as military and rescue forces. These specialized units in critical situations might be overwhelmed with search & rescue activities and numerous requests for saving lives, assets, it is most likely that they will not be able to promptly give information such as specific person location, field and weather conditions, or to communicate with all affected individuals. Thus for our intention for documenting our findings is to propose solutions which can provide useful visual information, support alternative communication channels, and location mapping with support of radio amateur communications and radio amateur operators (hams).

II. RATIONALE FOR RADIO AMATEUR FREQUENCY SPECTRUM AND APPLICATIONS USAGE

Our intention is to propose compact housing for small microcomputer and other components which needs Internet connection able to communicate versatile type (video, voice, and data) of communication. Innovations in radio communications has great history and in last two decades radio spectrum became available to broader usage and in use by large part of global population through a usage of WIFI networks. While WIFI spectrum (2.4 and 5 GHz) as radio communication became available to broader usage this is just small part of radio spectrum which is available for testing and research activities. Because of its broad usage, sometimes researchers can find mentioned spectrum overwhelmed with traffic that different users can produce in the same time on WIFI networks. From the past, and lessons learned we may conclude that radio spectrum allocated for radio amateur spectrum could be used when all other means of traditional communication fails. Services that radio amateurs or Hams provide in every day and especially in critical situations [1] is crucial for critical infrastructure which is important for:

"health, safety, security or economic well-being of citizens or the effective functioning of governments" stated by [1].

This is because they provide information and technical support with communication services usually over HF radio spectrum and other services such as services that we presented in this paper: APRS-IS [2], Echolink / Svxlink [3], in cases of emergency when other means of communications does not work.

A. Rationale for position tracking using APRS-IS

Automatic Packet Reporting System (APRS) is real time radio amateur service able to transmit position reports, weather reports, and messages between users which are processed and visualized. Position reports supported by APRS or more precisely APRS-Internet Service (APRS-IS) use GPS data. One of Internet places which visualize APRS data is aprs.fi web site which updates data in real time as they are sent over Internet via nodes which are able to communicate in the same time with Internet and radio amateur nodes and networks. Because APRS position data are logged and visualized in real time APRS-IS can be used to support community in cases of emergencies [4, 5, and 6].

Using APRS-IS it is possible to achieve multiple goals [7]:

"Accurate position information about mobile & fixed stations. Tracking a mobile station. Monitoring of 'radio to radio digital text messages' on an Internet browser. Non-ham family members of a ham can track him/her and read

messages. Monitoring of real-time weather information originated from ham radio weather stations like Wind Speed, Wind Direction, Temperature, Barometric Pressure, Humidity, Rainfall etc. are available on this map with just the click of a button. Plotting of objects on the map. For example an accident site, crash site, intimation about a traffic jam for route planning, fire site, flood, cyclone, land-slide, road blockage etc. Allows important information to be exchanged without human intervention"

To use radio frequency (RF) and APRS-IS and VOIP Svxlink for testing and educational purposes, individuals and organizations have to have radio license within national associations. Each licensed radio amateur must have assigned call sign number to participate in APRS-IS. One of reasons for experimenting with APRS-IS is trend of new ham licensees issued in United States [8] and ARRL [9]. For the interest growth of RF usage, there could be reasons such as restructuring, new frequency band plan [8], [9], but available and low cost technology play important role as well. In this paper we presented low cost and available technology for sustainable APRS exploitation and infrastructure scalable for future developments. All mentioned could be reasons that trend shown in Figure 1, will be followed in other parts of globe.

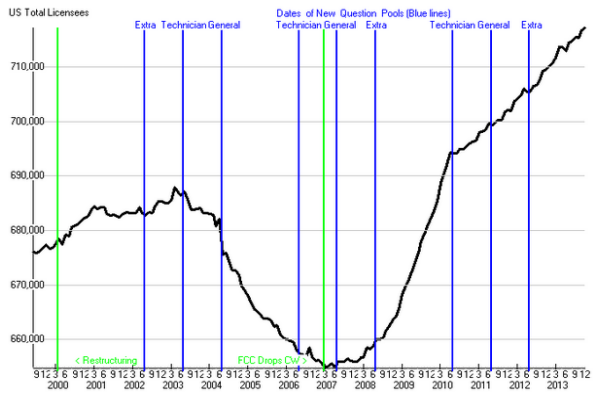


Figure 1. Steady and constant growth of new ham licensees in the US [8], [9]

Inventor of APRS is Bob Bruninga [10] in 1980's, and this system in that time was called Connectionless Emergency Traffic System (CETS) and was developed for National Disaster Medical System [11]. In the 1990's there were no available geographical maps like today's google maps, and maps have to be drawn manually. These maps were able to support smaller geographical areas compared with today's available mapping mechanisms which can span whole world. Today's technology allows usage of APRS-IS to transfer locally collected data, or local data collection point via TCP/IP using UDP, TCP, or HTTP directly to APRS-IS, or indirectly via RF and APRS Internet gateways [12] which is our choice since it eliminates need of operational GSM / GPRS or other network that would support TCP/IP communication between collection point and IGates [13]. An IGate [13] is a gateway APRS station. The IGate's task is to allow packets to cross from the local RF network to the APRS-IS as it is shown in Figure 2 where we presented our test bed infrastructure. APRS-IS is used as infrastructure to map location using commercially available radio APRS-IS is technology which brings different technological and open source advances at one place such as Linux operating system and applications which supports APRS-IS, TCP/IP infrastructure, low cost microcomputer hardware. One of indicators of how APRS-IS is used, can be seen at the aprs.fi web site. Referenced web APRS site is home for different information such as sensor readings, weather station reports. Different sensors can be used to collect data which might be crucial for early warnings on flood, fire, mudslides, and other potential nature disasters. These sensors can deliver information directly to Internet or as backup option via hamnet [14,15] network for rising alarms allowing proactive actions by those potentially affected and to rescuers. In recent years low cost and energy efficient microcomputer devices such as Raspberry Pi (RPi) [16] microcomputer became available which are suitable for education purposes [17, 18, 19, 20] which we used in our project to establish infrastructure for APRS reporting. Together with RPi it is possible to use software (in our case open source software) called software defined radio (SDR). The SDR Forum, in collaboration with the Institute of Electrical and Electronic Engineers (IEEE) P1900.1 group defined SDR as:

"Radio in which some or all of the physical layer functions are software defined" [21]

Today there are numerous commercial SDR solutions (hardware and software) with prices in thousands of dollars. Together with SDR open source solutions such as pymultimonaprs [22] it is possible to use low cost hardware to capture analogue and digital signals and use them to process captured signals on microcomputers such as RPi and build low cost SDR solutions which can be used for APRS-IS infrastructure. Core hardware component for this solution is RTL2382U [23] chipset which is used as DVB-T tuner for reception digital signals. Functionality of this chipset to capture radio frequency range from 22 to 1100 MHz [24] this chipset functionality which can be found in USB DVB-T dongles was discovered by Antti Palosaari and other researchers [25].

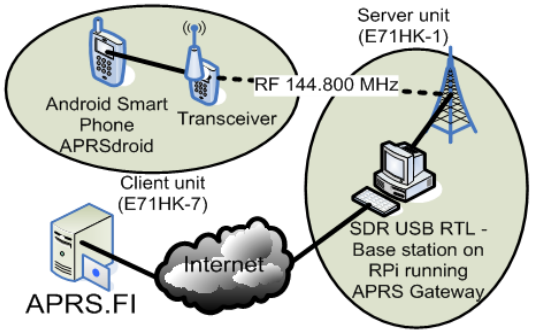


Figure 2. Infrastructure for APRS-IS

This type of infrastructure shown in Figure 2. allows location mapping and updates on APRS web site over Internet as other Internet services with difference that remote transceivers can reach transceiver base station from up to tens of kilometers (or more) when other infrastructure is not available.

B. Rationale for establishing VOIP communication using Echolink – Svxlink nodes

We saw potential of VOIP communication using open source svxlink [26] (variant of Echolink [27] software

built for linux distributions) software, in disaster relief operations. This software allows radio communication even if radio amateur does not have radio transceiver, to learn and educate other interested individuals and organizations, to test, research and to build infrastructure for other project and for potential disaster situations where radio communication was used [28] and can be used. Well educated individuals in alternative communication can play crucial role in any post disaster situation.

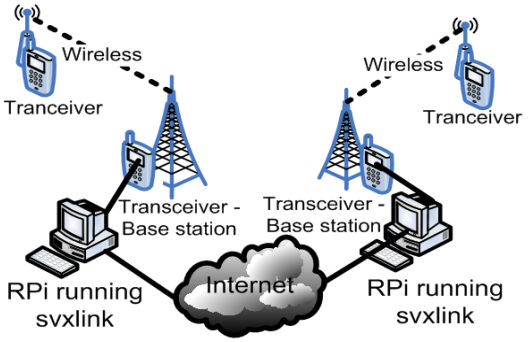


Figure 3. Infrastructure for APRS

C. *Rationale for establishing Motion video server*

Motion is open source surveillance video capturing software [29] able to capture video from one or more video cameras. Video can be monitored in real time, or pictures can be created in cases of detected motion. This is well documented software used on RPi microcomputers with RPi HD, and generic web cameras. Motion software has web based configuration management user interface and by accessing different ports to RPi with installed Motion it is possible to access live streaming or manage Motion software. Here motion is used to capture movement in near proximity of VOIP and GPS node that will visualize area around node for which we now exact GPS location.

D. *Expected goals and outcomes from establishing APRS, Echolink and Motion using on Raspberry Pi*

First goal was to use lessons learned, to build and document study case and to propagate practical and hands-on knowledge for education purposes first to student at University of Sarajevo, Bosnia and Herzegovina than to other local and other country universities.

Secon goal was to promote idea of using and providing infrastructure such as hamnet and svxlink VOIP as open source and free to use technologies in emergency situations.

Third goal was to build initial infrastructure with two svlxlink VOIP and iGate nodes located at two sites in Sarajevo city valley. This two svlxlink VOIP iGates would be able to accept local connections and transfer data to APRS-IS (Internet). This would open possibility for test and research activities. With established infrastructure, it is possible to measure usage and frequency of pymultimonaprs [22] application usage by users. Other test and research activities are possible such as different antenna design and usage of different frequencies.

E. *Design of experiment issues and scenario description*

As a practical part of this paper, our aim was to carry out performance measurements of number connections changing input radio frequencies. According to the theory design of experiment (DOE) [30] we deliberately change variables such as hardware components (transceivers), smart phones, antennas, configuration options for communication with svxlink [3], pymultimonaprs [22] and aprsdroid [31], Motion [32] applications.

III. Required Software and Hardware Components

For this project, different low cost versatile hardware and open source software components are needed. While software is open source hardware is needed but for which budget is less than 170 Euros, if Internet connection exist with appropriate access device.

A. *Hardware components*

In order to setup fully functional svxlink, APRS, and video capturing node there is need for hardware components plus cables, connectors stated below:

1. Microcomputer, Raspberry Pi B with following hardware specifications [16]:
   - Processor (CPU): ARM1176JZF-S (700 MHz)
   - Memory (RAM): 512 MB
   - Graphic processor (GPU): Broadcom VideoCore IV
   - System-on-Chip (SoC): Broadcom BCM2835
   - Communication: 1 × Ethernet 10/100 Mbps, 2 × USB, Audio-Out,
   - HDMI
2. Power: Micro-USB, ca. 300…700 mA
3. SD memory card minimum 2GB, and minimum class6, supported SD cards can be found on this link [33].
4. USB RTL DVB-T dongle
5. PMR or radio amateur transceiver such as Baofeng UV-5RE, 1 / 5 W, output [34].
6. USB hub with external power supply
7. USB Web camera or RPi HD camera
8. Smartphone with suitable Android OS

For initial system setup it is good to have USB Keyboard, monitor and HDMI / VGA convertor if needed because Raspberry Pi has only HDMI connector and not a VGA connector that most monitors are equipped with.

B. *Software components*

Five software components are needed for installing local APRS infrastructure are available on referenced web sites Raspbian Debian (Wheezy) distribution [34], pymultimonaprs [22], svxlink [3], Motion [32], and aprsdroid [31].

C. *Budget estimation*

As stated above budget for hardware is less than 170 Euros and software is open source solution, and hardware is only expense of this project.

TABLE 1. HARDWARE COMPONENTS PRICE LIST

Hardware component	**Price in Euros**
Raspberry Pi model B	32,50
HDMI / VGA converter	20
SD memory card	8
Baofeng transceiver	30
USB DVB-T dongle with RTL 2832U chipset	40
USB hub wit power supply	20
WEB camera	10
Cables an connectors	5
TOTAL	**165,50**

IV. INTERNET, PHYSICAL CONNECTIONS, AND RADIO INTERFACES

Prerequisites for establishment pymultimonaprs, and svxlink node are to have Internet connection.

A. *Internet connection*

There is no need for special port forwarding on firewall for pymultimonaprs [22] to operate and send GPS and other APRS data to APRS-IS. On the other hand Svxlink requires that router or firewall allow inbound and outbound UDP to ports 5198 and 5199, and outbound TCP to port 5200. Router has to be configured to "forward" UDP ports 5198 and 5199 to the Raspberry Pi on which svxlink is running [34]. For Motion software to be able to provide online capturing there is need to enable to port forwarding for port 8081 which is default port for Motion video broadcasting. There is no hardware requirements other then web camera has to be connected to RPi via powered USB hub.

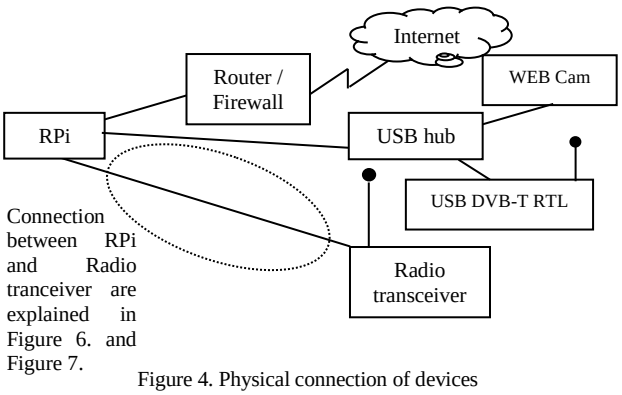


Figure 4. Physical connection of devices

Above-mentioned hardware has to be physically connected so that RPi is connected to router or firewall directly through UTP cable or via WIFI to switch or other communication device as it is shown in figure below.

B. *Physical connections for APRS*

SDR DVB-T RTL receiver which has task to receive GPS data over radio to be cone connecter to RPi via USB hub.

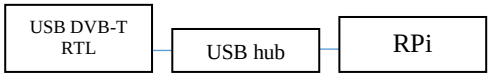


Figure 5. Block scheme of server side

C. *Audio physical connections for Svxlink*

RPi and radio transceiver have to be connected through audio cables, appropriate connectors, and using electronic components specified in Figure 7.

Kenwood, Wouxun, Baofeng [34] transceivers have same headset pin outs, and in Figure 6. is shown our connection scheme setting for mentioned types of devices.

In Figure 7. is shown scheme for Yaesu or Baofeng transceivers (jumper selection) for which sources [36, 37] were used as reference points. Idea for testing below described two options is to prove feasibility of using first and cheaper option for which only two condensers, cables and connectors were used.

D. *Phisical audio connection using VOX transceiver functionality*

For this option, transceiver has to be configured to use VOX option which is functionality to initiate transmission once it is recognized that sound (voice) is coming to microphone what simulates push to talk (PTT) button.

For this particular implementation voice is coming from USB sound card what has to be recognized as request for transmission. For this particular implementation only two condenser of 5 µF where used and appropriate audio connectors and conductors.

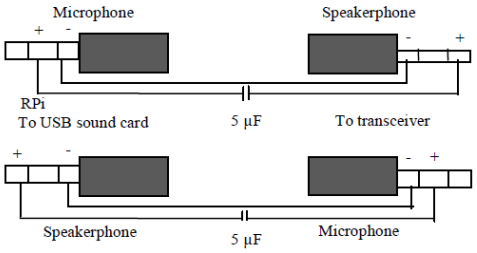


Figure 6. Physical connection of headset pin outs

E. *Phisical audio connection using USB/serial cable*

For this option, transceiver has to be used with USB conversion cable into serial cable which is used to capture voice and simulate PTT button via CNY17 photocoupler with appropriate audio transformers, connectors and conductors which is built based on [36, 37] options for Yaesu and Baofeng transceivers. Implementation is performed so that manual jumpers (JP) exists which can be used to choose different transceiver option (Yaesu or Baofeng) selection.

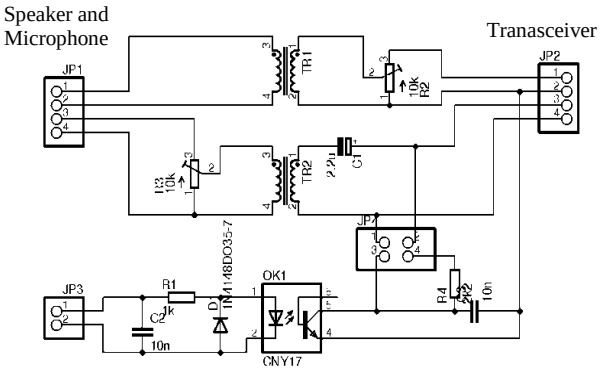


Figure 7. Physical connection of headset pinouts with USB / serial cable

V. BRIEF SETUP AND OPERATION INSTRUCTIONS

In order to operate pymultimonaprs [21], svxlink [3], Motion [32], node with Rasbian operating system it is important to install, configure and run it as defined in manual web pages referenced below.

A. *Setup instructions*

Installation image of operating system on RPi is described on web site [36]. For initial Raspbian Debian (Wheezy) implementation on RPi it is advisable to use Win32 Disk Imager [40] to install Raspbian Debian (Wheezy) operating system into SD card from PC which has SD card drive. To prepare Raspbian Debian (Wheezy) distribution for pymultimonaprs [21], svxlink [3], and Motion [32] installation and to install pymultimonaprs it is advisable use these installation procedures: for pymultimonaprs [41, 42], for svxlink [43], and Motion [44].

B. *Operation instructions*

After successful installation of pymultimonaprs, application can be started by executing command pymultimonaprs, or it can be configured to be started after each reboot by configuring /etc/rc.local file.

To start svxlink as a daemon command svxlink –daemon shall be executed and to start it at console is possible by typing command svxlink.

To start Motion application as daemon command `service motion start` shall be executed.

VI. APRS, MOTION, SVXLINK DATA AVAILABILITY AND OPERATION RESULTS

To enable APRS update information changes (received / transmitted packets, load, and possible other statistics and information) which are updated on web site www.aprs.fi in real time as it is shown in Figure 8. And Figure 9. This APRS online view can be used as a dashboard for monitoring activities for specific node or group of nodes and to monitor their available telemetry readings.

Figure 8. APRS shows two iGate nodes in Sarajevo valley

In Figure 9. are shown tracking results from aprsdroid application sent from smartphone via transceiver to APRS IGate.

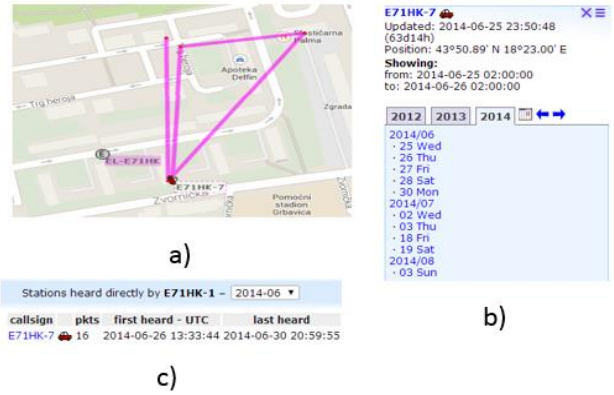


Figure 9. APRS information for E71HK (call sign of the first author of this paper) and neighboring stations

It is possible to APRS activities such as distance for specific date, using aprs.fi web site as it is shown in Figure 10.

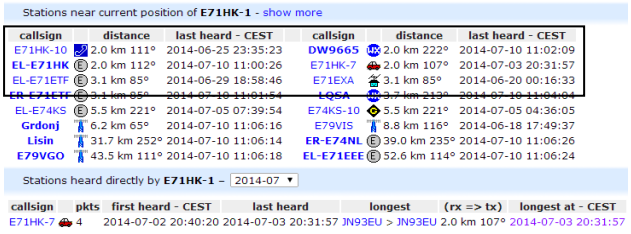


Figure 10. APRS information for E71HK and neighboring stations

It is possible to monitor traffic load on svxlink node in Erlangs for different period of time such as daily, weekly, monthly, and annually as it is shown in Figure 11. ER-E71ETF svxlink node which is node located at the Faculty of Electrical Engineering, University of Sarajevo. This node works as local repeater.

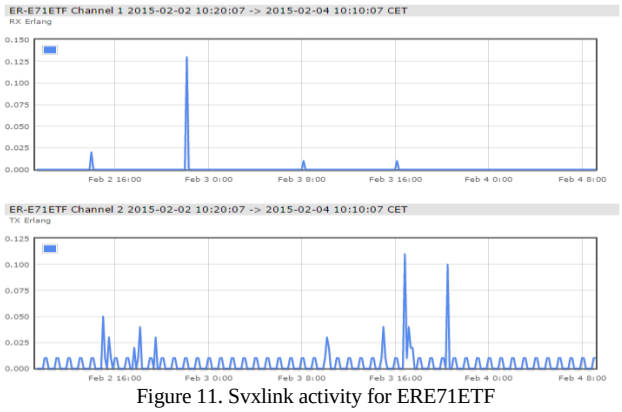


Figure 11. Svxlink activity for ERE71ETF

VII. CONCLUSIONS

From the preliminary results we may conclude that goals are reached because low cost infrastructure for education purposes and for disaster relief situations is built. This infrastructure is able to carry VOIP, reports APRS data and broadcast video with support of Internet or hamnet using low energy consumption hardware such as RPi. It is clear by APRS reports that number of radio amateurs were connected to already established svxlink (Echolink) nodes and initial awareness in local community

is reached by involving new radio amateur licensees (30 new radio amateur licenses for students of University of Sarajevo were issued during education process and during process of writing this paper), and involving other educational institutions to participate in radio amateur activities supported by University such as initial building hamnet infrastructure which is another project in progress and new svxlink nodes across Bosnia and Herzegovina what can be monitored in real time over aprs.fi web site.

VIII. REFERENCES

[1] R. Leszczyna I.N. Fovino M. Masera, Approach to securit assessment of critical infrastructures' information systems, Vol. 5, Iss. 3, pp. 135–144, Published in IET Information Security, doi: 10.1049/iet-ifs.2010.0261, The Institution of Engineering and Technology 2011

[2] APRS-IS Available from: http://www.aprs.fi/ [Accessed on 13.01.2015]

[3] Echolink. Available from: http://www.echolink.org/ [Accessed on 13.01.2015]

[4] Weinert, A.J. ; Breimyer, P. ; Devore, S.M. ; Miller, J.M. ; Brulo, G.S. ; Teal, R.S. ; Zhang, D. ; Kummer, A.T. ; Bilen, S.G. Providing communication capabilities during disaster response: IEEE Conference on Technologies for Airborne remote communication (ARC) platform, Homeland Security (HST), 2012 Publication Year: 2012 , Page(s): 395 - 400

[5] Rollema, D. Amateur radio emergency network during 1953 flood Proceedings of the IEEE Volume: 92 , Issue: 4 Publication Year: 2004 , Page(s): 759 – 762

[6] Uchida, N. ; Kawamura, N. ; Shibata, Y. Evaluation of Cognitive Wireless Based Delay Tolerant Network for Disaster Information System in a Rural Area, Seventh International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), 2013

[7] APRS benefits, Available from: http://www.riverdevil.org/APRS/vu2mue/ [Accessed on 13.01.2015]

[8] R. Lindquist, New Licensees, 98 QST Journal, Vol 98. No3, The National Association for Amateur Radio, March 2014

[9] Amateur Radio Showing Steady Growth in the US , 08.01.2014 Available from: http://www.arrl.org/news/amateur-radio-showing-steady-growth-in-the-us, [Accessed on 25.04.2014]

[10] Bruninga B., A brief history and bibliography of APRS, Available from: http://www.aprs.org/APRS-docs/ARTICLES.TXT [Accessed on 25.06.2014]

[11] Bruninga B.,CONNECTIONLESS EMERGENCY TRAFFIC SYSTEM, , Packet Radio Magazine, pp 4-5, July 86.

[12] Dimse S, Setting Up an APRS/Internet Gateway, Available from: http://www.aprs.net/vm/DCC97/internet/ [Accessed on 25.06.2014]

[13] APRS iGate, Available from: http://info.aprs.net/index.php?title=IGate [Accessed on 26.08.2014]

[14] Hamnet Information et actions engages, Available from: http://www.ref-union.org/index.php?view=article&catid=85:hamnet&id=706:informations-et-actionsengagees&format=pdf&option=com_content&Itemid=413, [Accessed on 11.01.2015]

[15] The Hamnet-Database – Interactive map, Available from: hamnetdb.net [Accessed on 05.01.2015]

[16] Raspberry. Available from: http://www.raspberrypi.org/help/faqs/[Accessed on 05.05.2014]

[17] Ali, M. ; Vlaskamp, J.H.A. ; Eddin, N.N. ; Falconer, B. ; Oram, C. Technical development and socioeconomic implications of the Raspberry Pi as a learning tool in developing countries, 5th Computer Science and Electronic Engineering Conference (CEEC), 2013, Page(s): 103 – 108

[18] Mitchell, G.The Raspberry Pi single-board computer will revolutionise computer science teaching [For & Against] Engineering & Technology Volume: 7 , Issue: 3, 2012 , Page(s): 26

[19] Srinivasan, M. ; Anand, B. ; Antony Venus, A.J. ; Victor, A.N. ; Narayanan, M. ; Sree Rakshaa, S.P. ; Vijayaraghavan, V. GreenEduComp: Low cost green computing system for education in Rural India: A scheme for sustainable development through education, Global Humanitarian Technology Conference (GHTC), 2013 IEEE, 2013 , Page(s): 102 – 107

[20] Skoric, M.M. Amateur radio communications, software and computer networks in education, International Conference on Signal Processing, Computing and Control (ISPCC), 2013 IEEE

[21] SDRF Cognitive Radio Definitions, Available from: http://www.sdrforum.org/pages/documentLibrary/documents/SDRF-06-R-0011-V1_0_0.pdf, [Accessed on 26.08.2014]

[22] pymultimonaprs, Available from: http://asdil12.github.io/pymultimonaprs/, [Accessed on 26.08.2014]

[23] RTL2832U, DVB-T COFDM Demodulator + USB 2.0 Available from: http://www.realtek.com.tw/products/productsView.aspx?Langid=1&PFid=35&Level=4&Conn=3&ProdID=257, [Accessed on 26.08.2014]

[24] Specifications for RTLSDR, Available from: http://sdr.osmocom.org/trac/wiki/rtl-sdr, [Accessed on 26.08.2014]

[25] History and Discovery of RTLSDR Available from: http://rtlsdr.org/#history_and_discovery_of_rtlsdr, [Accessed on 26.08.2014]

[26] Svxlink, Available from: http://www.svxlink.net/ [Accessed on 05.05.2014]

[27] Echolink. Available from: http://www.echolink.org/ [Accessed on 05.05.2014]

[28] Uysal, M.; Heidarpour, M.R. Cooperative communication techniques for future-generation HF radios, Communications Magazine, IEEE Volume: 50, Issue: 10 2012 , Page(s): 56 - 63

[29] Motion, Available from: www.lavrsen.dk/foswiki/bin/view/Motion/WebHome, [Accessed on 14.01.2015]

[30] DoE What is experimental design, Available from: http://www.itl.nist.gov/div898/handbook/pri/section1/pri11.htm [Accessed on 03.05.2014]

[31] APRSDROID Available from: https://aprsdroid.org/ [Accessed on 03.05.2014]

[32] Motion, Available from: http://www.raspberrypi.org/forums/viewtopic.php?t=18314, [Accessed on 14.01.2015]

[33] RPi SD cards, Available from: http://elinux.org/RPi_SD_cards [Accessed on 05.05.2014]

[34] Baofeng UV-R5E, Available from: www.baofengradio.com/ [Accessed on 05.01.2015]

[35] Firewall Solutions, Available from: http://www.echolink.org/firewall_solutions.htm [Accessed on 12.05.2014]

[36] Yaesu, Available from: http://www.qsl.net/ta7i/yaesu%20ISOKIT.htm [Accessed on 13.05.2014]

[37] Echolink interface, Available from: http://hamradio.etf.unsa.ba/2014/05/30/echolink-interfejs/, [Accessed on 01.06.2014]

[38] Raspberry Operating system distributions, Available from: http://www.raspberrypi.org/downloads/ [Accessed on 12.05.2014]

[39] Raspbian installation process, Available from: http://www.raspberrypi.org/documentation/installation/installing-images/linux.md, [Accessed on 12.05.2014]

[40] Win32 Disk Imager, Available from: http://sourceforge.net/projects/win32diskimager/ [Accessed on 13.05.2014]

[41] APRS iGate using Raspberry Pi Board, Available from: http://yd0nxx.wordpress.com/2013/04/30/aprs-igate-using-raspberry-pi-board/ [Accessed on 26.08.2014]

[42] APRS Rx-only IGate with Raspberry Pi and DVB-T dongle, Available from: http://www.kubonweb.de/?p=130 [Accessed on 26.08.2014]

[43] Installation Instructions, Available from: https://github.com/sm0svx/svxlink/wiki/InstallationInstructions [Accessed on 26.01.2015]

[44] Motion Guide – Installation, Available from: http://www.lavrsen.dk/foswiki/bin/view/Motion/MotionGuideInstallation, [Accessed on 26.01.2015]

Evaluation of Software Defined Radio Systems for Smart Home Environments

I. Vitas, D. Šimunić, P. Knežević

Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
igor.vitas@fer.hr

Abstract – Software Defined Radio systems are predicted to be radio access networks of next generation in all kind of wireless communications systems. Their use is not limited to radio front end that is first chosen for system operation. They can be upgraded by time as system needs changes or evolves which gives them radio front end flexibility never seen before in radio systems. Smart home environments such as eWALL system and other uses numerous of wireless technologies to control lot of appliances in them or to monitor specific health state of the people living there etc. Therefore, lot of wireless technologies have been developed solely for that purpose, leaders among them are ZigBee and Z-Wave technologies which are today standard for wireless home automation systems that are energy efficient and secure. Software Defined Radio is certainly very useful for future mobile communication systems. In this work Software Defined Radio system will be evaluated with a purpose of introducing local radio networks that are used in smart home environments based of Software Defined Radios. Benefits and shortcomings of using Software Defined Radio in this systems will be shown and evaluated.

Keywords **–** Software Defined Radio, smart home environments, local area networks, ambient assisted living, smart living, ubiquitous computing

1. INTRODUCTION

Software Defined Radio (SDR) is enabling technology for future wireless communication systems. Software Defined Radios are used today primary for research purposes but devices that will have the same or similar functionalities as Software Defined Radios will be used for future wireless communication systems. Re-configurability and re-programmability are key aspect of this kind of wireless communication systems. All technical systems that are made to be reconfigurable and reprogrammable are showing good perspective in future high-tech systems, such as Field Programmable Gate Array (FPGA) devices that are also key components of modern Software Defined Radio architecture.

Ambient assisted living (AAL) is one the focuses of modern European community as trends in last years show increase in population that is aged over 65 years. In year 2035 there will be 1/3 of the people that are older than 65 years and 1/9 of people that are older than 85 years.

Ambient assisted living is a wide term because it implies many research fields such as:

- Smart home environments
- Ambient intelligence
- Context aware computing
- Multi agent systems
- Robotics
- Internet of Things
- Wireless Sensor Networks
- Artifical Intelligence
- Human to Computer Interaction

As it is related to many research areas Ambient assisted living is also related to social economic trends in European countries. European union is community of different member states but its laws and economic practices in field of healthcare can still vary from state to state. There is a same challenge in development of AAL system both on technological and socio-economic level. This trend can also be seen in field of assisted living for elderly people. For example elderly people from Sweden often go to elderly residence in the south of Spain where residence for elderly people is cheaper and weather is warm. There is more than 5000 nursing homes in Spain and they are mostly private owned. For people who are staying in Sweden when they are in elderly age there are senior housing facilities which can be adapted to senior wishes regarding domotics, health care support and Ambient assisted living in general. Denmark is known for do-it-yourself retirement community because there groups of friends or like-minded people develop their own villages or apartments for living at elderly age. In every country pension schemes are different in amount of money that seniors receive per month as well in amount in healthcare support and other benefits in general for seniors that can be state financed. Technological trends in Ambient assisted living must conform to newest trends from medicine, economy, law and psychology. Developing of Ambient assisted living system that will be useful, affordable and admissible in all the EU is therefore complex task [1]. eWALL is Ambient assisted living system that is being developed by consortium financed by EU FP-7 research program. Architecture of eWALL system is shown on Figure 1.

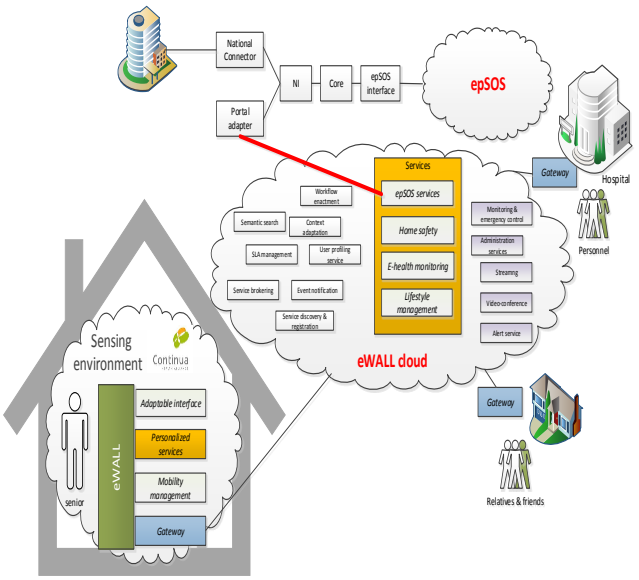


Figure 1 Architecture of eWALL - ICT system for AAL living.

Continua Health Alliance [2] announced that lifestyle management can address 60-80% of all cases, that happen to senior citizens, of unpredictable cases in the wellness and pre-illness conditions. It is obvious that the costs for the seniors and the National Departments of Health thus can be significantly reduced with introduction of AAL systems to environments in which elderly people spent most of their time. Large number of researchers and initiatives are trying to cope with the task of introducing AAL systems to everyday life of elderly people.

2. SENSOR TO GATEWAY WIRELESS CONNECTION IN SMART HOME/AAL ENVIRONMENT

Sensors are very important for AAL systems because they are producing low level metadata that is used later in AAL system for smart and intelligent reasoning. Metadata that is coming from these sensors need to be measured and transmitted safely to central gateway in AAL or smart home environment. Hardware and system level connection between sensors in smart home or AAL environment to central home gateway is given in Figure 2 and Figure 3.

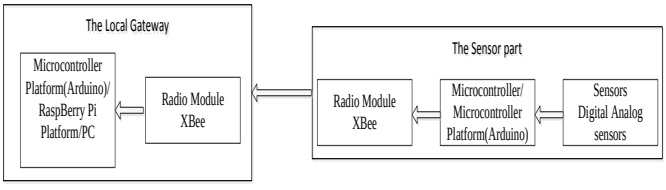


Figure 2 Possible hardware setup of sensor to gateway connection in smart home or AAL environment.

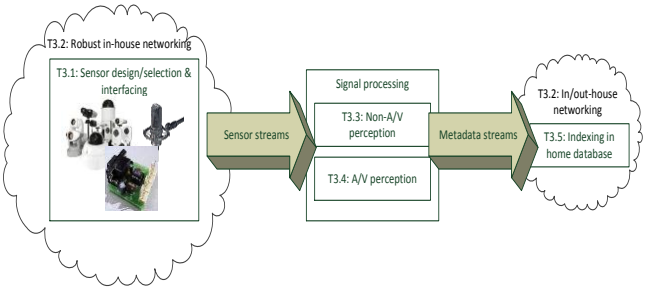


Figure 3 Sensors to gateway connection in smart home or AAL environment from system level perspective.

Connection between sensors, actuators and home gateway [3] can be done on wireless or wired way. Wireless connectivity offers freedom of nodes positioning all around the smart home/AAL space of interest but can be problematical when wireless connected sensor or actuator is on the edge that wireless signal can reach. Nevertheless this method of connecting sensors or actuators in this kind of environments is more usual than wired connections. Power Line Communications are becoming more and more represented for networking of various devices in smart home as well as in AAL environments. This market is rapidly growing because technology, primary chipsets, that enables this kind of communication are becoming more accessible and affordable. Most notable alliance of companies in field of Power Line Communications is Home Plug Alliance and CSEP – consortium for SEP2 interoperability. Vision of Home Plug Alliance connected home is shown on Figure 4.

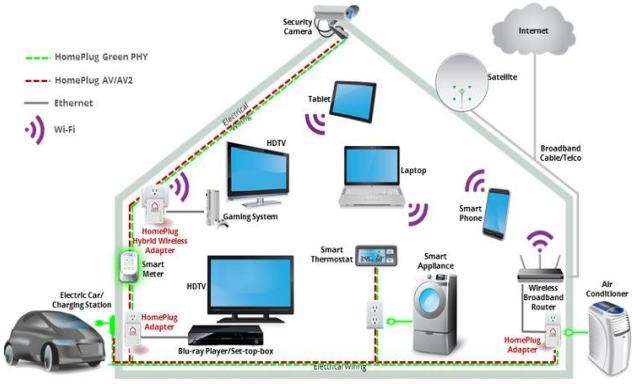


Figure 4 Home Plug Alliance vision of connected home environment.

Home Plug Alliance is committed to providing standards and set of guidelines for devices that are used for Power Line Communications. Home Plug Alliance defines Home Plug AV standard that is used to deliver an energy efficient and cost effective networking with optimized throughput for applications such as devices automation/control in smart home environment, security monitoring, energy management, device automation and all other machine to machine communications that are tolerant to communication latency. Most important standards regarding this leading Power Line Communication group are IEEE 1901 and IEEE 1905.1 standards. IEEE 1901 standard describes PHY and MAC specifications for power communication devices

across the entire spectrum of In-Home, Transport and Smart Grid applications. IEEE 1905.1 standard describes how to make benefit from using wide adopted standards including Wi-Fi (IEEE 802.11), Ethernet (IEEE 802.3), HomePlug AV (IEEE 1901) and MoCA (coaxial cables) in a software abstraction layer. With this standard it is shown how combination of wireless and wired networks can be beneficial for high performance in house networking. IEEE 1905.1 basically combines performance and reliability of wired networks with the flexibility of wireless for mobile devices. Data rates that can be offered by todays Power Line Communication devices can go up to 1 Gbps. This data rate is sufficient to support modern multimedia services in whole smart living/AAL space of interest.

CSEP, consortium for SEP2 interoperability is very interesting because it combines both wireless standards such as ZigBee, WiFi, Bluetooth and Home Plug Alliance to make devices that will conform to IEEE 2030.5 standard. This standard defines applications that enables home energy management via wired and wireless devices that support IP.

Wireless connection is still the most used way of connecting sensors and actuators in smart home/AAL environments.
Most used wireless protocols for connection with sensors and actuators are:

- Bluetooth
- WiFi
- Z-wave
- ZigBee

Some technologies such as Bluetooth Low Energy and ZigBee have found special place in smart home/ambient assisted living environments because they are part of Continua Health Alliance. Continua Health Alliance is an international non-profit organization with mission to offer end-to-end, plug-and-play connectivity of devices and services for personal healthcare. ZigBee Alliance in document [4] defines ZigBee profile for wireless networks that are used to implement wireless communication in healthcare systems. Uses cases for ZigBee networks that are dedicated for healthcare applications are divided into next categories:

1. Disease management (DM)
    - Non-critical patient monitoring (episodic)
    - Non-critical patient monitoring (continuous)
    - Patient alarm monitoring (low acuity)
    - Drug administration (e.g. insulin pumps)
2. Personal Fitness Monitoring (PFM)
    - Monitoring / tracking fitness level
    - Personalized fitness schedule
3. Personal Wellness Monitoring (PWM)
    - Activity monitoring
    - Safety Monitoring
    - Living independently

It can be seen that this applications domain suits perfectly to case studies in smart home and ambient assisted living environments that are defined by Continua Alliance.

Table 1 ZigBee, Bluetooth and Bluetooth LE technology comparison.

	ZigBee RF4CE	Bluetooth	Bluetooth LE
Network architecture	Multiple Star + mesh	Point-to-point + Pico net	Point-to-point + Pico net
# Devices supported in network	Virtually unlimited	7-16	Virtually unlimited
Frequency band	2.4 GHz	2.4 GHz	2.4 GHz
Mac/PHY specification	IEEE 802.15.4	IEEE 802.15.1	IEEE 802.15.1
Latency	<5 ms	100 ms	<10ms
Battery life	+15 years	Days	Weeks
Data rate	250 Kbit/sec	1-3 Mbit/sec	1Mbit/sec
RF Range	+50 m indoor	10 m	10 m
Security	128 bit AES	128 bit AES	128 bit AES

3. SOFTWARE DEFINED RADIO ARCHITECTURE AND ITS POSSIBLE APPLICATIONS IN SMART HOME/AAL ENVIRONMENT

Software Defined Radio is wide term that can describe various reconfigurable radio systems. Software Defined Radio system is often associated with term cognitive radio that is just one, but most widespread, field of research around Software Defined Radio architecture and its applications. Software Defined Radio concept was introduced in 1999 by Joseph Mitola [5]. In this work Joseph Mitola explained the vision about reconfigurable and software components in radio front end communication chain. Goal was to make framework for reconfigurable wireless architecture so one radio device can work within multiple wireless communication systems. Most used Software Defined Radio system today is Universal Software Radio Peripheral (USRP) system firstly build by company Ettus Research which was then acquired by National Instrument. System design of USRP Software Defined Radio system is given on Figure 5.

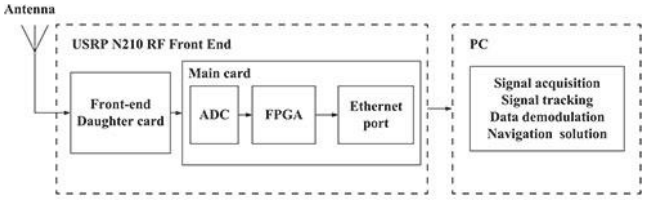


Figure 5 Architecture of NI USRP Software Defined Radio system.

One o of the key characteristics of USRP platform is radio frequency signal down and up conversion before ADC/DAC converters and Field Programmable Gate Array (FPGA) platform behind converters in radio front end chain. Field Programmable Gate Array (FPGA) platform is used for down converting sampled digital IF signal to digital baseband signal. After digital down conversion the signal is passed through digital filters and then sent to PC via USB 2.0 connection for further data demodulation. Similar thing is happening in transmit direction with difference of performing digital up conversion to digital IF signal instead of down conversion.

Front-end Daughter cards of USRP platform are used for implementation of radio front end circuits from antenna to the digital part of SDR system. The cards are offered in wide radio frequency bands from 10 MHz to 6 GHz so they can be

used for emulation of almost any radio frequency wireless communication system that is used today. There are numerous of projects conducted by university research groups and technological companies in which SDR USRP platform is used for emulation of modern high end wireless base station systems such as LTE, GSM, IEEE 802.11, IEEE 802.15.4. etc. All projects are implementing base stations using USRP platform in combination with PC. It is shown that all the issues that arise in this kind of wireless communication system implementation can be solved by modifying hardware or the software part of the system. Most of the systems are programmed using GNU Radio framework for USRP Software Defined Radio programming. One device can theoretically be used for implementation of different radio communication protocols with just change of the software. There are also applications of Software Defined Radio systems that are used for purposes such as radio astronomy [6].

Architectures of smart home environments, especially ambient assisted living environments rely on complex ontologies that define ambient assisted living environment architecture from data models point of view [7]. Ontologies aim to add semantics, i.e. meaning, to information to allow for reasoning over the data to derive further knowledge. Further reasoning in smart home environment can be implemented on a central home computer [8] and can also include ontologies for deployed sensor networks. In the past few years there has been work in order to define ontologies for sensors and sensor networks. As a result a number of ontologies have been defined that serve different purposes, applications and domains. Ontology that is developed by W3C consortium for sensor networks can be seen in Figure 6. This field of research is especially important for addressing future Internet of Things systems that will deliver a lot of data from different sensor sources in near future.

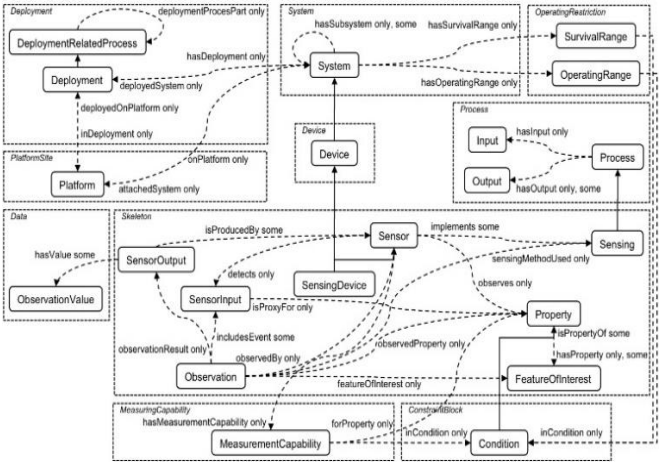


Figure 6 Ontology of W3C semantic sensor network.

Software Defined Radio can be used as a wireless hub in smart home/ambient assisted living environments to support unification of wireless links used in environment of that kind through one device. One Software Defined Radio system device can be used to control various wireless controlled domotic systems as well as other wireless multimedia systems that are used in environment. Wireless controlled domotic systems are usually not working full time so their reduced duty cycle is ideal for using one Software Defined Radio system device to control lot of wireless devices that can be supporting different wireless protocols on different frequencies made by different electronic equipment manufacturers.

4. CONCLUSION

In this work concept of smart home environments and ambient assisted living environments was presented through various methods and technologies that are used for wireless and wired connectivity in them. This connectivity is used for different purposes from controlling of home automation systems to streaming of multimedia data to user devices such as smartphones or tablets. Key aspects of current standards and technologies in this field is shown and concept of Software Defined Radio system as a general purpose wireless hub for controlling all wireless related device in smart home/ambient assisted living environment is given. In future work concept of controlling devices that utilize different wireless communication protocols with one Software Defined Radio will be implemented using National Instruments USRP radio platform.

REFERENCES

[1] S. Davidoff, M. K. Lee, J. Zimmerman, and A. Dey. Principle of Smart Home Control. In *Proceedings of the Conference on Ubiquitous Computing*, pages 19–34. Springer, 2006.

[2] http://www.continuaalliance.org/

[3] Home gateway technical requirements: Residential profile. Technical report, Home Gateway Initiative, 2008.

[4] ZigBee Health Care Profile Specification, ZigBee Document 075360r15, ZigBee Alliance March 2010

[5] Mitola, J. ”Software Radio Architecture: A Mathematical Perspective” IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 17, NO. 4, APRIL 1999.

[6] Prabaswara, A; Munir, A; Suksmono, AB., ”GNU Radio based software defined FMCW radar for weather surveillance application,” Telecommunication Systems, Services, and Applications (TSSA), 2011 6th International Conference on , vol., no., pp.227,230, 20-21 Oct. 2011.

[7] D. Bonino and F. Corno. Dogont - ontology modeling for intelligent domotic environments. In *7th International Semantic Web Conference*, 2008.

[8] S.-L. Chung andW.-Y. Chen. MyHome: A Residential Server for Smart Homes. *Knowledge-Based Intelligent Information and Engineering Systems*, 4693/2007:664–670, 2007.

Smartwatch as a Remote Server Monitoring Device: Implementation and Interaction Design

Armando Vega and Sandi Ljubic
University of Rijeka, Faculty of Engineering, Rijeka, Croatia
armando.vega@riteh.hr, sandi.ljubic@riteh.hr

Abstract - Present day computer ecosystem offers advanced connectivity of numerous devices, systems and services, providing the basis for the Internet of Things paradigm implementation. What most people perceive of its structure often lacks acknowledgement of the smallest parts of it, albeit as of recently the most used ones, namely smartphones and smartwatches. Being recognized as a prominent member of the wearable/ubiquitous computing family, smartwatches have been gaining in popularity proving themselves valuable in various contexts, ranging from providing fitness applications and general health monitoring, to always-on assistive technology for people with special needs. This paper focuses on an alternative approach to using smartwatches, considering them monitoring tools for system and network administrators able to provide fast and readily available information on various network equipment and services. For this purpose a proof-of-concept application suite was built, used to provide the status of important services on a remote Linux-based server, as well as the status of the server system itself. We show how smartwatches can offer different, and some unique, previously considered impractical ways to convey information. Supporting HCI aspects are specifically discussed, with respect to the Pebble smartwatch basic I/O capabilities limited to four hardware buttons and 1.26" monochromatic display.

I. INTRODUCTION

Recent advances in miniaturization and diminishing costs in battery, processor, sensor, and communication technologies are now enabling smartwatches to move from the specialist market to the mainstream [1]. Most of known smartwatch design efforts have seen them built by companies as wearable companions to their customary products, namely various smartphones and tablets. This made them dependent on existing same-manufacturer equipment, and usually limited in capabilities to conform to their counterparts. The smartwatch products sported custom OEM operating systems, proprietary interconnection interfaces and development processes tailored to a particular device or a specific generation of devices. The challenges of such vertically integrated design needed to be overcome for the smartwatch application development ecosystem to strive.

The Pebble watch, developed in 2012 as a result of one of the most successful campaigns in the history of crowd funding at the time, can be considered the first fully independent smartwatch [1]. It is very much still a device with a custom tailored operating system, however, it provides the developers with an easy to use SDK and open application programming interfaces, thus enabling easy integration and communication with most popular mobile operating systems running on present day smartphones and tablets.

Regarding the practical advantages of a device that is worn directly on the skin, various different approaches to interaction with target applications can be considered. Most current day applications are implemented leveraging the various built-in sensors, which provide information about user's movement and position in space. A plethora of existing mHealth applications provide general health and fitness monitoring and advisement, as well as remote tracking of said information, by making use of the paired devices' connectivity capabilities.

Such a device itself does not come without limitations, observed mostly in the relatively modest computing capabilities and a small screen size, as well as a seemingly small number of ways to interact with it. This provides both developers and researchers with certain challenges when it comes to design of applications for the platform, that need to be met with general usability guidelines in mind. Considering the device and its current application ecosystem, we decided to apply ourselves to research the possibilities of interaction, extending it by using sensor data in addition to buttons for input, and haptic feedback in addition to using the display for visual representation of information.

II. RELATED WORK AND MOTIVATION

Considering the focus of our work, related research efforts from the smartwatch domain can be divided into two main categories: (i) smartwatch application context of use, and (ii) interaction design for smartwatch devices. While context of use studies usually highlight the utility of particular smartwatch application in a given surroundings, HCI-oriented research specifically investigates ways of interacting with such a small wrist-worn device.

Smartwatches usually play a role of Bluetooth-tethered peripheral interfaces for smartphone devices, supporting common tasks such as message and phone call notifications checking [2]. As their onboard computational power, networking, and displays become more enhanced, game-based entertainment starts to be popular in this small form factor domain as well. Moreover, most modern smartwatch devices come with integrated motion sensors (e.g. accelerometer and magnetometer), which enables

physical activity recognition through analysis and classification of wrist movement patterns [3][4]. Accordingly, there exists an emerging trend in providing advanced smartwatch applications for fitness tracking and general health monitoring. Android Wear [5], Google's Android-based platform designed for smartwatches and other wearables, involves features such as step counting and heart rate monitoring, as well as GPS-based tracking of walking/running/cycling distance and speed. Smartwatches are also recognized as valuable tools within always-on assistive technology, since they can provide help for elderly people (e.g. in fall detection scenarios) and for users with special needs (e.g. people with visual impairments [6]). Additionally, implementations for general navigation within geo-map-oriented applications [7] and turn-by-turn traffic-based navigation are also available. Indeed, since the introduction of the first watch running a full operating system [8], smartwatch devices and related applications are being constantly improved, demonstrating usefulness in various contexts of use.

When it comes to the HCI aspects of smartwatches, we can straightforwardly claim that the form factor restricts both input and output capabilities. The predominant input methods in this domain still rely on physical buttons and touchscreens (where available). Since watches lack the surface area needed for numerous hardware switches, traditional keypad concept is upfront abandoned, thus making designs with just a few physical buttons a preferred option. However, as is the case with smartphones, text entry can be successfully performed using smartwatches, either by making use of touchscreen-based soft keyboards with specific character layouts [9], or by utilizing advanced skin-based and around-device interaction techniques [10]. Touchscreen displays are incorporated in today brands' flagship products (e.g. Samsung Gear S with a curved screen; Motorola Moto 360 with a circle-shaped display), with *tap*, *long press*, and *swipe* being the main interaction modalities for usual tasks such as pointing, selecting, and scrolling. Nevertheless, handling such small displays (less than 2") using direct touch imposes an interaction burden which can be attributed to the well known fat finger problem. Sensor-based interaction represents an additional benefit in smartwatch usage, since 3D mid-air gestures can be used, beside activity recognition, for application interface control as well. Various interaction sets can be derived in order to digitally augment commonly used gestures, as shown in the *Office Smartwatch* application [11], wherein knock, lock, and return gestures are implemented and afterwards used for controlling the office environment. Smartwatch input control, however, can be further empowered by using the watch face itself as a multi-degree-of-freedom mechanical interface. Although such a product cannot be found in the market, a proof of concept smartwatch that supports continuous 2D panning and twist, as well as binary tilt and click is already provided [12]. Finally, contemporary smartwatch devices have voice recognition enabled, thus providing the possibility to control various services with voice commands. Android Wear platform supports making voice memos and reminders, using voice web search, as well as voice control of navigation and multimedia streaming. From the output capabilities standpoint, smartwatches usually rely only on display technology (for presenting visual information), and acoustic/vibrotactile feedback (for invoking notifications).

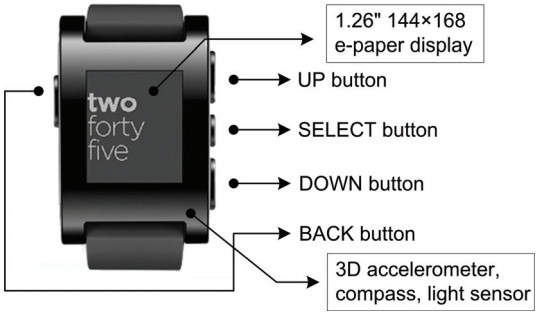


Figure 1. The front face of the Pebble smartwatch used in our proof-of-concept solution

Our proof-of-concept application suite contributes to an ubiquitous computing ecosystem by providing computer network administrators a wearable monitoring tool in a form of a dedicated smartwatch application. To the best of our knowledge, there exists no reported work regarding this specific context of smartwach usage. In addition, we show how good interaction design, involving multimodal operations and feedback control, can overcome rather limited I/O capabilities of the used smartwatch device. Namely, our application, intended to be run on the non-touchscreen-enabled Pebble smartwatch (Fig. 1), can be controlled with built-in hardware buttons, as well as with wrist movements only – using tilt and shake gestures. Incoming notifications are also presented in a multimodal way: along with visual information on the smartwatch display, various vibrating patterns are used as context-specific feedback. Consequently, the proposed design allows both touch-based and touchless interaction, as well as application handling without the explicit need for visual attention.

III. Application Suite Implementation

The design and implementation of the TOSS (*Timely Overview of Server Status*) application suite is divided into several parts (as shown in Fig. 2), consisting of the: (i) backend server-side application, (ii) Android-based intermediary service, and (iii) the Pebble smartwatch frontend application itself. A parallel can easily be drawn with the classic MVC (*Model-View-Controller*) software architectural pattern [13], as the respective design conforms to the exact pattern in all of its elements.

The server-side application, conforming to the *Model* part of the architectural pattern, manages the collection and provision of data for the suite. The intermediary Android-based service, related to the *Controller* part of the architectural pattern, in turn takes the data provided by the *Model* (the server-side application), reformats it according to the needs of the *View*, and leverages push-style communication to deliver data to the *View*. Finally, the Pebble smartwatch application, reflecting the *View* part of the architectural pattern, provides the user interface which conveys the information provided by the *Controller* part of the suite (the Android-based intermediary service).

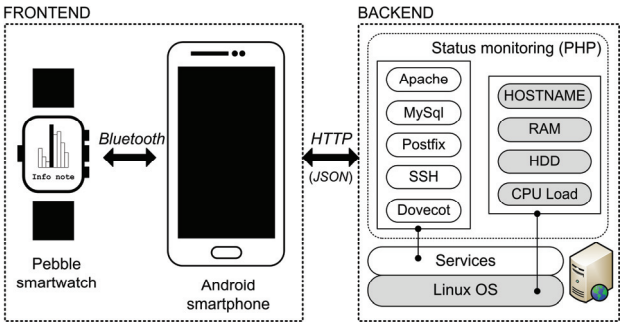


Figure 2. TOSS application suite design schematic

As each of the aforementioned elements required a completely different implementation approach, the design and implementation details of every system component will be further discussed separately.

A. Server-side monitoring application

The server-side monitoring application is implemented using an existing HTTP serving application (e.g. Apache HTTP Server) that supports execution of the PHP server-side scripting language. The application itself is written using PHP, and is built as a modular entity capable of dynamically loading supplementary modules which in turn provide additional data to the system. Each module is responsible for a single data point. The data retrieval part of a single module can have different means of getting the required data. However, all modules make use of a common interface for communication with the central data collection component. Provided data format is based on key-value pairs, consisting of the data identifier (e.g. mail service process name, system property name) and the data pertaining to the mentioned feature (e.g. service or system property status).

For the purposes of our proof-of-concept application suite, the server-side modules have been built to provide data on the different properties of a server running a GNU/Linux operating system and services most commonly found in the corresponding environment. As such, they are not in and of themselves portable to other platforms, as they leverage standard GNU/Linux interfaces for acquiring system and process information. However, since data collection modules can have different means of implementation, new modules can be built that would interact with other operating systems and platforms. Table I presents the list of available server-side modules, including a short description of the data they provide.

TABLE I. AVAILABLE SERVER-SIDE MODULES PROVIDED IN THE APPLICATION SUITE

Module name	Description (data provided)
mod_apache	Running status of the Apache HTTP Server service
mod_disk_usage	Disk usage percentages for all currently mounted volumes
mod_dovecot	Running status of the Dovecot MDA (*Mail Delivery Agent*)
mod_hostname	Configured hostname of the serving machine
mod_mysql	Running status of the MySQL database server
mod_postfix	Running status of the Postfix MTA (*Mail Transfer Agent*)
mod_ram_usage	Amounts of total and used system RAM
mod_ssh	Running status of the OpenSSH secure remote access server
mod_system_load	Load averages of the system's CPU for the last 1 minute, last 5 minutes and the last 15 minutes


```
{   "apache":"OK",
    "disk_usage":
    {   "rootfs":"6%",
    "/dev/root":"6%",
    "devtmpfs":"0%",
    "tmpfs":"0%"
    },
    "dovecot":"OK",
    "hostname":"isis",
    "mysql":"OK",
    "postfix":"OK",
    "ram_usage":
    {   "Used":"248",
    "Total":"989",
    },
    "ssh":"OK",
    "system_load":
    {   "1min":"0.00",
    "5min":"0.01",
    "15min":"0.05",
    }
}
```


Figure 3. Example of server-side generated JSON formatted data

The central server-side application, encompassing all the built-in modules, is itself completely platform independent, requiring only a PHP-interpreting capable HTTP server application. The collected data is provided to the client over a HTTP or ideally over a secured HTTPS connection, depending on the server configuration. Data transfer utilizes a JSON (*JavaScript Object Notation*) formatted text (Fig. 3) – the format supported in commonly used programming languages which usually provide readily available JSON parsing functions.

B. Android intermediary service

The intermediary service for the Android mobile operating system, used to collect the server-side generated information and push it to the Pebble smartwatch application, is implemented following a recommended approach by making use of the *SyncAdapter* framework [14]. As opposed to a simple *Activity* or *Service* application component, a *SyncAdapter* is handled in a special way on an Android operating system. This framework is specifically designed for tasks scheduled to repetitively run at specified time intervals. It provides, among other things, automated network status checking – that pauses execution until a network connection becomes available, and automated queuing of different (concurrently scheduled) sync adapters – which in turn provides improved battery performance. Moreover, *SyncAdapter* nicely incorporates into the standard Android accounts manager, enabling easy access to out-of-schedule manual invocation, as well as disabling/enabling its function altogether. The pluggable architecture of the framework allows for easy execution of callable components, thus providing us a possibility to easily implement the needed communication procedures. The officially provided PebbleKit Android Java library was

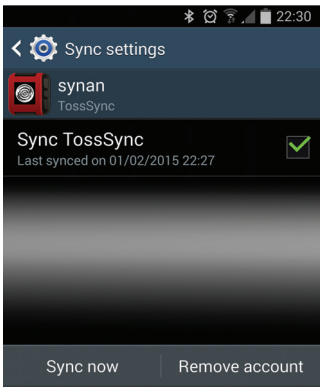


Figure 4. *SyncAdapter* component integrated into the Android OS

used for establishing Bluetooth communication with the Pebble smartwatch.

The Android intermediary service which incorporates our sync adapter, called *TossSync* (Fig. 4), is configured to synchronize with the server-side application at a specified time interval. Given the nature of system administration work, in our pilot testing sessions we found the most reasonable time interval (for such a fast response requiring purpose) to be 1 minute. Such a short interval of recurring data retrieval should provide a system administrator with a timely information that can quickly be responded to. Once new data is received on the Android smartphone, the JSON format is parsed and converted into a simpler key-value format counterpart appropriate for the Pebble communications API. Key-value pairs are then pushed over Bluetooth to the Pebble smartwatch OS, which in turn delivers the data to the target application. The Android component of the suite stores the latest received information from the server, and invokes the data push to the Pebble smartwatch only when some status change has been detected. The idea behind this is to improve battery performance on both the Android device and the smartwatch.

C. *Pebble smartwatch application*

The application running on the Pebble smartwatch, named *PebbleToss*, is used to convey information to the user wearing it by visual and haptic feedback. The application was implemented using the C programming language, leveraging the officially provided Pebble SDK for gathering available sensor information, button press handling and establishing Bluetooth communication with the Android intermediary service running on the paired smartphone.

Given the embedded nature of the smartwatch device, there were a few challenges that needed to be overcome during both the design and implementation phases of development. One notable characteristic is the previously mentioned small 1.26" screen, providing a viewable area of only 144×168 pixels with 1-bit black and white graphics. This constraint was especially important during the design phase, as it provided hard limits on the ways of conveying the needed information. Another challenge came in the form of computing capabilities, as the smartwatch provides only 128KB of RAM, of which only 24KB is available to applications, the rest being used for the function of the operating system itself. It was therefore important to make the application as efficient as possible with regards to memory usage.

The application's user interface consists of several parts, notably the main application window (MW) and the service status window (SSW), example snapshots of which can be seen in Fig. 5. Each window was specially designed given what it would be used for, combining both graphics and text to provide the needed information.

The main window can visually be separated into four sections, split horizontally. The topmost section displays the name of the server that is currently being monitored, taken as the defined hostname provided in the server-side generated data. The following section is the only part that doesn't depend on the data provided from the outside, as it displays the current day of the week, date and time. As seen on the example snapshots in Fig. 5, the part of this section displaying time has inverted colors with regards to the rest of the screen (white text on a black background), as it is considered most used and should easily be distinguished from the rest of the information given only a quick glance. The subsequent display section shows status of server system memory (RAM) usage, as well as used file system space (HDD). It combines graphics and text, providing graphical representation of the ratio of used and available resources. The information is shown using bar charts, with identifiers on each side of the chart specifying to what each bar corresponds to. This provides easily understandable and universal feedback that does not depend on the actual units used in each of the server properties. Finally, the bottommost section of the main window displays the system CPU load information, using the standard format of showing the load average for the past 1 minute, past 5 minutes and the past 15 minutes.

The service status window is used to display current status information on selected system services. This

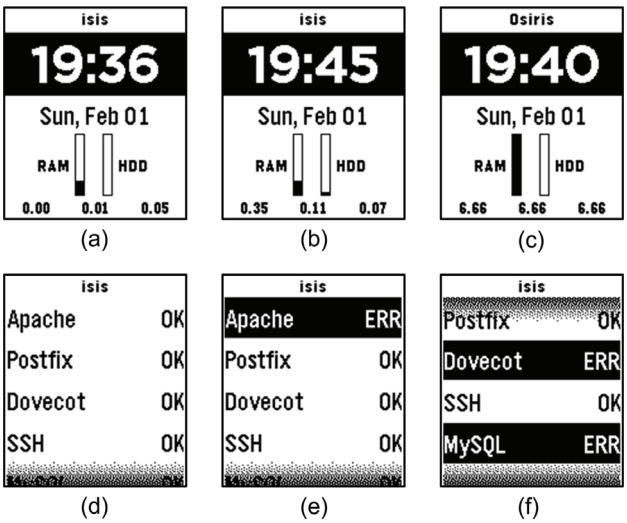


Figure 5. Smartwatch application snapshots. (a) MW (no vibrating alert): RAM and HDD are OK, system load is recently decreasing (5%→1%→0%); (b) MW (no vibrating alert): RAM and HDD are OK, system load is recently increasing (7%→11%→35%); (c) MW (vibrating alert - RAM pattern): Out-of-memory issue, system load is recently stagnating at 6.66%; (d) SSW (no vibrating alert): all services are OK; (e) SSW (vibrating alert - Apache pattern): www issue; (f) SSW (vibrating alert - Dovecot pattern followed by the MySQL pattern): email issue and database issue.

window is visually separated into two sections. The topmost section contains the name of the server that is currently being monitored, corresponding to the topmost section of the main window, providing UI consistency. The second section contains a scrollable list of selected system services and their last known status. As can be seen in Fig. 5, a service that has a status considered without problems shows the corresponding "OK" message, while services with reported problems have their status shown as "ERR" and are additionally highlighted (inverting the colors of the text and the background) providing easier discernment.

IV. INTERACTION DESIGN

In the smartwatch application development process, we put a special emphasis on simplifying both the user interface and related interaction, by taking advantage of Nielsen's usability heuristics [15]. Specifically, our goal was to provide simple and natural dialogue ("less is more" concept), consistent screen design, comprehensible metaphors (both icons and text abbreviations), and continuous feedback. Furthermore, we opted for multimodal interaction, thus giving the users possibility to handle the application in a way that serves given context of use the most. Following requirements for interaction design were settled:

- a quick and simple switch between the system status window and services status window
- one press forced-update invocation
- direct scrolling of services status window content
- different vibration patterns for different server-side notifications
- alarm timeouts (no forced deactivation)
- ability to perform basic application actions with and without pressing buttons on the smartwatch

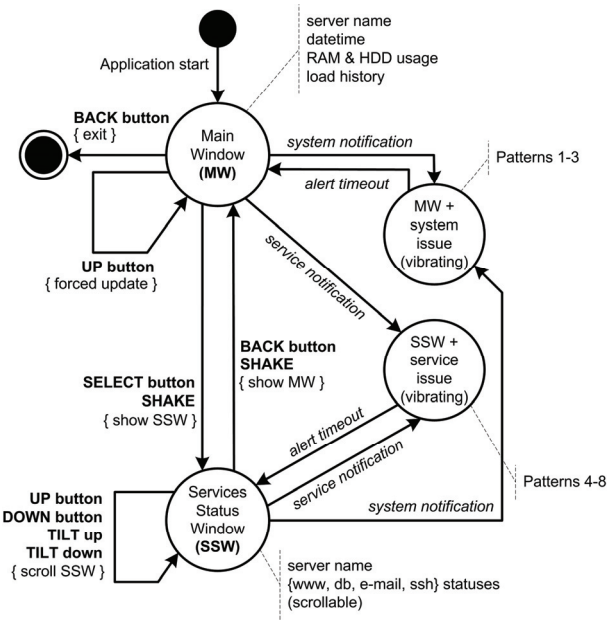


Figure 6. State diagram of interaction design used in our prototype

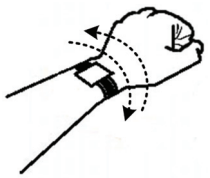


Figure 7. Touchless interaction: invoking tilt actions (for scrolling) and shake gesture (for windows switching) using wrist movements only

Fig. 6 presents the final version of interaction design incorporated in our solution, in a simple state diagram form. State transitions are either user-driven (representing user's explicit actions) or system-driven (automatically fired by server-side notifications).

To facilitate a quick and simple switching between two display screens (MW and SSW), SELECT and BACK buttons are chosen, making this basic window navigation consistent with other smartwatch applications (with respect to going down/up in the window hierarchy stack by pressing SELECT/BACK). The same action is additionally enabled by a SHAKE gesture, consisted of double tilt up/down wrist movements. While SHAKE gesture is thus used for windows toggling, tilting the smartwatch device in a single direction results with TILT UP or TILT DOWN – actions dedicated to scrolling activity within SSW display (Fig. 7). Scrolling can be concurrently achieved using UP and DOWN physical buttons. Within TILT and SHAKE gestures implementation, we defined a *neutral position zone* – an angle-based offset wherein the smartwatch is assumed to be motionless. By doing so we were able to filter out involuntary wrist rotations, i.e. to augment the disambiguation between spontaneous and deliberate movements (cf. [16]).

Although data synchronization with the server is set to be frequent (1 min interval), we nevertheless provided an update-on-demand possibility. Pressing the UP button when MW is displayed will immediately check for possible system/services status changes. Since forced update will likely be used on rare occasions, button press in this context does not have its tilt-based counterpart. All other interface actions allow for multimodality, providing the base for concurrent use of both touch-based and touchless interaction (Table II). Consequently, user is able to interact with the application even if her/his hand (not wearing the smartwatch) is busy.

TABLE II. INTERACTION MODALITIES AND ASSOCIATED INTERFACE ACTIONS

Interface action	Touch-based interaction	Touchless interaction
Show SSW	SELECT button press	SHAKE gesture
Show MW	BACK button press	SHAKE gesture
Scroll up SSW	UP button press	TILT UP gesture
Scroll down SSW	DOWN button press	TILT DOWN gesture
Forced data update	UP button press	-
Alert (de)activation	-	-

One of the main interaction aspects of our smartwatch application is providing feedback on particular system/service notification. When status change is detected, a sequence of interface actions is accordingly performed: (i) bringing front the related window (MW/SSW), (ii) updating GUI elements (string values and/or bar charts), and, if necessary, (iii) starting a notification alert with a particular vibration pattern. In general, a vibration pattern is represented by a sequence of vibration and stillness intervals with previously defined time length. We devised altogether eight vibration patterns: three for system notifications, and five for service notifications. For example, *low-disk-space* issue is notified by 8-seconds long continuous vibration. *Out-of-memory* problem is represented by 8 consecutive short vibrations (0.5s) separated with 0.5s stillness intervals (i.e. "beeping" for 8 seconds). Vibration pattern dedicated to the *high-CPU-load* starts with 4s "beeping", and is followed by 4s continuous vibration. On the other hand, all service notification patterns are constructed using Morse codes, reflecting only the first letters of service names (e.g. A for an Apache error, M for a MySQL error, and so on). We used a 1s base interval for coding dots, dashes, and silences. Initial testing in the development phase showed vibration patterns to be sufficiently diversified and suitable for easy recognition in eyes-free scenarios. By using vibration feedback alongside standard visual display, interaction design was empowered with multimodality on the output side. As a final result, the prototype application allows not only touch-based and touchless interaction, but full "blind-handling" as well.

V. Conclusion And Future Work

In the presented application suite prototype implementation, the Pebble smartwatch is used as a client interface for remote server monitoring. Owing to the smartwatch continual connection to the skin and standard mount location (wrist), target users can be provided with essential information in a timely fashion, irrespective of the current context of use (e.g. while driving, at a meeting, in a cinema). We have demonstrated that with a careful interaction design it is possible to overcome rather limited I/O capabilities of the used device. Namely, standard interaction modalities such as button presses and visual feedback are augmented by making use of the device's motion sensors and vibration motor. Hereby simple mid-air gestures (tilts and shake) were designed and initially tested for a suitable touchless interaction. Different vibration patterns are dedicated to the respective server issues, enabling the user to both disambiguate and comprehend the receiving notifications even without looking at the smartwatch display.

For future work, we plan to carry out formal usability testing of the proposed application. Specifically, we want to evaluate the intuitiveness of our interaction design, as well as get a more detailed insight into the efficiency, learnability, and users' satisfaction attributes. Finally, required physical efforts (wrist rotations) and mental efforts (disambiguating vibration patterns) have to be tackled in order to get a better understanding of ease-of-use constraints.

Acknowledgment

The work presented in this paper is supported by the University of Rijeka research grant Grant 13.09.2.2.16.

References

[1] R. Rawassizadeh, B. A. Price, and M. Petre, “Wearables: Has the Age of Smartwatches Finally Arrived?,” Communications of the ACM 58(1), ACM, pp. 45-47, January 2015.

[2] J. Knibbe et al., “Extending Interaction for Smart Watches: Enabling Bimanual Around Device Control,” Proc. Extended Abstracts on Human Factors in Computing Systems (CHI EA'14), ACM, New York, 2014, pp. 1891-1896.

[3] G. Bieber, T. Kirste, and B. Urban, “Ambient Interaction by Smart Watches,” Proc. 5th Int'l Conf. PErvasive Technologies Related to Assistive Environments (PETRA'12), ACM, New York, 2012, pp.1-6.

[4] G. Bieber, M. Haescher, and M. Vahl, “Sensor Requirements for Activity Recognition on Smart Watches,” Proc. 6th Int'l Conf. PErvasive Technologies Related to Assistive Environments (PETRA'13), ACM, New York, 2013, pp.1-6.

[5] Google: Android Wear, http://www.android.com/wear/

[6] L. Porzi, S. Messelodi, C. M. Modena, and E. Ricci, “A Smart Watch-based Gesture Recognition System for Assisting People with Visual Impairments,” Proc. 3rd ACM Int'l Workshop on Interactive Multimedia on Mobile & Portable Devices (IMMPD'13), ACM, New York, 2013, pp. 19-24.

[7] F. Kerber, A. Krüger, and M. Löchtefeld, “Investigating the Effectiveness of Peephole Interaction for Smartwatches in a Map Navigation Task,” Proc. 16th Int'l Conf. Human-Computer Interaction with Mobile Devices & Services (MobileHCI'14), ACM, New York, 2014, pp. 291-294.

[8] M. T. Raghunath, and C. Narayanaswami, “User Interfaces for Applications on a Wrist Watch,” Personal Ubiquitous Comput. 6(1), Springer-Verlag London, pp. 17-30, January 2002.

[9] M. D. Dunlop, A. Komninos, and N. Durga, “Towards High Quality Text Entry on Smartwatches,” Proc. Extended Abstracts on Human Factors in Computing Systems (CHI EA'14), ACM, New York, 2014, pp. 2365-2370.

[10] M. Funk, A. Sahami, N. Henze, and A. Schmidt, “Using a Touch-Sensitive Wristband for Text Entry on Smart Watches,” Proc. Extended Abstracts on Human Factors in Computing Systems (CHI EA'14), ACM, New York, 2014, pp. 2305-2310.

[11] Y. Bernaerts, M. Druwe, S. Steensels, J. Vermeulen, and J. Schoning, “The Office Smartwatch: Development and Design of a Smartwatch App to Digitally Augment Interactions in an Office Environment,” Proc. Companion Publication on Designing Interactive Systems (DIS Comapanion'14), ACM, New York, 2014, pp. 41-44.

[12] R. Xiao, G. Laput, and C. Harrison, “Expanding the Input Expressivity of Smartwatches with Mechanical Pan, Twist, Tilt and Click,” Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI'14), ACM, New York, 2014, pp. 193-196.

[13] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns: Elements of Reusable Object-Oriented Software. Boston, MA: Addison-Wesley Longman Publishing Co., 1995.

[14] Android Developers: Transferring Data Using Sync Adapters, http://developer.android.com/training/sync-adapters/index.html

[15] J. Nielsen, Usability Engineering. San Francisco, CA: Morgan Kaufmann, 1993.

[16] S. Ljubic, V. Glavinic, and M. Kukec, “Predicting upper-bound text entry speeds for discrete-tilt-based input on smartphones,” Journal of Interaction Science 2(3), SpringerOpen, pp. 1-15, September 2014.

Improving city infrastructure resilience capacity: water-supply network case study

O. Jukić, I. Heđi, I. Špeh
Virovitica College, Virovitica, Republic of Croatia
{oliver.jukic | ivan.hedi | ivan.speh}@vsmti.hr

Abstract - Main goal of research in this paper is to improve city infrastructure resilience capacity as well as management process of water supply network. Namely, water supply network is located over the wide city area, typically managed by company established for that purpose owned by city government. Water supply network is very important part of the city infrastructure. Hence, it is necessary to ensure high-quality and reliable water supply network management process. Our research is focused on mapping of telecommunications network management principles into water supply network management. Complete network view is graphically and geographically oriented using GPS based network elements coordinates. Real-time network control and management is based on network inventory management system which is presented in our previous work. Data from the real network are included in network management process using telemetry techniques and will enhance network management process. Finally, network model and relevant real-time data are integrated into the complete system that will cover all city infrastructure, acting as infrastructure data interchange portal usable especially in the case of disasters such as flooding, earthquakes, volcanos, etc. Indirectly, citizens will communicate with this platform through different users or services using telecommunications infrastructure.

I. INTRODUCTION

A. Water supply network

Every water supply network is located over the wide city area, typically managed by company established for that purpose owned by city government. Water supply network is very important part of the city infrastructure. In network theory, there are four general classes of networks: technological networks, social networks, information networks and biological networks. The division into classes is a useful one, since networks in the same class are often treated using similar techniques, ideas and algorithms [1]. Both telecommunication and water supply network belong to the class of technological networks. In water supply network almost all network elements are invisible, located in the ground. In most cases, network elements are not accessible through any communication infrastructure, and they are "software-free" elements. Hence, there is no possibility to implement self-detection or auto-discovery features. Hence, it is rather important to implement and maintain network configuration database. Water supply network operators usually have "database" as a collection of different type documents describing network, even in the electronic form. However, that is not enough for maintenance and management activities. There is very strong need for information system for network configuration management, which we call "inventory" management (in water supply network we use term "network inventory" for all elements building water supply network). As a first step in our research we have focused on related work and developing of our own water supply inventory management system [2].

B. Inventory management system architecture

Inventory management system architecture is shown on figure 1.

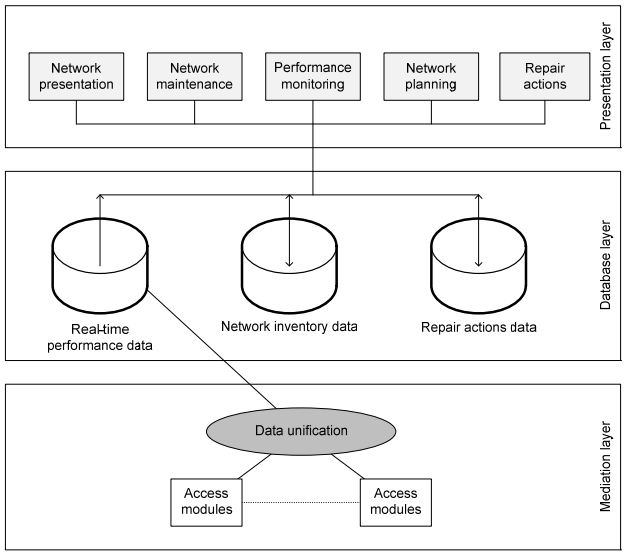


Figure 1. Inventory management system architecture

The system was developed and launched during the project "Knowledge-based water supply management" sponsored by company "Komrad d.o.o.", Slatina. System core is located at database layer. There are three main databases: network inventory database, repairing action database and real-time performance database. Central database in the context of network inventory system is network inventory database. It can be assumed as configuration database. Since there are no auto-discovery or self-detection features in a "software-free" network,

network inventory data must be imported into the database through graphical network presentation interface. Data collection process includes consolidation of all possible sources within the company, such as: network schemas, network descriptions, project plans, repairing logs, internal human knowledge, informal notes etc. Generally, there are two types of network elements: basic elements, such as valves or pipes, that cannot be divided into the smaller parts; compound elements, such as manhole, are compound of number of basic elements. GPS coordinates are a very important part of network inventory data model, since they allow network presentation on any GPS supporting platform [2]. Main functions of mediation layer are:

- data collection from water-supply network
- unification of all data formats
- aggregation of collected data on low level, preparing it for storage in real-time performance data warehouse

All data are collected by specific applications called access modules. Every access module uses specific techniques and performs specific actions in order to obtain data from the water-supply network. For instance, data can be collected through any kind of communication protocol (usually using GPRS for data transport), from existing databases or industrial measurement solutions [6]. Topology for collecting real-time data is shown on figure 2.

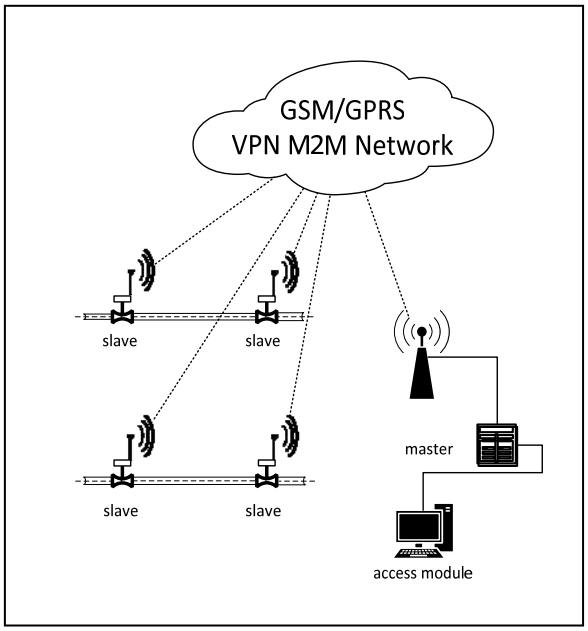


Figure 2. Topology of communication for collecting real-time data

Network presentation GUI is web interface located on the top of database layer. At presentation layer there is performance monitoring module. This is graphical interface which is connected to real-time performance data database and reading last measurement value for each node which consists of slave flow measuring module and serves as "performance cockpit". Basis for this graphical interface is Network inventory data presentation on GPS supporting platform (Figure 3.).

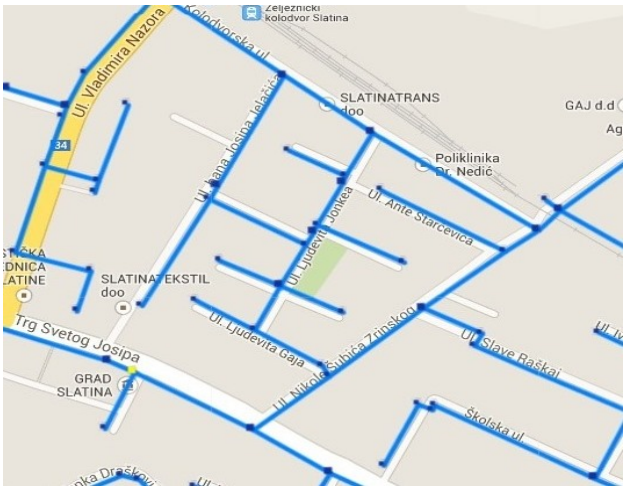


Figure 3. Network inventory data presentation on GPS supporting platform

C. *Related work*

Vienna Waterwork's approach to network maintenance and rehabilitation is described in [3]. Network information system of Vienna Waterwork is briefly presented. System is primarily used to provide assistance platform for planning process. For predicting pipe failures the proportional hazards model (PHM) is used.

GIS-based water supply information system for an urban area is described in [4]. The basic aim of that water supply information system is to serve as a decision support tool for Waterwork's management by helping maintain records of leaks and repairs.

Both systems are focused on inventory management, but currently without GPS coordinates. Real-time data integration is also not considered.

In [5] system using telemetry for obtaining real-time data from water-supply network is described. However, all data are propagated to SCADA application; inventory management is not considered.

All facts mentioned above have encouraged us to continue research and development activities leading to the integration of real-time collected data with network inventory system based on GPS coordinates.

II. PLATFORM TO IMPROVE CITY INFRASTRUCTURE RESILIENCE CAPACITY

A. *Water supply area*

In the water-supply network covered by our inventory management system, water is pumped into the two so-called zones: Low zone covers city area. There is clean water tank above the city - unused water from distribution network is pumped to this clean water tank. If necessary (for instance, if pumps are out of order, or capacity provided is less then capacity needed), water from tank is distributed by gravity to the city. High zone covers neighboring municipalities, with couple of alternative drinking water sources with limited capacity.
Both high and low zone have its own distribution network and also its "access" network where customers are

connected to the network. All data about network inventory are handled by information system described in previous section, providing real-time data collection, therefore enabling not only inventory data collection, storage and visualization, but also real-time network management.

B. Water supply network general architecture

The general water flow in the architectural system shown in Figure 4 is defined as follows:

1. Water is extracted from the sources under the earth. Physical and chemical processing is done.

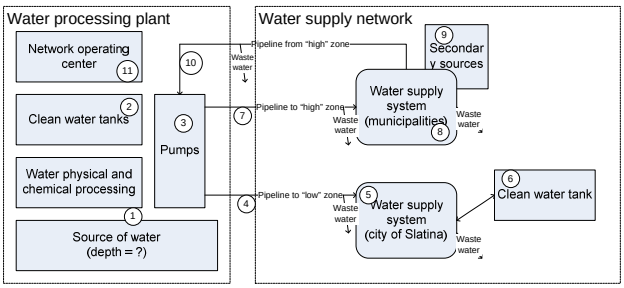


Figure 4. Water supply network general architecture

2. Water is stored into the clean water tanks (3 tanks).
3. Pumps are pumping clean water from the tanks in 2 directions called "Low" and "High" zone. "High" zone is actually city while "Low" zone are municipalities around. Actually, this is the point where water-flow enters the water distribution network.
4. Main pipeline to "high" zone is couple of kilometers length. There is the possibility to waste larger amount of water in short time period, because of huge diameter.
5. Water is distributed all over the city. Unused water is pumped to clean water tank located above the city.
6. If necessary (for instance, if pumps are out of order, or capacity provided is less then capacity needed), water from tank is distributed by gravity to the city.
7. Main pipeline to "low" zone is linking number of municipalities. There is the possibility to waste larger amount of water there, because of huge diameter also.
8. Every municipality has its own distribution network.
9. There are several small secondary sources of water located in municipalities. This water is injected in the general pipeline.
10. Pipeline for "low" zone forms a loop returning water to the water processing plant.
11. Whole process is monitored by Network operating center.

C. Network operating center

The network operating center is located at the "Water processing plant". Company is always guaranteeing personnel in these premises.

The General network operating center architecture is shown in Figure 5.

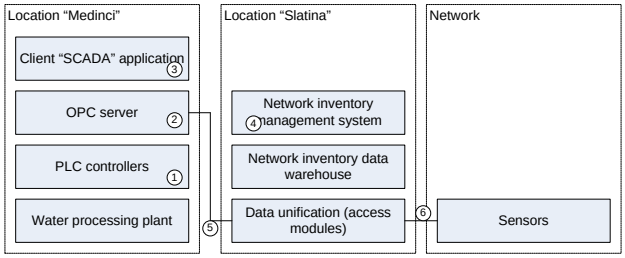


Figure 5. General architecture of network operating center

The network monitoring system includes:

1. Number of PLC controllers collect data from the plant. These data are mostly related to water processing, but there are number of parameters related to water distribution. For instance:
   - Pressure of water flow toward the "Low" or "High" zone
   - Flow of water stream toward the "Low" or "High" zone
   - Current amount of water in clean water tanks
   - Current pump status
   - Current pump power
2. Data from OPC controllers are stored in OPC (Ole for Process Control) server. Data are obtainable by any OPC client (standardized access).
3. The SCADA application, actually it is "RSView", OPC client. There are number of graphical presentations of water processing with integrated real-time data measured by PLC controllers.
4. Network inventory management system is located in the City of Slatina (web application with MSSQL server database).
5. Data from OPC server can be integrated in Inventory management system also, allowing tracking of important KPIs
6. Data from sensors located in network are integrated into Inventory management system also.

D. Central IT platform

Generally Central IT platform is used to collect whole relevant data from different sources such as water supply network operating center, electricity network operating center, gas transmission network operating center, telecommunication network operating center, citizens, 112 service and so on in emergency situation like flooding or some other type of disaster and triggers some relevant actions such as crisis emergency alerting, crisis management actions etc. Currently in test phase we collect relevant data which are connected to distribution of clean water which is very important when flood occur. Next phase of this research is to include all other relevant source and get complete central IT platform to improve city infrastructure resilience capacity.

One potential scenario of using that platform is when floods occur. This platform will be used (not exclusively) by (see Figure 6.): first responders such as police officers, firefighters, emergency medical technicians (EMT), infrastructure operators (water supply, electricity, gas transmission and telecommunications network operators),

general population, city council and its services, service #112, local electronic media, other services, NGOs, etc.

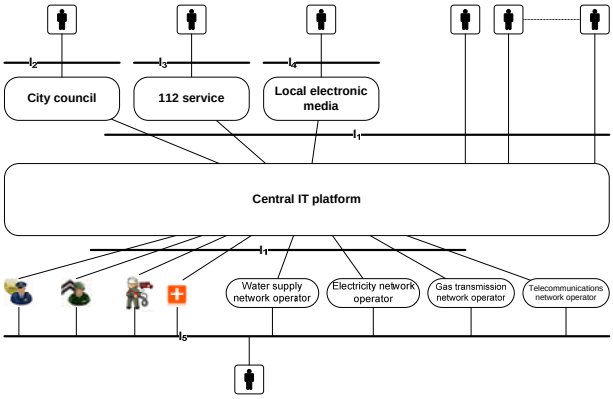


Figure 6. Central IT platform participants (logically)

Users from 1-5 will approach platform in both directions through a platform front-end (**I1**). Term "both directions" means they will be able to trigger some actions, or to receive guidelines, triggers, data etc. from platform. Some data are collected in permanent regime, while other are collected on demand (when something is "triggered" from the environment). Citizens will approach platform directly (**I1**) or indirectly. It is expected that more citizens from urban area will approach platform directly. Citizens from rural area probably will use another channels to send/obtain relevant information related to current disaster situation. Indirectly, citizens will communicate with platform through different users/services using telecommunications infrastructure: by contacting (phone, SMS, e-mail,...) to city council (**I2**), by contacting (phone, SMS, e-mail,...) to service #112 (**I3**), by contacting (phone, SMS, e-mail,...) to local electronic media and by receiving information from media (**I4**), by contacting (phone, SMS, e-mail,...) first responders and/or infrastructure operators directly (**I5**). One possible scenario when floods occur are shown on Figure 7.

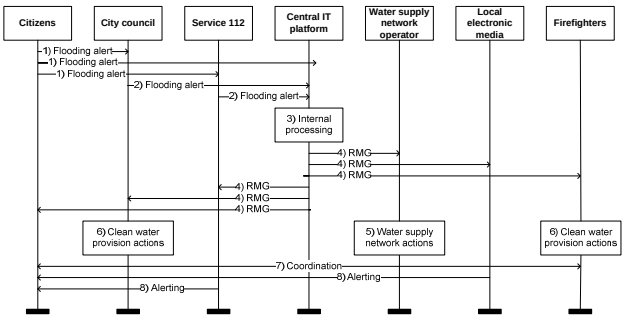


Figure 7. Flooding scenario

The workflow acts as follows: flooding alert is sent from citizens via different channels to city council, Service 112 or Central IT platform. Flooding alert is received and forwarded from the city council and Service 112 to the Central IT platform. Central IT platform performs internal processing, based on data collected on permanent regime, data from other sources, internal algorithms, rules etc. Resilience management guidelines (RMG) are launched and delivered/published to all platform users. For instance, Citizens can approach RMG by web or mobile application. According to RMG received, Water Supply Network Company will perform specific actions on water supply distribution network. For instance: detect areas where clean water distribution is disrupted increase/decrease pump power, switch on/off specific valves in order to cause re-routing of water-flow, fill up clean water tanks, generate specific alerts to citizens and other. At the same time, city council and firefighters will organize clean water provisioning to the areas where water clearance is disturbed. Actions mentioned above are coordinated. Local electronic media will inform citizens about current situation (e.g. special news). Service #112 will generate some public alerts (e.g. public city alarm – siren).

E. Legislation at the area of emergency management

System of protection and rescue of citizens, material and other goods in catastrophes and big accidents, way of controlling, handling and coordinating, tasks and constitutions of bodies for administration, way of alarming and informing in activities of protection and rescue in Republic of Croatia is regulated by the Law on Protection and Rescue. In catastrophes and big accidents district head (city council), mayor and district prefect directly command operative forces of protection and rescue of units from local and regional government. Each person has right to complete and timely information about all threats from the beginning of catastrophes and possibilities, ways, measures and activities of protection and rescue and if they notice threat from beginning of accident, big accident or catastrophes or its beginning, it is their duty to inform Service 112 without any delay, using the most convenient way and the fastest way of information transfer [7]. According to the Law on Critical Infrastructures, national critical infrastructures are systems, networks and objects of national importance whose disruption of functioning or disruption of goods or services supply can have serious consequences on national security, health and lives of people, possessions and environment security and economic stability and undisrupted functioning of government (water supply network, telecommunication network, electricity supply network etc.) [8]. By–law about standard operative procedures regulates the way of common action of operative forces (standard operative procedure) and Service 112, procedure for transfer of all available information connected to the emergency situation, accident or catastrophe after receiving information through number 112, procedure of processing and forwarding received information to operative forces and other participants of protection and rescue using most suitable and the fastest way of information transfer. So, there are some possibilities to implement Central IT platform to Service 112.

III. CONCLUSION

In this paper we have presented information system which is used to improve city infrastructure resilience capacity and it is based on inventory management system in water-supply network. Based on existing experience in telecommunications management area, we have compared basic aspects of management in water-supply and telecommunication networks. Encouraged by the results, we have developed inventory management system for water-supply network. Some of system components are still subject of research efforts; it will be subject of future work and will be presented in future papers. Main goal of research in this paper is to improve city infrastructure resilience capacity as well as management process of water supply network. Namely, water supply network is located over the wide city area, typically managed by company established for that purpose owned by city government. Water supply network is very important part of the city infrastructure. Hence, it is necessary to ensure high-quality and reliable water supply network management process. Our research is focused on mapping of telecommunications network management principles into water supply network management. Complete network view is graphically and geographically oriented using GPS based network elements coordinates. Real-time network control and management is based on network inventory management system which is presented in our previous work. Data from the real network are included in network management process using telemetry techniques and will enhance network management process. Finally, network model and relevant real-time data are integrated into the complete system that will cover all city infrastructure, acting as infrastructure data interchange portal usable especially in the case of disasters such as flooding, earthquakes, volcanos, etc. Indirectly, citizens will communicate with this platform through different users or services using telecommunications infrastructure.

REFERENCES

[1] M.E.J. Newman, "Networks – An introduction", Oxford University Press, Oxford, 2010.

[2] O. Jukić, I. Heđi, "Inventory management system for water supply network", Proceedings of the 37th International Convention MIPRO 2011, pp. 664-667, MIPRO, Opatija, 2014.

[3] D. Kasses, D. Fuchs-Hanusch, "Vienna Waterwork's approach to network maintenance and rehabilitation" Water 21 – Magazine of the Internationa Water Association, United Kingdom, pp. 33-35, February 2013

[4] A.K. Dikshit, P.V. Krishnarao, "A GIS-based Water Supply Information System for an Urban Area", Managing Environmental Knowledge, EnviroInfo, Graz, 2006.

[5] M. Danijel, T. Lisak, G. Malčić, "Telemetrija i upravljanje vodoopskrbnim sustavom", Proceedings Vol. III MIPRO 2009, Computers in technical systems & Intelligent systems, pp. 163-167, MIPRO, Opatija, 2009.

[6] O. Jukić, I. Heđi, "Consideration on real-time network management for water supply network", Proceedings of International conference on applied internet and information technologies, pp. 48-52, Zrenjanin, 2014.

[7] Republic of Croatia, Law on Protection and Rescue, NN 174/4, addition 79/07, 38/09, 127/10, 2004.

[8] Republic of Croatia, Law on Critical Infrastructures, NN 56/13, 2013.

Elektronička komunikacijska infrastruktura u zgradama

Dr.sc. Justina Bajt
Hrvatski Telekom d.d.
Zagreb, Republika Hrvatska
justina.bajt@t.ht.hr

Sažetak - Europska unija puno pažnje u svojim propisima poklanja razvoju elektroničkih mreža velikih brzina. Prema hrvatskim propisima o gradnji zgrada, svaka zgrada mora imati primjerenu elektroničku komunikacijsku infrastrukturu (EKI), koju mora izgraditi investitor prilikom gradnje zgrade, odnosno kasnije (su)vlasnici zgrade. Međutim, brzi razvoj elektroničkih komunikacija i povećanje broja korisnika elektroničkih komunikacijskih usluga u stambenim i poslovnim zgradama zahtijeva da, radi pružanja usluga svome korisniku, operator mora postaviti dodatnu EKI u zgradu. Postavljanje dodatne EKI zahtjeva uređenje odnosa između (su)vlasnika zgrade i operatora po pitanju ulaganja u EKI u zgradama i vlasništva na EKI. Budući da propisi iz područja elektroničkih komunikacija i gradnje ne uređuju u potpunosti ovo područje, pravna rješenja za investiranje u dodatnu EKI traže se u okviru općih propisa. S obzirom na značaj EKI u zgradama za budući razvoj elektroničkih komunikacija, zaključujemo da je ovo područje podnormirano i da je pravnom uređenju EKI u zgradama potrebno posvetiti posebnu pažnju u propisima o gradnji i elektroničkim komunikacijama. Bez omogućavanja operatorima nesmetanog pristupa infrastrukturi u poslovnim i stambenim zgadama radi pristupa do krajnjeg korisnika usluga, nije moguće očekivati značajan napredak u pružanju usluga korisnicima.

I. UVOD

Države članice Europske unije (EU) prihvaćanjem Digitalne agende za Europu (DAE) preuzele su obvezu omogućiti osnovni širokopojasni pristup svim Europljanima do 2013. i osigurati da do 2020. svi Europljani imaju pristup internetskim brzinama većim od 30 Mbit/s te da najmanje 50 % kućanstava u EU bude pretplaćeno na internetske veze brzine iznad 100 Mbit/s. Slijedom postavki DAE, kao sveobuhvatnog plana Europske komisije za poticanje gospodarskog rasta kroz stvaranje konkurentnije i digitalno modernije Europe, Europski parlament i Vijeće donijeli su Direktivu 2014/61/EU o mjerama za smanjenje troškova postavljanja elektroničkih komunikacijskih mreža velikih brzina (SL L 155/11, 23.5.2014.). Na europskoj razini time se osigurava stabilan pravni okvir radi stimuliranja ulaganja u otvorenu i konkurentnu brzu internetsku infrastrukturu i pripadajuće usluge.

Cilj je Direktive 2014/61/EU olakšati i potaknuti uvođenje elektroničkih komunikacijskih mreža (dalje u tekstu: EKM) velike brzine promičući zajedničko korištenje postojeće fizičke infrastrukture i omogućavajući učinkovitije postavljanje nove fizičke infrastrukture[1] kako bi trošak uvođenja takvih mjera bio što niži (čl. 1. Direktive). Direktiva polazi od toga da bi smanjenje troškova postavljanja EKM-a doprinijelo ostvarivanju digitalizacije javnog sektora šireći učinak digitalne financijske poluge na sve sektore gospodarstva uz smanjenje troškova za javnu upravu i pružanje učinkovitijih usluga građanima. Ostvarivanje ciljeva DAE i time ostvarenja općeg interesa države, njezinih građana i gospodarskih subjekata, zahtijeva da uvođenje EKM-a bude čim bliže lokaciji krajnjih korisnika uz istodobno potpuno poštovanje načela proporcionalnosti u pogledu ograničenja prava vlasništva vlasnika zemljišta i zgrada te vlasnika EKI.

Radi smanjenja troškova postavljanja širokopojasnih mreža na cijelom području EU, države članice trebaju stvoriti nacionalni pravni okvir za:

- Učinkovito planiranje i koordiniranje aktivnosti gradnje fizičke infrastrukture, uključujući i fizičku infrastrukturu unutar zgrada te smanjenje administrativnih tereta.
- Omogućavanje pružateljima usluga javnih komunikacijskih mreža pristupa fizičkoj infrastrukturi, uključujući i infrastrukturu unutar zgrada.
- Učinkovito dijeljenje fizičke infrastrukture u koju je, osim EKI uključena i infrastruktura drugih sektora javnih usluga (struja, plin, nafta, odvodnja otpadnih voda, topla voda te voda osim vode za ljudsku potrošnju), a koja je pogodna za postavljanje telekomunikacijskih kabela.

Države članice moraju do 1. siječnja 2016. godine donijeti i objaviti zakone i druge propise potrebne za

[1] Direktiva 2014/61/EU u čl. 2. definira „fizičku infrastrukturu" kao element mreže koji je namijenjen za smještanje drugih elemenata mreže, a da pritom i sam ne postane aktivni element mreže, poput cijevi, stupova, vodova, kontrolnih soba, zdenaca, ormarića, zgrada ili ulaza u zgrade, antenskih instalacija, tornjeva i stupova namijenjenih za pružanje javnih usluga (elektroničke komunikacije, struja, plin, nafta, odvodnja otpadnih voda, topla voda i voda). Međutim kabeli, uključujući svjetlovodne niti koje se ne koriste, kao i elementi mreže koji se koriste za opskrbu vodom namijenjenom za ljudsku potrošnju, nisu fizička infrastruktura u smislu te Direktive.

usklađivanje s tom Direktivom i o tome obavijestiti Europsku komisiju.

II. TEMELJNI ZAHTJEVI DIREKTIVE 2014/61/EU KOJE DRŽAVE ČLANICE MORAJU PRENIJETI U NACIONALNO ZAKONODAVSTVO

U cilju smanjenja troškova i olakšavanja uvođenja EKM-a velike brzine do samih krajnjih korisnika i istodobnog osiguravanja tehnološke neutralnosti, RH mora uskladiti svoje propise sa sljedećim temeljnim zahtjevima Direktive 2014/61/EU koji se odnose na fizičku infrastrukturu u zgradama i pristup toj infrastrukturi:

- Sve novoizgrađene zgrade na lokaciji krajnjeg korisnika usluga, uključujući njihove elemente u zajedničkom vlasništvu, a za koje su zahtjevi za građevinske dozvole podneseni nakon 31. prosinca 2016., moraju biti opremljene fizičkom infrastrukturom, tj. EKI-em unutar zgrade prilagođenom EKM-u velike brzine sve do zaključnih točaka mreže, a sve novoizgrađene višestambene zgrade i pristupnom točkom. Ista se obveza primjenjuje i u slučaju opsežnih radova na obnovi zgrade za koje su zahtjevi za građevinske dozvole podneseni nakon 31. prosinca 2016.
- Ako je zgrada već opremljena EKM-om velike brzine, svakom pružatelju usluga javnih komunikacijskih mreža koji ima aktivnog pretplatnika u zgradi pristup EKM-u zgrade mora biti osiguran pod transparentnim, razmjernim i nediskriminirajućim uvjetima.
- Svaki pružatelj usluge javne komunikacijske mreže ima pravo o svom trošku uvesti svoj EKM do pristupne točke.[2]
- Svaki pružatelj usluge javne komunikacijske mreže ima pravo pristupiti bilo kojoj postojećoj fizičkoj infrastrukturi unutar zgrade (EKI, cijevima za struju, plin, otpadne vode, toplu vodu) radi postavljanja EKM-a ako je udvostručavanje tehnički neizvedivo ili ekonomski neučinkovito.
- Gdje ne postoji ni pasivna ni aktivna infrastruktura prilagođena vezi velike brzine unutar prostorija krajnjih korisnika, niti alternativa za davanje na korištenje EKM-a pretplatniku, svaki pružatelj usluga javnih komunikacijskih mreža treba imati pravo dovesti svoju mrežu do privatnog prostora na vlastiti trošak, ponovno upotrebljavajući, ako je to moguće, postojeću fizičku infrastrukturu unutar zgrade (EKI te cijevi za struju, plin, odvodnju otpadnih voda, toplu vodu), pod uvjetom da utjecaj na privatno vlasništvo bude sveden na najmanju moguću mjeru.
- Svaki pružatelj usluga javne komunikacijske mreže ima pravo završiti svoju mrežu u prostorijama pretplatnika, uz suglasnost pretplatnika, pod uvjetom da se pritom na najmanju moguću mjeru svede utjecaj na privatnu imovinu trećih osoba.
- Svaki nositelj prava koje ga ovlašćuje koristiti pristupnu točku i fizičku infrastrukturu unutar zgrade, mora udovoljiti svim razumnim zahtjevima za pristup podnesenima od strane pružatelja usluga javne komunikacijske mreže pod pravičnim i nediskriminirajućim uvjetima.
- Pravo vlasništva vlasnika pristupne točke ili fizičke infrastrukture unutar zgrade ne smije biti dovedeno u pitanje u slučaju kada nositelj prava koje ovlašćuje na korištenje te fizičke infrastrukture ili pristupne točke, nije njihov vlasnik.
- Ne smije biti dovedeno u pitanje ni pravo vlasništva bilo kojih trećih osoba, kao što su vlasnici zemljišta i vlasnici zgrada, a države članice mogu utvrditi pravila o primjerenoj financijskoj naknadi osobama koje trpe štetu zbog izvršavanja navedenih prava.

Direktiva prepoznaje da ugradnja malih cijevi EKI-a tijekom izgradnje zgrade ograničeno utječe na povećanje troškova, dok naknadno opremanje zgrada cijevima EKI-a može biti značajan udio u trošku postavljanja EKM-a velike brzine. Stoga, od država članica zahtijeva se da sve nove zgrade ili zgrade koje je potrebno opsežno obnoviti[3] trebaju biti opremljene fizičkom infrastrukturom koja bi omogućila povezivanje krajnjih korisnika s EKM-om velike brzine. Također bi takve zgrade trebale biti opremljene pristupnom točkom putem koje bi pružatelj mogao pristupiti EKI-u i EKM-u unutar zgrade (dalje u tekstu za EKI i EKM zajedno koristi se skraćenica: EKMI). Ulagači u zgrade trebali bi ugraditi prazne cijevi od svakog stana do pristupne točke koja se nalazi u zgradi ili izvan nje da bi se omogućilo svakom operatoru javnih komunikacijskih mreža pristup do njegovih korisnika usluga u zgradi.

Budući da postojanje EKM-a velike brzine u zgradi bitno utječe na njezinu vrijednost, radi informiranja potencijalnih kupaca i stanara države članice trebale bi uvesti dobrovoljnu oznaku „pripremljeno za širokopojasni pristup" za zgrade opremljene takvom infrastrukturom i pristupnom točkom. Države članice mogu propisati i određena izuzeća od primjene te obveze, ali ta izuzeća moraju biti odgovarajuće obrazložena.

III. POSTOJEĆE HRVATSKO PRAVNO UREĐENJE ELEKTRONIČKE KOMUNIKACIJSKE INFRASTRUKTURE U ZGRADAMA

Na EKMI-i u poslovnim i stambenim zgradama odnose se sljedeći hrvatski zakoni i podzakonski propisi:

- Zakon o elektroničkim komunikacija (ZEK)[4] te podzakonski propisi doneseni na temelju ZEK-a:

[2] Pristupna točka je fizička točka koja se nalazi unutar ili izvan zgrade te koja je dostupna operatorima javnih komunikacijskih mreža, pri čemu je omogućena veza s infrastrukturom unutar zgrade prilagođenom EKM-u velike brzine.

[3] Opsežni radovi na obnovi prema definiciji iz čl. 2. Direktive znače izgradnju ili građevinske radove na lokaciji krajnjeg korisnika koji obuhvaćaju strukturne izmjene cijele fizičke infrastrukture unutar zgrade ili njezina značajnog dijela i za koje je potrebna građevinska dozvola.

[4] NN 73/08, 90/11, 133/12, 80/13 i 71/14

Pravilnik o tehničkim uvjetima za elektroničku komunikacijsku mrežu poslovnih i stambenih zgrada;[5] Pravilnik o načinu i uvjetima pristupa i zajedničkog korištenja EKI i druge povezane;[6] Pravilnik o svjetlovodnim distribucijskim mrežama (dalje u tekstu: SDM);[7]

- Zakon o vlasništvu i drugim stvarnim pravima (dalje u tekstu: ZV);[8]
- Zakon o gradnji[9] i Pravilnik o jednostavnim građevinama i radovima.[10]

ZEK uređuje dvije situacije gradnje, odnosno postavljanja EKMI-a u zgradama ovisno o tome postavlja li se takva infrastruktura i mreža u zgradi koja se tek gradi (novoj zgradi) ili u već izgrađenim zgradama koje nemaju odgovarajući EKMI (stare zgrade).

A. EKMI u novim zgradama

Investitor nove zgrade dužan je izgraditi EKMI u skladu s Pravilnikom o tehničkim uvjetima za elektroničku komunikacijsku mrežu poslovnih i stambenih zgrada, a EKMI mora biti obuhvaćen projektom zgrade. EKMI zgrade mora omogućiti svim (su)vlasnicima zgrade slobodan izbor operatora, a svim operatorima pristup zgradi uz ravnopravne i nediskriminirajuće uvjete (čl. 24. st. 5. ZEK-a). Iz navedene odredbe proizlaze obveze investitora gradnje zgrade (poslovne, stambene, poslovno-stambene), koje on ima tijekom gradnje, sve dok zgrada ne dobije uporabnu dozvolu. Koji je to minimum EKMI i povezane opreme potreban za namjenu zgrade određeno je glavnim projektom zgrade (dio koji se naziva elektrotehnički projekt), a u skladu sa standardima određenim Pravilnikom, koji moraju biti sukladni normama i tehničkim specifikacijama na koje upućuju obvezujuće zakonske odredbe čl. 24. st. 1., 2. i 3. ZEK-a. Budući da ZEK u čl. 3. utvrđuje da je „elektronička komunikacijska infrastruktura, ... gradnja, održavanje, razvoj i korištenje elektroničkih komunikacijski mreža i elektroničke komunikacijske infrastrukture i povezane opreme … od interesa za Republiku Hrvatsku," izgradnja EKMI u zgradama nije samo privatni interes investitora zgrade, nego je također interes RH, a koji se izražava kroz propisivanje minimalnih standarda EKMI-i određenih vrsta zgrada. Također, strateškim dokumentima RH zacrtan je razvoj širokopojasnih usluga i omogućavanje nesmetanog pristupa svakog građanina i pravnih osoba takvim uslugama, a nužni preduvjet tome je adekvatna EKMI u poslovnim i stambenim zgradama.

Koliki je značaj zakonodavac dao gradnji i korištenju EKMI-i zgrada govore i kaznene odredbe čl. 120. ZEK-a kojima su propisane novčane kazne za prekršaj[11] koji počini pravna osoba, odgovorna osoba u pravnoj osobi i fizička osobe koja ne poštuje odredbe čl. 24. st. 5. ZEK-a. Ako investitor zgrade ne izgradi EKMI primjerenu namjeni zgrade koja omogućava svim vlasnicima zgrade slobodan izbor operatora, a svim operatorima pristup zgradi uz ravnopravne i nediskriminirajuće uvjete, tada investitor čini prekršaj za koji može platiti novčanu kaznu navedenu u čl. 120. ZEK-a.

Dakle, sve zgrade koje su projektirane i izgrađene nakon stupanja toga Pravilnika na snagu 2009. godine, moraju imati svjetlovodni EKMI primjeren namjeni zgrade, a operatori mogu postaviti svoj EKMI samo za potrebe svojega korisnika ako im EKMI zgrade ne omogućava kvalitetno pružanje usluga.

B. EKMI u postojećim zgradama

U ranije izgrađenim zgradama vrlo često ne postoji EKI koji omogućava korištenje naprednih elektroničkih komunikacijskih usluga putem EKM-a velike brzine. Kada se u postojeću zgradu koja ima zastarijeli ili nedostatan EKMI, naknadno postavlja nova infrastruktura potrebna za pružanje novih usluga, primjenjuju se odredbe ZEK-a i ZV-a te Pravilnika o tehničkim uvjetima za elektroničku komunikacijsku mrežu poslovnih i stambenih zgrada. ZEK propisuje da vlasnici izgrađene poslovne ili stambene zgrade moraju omogućiti pristup svim operatorima u svrhu postavljanja, održavanja i razvoja EKMI-a, a izgrađen EKMI mora omogućiti svim vlasnicima zgrade slobodan izbor operatora, a svim operatorima pristup zgradi uz ravnopravne i nediskriminirajuće uvjete. Iz navedenog slijedi da su (su)vlasnici ranije izgrađenih zgrada, u kojima već postoji određen EKMI koji je sastavni dio zgrade i u suvlasništvu je svih (su)vlasnika zgrade, dužni omogućiti svim operatorima pristup toj zgradi, odnosno zajedničkim prostorima u/na/ispod zgrade, a u svrhu postavljanja, održavanja i razvoja EKMI-a (čl. 24. st. 6. ZEK-a).

Ako je u starim zgradama potrebno rekonstruirati, odnosno nadograditi postojeći EKMI zgrade ili postaviti tzv. prijenosni medij u premosnom sustavu, smatramo da je moguće primijeniti sljedeća pravila. Ako se rekonstruira postojeći EKMI zgrade tako da novi EKMI zamijeni dotadašnji EKMI zgrade, tada novi EKMI postaje konstruktivni, tj. neodvojivi dio zgrade. Međutim, ako se pored postojećeg postavlja dodatni EKMI zgrade, takav EKMI može biti odvojen od zgrade i predmet vlasništva pojedinog etažnog vlasnika, odnosno korisnika poslovnog ili stambenog prostora ili operatora elektroničkih komunikacija.

Vlasnici etažnih dijelova zgrade, odnosno korisnici poslovnih prostora (zakupci) i/ili korisnici stanova mogu imati svoj privatni EKM unutar zgrade. Takav EKM i za njega odgovarajući EKI može izgraditi o svome trošku svaki suvlasnik na čijem je suvlasničkom dijelu uspostavljeno vlasništvo određenog posebnog dijela nekretnine (etažni vlasnik) u skladu s odredbama čl. 82. st. 1. ZV-a. Dakle, suvlasnik ne mora za postavljanje takvog EKMI-a tražiti odobrenje od ostalih suvlasnika, nego može u skladu s građevinskim propisima, o svome trošku napraviti prepravke i zadiranje u zajedničke dijelove nekretnine vezano za postavljanje navedenih instalacija i to mu ne može zabraniti niti jedan od suvlasnika. Slijedom navedenog, suvlasnik (etažni

[5] NN 155/09
[6] NN 136/11, 44/12, 75/2013
[7] NN 57/14
[8] NN 91/96, 68/98, 137/99, 22/00, 73/00, 114/01, 79/06, 141/06, 146/08, 38/09, 153/09, 143/12 i 152/14
[9] NN 153/13
[10] NN 79/14
[11] U čl. 120. ZEK-a propisana je novčana kazna za pravnu osobu u iznosu od 50.000,00 do 500.000,00 kuna; za odgovornu osobu u pravnoj osobi u iznosu od 5.000,00 do 50.000,00 kuna; za fizičku osobu u iznosu od 2.000,00 do 20.000,00 kuna.

vlasnik) mogao bi postaviti EKMI za potrebe korisnika njegovog etažnog dijela.

Osim navedenog, dodatni EKMI u zgradi mogu postaviti i operatori elektroničkih komunikacija u skladu s odredbom čl. 14. st. 2. Pravilnika. Navedeno znači da u slučaju primjene različitih tehnologija EKMI-a (bakrene parice, koaksijalni kabeli, svjetlovodi), svaki operator ima pravo doći do svoga korisnika kroz zajedničke prostore u zgradi, postavljajući svoje kabele u zasebne nadžbukne kanalice ili zaštitne cijevi te da mu to (su)vlasnici zgrade (većina ili pojedinačno) ne smiju zabraniti. Dužnost vlasnika da omoguće pristup zgradi s pravnog aspekta ne znači da operator može slobodno, bez ikakvih pitanja i privola ući u tuđu nekretninu i raditi što ga je volja da bi postavio, održavao i razvijao EKMI. Pristup zgradi podrazumijeva pridržavanje određenih općih propisa o vlasništvu i drugim stvarnim pravima, kao i odredaba ZEK-a i Pravilnika.

C. Zakon o vlasništvu i drugim stvarnim pravima

ZV u čl. 82. st. 1. uređuje odnose među suvlasnicima u slučajevima kada bilo koji od suvlasnika želi napraviti nekakve promjene na zajedničkim dijelovima nekretnine, a koje se odnose na postavljanje vodova za svjetlo, plin, energiju, vodu, telefon i slične uređaje, radijske i televizijske antene, odnosno elektroničke komunikacije. ZV izričito određuje pravilo da suvlasnik na čijem je suvlasničkom dijelu uspostavljeno vlasništvo određenog posebnog dijela nekretnine (etažni vlasnik) ima ovlaštenje, ne tražeći za to odobrenje od ostalih suvlasnika, u skladu s građevinskim propisima, o svome trošku izvršiti prepravke i zadiranje u zajedničke dijelove nekretnine vezano za postavljanje navedenih instalacija i to mu ne može zabraniti niti jedan od suvlasnika. Dakle, suvlasnik ne mora za gradnju, odnosno postavljanje dodatnog EKI-a tražiti odobrenje od ostalih suvlasnika, nego može u skladu s građevinskim propisima, o svome trošku izvršiti prepravke i zadiranje u zajedničke dijelove nekretnine vezano za postavljanje navedenih instalacija i to mu ne može zabraniti niti jedan od suvlasnika. Dodatni EKI prema čl. 82. st. 1. ZV-a ne mora suvlasnik sam napraviti, nego to umjesto njega može napraviti operator koji mu pruža elektroničke komunikacijske usluge. Tako postavljen EKI u zgradi prema odredbi čl. 14. st. 2. Pravilnika o EKM poslovnih i stambenih zgrada smatra se dodatnim kabliranjem u zgradi (tzv. postavljanjem „prijenosnog medija u premosni sustav“) jer operator ima barem jednog korisnika u toj zgradi. Operator mora svoje nadžbukne kanalice i kabele postaviti tako da budu u skladu s građevinskim propisima i što manje narušavaju izgled zgrade, te da omogući i drugim operatorima postavljanje njihovih nadžbuknih kanalica. Predstavnik suvlasnika mora svakom operatoru koji ima najmanje jednog korisnika u zgradi dati izjavu, odnosno suglasnost kojom se omogućuje: postavljanje nadžbuknih kanalica ili zaštitnih cijevi i kabela u zgradi, te svakodobni pristup radi njihovog održavanja i korištenja. U slučaju prestanka pružanja elektroničkih komunikacijskih usluga u zgradi operator koji je napravio dodatno kabliranje, dužan je u roku 2 mjeseca po prestanku ugovora o isporuci elektroničke komunikacijske usluge: ukloniti dodatno kabliranje, dovesti dio EKMI-a zgrade u prvobitno stanje, sanirati sva pritom nastala oštećenja u zgradi. Ako operator ne učini navedeno, dodatno postavljene instalacije (dodatno kabliranje) postaju vlasništvo (su)vlasnika zgrade u skladu s odredbom čl. 14. st. 2. navedenog Pravilnika[1].

D. Pravilnik o tehničkim uvjetima za elektroničku komunikacijsku mrežu poslovnih i stambenih zgrada

Pravilnikom je utvrđen minimum EKMI-a koji bi investitori morali zadovoljiti prilikom gradnje poslovnih i stambenih zgrada. Taj minimum EKMI-a zgrade morao bi biti obuhvaćen glavnim projektom zgrade te bi slijedom toga taj EKMI, u skladu s općim propisima o gradnji, bio konstruktivni dio zgrade koji je od nje pravno neodvojiv. Iz općih propisa o gradnji i odredaba ZV-a o originarnom stjecanju vlasništva građenjem proizlazi da bi taj EKMI, kao konstruktivni dio zgrade, nužno pravno bio u suvlasništvu (su)vlasnika zgrade.

Investitor može glavnim projektom zgrade predvidjeti vrlo složen i skup EKMI koja udovoljava najvišim standardima za EKMI određene vrste zgrade. U tom slučaju investitor bi trebao u skladu s glavnim projektom i sagraditi takav EKMI koji također postaje sastavni, tj. pravno neodvojivi dio zgrade. Međutim, iz Pravilnika nije jasno može li investitor u glavnom projektu imati projektiran EKMI koji udovoljava minimalnim standardima propisanim Pravilnikom, a u izvedbi projekta napraviti EKMI koji udovoljava višim standardima. Smatramo da Pravilnik daje dovoljno mogućnosti da se već prilikom gradnje zgrade pored projektiranog EKMI-a izgradi i dodatni EKMI-i (čl. 14. st. 2.). Takav EKMI-i ne mora biti konstruktivni dio zgrade i slijedom toga može imati zasebnu pravnu sudbinu. Nadalje, mišljenja smo da za onaj dio EKMI-a koji investitor može postaviti preko određenog minimuma za određenu vrstu zgrade koju gradi tj. dodatni EKM i/ili EKI koji gradi, odnosno postavlja o svome trošku ili trošku nekog drugog subjekta (zakupca, ortaka, operatora, pojedinog etažnog vlasnika stambenog ili poslovnog prostora) kao tzv. „privatnu mrežu“, ne bi smjelo postojati ograničenje vlasništva na tom EKM-u i/ili EKI-u kako je određeno čl. 14. st. 2. i 3. Pravilnika. Takav EKM i/ili EKI mogli bi biti premosni sustav[12] u skladu s čl. 14. st. 2. Pravilnika koji je u vlasništvu operatora sve dok on pruža elektroničke komunikacijske usluge barem jednom svome korisniku u zgradi, a nakon toga operator mora ukloniti EKMI iz zgrade ili na valjanom pravnom temelju prenijeti EKMI u vlasništvo (su)vlasnika zgrade. Niti ZEK, a niti Pravilnik ne određuju da je premosni sustav u zgradi, predmet zajedničkog korištenja infrastrukture od strane drugih operatora tj. da operator koji je vlasnik EKMI-a u zgradi mora omogućiti korištenje svojeg EKMI-a i drugim operatorima.

Nadalje, iz odredbe čl. 14. st. 3. Pravilnika proizlazi da investitor u EKM-i i/ili EKI i/ili pristupnu kabelsku kanalizaciju zgrade može biti i operator elektroničkih komunikacija. Ako operator investira u gradnju kabliranja zgrade, odnosno EKMI-a zgrade i/ili pristupne kabelske kanalizacije zgrade, pitanje je tko postaje vlasnik EKMI-a i/ili pristupne kabelske kanalizacije. Smatramo da bez

[12] Premosni EKMI u zgradama s više korisnika poslovnih i/ili stambenih prostora definiran je čl. 6. st. 2. t. b. Pravilnika kao: „dio EKMI-a zgrade za potrebe samo određenog korisnika prostora u zgradi izveden premoštenjem zajedničkog EKMI-a“.

obzira što Pravilnik ne uređuje tko postaje vlasnik EKMI-a zgrade, operator ne bi mogao steći vlasništvo na onom dijelu EKMI-a koji je prema glavnom projektu zgrade neophodni minimum primjeren namjerni zgrade i time sastavni dio zgrade. Dakle, operator kao investitor gradnje EKMI-a i pristupne kabelske kanalizacije, mogao bi postati vlasnik samo onog dijela EKMI-a i pristupne kabelske kanalizacije koji prelazi okvire zadane glavnim projektom zgrade. Međutim, tada bi operator kao investitor EKMI-a i pristupne kabelske kanalizacije i (su)vlasnici zgrade svoje odnose morali međusobno urediti odgovarajućim pravnim poslom.

E. Preporuke Hrvatske agencije za mrežne djelatnosti

Hrvatska regulatorna agencija za mrežne djelatnosti (HAKOM) u suradnji s KUSRH – Koordinacijom udruga stanara RH, objavila je 2014. godine na svojim web stranicama brošuru „Svjetlovod dolazi u zgradu…“. Brošura je namijenjena (su)vlasnicima, predstavnicima (su)vlasnik i upraviteljima zgrada i odnosi se na svjetlovodne distribucijske mreže (SDM) te s njima povezane kućne instalacije. HAKOM posebno naglašava da se objavom brošure ne preferira svjetlovodna tehnologija u odnosu na neke druge tehnologije na koje se ne odnose propisi o SDM-u. Dakle, brošura HAKOM-a ne odnosi se na elektroničke komunikacijske instalacije u zgradama koje se sastoje od bakrenih i/ili koaksijalnih kabela, kabela za antenske prihvate, antenskih prihvata. Brošura se ne odnosi niti na one svjetlovodne instalacije u zgradi čiji je model postavljanja u zgradi uvjetovan tzv. zatvorenim modelom svjetlovodne mreže koja je postavljena do ulaza u zgradu ili u blizini zgrade. Naime, mreže koje su građene po modelu zatvorenih (krutih) svjetlovodnih mreža koji se primjenjivao prije stupanja na snagu Pravilnik o SDM-u, trebaju se po tom modelu završiti do stana, odnosno poslovnog prostora korisnika.

F. Temeljne postavke hrvatskih propisa o elektroničkoj komunikacijskoj infrastrukturi i mreži u zgradama

Iz gore navedene analize hrvatskih propisa o EKMI u zgradama koje proizlaze sljedeće temeljne postavke:

- EKI u zgradama može biti sastavni, konstruktivni dio zgrade i tada je u vlasništvu vlasnika, odnosno suvlasnika zgrade. Međutim neki dijelovi EKI-a kao što su npr. nadžbukne kanalice i kabeli u njima mogu biti u vlasništvu operatora elektroničkih komunikacija koji su ih postavili u zgradu radi pružanja usluga svojim korisnicima.
- U skladu s propisima o gradnji, za svaku rekonstrukciju, nadogradnju ili zamjenu postojećih, u pravilu, dotrajalih ili neadekvatnih bakrenih telekomunikacijskih instalacija u zgradi koje su postavljene u zgradu prilikom njezine gradnje i obuhvaćene su projektnom dokumentacijom zgrade, (su)vlasnici zgrade, kao investitori tih radova moraju dobiti izmjenu, odnosno dopunu građevinske dozvole zgrade u dijelu koji se odnosi na elektroenergetske i elektroničke komunikacijske instalacije zgrade (dio glavnog projekta zgrade pod nazivom „elektroprojekt“), tj. novu građevinsku dozvolu zgrade te se nakon njezine pravomoćnosti može početi graditi EKI. Takve nove instalacije konstruktivni su dio zgrade i, u skladu s propisima o vlasništvu i gradnji, vlasništvo su (su)vlasnika zgrade.
- Postavljanje dodatnog EKI-a u zgradi, bez zadiranja u postojeći EKI (ispravne i/ili dotrajalne ili neadekvatne instalacije), ne zahtijeva ishođenje građevinske dozvole zgrade (izmijene glavnog projekta zgrade u dijelu koji se odnosi na elektroenergetske i elektroničke komunikacijske instalacije zgrade).
- Pravilnik o tehničkim uvjetima za EKM poslovnih i stambenih zgrada, koji je na temelju ZEK-a donio HAKOM, detaljno uređuje tehničke uvjete gradnje, odnosno postavljanja EKMI-a u zgradama (novim i postojećim), kao i obveze (su)vlasnika zgrada i operatora koji pružaju elektroničke komunikacijske usluge korisnicima u zgradama. Projektanti zgrada, investitori gradnje poslovnih i stambenih zgrada i (su)vlasnici tih zgrada dužni su u procesu planiranja, projektiranja, gradnje, rekonstrukcije, dogradnje, uporabe i održavanja EKMI-a poslovnih i stambenih zgrada postupati u skladu s Pravilnikom [2].

IV. PRIMJENA RAZLIČITIH TEHNOLOGIJA GRADNJE, ODNOSNO POSTAVLJANJA EKMI-A U ZGRADAMA

Raniji propisi EU-a te najnovija Direktiva 2014/61/EU o mjerama za smanjenje troškova postavljanja EKM-a velikih brzina omogućavaju primjenu različitih tehnologija gradnje, odnosno postavljanja EKMI-a u zgradama (bakrene parice, koaksijalni i svjetlovodni kabeli) te slijedom toga ne diskriminiraju niti jednu od navedenih tehnologija u odnosu na druge tehnologije. I hrvatski propisi omogućavaju primjeni različitih tehnologija gradnje, odnosno postavljanja EKMI-a u zgradama, premda su dosadašnje aktivnosti HAKOM-a pretežito bile usmjerene na razvoj EKM-a po otvorenom modelu SDM-a. HAKOM je pravilnicima uredio da u zgradama koje nemaju izgrađen EKMI zgrade te pristupnu točku, operator koji je vlasnik SDM-a mora svoju mrežu završiti u zgradi i to do svakog stana, odnosno poslovnog prostora, tj. postaviti sve horizontalne i vertikalne instalacije u zajedničkim prostorima zgrade i to po otvorenom modelu postavljanja takvih mreža.

Radi pružanja novih usluga telekom operatori uvode pristupne svjetlovodne i koaksijalnih mreže kao NGA mreže (mreže nove generacije). NGA omogućavaju korisnicima brzine pristupa iznad 30 Mbit/s (a neke od njih omogućuju i pristup brzinama većim od 100Mbit/s). Takve pristupne mreže zahtijevaju uvođenje iste takve tehnologije i u zgrade, kako bi novi medij došao do krajnjeg korisnika. U RH postoji više operatora koji grade/postavljaju nove pristupne mreže i koji biraju tehnologiju postavljanja EKMI-a u skladu sa svojim poslovnim odlukama. Svaka od navedenih tehnologija ima svoje tehnološke posebnosti u smislu postavljanja u zgrade. Prema europskim i hrvatskim propisima niti jedna ne smije biti diskriminirana, odnosno stavljena u

privilegirani položaj u odnosu na druge tehnologije. Stoga u zgradama može biti istodobno primijenjeno više vrsta tehnologija (bakrene parice, koaksijalni kabeli, svjetlovodni kabeli/niti) ovisno o tome koju tehnologiju koriste pojedini operatori koji moraju udovoljiti zahtjevima svojih korisnika usluga. Budući da različite tehnologije postavljanja EKMI-a u zgrade imaju svoje posebne tehničko-tehnološke zahtjeve, ne mogu se kabeli različitih vrsta postavljati u istu nadžbuknu kanalicu ili zaštitnu cijev, nego je u pravilu potrebno za svaku vrstu kabela postaviti zasebne kanalice ili zaštitne cijevi.

Bakrena parična infrastruktura desetljećima je bila jedina infrastruktura koja se nalazi u zgradama i koja je bila prvenstveno prilagođena pružanju govorne usluge. Ta infrastruktura, u pravilu, obuhvaćena je projektom same zgrade te je stoga, sukladno hrvatskim propisima o gradnji i o vlasništvu, sastavni dio zgrada i u vlasništvu je (su)vlasnika zgrade. U RH bakrene parice i danas su najrašireniija infrastruktura koja je dostupna gotovo do svakog kućanstva/korisnika (cca. 1,6 mil.), a u većini stambenih zgrada to je jedina EKI. Nove usluge, koje se u pravilu pružaju u različitim paketima (širokopojasne usluge koje uključuju pristup internetu, usluge distribucije televizijskog signala), zahtijevaju znatno veće brzine prijenosa od onih koje su bile potrebne isključivo za prijenos telefonskog signala, odnosno govorne usluge. Stoga, da bi operatori mogli pružiti kvalitetne usluge svojim korisnicima, ZEK i ZV omogućavaju im postavljanje potrebnih instalacija do svakog njihovog korisnika.

Koaksijalni kabel je prijenosni medij „kabelske televizije“ koji korisnicima omogućuje korištenje svih multimedijskih usluga kao što su HD/3D televizijski programi, korištenje najbržeg interneta te usluga fiksne telefonije koji su sastavni dio modernih širokopojasnih mreža. Koaksijalni kabel danas je vrlo raširen prijenosni medij jer ima vrlo dobre prijenosne karakteristike i nisku cijenu. Najčešće se koristi kao završni prijenosni medij u sustavima kabelske televizije i privatnim kućanstvima povezivanjem satelitskih i zemaljskih antena s televizijskim prijemnicima i računala. Također, koristi se kao medij za analogni i digitalni prijenos podataka i različitih vrsta sadržaja. Koaksijalni kabeli su prijenosni medij u HFC mrežama (Hybrid Fiber Coax).

Svjetlovod je prijenosni medij sastavljen od više vrlo tankih svjetlovodnih niti načinjenih od stakla ili plastike koje provode svjetlosni signal te se omogućuje 100 puta veće brzine u usporedbi sa sadašnjim bakrenim mrežama (primjenom ADSL-a). Svjetlovodne mreže omogućuju izuzetno visoku kvalitetu prijenosa podataka na velike udaljenosti, gotovo bez slabljenja signala i sa znatno manjom potrebnom energijom te su neosjetljive na elektromagnetske smetnje. Na području RH postoje dva osnovna tipa svjetlovodnih pristupnih mreža: mreže koje su postavljene prije stupanja na snagu Pravilnika o SDM-u, te mreže koje se postavlja prema Pravilniku o SDM-u. Funkcionalna pravila za svjetlovodne mreže koje se postavljaju prema Pravilniku o SDM-u, opisana su u HAKOM-ovoj brošuri. Mreže koje su postavljene prije stupanja na snagu Pravilnika o SDM-u, tzv. kruti P2MP (točka-više točaka) model trebaju se završiti u zgradama po starom krutom modelu jer se ne može promijeniti konfiguracija izgrađene pristupne mreže od centralnih lokacija do zgrada korisnika te se u skladu s time moraju postaviti i kućne instalacije (EKMI zgrade). Model izgradnje kućne instalacije (premoštenja-vertikala) na postojećoj FTTH mreži, podrazumijeva izradu kućne instalacije do ispred svakog stana, odnosno ugradnju i vertikalnih i horizontalnih kanalica u zgradi, ugradnju glavnog razvodnog ormara i etažnih razvodnih ormara, ugradnju splitera, povlačenje instalacijskih kabela u vertikalama te priključnih kabela do ispred svakog stana. Izgradnja opisane kućne instalacije zahtjeva zasebni premosni sustav (sustav kanalica i glavnog razvodnog ormara), koji nije pogodan za smještaj drugih tehnoloških rješenja (druge vrste kabela).

V. ZAKLJUČAK

Prema odredbama ZEK-a, Pravilnika o EKM u poslovnim i stambenim zgradama i ZV-a moguće je, osim EKMI koji je sastavni i neodvojivi dio zgrade, imati i EKMI koji je „privatna mreža“ određenog korisnika usluga kao premosni sustav u vlasništvu operatora koji pruža usluge u toj zgradi. Dakle, „privatne mreže“ korisnika prostora, odnosno operatora mogu biti u vlasništvu osoba različitih od (su)vlasnika zgrade. Iz Pravilnika nije potpuno jasno koji dijelovi EKM-a i/ili EKI-a su obvezatni dijelovi infrastrukture zgrade, tj. moraju biti dio glavnog projekta zgrade te su kao sastavni dio zgrade vlasništvo (su)vlasnika zgrade, a koji se dijelovi EKM-a i/ili EKI-a mogu naknadno postaviti i biti kao „privatna mreža“, odnosno premosni u vlasništvu drugih subjekata na temelju nekog stvarnopravnog ili obveznopravnog odnosa.

Radi smanjenja troškova za postavljanje EKM-a velike brzine u zgradama Direktiva 2014/61/EU omogućava korištenje i drugih vrsta fizičke infrastrukture u zgradi, a što nije uređeno hrvatskim propisima. Stoga bi trebalo mijenjati, odnosno donijeti odgovarajuće propise kojima bi se hrvatske propise uskladilo s navedenom Direktivom tako da se omogući korištenje EKI i druge fizičke infrastrukture u zgradama, uz poštivanje načela neutralnosti tehnologije i nesmetanog pristupa operatorima do korisnika usluga.

LITERATURA

[1] J. Bajt, “Građanskopravno uređenje elektroničke komunikacijske infrastrukture u zgradama,” Liber amicorum in honorem Vilim Gorenc, Pravni fakultet Sveučilišta u Rijeci, Zavod za građansko pravo, 2014., pp. 298–318, (izvorni znanstveni rad)

[2] J. Bajt, „Elektroničke komunikacijske instalacije u zgradama“, Novi informator, broj 6335 od 22. studenog 2014., pp.8.

ALGORITHMS & PROTOCOLS

Integrating Extreme Programming with Software Engineering Education

Jason Jen-Yen Chen* and Mike Mu-Zhe Wu*
* National Central University, Jhong-Li, Taiwan
jasonjychen@gmail.com

Abstract - This paper presents an 11-step "myAgile" method that integrates the extreme programming (XP) with existing university courses, attempting to build a modern software engineering education. Compared with XP, "myAgile" covers more phases, such as requirement engineering, detailed design, and reverse engineering. There were 288 students in 4 universities in Taiwan who participated in this effort. The students rigorously applied the method to develop, and to maintain, an object-oriented Grade System project in Java. The results showed that, objectively, the students produced maintainable software, which was clearly shown in another maintenance project by a different team at a distant time after the development. And, subjectively, students reported high job satisfaction that is so desperately needed in this profession.

I. INTRODUCTION

The extreme programming (XP) [1] is really a powerful agile method, which has been applauded by software engineers around the world, especially after the announcement of the agile manifesto [2] in 2001. Its 12 XP practices [3] bring us very useful engineering techniques. Note that software development involves both engineering and management aspects. The most widely-used agile method SCRUM [4], on the contrary, provides good management techniques. Because this work focuses on software engineering, rather than software management, SCRUM will not be further discussed here.

On light of this trend, we see an urgent need to integrate XP with existing courses, such as Object-oriented Analysis, Data Structure, Algorithm, Java Programming, Software Engineering, Software Testing, Software Engineering Environment, and further, to build a modern software engineering education. The endeavor brings about an 11-step XP-based "myAgile" method, which seems to be a simple and elegant way to do it.

This paper reports this "myAgile" method and its recent usage experience in Taiwan. The paper is organized as follows. First, the steps of the XP-based "myAgile" method are described. Next, a Grade System project is depicted. The students rigorously applied the method to develop the project. Then, a project questionnaire is used to collect students' feedback. Last but not least, a maintenance project highlighted this method.

II. "MYAGILE" METHOD

" myAgile" stresses face-to-face communication, which is of paramount importance in any agile method. All the steps of "myAgile" are thus conducted in "pair programming". That is, two developers sit side by side, one being the driver, while the other the reviewer. And they can switch role at any time. This ensures that very intense inter-human communication occurs in every single step of software development in order to obtain the highest possible software quality.

The 11 steps of the "myAgile" method are depicted below in Sections A to K. The input and output of every single step is described through detailed examples. Note that steps 5, 6, 9, and 10 are the original XP steps. Steps 0, 1, 2, and 3 are adopted from Software Engineering, especially Object-oriented method. And steps 4, 7, and 8 are specially designed by author Professor Chen. Overall, this method covers 3 phases of software process: 1) requirement engineering (steps 0, 1, 2), 2) architectural design (steps 3, 4, 5, 6), and 3) detailed design (steps 7, 8, 9, 10).

A. Step 0: Exploring Requirements

This step gathers the features (user stories) of the system by a pair of two customers. Since this step will be conducted in the customer site, rather than in the development site, it is omitted from the student project that will be reported in this paper.

B. Step 1: Scenarios

aUIpromptID or quit
The user wants to quit
aUIshowFinishMsg

Figure 1. (a) A simple scenario

aUIpromptID or quit
The user inputs ID
aUIshowWelcomeMsg

aUIpromptCommand:
1)show grade 2)show rank 3)update weights 4)exit
The user wants to exit

aUIpromptID or quit
The user wants to quit
aUIshowFinishMsg

Figure 1. (b) A complex scenario that extends the simple scenario

For each user story, an "on-site customer" and a developer sit side by side, jointly develop various scenarios - from simplest case gradually extending to complex cases in a step-wise refinement manner. For the

user story “manage grade systems”, Fig.1 (a) shows a simple scenario, while Fig. 1 (b) shows a complex one, which extends the simple one. In Fig. 1 (a), “aUI” is an object of the UI class. And, “promptID” and “showFinishMsg” are two public methods in the UI class. Note the boldfaces in Fig. 1 (b) shows that they originally come from Fig. 1 (a). If the developers are quite familiar with this, the step-wise refinement can be relaxed, meaning that the number of scenarios, as well as the development effort, can be reduced.

C. Step 2: Acceptance Test Cases and User Manual

For each scenario, add the exact data (expected system output data and user input data) to it. By doing so, a scenario turns into a testable acceptance test case. After all the cases are done, they are summarized to obtain a simple user manual that quickly gives users a feel of what the software is all about. Fig. 2 shows the acceptance test case related to the scenario in Fig. 1 (a). In Fig. 2, the system outputs “Enter ID or Q (Quit)?” and the user inputs “Q” to quit (note that “$” in “$Q” denotes it is a user input). After that, the system outputs “Finished” meaning that the system is finished.

```
aUIshows : Enter ID or Q (Quit) ?  $ Q
aUIshows : Finished
```

Figure 2. An acceptance test case

D. Step 3: CRC Session

For each acceptance test case, the on-site customer and 4 developers (2 pairs), totally 5 people, sit around a table to conduct a class-responsibility-collaborator (CRC) session [5] that traces a case to extract objects’ names and object interactions’ names. The objects’ names can easily deduce their class names, while object interactions names become public method names (responsibilities) for the class. Also, collaborator classes are identified during the trace. A class name along with its public method names make up a class interface. And the system architecture is composed of all these class interfaces. This really is an agile way to do Object-oriented analysis. Last but not least, don’t forget to add a class header on top of each and every class interface for documentation purpose. Fig.3 shows an example of class interface with class name C1 and two method interfaces, namely, a constructor C1 and another method m1 with two parameters p1 and p1 and a return class C2. Of course, there is a class header on top of it to document it.

```
/*-----------------------------
    Here is the class header
-----------------------------*/
public class C1 {
        publicC1 ();
        public C2  m1 (p1, p2);
} //end class C1
```

Figure 3. A class interface C1

E. Step 4: Reverse Engineering Tool

The Java class interfaces can be automatically processed by a reverse engineering tool to get graphical documents, such as UML class diagram [6]. Currently, the tool eUML2 [7] is used for the project. The developers really do not need to spend time anymore to draw the diagrams and further to verify them. It is guaranteed that the diagrams are consistent with the source code. What a relief!

This step is much more important for maintenance projects than for from-scratch projects. In this case, the maintenance engineers should first get a hard copy of the class diagram by using the tool. Then, carefully use pencil and eraser to modify the diagram and to trace the modified diagram against acceptance test cases through pair programming (1 pair) - or group review (2 pairs) if a case is a bit complicated. After the modification is done, modify the class interfaces and class headers next. This maintenance process has been field verified by the maintenance project in the Central University, as described in ”the maintenance” section later in this paper.

F. Step 5: Dispatching and Scheduling

After step 3 CRC session, the class interfaces are available and are posted on the white board in the software company. Each class interface is “picked-up” by a pair of developers based on the pair’s experience and preference. To be exact, this is not dispatching work. Instead, it is picking-up work. About scheduling, a pair themselves decides the number of days they need to finish a class, again, based on their experience plus a time allowance. Like XP, we use whiteboard to record and control the dispatching and the scheduling of programming tasks.

G. Step 6: Unit Test Code

Every public method (called a unit in XP) in a class interface needs to develop a number of test cases. And, each case is a combination of input parameter values of the method. For example, if a method “m1” has two input parameters, each with 2 values. Then, there are (2 times 2) 4 cases to test for that unit. Incidentally, it is normal as well as healthy to test quite a few cases for a unit. For example, if the first parameter has 5 values and the second 8, then there would have been (5 times 8) 40 cases to test. Next, develop JUnit [8] test code for each test case. Briefly, what test code does is to assert equality of: 1) hand-calculated expected result and 2) the actual result by executing the method. Fig. 4 illustrates the test cases and the simplified JUnit test code for method “m1”.

```
m1 test cases
test case 1: p1 with value1   p2 with value1
test case 2: p1 with value1   p2 with value2
test case 3: p1 with value2   p2 with value1
test case 4: p1 with value2   p2 with value2

public m1Test1 () {
  1. construct an object of C1 called obj1
  2. hand-calculate expectedResult of executing m1(value1, value1)
  3. actualResult = obj1.m1(value1, value1)
  4. assertEqual (expectedResult, actualResult)
public m1Test2 () { ……}
public m1Test3 () {…….}
public m1Test4() {…….}
```

Figure 4. An illustration of unit test code

H. Step 7: Data Structure Design

In object-oriented programming (OOP), presumably the public methods of a class work on the common data structure of that class. Furthermore, the data and the methods are subtly inter-related. That is, if a data structure is rather abstract, then the algorithms will be relatively simple. On the contrary, if low-level concrete data structures, such as arrays, are used, then the algorithms will get complicated. In this case, the source code will get longer, more complex, and thus more difficult to maintain.

We use a graphical document called "design sketch" to record the data structure along with a simple set of input data. The design sketch helps the developer to figure out how to get the output data from the input data, thus help develop build the problem solving thinking. It is required in this method to use as high-level data structures as possible, such as those in the Java collection framework (JCF).

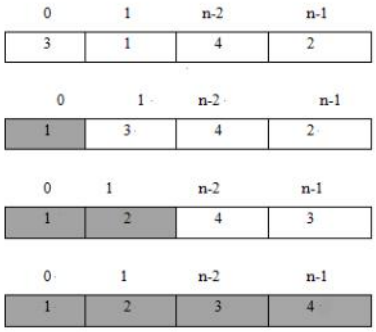


Figure 5. A design sketch of SORT [9]

For illustration purposes, however, we use a low-level data structure "array" here. Fig. 5 (reproduced from [9]) shows a design sketch with an array data structure, along with the input data "3, 1, 4, 2". And the task is to sort "3, 1, 4, 2" to "1, 2, 3, 4" . Note that the task is to sort n data. But, the design sketch intentionally shows only 4 data. The rationale behind this is to simplify problem solving thinking and pseudo-code tracing that follows next.

The sketch shows that we select the minimal number or "min" (1), from a list "3, 1, 4, 2" and swap "min" with the leftmost number (3). Next, we do the same to a smaller list without the leftmost number, that is "3, 4, 2". Repeat this until the list has been reduced to only two numbers in it, that is "4, 3". Amazingly, we now have a sorted list "1, 2, 3, 4". Notice that the array index 0, n-2, n-1 in the sketch will help the developer a lot in the algorithm design next. For example, how many iterations are needed? This question is easy to answer with the help of the sketch. The sketch in Fig. 5 clearly shows that the first iteration works on the list starting at index 0, and that the last at index n-2 (see the lists in white background). This gives us the idea of designing a loop from 0 to n-2, as shown in the pseudo-code in Fig. 6 (b).

When it comes to Java Programming, we use the Java data structure classes of the Java collection framework (JCF) [10] such as ArrayList, LinkedList, TreeMap, TreeSet, HashMap, HashSet, and PriorityQueue. Notably, the methods of these classes are all equipped with Big O time estimate. This enables development teams to get worst time estimates for all the methods they developed, which is so desperately needed by the customer!."

I. Step 8: Algorithm Design

Following the previous step, we write down the problem solving thinking in the natural language text called "pseudo-code". This usually takes several rounds from a very abstract level down to the algorithm level. Fig. 6 (a) shows a high-level abstract pseudo-code of SORT, while Fig. 6 (b) shows its low-level pseudo-code that is the algorithm. As just mentioned, the loop indexes specified in Fig. 6 (b) such as 0 and n-2 can be traced all the way back to the design sketch in Fig. 5!

```
1.select minfrom initial list,
  and then swap it with the leftmost number.
2.select min from smaller list (without the leftmost number),
  and then swap it with the leftmost number
3. repeat (2) until the smaller list contains only two numbers
```

Figure 6. (a) A high-level pseudo-code of SORT

```
fori from 0 upton-2
    1.min points to leftmost number of array[i..n-1] (that is i)
    2select min from array[i..n-1] (that is smaller list)
        forj fromi+1 upton-1
            if the number at j is less than that at min, let it be min end if
        end for
    3.swap the number at min with the leftmost number (that is array[i])
end for
```

Figure 6. (b) A low-level pseudo-code of SORT

One word about functional decomposition seems worthy of mentioning here. Each public method usually contains a big chunk of functionality, which should be decomposed into smaller, simpler lower-level private methods. This is actually the technique of structured programming (SP) [11]. SP says that, by using the 3 structured constructs of: 1) sequence, 2) selection, and 3) repetition, go to-less method can be constructed. Such a method has only one entry and one exit with subroutine calls in it. In reading such methods, there is thus no jumping around with undisciplined go to statements. Therefore, each method can be understood as one independent conceptual unit. If needed, you can read the details of a subroutine, which is another conceptual unit in its own right that can also be separately read and understood. Ideally, at the end, a subroutine should be less than 10 source lines, which makes code very readable, and thus highly maintainable.

One subtle thing about pseudo-code writing is that, for non-English speaking students, it is not easy at all to write pseudo-code in simple, correct English. Our solution to this problem is a technique called "Chinese sentence with English terms". That is, the students embedded English terms, which later are likely to become class names, object names, methods names, and so on, inside Chinese sentences. Students thus only need to develop right English terms, they no longer have to worry about English writing. This technique really helped the students.

One last thing about this step is that once the pseudo-code is done, the students need to manually trace it against a simple set of input data. With the abstraction provided by pseudo-code, the tracing would not be as difficult as tracing source code. And, if tracing is done in a "slow but firm" manner, all bugs will go away, even before coding starts! Nevertheless, it surely takes a lot of practice and quality-consciousness to do it right."

J. Step 9: Coding

Coding becomes an easy job in the "myAgile" method. All you need to do is just follow the logic of pseudo-code, and add Java source code beneath it.

K. Step 10: Unit Testing and Acceptance Testing

Unit testing refers to running the test code of a unit to test the source code of the unit. In XP, unit testing is equivalent to integration testing as will be explained next. The reason for this is that we continuously (one unit at a time) unit-test and integrate the units into the system in a bottom-up manner, thus making the low-level units naturally integrated with the unit currently being tested. This is called "continuous integration" [12], which turned out to be a powerful feature of XP.

When all the units of a user story (feature) are integrated, the on-site customer manually tests the user story against the acceptance test cases. That is, the on-site customer reads the test cases one by one and manually run the system accordingly. This is the acceptance testing of XP. The Software Testing course is thus integrated into this method. All the 11-steps of "myAgile" have been covered. Next, we will compare XP with "myAgile".

L. Comparing XP with myAgile

Compared with XP, "myAgile" shows 3 merits:

- Requirement engineering

  Trying to be as light-weighted as possible, XP pays little attention to requirement documentation. This certainly hurts acceptance testing, which needs to execute acceptance test cases to make sure that the system features can be accepted. To remedy that, " myAgile" requires on-site customer to develop scenarios from simple ones to complex ones in a step-wise refinement manner. Further, with exact data being added to the scenarios, they gradually evolve to become acceptance test cases. And, all these documents are in electronic file. All these measures greatly facilitate the evolution of requirement documents. Requirements change so rapidly that they need to be well documented and managed.

- Detailed design

  XP does not cover detailed design of classes at all. XP relies on qualified developers to design good classes. This is not enough! Instead, "myAgile" uses design sketch and pseudo-code to design data structure and then algorithm in a step-by-step way. By doing so, multiple abstraction levels of software can be constructed, which facilitates maintainers to quickly move from one abstraction level to another during maintenance process. Incidentally, this approach naturally integrates two required computer science courses, namely, "Data Structure" and "Algorithm", into this method.

- Reverse engineering

  One drawback about traditional software engineering is that design charts are usually inconsistent with the ever-changing source code. This leads to great difficulties in maintenance. Thanks to reverse engineering tools, we now have solution for it. You don't need to hand-draw design charts in the first place. Whenever you need to read design charts, just get them automatically by using a reverse engineering tool. Furthermore, it is guaranteed that the design is consistent with the code all the time. Next, we will describe the Grade System project.

III. THE PROJECT

We have designed a Grade System project for the students to apply the "myAgile" method. The reason why we chose this project is that a developer had to be very familiar with the problem domain, and this project indeed works on the domain that students are most familiar with.

This Grade System allows a user (student) to get his or her total grade and rank. Total grade is based on the weights, which can be updated. Rank denotes the order of the total grade of a student in the whole class.

Input file: The scores of all the students in a class. For example:

962001044 Lin, John 87 86 98 88 87

962001051 Lee, Wen 81 98 84 90 93

Note that the data field names for the above are:

ID name lab1 lab2 lab3 midTerm finalExam

The user menu to input command looks like this:

Input command 1) G (show Grade)
2) R (show Rank)
3) W (update Weight)
4) Q (Quit)

The students were given a 15-page project document [13] that includes:

- Scenarios (4 normal ones plus 2 exceptional ones)
- Simplified input file for tracing (two lines only)
- Acceptance test cases
- CRC design
- Architecture design in class interface, along with detailed design in design sketch and pseudo-code
- Real input file (about 60 lines)
- XP practices used in the project

As just mentioned, we gave the students 6 scenarios along with 6 acceptance test cases. And we ask them to add 2 more cases. We did CRC session for the students so that all of them work on the same architectural design, which is composed of 4 classes, namely, "Main", "UI", "GradeSystems", and "Grades". The data structure used in class "GradeSystems" is Java "LinkedList". The pseudo-codes of all the methods are given to students. And, an exemplary JUnit test code of a method is given too. Also, a simplified input file with just two lines (see the input file above) is provided for students to trace the algorithms. One-line input file cannot do it because executing a loop requires two lines at least. Of course, a real input file with about 60 lines is provided later for acceptance testing.

Regarding the development process, the students are required to use pair programming to develop this system in Java. With regard to the software engineering environment, the students must use:

- Eclipse [14] to develop easy-to-read Java source code, including header and pseudo-code
- JUnit to develop Java unit test code
- eUML2 to automatically generate UML class diagrams

Every method should not exceed 10 source lines. If too long, break it down into low-level private method. Every public method is required to have Big-O worst time estimate. At the end, turn in 3 documents:

- User manual & Acceptance test case in .doc file
- Source code in .java file
- Unit test code in .java file

The project grading will be based on readability of the documents just mentioned. And, a questionnaire will be distributed to collect students' feedback about using the "myAgile" method. Currently, we are particularly interested in how "Pair programming" and "Continuous integration" two important XP practices were going in this project.

The project has been successfully conducted in 4 universities in Taiwan, namely, Central University, Tsing-Hua University, Tam-kang University, and Taichung University of Education. The project duration varies from one university to another. In Central University, first, it is a one-month development project, followed by about one month "unlearn" period during which the students were told to forget about the project, and then it is a one-week maintenance project that maintains a project developed by other classmates. Next, we will describe the project questionnaire and its results.

IV. QUESTIONNAIRE

The questionnaire merely collects students' feedback about the method through the project. Category A question checks how well the students understand every single step of the method. Category B checks how well they use two XP techniques: Pair Programming [15] and Continuous Integration. And, Category C asks the students what impresses them most, and how their teams overcome communication problems, especially with foreign students in the team.

The project was conducted in 2013 in 4 universities in Taiwan. For administrative reasons, only 2 universities joined in 2014. There were totally 288 students who participated in it. All the students are new to "myAgile". The result of the questionnaires is summarized in Tables 1 and 2. As mentioned before, step 0 was omitted. Table 1 shows that Step 4 "reverse engineering tool" is the step that most students do not understand, regardless of university, with low scores 21, 73, and so on. This may result from the lack of free academic license of the tool. About the two focused XP techniques (Pair Programming and Continuous Integration), the students of Central University understand them better than those in other universities. Perhaps this is due to the teaching effort by Professor Chen at Central University.

TABLE I. THE PERCENTAGE OF STUDENTS WHO UNDERSTAND STEPS 1 THROUGH 10

	Step	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2013	TsingHwa U.	95	98	48	21	68	77	98	98	97	90
	Central U.	98	100	81	73	90	96	90	83	100	98
	TamKang U.	87	67	42	22	60	56	76	71	69	73
	Taichung U.	100	96	71	31	43	78	84	63	80	86
2014	TsingHwa U.	100	95	65	40	90	75	90	90	100	70
	Central U.	100	96	89	79	89	96	96	92	94	96

TABLE II. THE PERCENTAGE OF STUDENTS WHO UNDERSTAND THE TWO XP PRACTICES

	XP practice	Pair programming	Continuous integration
2013	TsingHwa U.	76	63
	Central U.	90	85
	TamKang U.	67	62
	Taichung U.	75	69
2014	TsingHwa U.	77	70
	Central U.	83	78

Regarding the communication problems in a team, as quoted by a foreign student in Tsing-Hwa University who is from Central America, "There is no communication problem in our team at all!" It is a relief to know that language is not really a big problem, as long as everyone in the team is capable of speaking simple English.

V. THE MAINTENANCE

As said earlier, the project was conducted in 4 universities in Taiwan. Among them, the Central University finished the development project earlier than others, around mid-term exam. Therefore, Prof. Chen decided to conduct another maintenance project just before final exam.

The idea is that there are so many quality attributes, such as efficiency, usability, reliability, or maintainability. The authors strongly feel that maintainability is the most important one. The reason for this is that under the assumption of agile culture, the world is changing rather than static - and changing fast. With this, software

absolutely should be highly maintainable to cope with it. Further, with good maintainability, other quality attributes can be improved. For example, if usability is not good enough, then the user interface part should be maintained to fix this. Here, maintainability comes in. Therefore, the authors decided to run another maintenance project to examine the maintainability of the developed software.

The maintenance project is conducted as below. Right after the students finished the development project around mid-term exam, they were asked not to read, talk, or think about the project. This is the "unlearn" period trying to erase all the human memories about the project. This simulates the real world situation that a maintenance project usually comes many years after the original development project. Until two weeks before the final exam, Professor Chen randomly assigns each project to another different team to start the maintenance project.

The maintenance task mainly adds a user command "M (modify grade)" to the project. The maintenance team needs to read all the documents to get a full understanding of the architecture of the project. With the help of headers and reverse engineering tool, this can be quickly done. Then, the pair has to carefully modify the document (including source code) in a top-down manner. Along the way, they will use the original test code (some might need maintenance too) to assist continuous integration of the new system.

On the final exam day, the students stood up in front of the entire class to report their maintenance experience. And, the result was astonishing! Many pairs reported that they at first expected to finish the job in a week. But, it turned out that they finished it in one night! Because all the projects follow the same "myAgile" method to develop, it thus is easy for maintainers to read all the documents. They enjoyed the maintenance very much! There was a sense of high job satisfaction in the classroom that day. What a feeling! The authors sincerely hope that this passion can help the students through the inevitable tough times in their careers ahead.

Further, this is the first time for most students to read other people's code. They showed appreciation of other people's work. That is, they learned from others. This can lay foundation of the "respect" value of the agile culture.

VI. CONCLUDING REMARKS

The agile method, especially the engineering aspect provided by extreme programming (XP), has brought about a profound change to the software engineering profession, making it more enjoyable and more quality-conscious. The XP-based "myAgile" method integrates XP with existing courses to build a modern software engineering education. There are totally 11 steps in this method. Compared with XP, "myAgile" shows 3 merits:

- Requirement engineering

  It requires on-site customer to develop scenarios from simple cases to complex cases in a step-wise refinement manner.

- Detailed design

  It uses design sketch and pseudo-code to design data structure and then algorithm in a step-by-step way.

- Reverse engineering

  Design charts are automatically obtained, and guaranteed to be consistent with code all time.

A Grade System project is designed for students to practice the method in 4 universities in Taiwan. The post-project questionnaire showed that the students understood most steps, except for step 4 "reverse engineering tool", probably due to the lack of academic license. Also, it showed two focused XP practices, "pair programming" and "continuous integration", were well-understood by most students. In Central University, another maintenance project was conducted, which revealed that the developed Grade Systems were highly maintainable. Amazingly, some students showed high job satisfaction, which is so rare, yet so desperately needed, for software engineering profession.

REFERENCES

[1] Beck, K. (1999). Extreme Programming Explained: Embrace Change. Prentice Hall.

[2] Fowler, M., Highsmith, J. The Agile Manifesto [Online] Available: http://www.pmp-projects.org/Agile-Manifesto.pdf/.

[3] Jeffries, R. What is Extreme Programming? [Online] Available: http://www.xprogramming.com/.

[4] Schwaber, K., Beedle, M. (2002). Agile Software Development with Scrum. Prentice Hall.

[5] Beck, K. (1989). A Laboratory for Teaching Object-Oriented Thinking, In: Proceedings of OOPSLA'89.

[6] Booch, G., Rumbaugh, J., Jacobson, I. (2005). The Unified Modeling Language User Guide, 2nd ed. Amazon.

[7] Soyatec .eUML2. [Online]. Available: http://www.soyatec.com/.

[8] GitHub. JUnit. [Online]. Available: http://www.junit.org/.

[9] Chen, J. J-Y., Wu, M. M-Z. (2013). myAgile: A XP-based Method for Modern Programming Education, In: Proceedings of International conference on Advances on Computer Science and Information Technology, p. 6-10.

[10] Collins, W. J. (2011). Data Structures and the Java Collections Framework. 3rd ed. Wiley.

[11] Structured Programming. [Online]. Available: http://www.en.wikipedia.org/.

[12] Duvall, P. M., Matyas, S., Glover, A. (2007). Continuous Integration: Improving Software Quality and Reducing Risk, 1st ed. Addison-Wesley.

[13] Chen J. J-Y. Grade System Project. [Online] Available: http://www.agilemethod.csie.ncu.edu.tw/.

[14] The Eclipse Foundation. Eclipse. [Online]. Available: https://eclipse.org/.

[15] Lui K. M., Chan K. C. C., Nosek J. T. (2008). The Effect of Pairs in Program Design Tasks, IEEE Transactions on Software Engineering 32(2) 197-211./.

Sensors Definitions and Positioning for Load Balancing Algorithms

Branko Radojević*, Mario Žagar**, Ivana Golub*
* Croatian Academic and Research Network (CARNet), Zagreb, Croatia
** Faculty of Electrical Engineering and Computing (FER), Zagreb. Croatia
Branko.Radojevic@CARNet.hr, Mario.Zagar@fer.hr, Ivana.Golub@CARNet.hr

***Abstract*- Building a complex load balancing mechanism requires good knowledge of network and networking metrics, all parts of computer systems that participate in the service, and especially sensors and sensor metrics that are providing important inputs to the mechanism algorithm. It is essential to identify potential sensors that can be used together with their specific metrics, and to understand on what degree those sensors and metrics can and may influence decisions of load balancing algorithms. The analysis of potential candidates, expected, and observed data that those sensors are returning together with their correlations with final decisions will be presented in this paper. To have the right data, and to understand the crucial correlations is essential in building a new advanced load balancing algorithm that can proactively influence load balancing decisions or reactively change decision in handling critical situations.**

I. INTRODUCTION

As discussed in previously published work [1] the intention is to define a new load balancing algorithm that will have better insight of shared resources, not only based on ingress (input) and egress (output) network traffic. Instead it will have much thorough overview of the entire computer system that is serving clients in order to avoid unwanted non-optimal balancing decisions leading to undesired outages. The load balancing decisions will be made in centralized software module called CLBDM (Centralized Load Balancing Decision Module), external to hardware load balancers, which will collect complete set of data, run load balancing algorithm, and feed the load balancers with balancing decisions.

Beside the CLBDM, there is also need to install agents on all servers in system, which will collect raw data from sensors, adjust them based on certain specifics of the local system, and send data to CLBDM.

In this case load balancers are barely routing the network traffic, not making any decisions.

It was also discussed in [1] that some of possible sensors beyond network traffic might be CPU usage and memory. The novel approach of this paper is to identify most indicative sensors together with their placement and identification of the expected numerical results of the sensors and their relevance or impact in building load-balancing decisions.

After identification of possible sensors a number of simulations were done in lab, which showed that there are correlations between sensors and final results. Moreover, impact factors of all discussed sensors were identified, quantified and built in the final load-balancing algorithm.

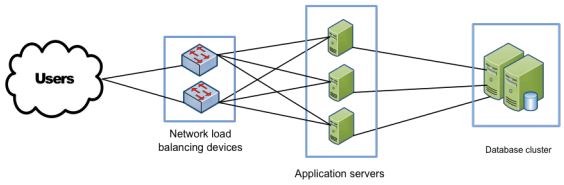


Figure 1: Schematics of typical high-availability computer system with hardware load balancers.

Computer system on which the tests were run is shown on Figure 1., and it is considered to be one of the most common setups that can be found nowadays.

II. DEFINITION OF SENSORS

The obvious starting point in identifying parts of computer system that can provide valuable information for load-balancing decisions was to observe statistics from network side of the system. As discussed in [1][2] it was already proven that data collected from network sensors, although exclusively used by other algorithms, could not be the only source of information for the advanced load-balancing system.

Beyond networking, sensors like CPU usage, memory usage, disk subsystem utilization, and number of concurrent users were considered.

All sensors will work on 5 seconds averages, to remove possible spikes, and not to overload the system with too much sensor reading activities.

Networking

Network sensors are obvious choice for all load-balancing algorithms, but should be used together with other sensors to overcome possible unwanted load-balancing decisions.

Beyond usual data that can be collected about the ingress and egress traffic load, other type of sensors can also be used.

The choice of sensors depends of the role of monitored server in computer system. For example front facing web/application servers are having one set of important sensors, while on the backend database servers completely different sensors and values will trigger our load-balancing decisions.

Another good candidate for sensor is the network latency. For example, if latency is detected by the software agent residing on web server on it's connection toward the database server, it might indicate issues on the database server(s) if the networking issues are eliminated.

All data that is produced by the agent and reported to the CLBDM must be adjusted to common set of values.

CPU usage

One of the most indicative sensors is CPU usage gathered from all components of the system. Values received from this sensor can clearly indicate the unwanted situations. One of the toughest decisions is to set triggering levels in configuration. It was observed that depending of the CPU type and installed operating system this value can differ, and at certain combination 90% CPU usage is acceptable, but 92% becomes an issue. On other systems this value can be at 95%. It is even more difficult to detect an issue where there are more cores in the CPU (and that is usual today), and received data even more depends on the operating system that is reporting it. So those data should also be adjusted to common set of values.

Memory

Memory sensor is somewhat similar to CPU usage sensor, and it also heavily depends on the operating system in use. Some operating systems (like Linux) cache parts of storage in memory, and it reports almost always that the memory is completely used, while the actual usage is different. Sensor should be aware of the real memory usage by having the knowledge of the host operating system and the way how it reports real memory consumption.

Disks

Sensors that observe disk status are also relevant and important for best load-balancing decisions, especially on database servers. There are many values that can indicate overload on the server, like the unusual traffic patterns, high disk usage and increased latency in disk responses. All of this can indicate many different situations, for example that disk subsystem is really overloaded or that disk system has imminent failure approaching. This sensor should be written in such a way that these possibilities are identified and observed. It also depends on the type of database setup if load-balancer can do anything to mitigate situation when disk subsystem becomes overloaded.

Number of http requests

One of the good indicators that something has gone wrong is unexpected rise of the http requests. This sensor might also be considered as one of the networking sensors, but it was isolated because it is better to pull this information from web server than from a networking subsystem where http requests can be aggregated (some network accelerators aggregate multiple http request to a single one to offload traffic from web servers). While the number of http request on the web server might still be aggregated, for the sake of sessions they must be uniquely identifiable – and counted for.

Number of users and sessions

One of the most observed sensors is a number of users and sessions. In ideally balanced system this number should be equal across all front facing web/application servers. Although it might not be a sensor on which to base decisions, it is important input for CLBDM to track current status of users and servers.

III. Adjusting Sensor Values on Agent Modules

As discussed before there is a need to make adjustments on agent (sensor) side in order to provide more unified data to CLBDM. It was generally decided that all sensors indicating utilization will be adjusted to fit range of values from 0 to 100 for a 5 second averages.

One of the most difficult sensors to adjust is a CPU usage sensor. In practice, it depends on the underlying operating system, number of processors and cores installed in the server.

Since most of modern applications and modern operating systems are handling multiprocessor and multicore environments well, it was decided that the overall CPU utilization will be calculated as aggregated usage of all cores divided by the number of cores to get the uniform numbers.

The aggregated number that have just been computed in this way will be then adjusted by the local agent, aware of the local operating system and caveats of its CPU reporting and handling typical values. Their values are adjusted in a way that ranges from 0 to 69 indicate normal operation, from 70 to 89 indicate intensive CPU operation and above 90 indicate overloaded processor.

Figure 2. shows how the typical values are adjusted by the agent.

Adjustment is calculated by the relatively simple formula that takes as one of the inputs a coefficient of CPU saturation (C_{cs}) that has been measured in lab, and then it is adjusted for sending to CLBDM. Here is the equation used:

$$U_c = \frac{U_r C_a}{C_{cs}} \qquad (1)$$

In this equation, C_a is a coefficient for adjustment, in this case we have decided that it is fixed to 90% to match previously decided scale. The C_{cs} is coefficient that depends on operating system and processor types. U_r is utilization as it is read from the system. It was observed in lab and the values are shown in Figure 3.

OS	CPU#	Cores	Per core utilization	Aggregated utilization	Adjusted value
Windows Server 2008	1	1	60	60.00	58.06
	1	4	38/26/33/41	34.50	33.39
	2	4	77/78/77/68 78/56/66/65	70.62	68.34
Linux 3.12	1	1	60	60.00	54.55
	1	4	38/26/33/41	34.50	31.36
	2	4	77/78/77/68 78/56/66/65	70.62	64.20
Windows Server 2012 R2	1	1	60	60.00	56.25
	1	4	38/26/33/41	34.50	32.34
	2	4	77/78/77/68 78/56/66/65	70.62	66.20

Figure 2: Adjusted values for some typical CPU setups and readings.

OS	CPU Saturation level
Windows Server 2008	93%
Linux Kernel 3.12	99%
Windows Server 2012 R2	96%

Figure 3: Values for CPU saturation for listed operating systems.

Beside the adjustment of CPU Usage values, it is also important to get correct information about real memory usage, since some of modern operating systems use the rest of the free memory as a dynamic cache, so the reported free memory might seem to be consumed, while in fact it is still available. The AM (Agent Module) needs to calculate exact amount of consumed memory and report to CLBDM in a percentage of total memory. Partial memory listing from a Linux based web server in Figure 4 shows that of 1.8 Mbytes of memory, only about 0.2 Mbytes is free, while in fact only about 1 Mbytes is consumed by the system and applications, and the 0.6 Mbytes is consumed by cache.

```
MemTotal:        1737364 kB
MemFree:          214568 kB
Buffers:          312484 kB
Cached:           653208 kB
SwapCached:         2444 kB
Active:          1041344 kB
Inactive:         354548 kB
Active(anon):     359332 kB
Inactive(anon):    70944 kB
Active(file):     682012 kB
Inactive(file):   283604 kB
Unevictable:           0 kB
Mlocked:               0 kB
```

Figure 4: Part of memory listing from a Linux based web server

When it's time to calculate the free memory for Linux systems, it is correct to sum consumed memory and deduct sum from the total available memory, than to blindly believe in number that system reports as a free memory. Free memory can be calculated with following formula:

$$M_{free} = M_{max} - (M_{sys} + M_{cache} + \cdots + M_{user}) \qquad (2)$$

From the disk subsystem it is possible to get a number of information about the current status, like the number of read/write operations, average read/write times, number of IOPS (I/O Operations), etc. Figure 5 shows a typical report from the Linux kernel.

```
202   0 xvda 37765 30651 1276141 151404 8059257
26650486 277721941 684447000 6 18829820 684595580

202   1 xvda1 3144 12263 15741 6068 1389 119 50621
125440 0 10568 131508

202   2 xvda2 33951 18348 1254720 141604 8057223
26648766 277653352 684290148 0 18820160 684428928

202   3 xvda3 642 40 5456 3700 645 1601 17968
31412 0 5700 35112

202  80 xvdf 1512 1622 25162 5476 31513 11837
346816 77940 0 23144 83416

202  96 xvdg 4813 4292 75194 13376 53535 171714
1801992 2168668 0 50712 2182036

202 112 xvdh 101 340 798 100 0 0 0 0 100 100

202 113 xvdh1 74 340 582 60 0 0 0 0 60 60
```

Figure 5: Part of disk subsystem report from a Linux based web server

It is a little bit more difficult to adjust this type of raw data to fit need of CLBDM, but it is possible if data are tracked through the time. One of the important information here is the number of IO operations, bolded and underlined in each row. This number shows the number of operations in last second. As discussed in [4] the number of maximum IOPS for single hard disk depends on several factors (block size, disk type, …), but generally following values can be found in Figure 6.

Disk Type	Max IOPS
HDD SATA, 3 Gbit/s, 7200 RPM	≈150
HDD SAS, 6 Gbit/s, 15000 RPM	≈350
SSD SATA, 3 Gbit/s	≈1770

Figure 6: Maximum number of I/O operations on typical drives

There is additional difficulty if the drives are in the RAID array, or if the server is attached to external storage, then the number of maximum IOPS must be carefully calculated and fed into the Agent Module. For final normalization before reporting to CLBDM, following equation is used:

$$U_d = \frac{n_r}{n_{max}} * 100 \qquad (3)$$

where n_r is number of actual IOPS, and the n_{max} is a estimated maximum of IOPS for the attached disk subsystem.

IV. Results from Simulations

After identification of potential sensors, their normalization in the Agent Modules, a number of simulations were conducted in lab to prove correlations between sensors and the real status of the system.

The lab was setup as a virtual copy of the typical computer system setup as shown in Figure 1. Additional resources were added to play the role of the users (external to this system), and additional tools were used to simulate overloaded web/application servers. All simulated problems were conducted on one of the application servers, while others were not affected. Figure 7 shows the type, and number of simulations conducted.

Simulation type	Test ID	Seq. number
CPU overload, >99%	C1	1 – 200
Network overload, >99%	N1	201 – 400
Disk overload, >99%	D1	401 – 600
Combined, CPU + DISK	CD	601 – 800
Combined, all	CA	801 - 1000

Figure 7: Type of simulations conducted together with seq. numbers

Figure 8 shows the result of conducted tests and correlations with identified sensors.

	Simulations conducted				
Sensors observed	C1	N1	D1	CD	CA
Network traffic	0.18	99.9	0.27	4.3	97.7
Network latency	98.8	97.8	0.02	88.5	98.4
CPU usage	100.0	67.3	58.8	88.3	100.0
Memory usage	48.4	99.4	73.4	99.2	97.4
Disk usage	38.1	42.1	100.0	95.3	99.3
Number of http req.	0.0	99.9	0.2	0.0	34.3
Number of users and sessions	0.0	99.9	0.0	0.0	42.1

Figure 8: Correlation table for identified sensors and conducted lab simulations

The correlation table shown in Figure 8 shows correlations in percentage, meaning that 100% has perfect correlation between the sensors and the conducted tests, while the number close to 0 shows that there is no direct correlation detected for that sensor/test pair. Correlations were calculated by increasing overload factors during tests and observing results in sensor values.

First conducted test is the CPU overload test (C1) and its impact on the sensors on the web server. Test was repeated 200 times. To simulate high CPU usage command line tool called `stress` was used, and for reading of the results a number of tools like `vmstat`, `mpstat` and `htop`. All command were invoked from scripts to easily handle repeating the test, and saving results in the file after each of iteration, together with average results. It was observed that during all iterations the results were repeating, so the averages taken were almost exact with a single iteration.

CPU overload test has shown that it has high impact on the CPU sensor, and Network latency sensor. It has medium impact on the Memory usage and Disk usage sensor, while it has no impact on other sensors. The results were inline with expectations, and it shows that it has high correlation between impact on the system and sensors.

All other tests were conducted in similar manner, and the results showed correlations between certain sensors and the impact on the system. Some test showed results that were more fluctuating between the iterations, but the fluctuations observed were about +- 10%, what was accepted as satisfactory and the averages were used.

VI. Related work

Load balancing in cloud and virtual environments is becoming increasingly important in recent years, and the number of authors working on the similar models is growing up [3].

Nuaimi et al. [5] presented a performance survey of several load balancing algorithms in cloud computing. Another study on dynamic load-balancing techniques was conducted by Randles, Lamb, and Taleb-Bendiab [7].

Kuhn and Sesum-Cavic in [6] proposed their solution called SILBA (Self Initiative Load Balancing Agents) which exchanges pluggable algorithms to select the best algorithm for a certain problem and condition.

Bhadani and Chaudhary [8] proposed solution similar in concept with [1], offering a central decision module called Central Load Balancing Policy for Virtual Machines (CLBVM) to balance the load evenly in a distributed VM/cloud computing environment and migrates particular VM to lightly loaded server if its performance is affected by other VMs.

T-Y et al.[9] proposed a method to find an algorithm to minimize the data duplication and redundancy. The algorithm proposed is called INS (Index Name Server) and it integrates de-duplication and access point selection

optimization. There are many parameters involved in the process of calculating the optimum selection point. Its feeble points are: No forecasting algorithm to identify the future behavior of the nodes and it is complex in terms of implementation.

Xiaona Ren et al.[10] considers the unique features of long-connectivity applications, which are increasingly popular nowadays in cloud computing. An improved algorithm is proposed based on the weighted least connection algorithm. In the new algorithm, load and processing power are quantified, and single exponential smoothing forecasting mechanism is added.

While a number of authors are conducting similar research, there are no papers that are directly related to novel approach presented in this paper.

VII. CONCLUSION

In this paper, basics of advanced centralized load balancing approach were shown. In order to get better overview of the entire computer system that is serving services to end-users, it is imperative to understand exactly key performance indicators from the entire system, before making load-balancing decisions.

Following that, a number of key sensors were identified, and for some of them it was shown how the values were adjusted before importing them in the CLBDM, because of the differences between operating systems and how they handle information about key performance indicators. With this additional information CLBDM will have better overview of the entire system, and by using it's own internal algorithms feed load-balancing decisions to the balancers. While the algorithms are beyond scope of this paper it is intended to have the ability to store multiple algorithms as CLBDM sub-modules and to choose most appropriate load-balancing algorithm on the CLBDM on case-by-case basis.

To check expected correlations between identified sensors and key performance issues in the system, exactly 1000 simulations were conducted in the lab environment, divided in five different simulation types. The results were shown in the figures, together with the correlations between sensors and simulations. The results from simulations confirmed that by observing a number of sensors across the system it is easy to identify a number of performance issues across the system, and to influence load-balancing decisions in accordance with received results. It also shows that observing only networking sensors cannot detect the impact that conducted tests made on the system.

All of these prove that this approach has capability to detect difficulties that were hard to detect before, and to influence load-balancing decisions through specialized centralized module that has more thorough overview of the entire system serving services to the end users.

LITERATURE

[1] B. Radojevic, M. Zagar, "Analysis of Issues with Load Balancing Algorithms in Hosted (Cloud) Environments", Proceedings of 34th MIPRO, pp. 125-129, 2011.

[2] M. Simjanonska, S. Ristov, G.Velkoski, M. Gusev, "L3B: Low Level Load Balancer in the Cloud", Proceedings of the IEEE EuroCon 2013, ISBN 978-1-4673-2232-4/13

[3] Zhong Xu, Rong Huang, "Performance Study of Load Balancing Algorithms in Distributed Web Server Systems", CS213 Parallel and Distributed Processing Project Report.

[4] White Paper: Fujitsu Primergy Server Single Disk Performance, http://globalsp.ts.fujitsu.com/dmsp/Publications/public/wp-single-disk-performance-ww-en.pdf

[5] K. Nuaimi, N. Mohamed, M. Nuaimi, and J. Al-Jaroodi, "A survey of load balancing in cloud computing: Challenges and algorithms," in Network Cloud Computing and Applications (NCCA), 2012 Second Symposium on, 2012, pp. 137–142.

[6] E. Kuhn and V. Sesum-Cavic, "A space-based generic pattern for selfinitiative load balancing agents," in Proceedings of the 10th International Workshop on Engineering Societies in the Agents World X, ser. ESAW '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 17–32.

[7] M. Randles, D. Lamb, and A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing", Proceeding 24th IEEE International Conference on Advanced Information Networking and Applications Workshops, pp. 551–556, 2010.

[8] A. Bhadani and S. Chaudhary, "Performance evaluation of web servers using central load balancing policy over virtual machines on cloud," in Proceedings of the Third Annual ACM Bangalore Conference, ser. COMPUTE '10. New York, NY, USA: ACM, 2010, pp. 16:1–16:4.

[9] T-Y., W-T. Lee, Y-S. Lin, Y-S. Lin, H-L. Chan and J-S. Huang, "Dynamic load balancing mechanism based on cloud storage" in proc. Computing, Communications and Applications Conference (ComComAp), IEEE, pp:102-106, January 2012.

[10] Xiaona Ren, Rongheng Lin & Hua Zou, "A Dynamic Load Balancing Strategy for Cloud Computing Platform Based on Exponential Smoothing Forecast", 978-1-61284-204-2/11, IEEE, 2011.

Design of the BCJR Decoding Algorithm with Reduced Space Complexity

Dejan Spasov
Faculty of Computer Science and Engineering,
Skopje, Macedonia
dejan.spasov@finki.ukim.mk

Abstract - Given an *M*-state (recursive) convolutional encoder, we show that, in theory, computing the forward alpha probabilities of the BCJR decoding algorithm can be done with 2*M* memory elements. Building on this idea we propose new design with reduced space complexity for the original BCJR algorithm. Initial experiments with rate-1/2 1025-bit-long Turbo Codes show possibility for compressing the size of memory for the alpha probabilities for about 97%.

I. INTRODUCTION

Given a length-*n* convolutional code with *M* states, the goal of the Bahl-Cocke-Jelinek-Raviv (BCJR) decoding algorithm is to provide Maximum A Posteriori (MAP) probability estimate L_i for each information bit y_i in the received sequence [1]. For a long time this algorithm has been considered of no practical importance due to its higher computational complexity compared to the Viterbi algorithm. However, the widespread popularity of Turbo codes sparked research interest on the decoder as well [2] - [4]. Many variations of the decoder have been proposed in attempt to lower the space or time complexities. One line of research that originated from [2], divided the trellis diagram in equally sized windows and then the BCJR algorithm is run within each window independently [2], [5]. Another popular algorithm that achieves memory reduction is the M-BCJR algorithm, in which certain states in the trellis diagram are considered dead an set to 0, thus reducing the memory requirement [6]-[8].

In this paper we addresses the issue of memory reduction in unoptimized BCJR algorithm. In Section II we start with the assumption that all forward alpha probabilities have to be stored and propose model with space requirements that do not depend on the code length *n*, but only on the number of encoder states *M*. Reduced space complexity comes at the cost of increased time complexity; thus, additional (but parallelizable) computations must be performed in the backward stage of the algorithm. In Section III we propose practically realizable BCJR decoder with reduced memory. However, due to the finite precision of computer arithmetic, this decoder cannot achieve the complexity of the theoretical model. In Section IV we use parallel concatenated Turbo Codes with RSC encoders to perform simulations with our modifications to the BCJR algorithm. We have observed memory reduction for 97% the forward α probabilities at the expense of acceptable 0.1 dB loss of the error correction capability of the Turbo system.

II. THEORETICAL FOUNDATIONS

In binary case, the Maximum A Posteriori (MAP) probability estimate at the time *i*, L_i, is given as logarithm of the ratio of two probability functions, i.e. when the observed bit is 1 and when the observed bit is 0 [9], [10]

$$L_i = \log \frac{\Pr\{y_i = 1 \mid observation\}}{\Pr\{y_i = 0 \mid observation\}} \qquad (1)$$

It is common that convolutional decoders are envisioned as to be moving through a trellis diagram and storing the necessary information in the vertices of the trellis. *Trellis diagram* is labelled *n*-partite graph, in which every path represents a valid codeword (fig. 1). Vertices of the *n* disjoint sets in the trellis represent all possible 2^l states of the encoder.

In this environment, the BCJR algorithm is envisioned as two stage process. In the first stage, known as *the forward α recursion*, the algorithm is moving through the trellis in left to right fashion and with each state *s* at the time *i* it associates probability function $\alpha_i(s)$ that is recurrently computed from its predecessors $\alpha_{i-1}(s')$ and the branch information $\gamma_i(s, s')$. In the second stage, known as the *backward β recursion*, the algorithm recurrently computes additional probability function $\beta_i(s)$ using different initial conditions and similar equations. Then, at the time *i* it combines $\beta_i(s)$

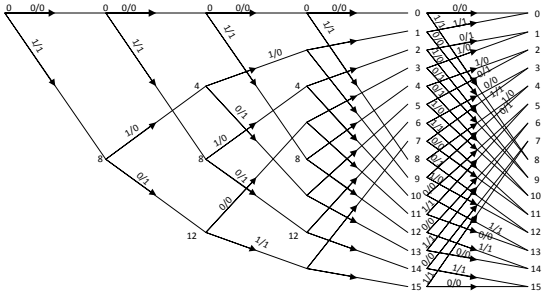


Figure 1. Trellis diagram of a 16-state convolutional code

with pre-stored $\alpha_i(s)$ and outputs (1).

The computation of the forward α probabilities starts from the initial conditions at the time i=0

$$\begin{cases} \alpha_0(s)=1 & s=0 \\ \alpha_0(s)=0 & s\neq 0 \end{cases} \tag{2}$$

and following the edges of the trellis with non-zero branch probabilities $\gamma_i(s,s')$, the decoder, at each iteration stores all $\alpha_i(s)$ and computes $\alpha_{i+1}(s)$, according to

$$\alpha_{i+1}(s')=\sum_s \alpha_i(s)\gamma_i(s,s'),\ c\ . \tag{3}$$

The computation of the backward β probabilities starts from the initial conditions at the time i=N

$$\beta_0(s)=1/N \quad \forall s \tag{4}$$

and following the edges with non-zero branch probabilities $\gamma_i(s,s')$, the decoder at each time i computes $\beta_i(s)$, according to

$$\beta_i(s')=\sum_s \beta_{i-1}(s)\gamma_i(s,s'),\ s'\in\{0,..,M-1\}\ . \tag{5}$$

During the backward stage our goal is to recompute $\alpha_{i+1}(s)$ from $\alpha_i(s)$ thus to eliminate the need to store all $\alpha_i(s)$ in memory. We will consider (3) as system of M linear equations with M unknowns, i.e $\alpha_i(s)$. Solving this system will give us the values $\alpha_i(s)$.

Let assume that with the forward recursion we have arrived at the time i=N-1 and we have computed all $\alpha_{n-1}(s)$. Next, we compute L_{n-1} and in parallel with the β recursion, we recompute all $\alpha_{n-2}(s)$ from $\alpha_{n-1}(s)$. We compute L_{n-2} and repeat this for all information bits.

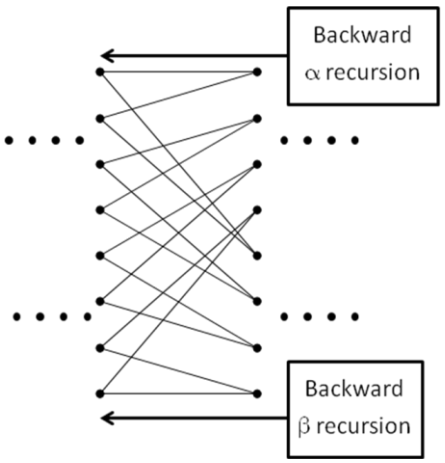


Figure 2. Backward computation of the $\alpha_i(s)$ probabilities

Thus with the forward recursion we only need to find all $\alpha_{n-1}(s)$, without storing any intermediary $\alpha_i(s)$. Then, during the backward computational stage we include additional processor (fig. 2) that backward-recursively computes $\alpha_i(s)$ from $\alpha_{i+1}(s)$.

We should note that this description of the BCJR algorithm can be applied on arbitrary block code.

A. BCJR Decoding of Convolutional Codes

In theory, any algorithm that solves system of linear equations can be used in the backward stage to solve (3). However, this approach may impose quadratic slowdown in the backward stage. Convolutional codes have certain regularity that helps in solving (3) faster. Convolutional codes over an alphabet of q elements have regular trellises that are made of M/q subgraphs known as butterflies. Thus, in the case of convolutional codes, (3) can be split into M/q independent systems of linear equations with q unknowns.

Example of a butterfly with corresponding channel and state probabilities for a convolutional code over binary alphabet is given on figure 3. The system of linear equations that corresponds to the butterfly is given with

$$\begin{aligned} \alpha_{i+1}(u)&=\alpha_i(x)\gamma_i(x,u)+\alpha_i(y)\gamma_i(y,u) \\ \alpha_{i+1}(v)&=\alpha_i(x)\gamma_i(x,v)+\alpha_i(y)\gamma_i(y,v) \end{aligned}\ . \tag{6}$$

Solving (6) for the unknowns $\alpha_i(x)$ and $\alpha_i(y)$, we obtain the expression that has to be computed in the backward decoding stage. Due to the symmetry in computations we use

$$\alpha_i(x)=\frac{\alpha_{i+1}(u)\gamma_i(y,v)-\alpha_{i+1}(v)\gamma_i(y,u)}{\gamma_i(x,u)\gamma_i(y,v)-\gamma_i(x,v)\gamma_i(y,u)}\ . \tag{7}$$

state x at the time i.

III. Design of the Decoder

In theory, forward α and backward α computation can be designed with memory size of just 2M. However, the finite precision of computer arithmetic introduces two types of errors that can deviate the decoder from the ideal behavior.

The first type of errors are *the round-off* errors. These errors occur during the many exponential computations of

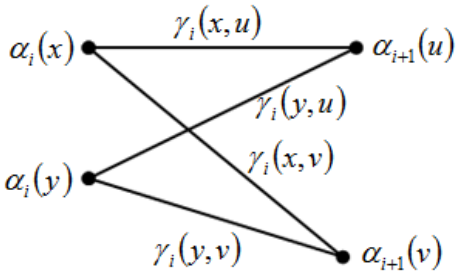


Figure 3. Binary butterfly of a convolutional code with corresponding metrics

the branch metrics $\gamma_i(s,s')$ in the forward and the backward stage. Round-off errors can cumulatively become so large that they will make the decoder catastrophic. To avoid this from happening, at this point of research, we use higher precision arithmetic – the IEEE's 754 double precision format. In addition we periodically store all $\alpha_i(s)$ at regular time intervals that are multiple of a constant T. Then we employ backward α computations between two successive storage points $\alpha_{iT}(s)$ and $\alpha_{(i+1)T}(s)$ (fig. 4).

The second type of errors are *the cancellation errors*. For example, in a butterfly it may happen that $\alpha_i(y) >> \alpha_i(x)$. Thus after the round-off stage, (6) will become

$$\begin{aligned}\alpha_{i+1}(u) &= \alpha_i(y)\gamma_i(y,u)\\ \alpha_{i+1}(v) &= \alpha_i(y)\gamma_i(y,v)\end{aligned}. \qquad (8)$$

We see that the value of $\alpha_i(x)$ is lost during the forward recursion. In the case of the BCJR algorithm we do not recover from these errors because they do not affect the performance.

In addition, it may happen $\alpha_i(x)$ to become negative. In this situation we can set $\alpha_i(x)$ to zero or we can take absolute value of (7).

The last issue that we have to resolve is the periodic normalization [5]. During the forward α computations, periodic normalization $\alpha_i(s)/t_{\alpha,i}$ is introduced in order to avoid overflow. However, during the backward α computation this normalization may create underflow. Hence we need to perform *denormalization*, i.e. to multiply $\alpha_i(s)$ with appropriate coefficients $t_{\alpha,i}$. Using the original normalization coefficients $t_{\alpha,i}$, would require to keep them all in memory. To avoid this we use normalization coefficients $t_{\beta,i}$, from the backward β recursion in the denormalization stage of the backward α recursion. Thus we eliminate the need to store the original normalization coefficients $t_{\alpha,i}$.

Next we give pseudo code for the backward α algorithm:

```
for each information bit i=n-1 downto 0
  if (i mod T == 0)
     restore α_i(s) from memory
  else
     for each state s<M
        using (5) compute α_i(s)
        α_i(s) = t_β,i * α_i(s) //denormalization
```

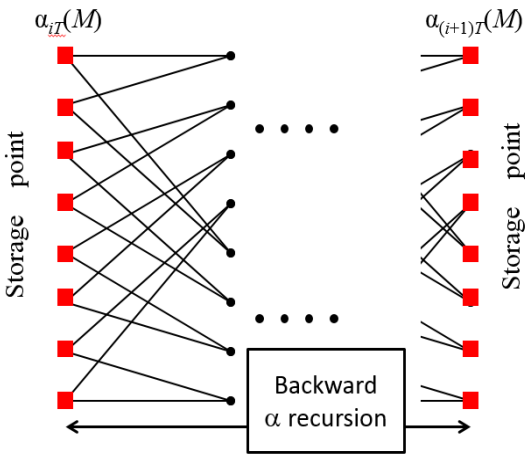


Figure 4. Periodical storage points in a real MAP decoder

IV. Practical Results

In testing our designs we have used parallel concatenated Turbo Codes (fig. 5). To be more precise we use rate-1/2 16-state 1025-bit long Turbo code over Gaussian channel. In addition, we use IEEE's 754 double precision format to store state and branch metrics, although we are fully aware that this format will not be very useful in implementing this algorithm on a chip.

The comparison between our algorithm and the traditional BCJR algorithm is given on figure 5. We can notice performance degradation when going from the regular BCJR decoder to the BCJR decoder with reduced memory size. For this simulation, we have chosen that the periodical storage points to be at distance $T = 35$ information bits. At this distance we clearly have observed performance loss of about 0.1 dB. However, careful research is needed in finding the largest possible acceptable T. We should note that at some point performance degradation becomes unacceptable. For example, we have noticed that this is clearly happening around $T = 50$. From the figure it can be concluded that for $T = 35$ performance loss is about 0.1 dB at BER around 10^{-5} – 10^{-6}. In our simulations round-off errors were dominant with the BCJR algorithm, the number of cancellation errors per frame was always less than 20 in a matrix of size 1025×16=16400 32-bit words.

V. Conclusion

We have demonstrated a possibility to achieve considerable memory reduction for the BCJR algorithm. We believe that this reduction technique is powerful enough so that other memory optimization techniques may become unnecessary. Moreover, with proper parallelization (like initiating the alpha and beta computations at the same time) we can build a decoder with space and time complexities that will come closer to the complexity of the Viterbi algorithm. This will open new horizons for the convolutional codes, like using

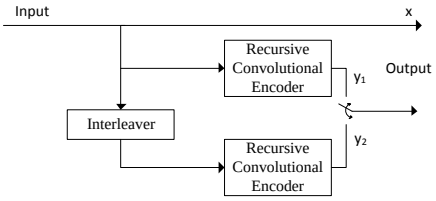


Figure 5. Binary butterfly of a convolutional code with corresponding metrics

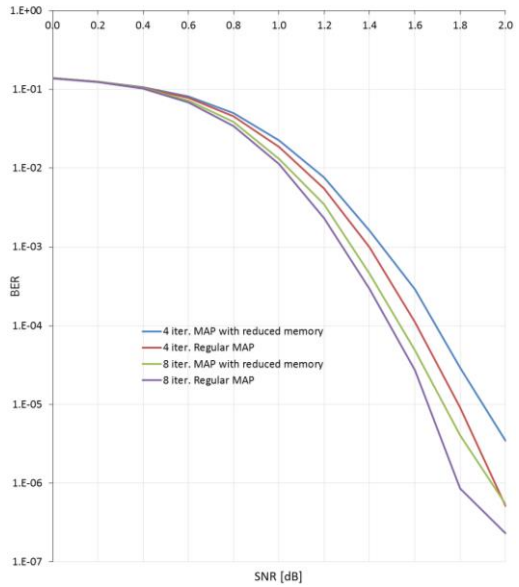


Figure 6. Performance comparison between MAP and reduced-memory MAP with periodical storage points at distance T=35

longer codewords or using convolutional encoders with 32 or even 64 states. On the hardware side, the benefit of memory saving is also important, because it will provide space reduction, less power consumption and increased decoding speed.

Popular and frequently implemented variant of the BCJR decoder is the Log-MAP algorithm. A question that naturally arises is if this reduction can be applied to the Log-MAP algorithm. While Log-MAP is often considered as MAX-Log-MAP with a correction term that is added to (7), in its nature, this algorithm is the BCJR algorithm with fixed-point arithmetic for the exponentiations. Thus in the case of the Log-MAP algorithm any memory reduction technique will increase the round-off errors, thus decreasing the interval *T* between two storage points (fig. 4).

REFERENCES

[1] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," IEEE Trans. Inform. Theory, vol. 20, no. 3, pp. 284–287, Mar. 1974.

[2] E. Boutillon, W. J. Gross, and P. G. Gulak, "VLSI Architectures for the MAP Algorithm," IEEE Trans. On Communications, vol. 51, no. 2, February 2003.

[3] M. Zhan, L. Zhou, " A Memory Reduced Decoding Scheme for Double Binary Convolutional Turbo Code Based on Forward Recalculation," in 7th International Symposium on Turbo Codes and Iterative Information Processing (ISTC), Gothenburg, Sweden, 2012.

[4] H.-M. Choi, J.-H. Kim, and I.-C. Park, "Low-power hybrid turbo decoding based on reverse calculation," ISCAS. 2006. pp. 2053–2056.

[5] Emmanuel Boutillon, Catherine Douillard, and Guido Montorsi, "Iterative Decoding of Concatenated Convolutional Codes: Implementation Issues," Proc. of the IEEE, Vol. 95, No. 6, June 2007, pp. 1201-1227.

[6] G. Colavolpe, G. Ferrari, and R. Raheli, "Reduced-State BCJR-Type Algorithms," IEEE Journal on Selected Areas in Communications, Vol. 19, No. 5, May 2001.

[7] M. Sikora, D. Costello, Jr. "Heuristic Survivor Selection for Reduced Complexity BCJR-type Algorithms," ISIT 2008, Toronto, Canada, 2008

[8] A. Ouardi, A. Djebbari, B. Bouazza, "Optimal M-BCJR Turbo Decoding: The Z-MAP Algorithm," *Wireless Engineering and Technology*, 2011, pp. 230-234

[9] J. Hagenauer, E. Offer, and L. Papke, "Iterative Decoding of Binary Block and Convolutional Codes," IEEE Trans. Inform. Theory, vol. 42, no. 2, pp. 429–445, Mar. 1996.

[10] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in Proc. ICC, Geneva, Switzerland, May 1993.

Towards a Self-organizing Network: An Algorithm

J. Seljan, D. Šimunić, G. Dimić, B. Vrandečić*
Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
{josip.seljan, dina.simunic, goran.dimic}@fer.hr
bernard.vrandecic@dok-ing.hr

Abstract – In this work we deal with the Achilles' heel of wireless sensor networks – energy efficient operation. By a combined MAC/physical layer attack on the issues contributing to energy overhead, we demonstrate the possibility of reducing the overall power consumption in a relatively straight-forward way without severe alteration of common communication protocols. Results of simulations of networks using the proposed scheme are shown, partially validating the crucial ideas behind it.

I. INTRODUCTION

The advent of large-scale wireless sensor networks (WSN) [1], [2] is well-nigh. This has been brought on by a simple fact gaining precedence as time goes on - with limited resources, it is not as important to amass more, but as to utilize the ones in possession as best as possible. Through constant automatized monitoring and administration of a given system, be it of agricultural, traffic or other kind, its performance can be significantly improved with tremendous savings in terms of needed resources. Moreover, the automatized administration is far less prone to miscalculation and error than human monitoring, thereby reducing the necessary manpower. Until recently, the monitoring and administration of a wide range of systems was infeasible due to a lack of affordable and energy-efficient sensing and controlling devices. At present, a number of formidable standards is at disposition, most notably the ZigBee platform. Hence, the rapid growth in application of large-scale WSNs is underway.

However, there is a number of open problems associated with large-scale networks, the crucial one being the lack of reliable, tested algorithms and protocols for network management, both from the traffic and energy-saving perspective. Though a large number of communication protocols and power-saving schemes have been devised that, in specific scenarios, outperform the standard ones found in e.g. WLAN, it seems there is no general consensus on the cornerstones of any wireless network design – network topology, shortest-path metrics, modulations, radio channel models, radiated power control, collision control etc. The reason for this dischord is the following. A „wireless sensor network“ (WSN) is an umbrella term referring to a large number of ad-hoc sensor networks, each designed with a specific purpose in mind and to be deployed in a specific enviroment. These range from battlefield monitoring, agricultural monitoring [3], patient monitoring in hospital enviroments etc. Hence, the requirements dictated by the application lead to radically different networks in terms of size, topology, data rates and, ultimately, energy consumption.

The aforementioned issues make obvious the fact that, if there is any progress to be made in the direction of a consistent scheme for sensor network optimization, efforts should be focused on the development of fairly general algorithms and schemes which could be applicable to a wide range of scenarios.

In this paper we propose several energy-saving strategies and a heuristic algorithm of MAC control, which could satisfy this requirement of generality. It is based on the combined probabilistic and quasi-deterministic administering of the network, both at the medium-access-control (MAC) and the physical (RF power) layer.

The paper is organized as follows. In section II, we define a general wireless sensor network and its constituents, along with the relevant parameters, reviewing the key problems in the reduction of power consumption. Then, we propose a scheme for mitigating these issues, in the form of general guidelines for power levels and data transmission speed. This is followed by a description of a power-saving algorithm based on the probabilistic relaxation of the stringent default settings of the 802.15.4 MAC protocol. We conclude that section with results of simulations carried out on networks operating according to the scheme proposed in the previous section. In addition, we discuss the potential drawbacks of the proposed scheme. We conclude the paper delineating further improvements which will be adressed in future work.

II. THE PROBLEM

A. *Definitions and initial assumptions*

A general wireless sensor network consists of a given number of nodes – devices collecting data generated by its adjunct array of sensors into a buffer, and transmitting it further. Each node represents a sensing „unit cell“ covering a given region of the enviroment to be monitored. In general, a node can have an arbitrary arrangenemt of heterogeneous sensors, each generating data at different rates. Here we shall not restrict ourselves

to a specific node architecture, or consider the impact of specific sensor arrangement in detail, but rather consider a node to be simply characterized by its data output. In other words, a node is a stochastic data source. We consider the network as centralized, i.e. there is a central node towards which all the data generated by sensing nodes is transmitted and is responsible both for data aggregation and network administration (e.g. a ZigBee network). This restriction excludes gossip-type networks, i.e. networks where the data generated by a node is disseminated towards all the nodes in the network.

Now, the sensors may be „sampling" a given dynamic physical variable at a constant rate, or the sampling may be event-triggered. As most commonly monitored processes obey the Poisson distribution, we may consider the total data output of a node in a given period of time to behave as a Poisson process [4]. This may aid in the estimation of characteristic transmission periods and the amount of data generated. In addition to being a data source, a node functions as a relay passing down data received from its neighbors to the destination node.

Seen from the perspective of energy consumption, the most significant process in a wireless sensor network is the radio transmission [2], contributing to between 50 to 70 % of the total energy overhead. Though largely dependent on the hardware node architecture, type of modulation and coding used, from manufacturer data sheets and on-site measurements it can be approximately evaluated that the act of transmission can consume anywhere from 10 to 200 mW, whereas sensing consumes approximately one to two orders of magnitude less. The CPU overhead, on the other hand, is comparable to the sensing consumption. In addition, in order for a network composed of sensing nodes to function reliably, one must ensure a stable data flow, i.e. avoid congestion.Then it can be seen that, in essence, one can swat two flies with one blow – ensuring a low data packet collision rate reduces the number of retransmissions, leading to a considerably lower power consumption. This is done by modifying a network's MAC layer. This is the direction a great number of authors take in mitigating the issue of energy efficiency, e.g. [5], [6], and is the direction we follow. However, we put greater emphasis on the interrelation between the MAC and the physical layer, as will be discussed in what follows.

Before we delve into the specifics of our consumption reduction strategies, we would like to address a commonly overlooked aspect of general wireless networks, not just sensor networks; its neglect leading to incorrect assumptions and, consequently, dubious MAC and routing algorithms. As is well known from mobile networks practice, the signal power in a wireless network fluctuates wildly, both in space and time due to strong coupling of a node's receiving/transmitting antenna to its enviroment. Obstacles, moving or stationary, irregular terrain and interefering signals from other sources all contribute to this predicament, rendering the network problem heavily stochastic. Unfortunately, a number of authors overly simplify the real-life scenario by various incorrect assumptions on the shape and size of a cell, neglecting potential link asymmetry and the influence of the terrain on fading. An excellent overview and rebuttal of these „myths" is given by Kotz *et al.* [7]. We base our work on the conclusions brought forth there in order to avoid typical pitfalls and enhance the generality of our approach.

B. Strategies, guidelines

As mentioned in the previous subsection, and thoroughly discussed in Kotz *et al.*, a realistic scenario in which a wireless sensor network operates is too complex to lend itself to simple models. Radio channels are prone to large variation in signal-to-noise ratio, and the mobility of obstacles and even nodes complicates analysis further. Though there are some radio channel models, extrapolated from a set of exhaustive measurements, which *may* give a more accurate picture of an actual scenario [8], they have not been tested extensively, and cannot be considered reliable. Sublimating the above discussion into a single conclusion, we state the first „ansatz":

In general, there is no reliable deterministic or quasi-deterministic metric applicable to the characterization of link quality between two nodes of a wireless sensor network. At best, a quasi-deterministic model may supplement a purely probabilistic metric.

In simple terms, how well a node transmits or receives data cannot be predicted by simple physical models, but can be quantified only by direct measurement based on reception acknowledgement. For example, one cannot predict if a data packet will be received by a node using the inverse-square power law, due to the large dynamics of the interfering signals. On the other hand, a node can request an explicit acknowledgement of reception from the destination node. On the basis of the ratio of successfully sent and dropped packets, a node can „deduce" whether it should increase the transmitting power or not, thus controlling the likelihood of successful transmission. Moreover, as shown by Blagojevic *et al.* [9], this line of thinking can be pushed even further to avoid explicit reception acknowledgement, which may actually degrade the performance of a WSN. Instead, a *probabilistic reception acknowledgement* may be used instead, where a node does not need to receive reception acknowledgements every time it sends a packet. It may ask for an acknowledgement sporadically, and based on the both the calculated reception ratio and its rate of change, decide which packets in the buffer should be sent/resent.

We build our approach upon this probabilistic reception acknowledgement paradigm. However, before we present the exact scheme, we would like to comment on general guidelines concerning the appropriate transmitted power and data rates leading to energy savings.

As most architectures suitable for implementation of WSNs grant the possibility of tuning the radiated power and varying the data rates, as well as of switching the nodes into various power-saving modes [1], it is paramount to use these degrees of freedom fully. The first rule of thumb applies to the radiated power. Specifically, whenever possible, a node should transmit the largest power possible, thus increasing the probability of reception; of course, provided it does not drown out the

signals from neighboring nodes. Therefore, when the package drop ratio, defined as

$$\eta = \frac{N_{dropped}}{N_{sent}} \quad , \tag{1}$$

increases above 0.4, the node should fire at maximum available power. For a ZigBee node, taken as an example, it is specified as 20 dBm. Otherwise, the power should be kept at approximately $0.7P_{\max}$ to reduce interference in neighboring cells; $P_{\max}$ is the largest power. As far as the transmission data rate is concerned, a larger data rate leads to a shorter transmission, thereby lowering the collision probability P_{col}. Another important observation is that if the node is in sleep-mode, as is the case when the sensor detection or innovation rates are low, the start-up time may be considerable once it goes back online. During this time no power is used for transmission, and hence one should be careful not to switch the node between the sleep and active state too often, especially when it generates a significant amount of data in consecutive frames. When this is the case, it is preferable to keep the node in the active state.

If these very general guidelines are adhered to, the power reduction factor with respect to the „out-of-the-box“ can be as high as 2. This has been confirmed empirically by network simulations carried out in OPNET.

C. A network algorithm

At this point we present the power-saving algorithm which partially mitigates the problem of long-term energy efficiency. Though there are plenty routing algorithms which use more complex score in choosing which nodes should be used as relays [10], [11], including , e.g., the effects of node power supply levels, interference etc., we opt to leave out such details, though they could be incorporated into our algorithm; this we leave for future work. In the present investigation, we assume a standard AODV routing algorithm as used in, e.g. ZigBee networks.

The algorithm presupposes the following statements to be true. Firstly, the nodes are static, with well-defined positions, as measured by the central node during the initial network setup. Each node knows its own position as well as of all the other nodes. If a node is added to the network, the central node calculates its position and passes it to all the other nodes. The communication is partitioned into time-frames, i.e. periods during which nodes can transmit and receive data, in turn partitioned into subintervals; a node picks an arbitrary subinterval and attempts to communicate. Secondly, during the initial setup, each node sends packets to its neighbors during a number of time-frames, requesting an acknowledgement reception. If the acknowledgement from the receiving node is not received within the next time-frame, the transmission is considered as failed. At the end of this “testing” period, each node has a cost table with packet reception ratios (transmission probabilities) for every node it can reach. Then, the network can start communicating according to routing tables built by the central node. During each round (time-frame), a sending node can update its cost table based on reception (or the lack of) acknowledgements. However, a more prudent strategy is to perform this every few rounds. In more detail, in general a WSN will not alter dramatically during several rounds of communication. Of course, fast-moving objects and short-burst noise sources may debilitate its functioning, but in most environments where such networks are deployed, this is not the case, or a relatively small proportion of nodes is affected. Therefore, cost tables may be updated once in a few rounds, which decreases the data overhead since a receiving node needs not send an acknowledgement in each one. This corresponds to a coarser “sampling” of the network statistics. It is possible if the statistics a node keeps does not change considerably in between two “sampling points”. In addition, it relaxes the condition of obligatory reception, necessitating retransmission until the packet is successfully delivered. Unfortunately, the implication is that there is no more guarantee of data delivery. However, most WSN networks do not transmit data which *must* be delivered. If a packet is not delivered, a newly generated one will replace it in the buffer. If this is the case, a simple strategy, ensuring that an “important” packet is delivered is to include a priority flag into the data packet – if a node receives a packet with the priority flag indicating an “important” packet, it should drop any other connection and relay the packet further. This will result in the interruption of a significant number of connections, which will need to be reestablished, increasing the energy overhead due to the increased number of new connection requests; this issue is the topic of ongoing research.

Now, as the cost table is updated, a node can assign a higher or lower transmission power depending on the probability of reception of the receiving node and its change between two transmissions. For example, if the probability of reception does not change significantly between two transmissions, while remaining above a certain threshold, the sending mode can lower the transmission power. Then, if the in probability reception does not drop below the threshold, it may maintain the chosen transmission power. This amounts to a dynamical adaptive power tuning scheme which can considerably decrease the overall power consumption.

As for the routing protocol itself, we assume the standard ad-hoc on-demand distance vectors algorithm (AODV) adopted in the ZigBee standard [12].

Essentially, the proposed algorithm blends the relaxation of the local-route cost calculation and package delivery with the AODV routing algorithm, adding the adaptive power tuning to form a flexible, simple and effective approach to power management.

To investigate the effect of the proposed acknowledgement relaxation, we compared the end-to-end delivery ratio of a network operating under the standard 802.15.4 protocol, and the same physical network operating under our proposed MAC algorithm. We have observed a 6 % degradation in the delivery ratio in the

latter case. This figure seems reasonable since a) the network dynamics is low (i.e. the generated data does not change abruptly over time periods on the order of tens of frames), b) the node RF power is, on average, increased with respect to the standard, out-of-the box network, and c) transmission times are shorter.

III. CONCLUSION

In this paper, we have argued that a probabilistic viewpoint should pervade any wireless sensor network design. The energy-saving algorithm presented is based on the relaxation of the explicit acknowledgement request, and entails an adaptive power tuning scheme, allowing one to control the power consumption in a simple way. Preliminary simulations have shown that the the reduction in power consumption with respect to a standard AODV-based network can be up to a factor of 2.6.

However, there remain open questions concerning the applicability of the algorithm to large-scale networks (having nodes on the order of tens of thousands and more), as well as the possible gains by partitioning the network into subnets, as is done, for example, when using the S-MAC protocol.

IV. ACKNOWLEDGEMENT

The research leading to these results has received funding from the project "Information and communication technology for generic and energy-efficient communication solutions with application in e-/m-health (ICTGEN)" co-financed by the European Union from the European Regional Development Fund.

BIBLIOGRAPHY

[1] H. Karl, A. Willig, Protocols and Architectures for Wireless Sensor Networks, John Wiley and Sons, 2005.

[2] W. Dargie, C. Poellebauer, Fundamentals of Wireless Sensor Networks, John Wiley and Sons, 2010.

[3] L. Bencini, D. Di Palma, G. Collodi, A. Manes and G. Manes , "Wireless Sensor Networks for On-Field Agricultural Management Process", Wireless Sensor Networks: Application-Centric Design, Yen Kheng Tan (Ed.), InTech, 2010.

[4] M. Haenggi, J.G. Andrews, F. Baccelli, O. Dousse, M. Franchescetti "Stochastic Geometry and Random Graphs for the Analysis and Design of Wireless Networks" , IEEE Journal on Selected Areas in Communications, Vol. 27, Nr. 7, September 2009.

[5] F. Chiti, M. Ciabatti, G. Collodi, D. Di Palma, R. Fantacci, G. Manes, A. Manes, I. Nelli, "D-STAR MAC Protocol: A Cross Layer Solution for Wireless Sensor Networks Endowed with Directive Antennas", Wireless Personal Communications, Vol. 47, Issue 1, 2008.

[6] S. Kwon, N. Shroff, "Energy-efficient unified routing algorithm for multi-hop wireless networks", IEEE Transactions on Wireless Communications, Vol. 11, Issue 11, 2012

[7] D. Kotz, C. Newport, C. Elliott, " The mistaken axioms of wireless-network research", Dartmouth College Computer Science Technical Report TR2003-467, 2013.

[8] K. Fall and K. Varadhan. *The ns Manual*, April 14 2002. , www.isi.edu/nsnam/ns/nsdocumentation.html

[9] M. Blagojevic, M. Nabi, M. Geilen, T. Basten, T. Hendriks, M. Steine, "A probabilistic acknowledgment mechanism for wireless sensor networks", 6th IEEE International Conference on Networking, Architecture and Storage (NAS), 2011.

[10] G. Yin, G. Yang, W. Yan, B. Zhang, W. Jin, "An Energy-Efficient Routing Algorithm for Wireless Sensor Networks", International Conference on Internet Computing in Science and Engineering, 2008.

[11] J. Guo, P. Orlik, J. Zhang, K. Ishibashi "Reliable routing in large scale wireless sensor networks", Sixth IEE International Conf on Ubiquitous and Future Networks (ICUFN) 2014.

[12] C. Perkins, E.M. Royer, " Ad-hoc on-demand distance vector routing", Second IEEE Workshop on Mobile Computing Systems and Applications, 1999.

Static Cluster and Dynamic Cluster Head (SCDCH) Adaptive Prediction-Based Algorithm for Target Tracking in Wireless Sensor Networks

Muath A. Wahdan[1], Mamoun F. Al-Mistarihi[1], and Mohammad Shurman[2]

(1) Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
muath.wahdan@gmail.com, mistarihi@just.edu.jo
(2) Jordan University of Science and Technology/ Network Engineering and Security Department, Irbid, Jordan
alshurman@just.edu.jo

Abstract-**Wireless sensor networks consist of large number of sensor nodes that have an ability to interact with the surrounding environment. The importance of finding the suitable protocol to estimate the trajectory for any moving target in a certain environment is a case of study for researchers. In this paper, new and accurate protocol for target tracking with minimum energy consumption is proposed. The proposed approach uses the static cluster and dynamic cluster head mechanism (SCDCH) to collect data from the active nodes, and then forwards it to the cluster head (CH), which in turn will send the collected data to the base station, without changing cluster border during the network's lifetime. Simulation results show that the proposed protocol has higher accuracy and save more energy compared with static clustering protocols which in turn increases the network lifetime.**

Keywords: ***Dynamic source routing (DSR), energy efficient, localization, static cluster and dynamic cluster head mechanism (SCDCH), target tracking.***

I. INTRODUCTION

Wireless sensor network (WSN) is an emerging technology which has a wide range of applications; included infrastructure protection, industrial sensing and diagnostics, battle-field and environmental monitoring, building's micro-climate controls system, biological and nuclear attack detection, context–aware computing (i.e., intelligent home and responsive environment) [1].

WSN inherits most of the features from the traditional networks. However, the unique nature of such networks adds more constraints that need to be addressed such as power consumption [2]. WSNs have several challenges such as restricted sensing and communication range as well as limited battery Capacity [3]. These limitations introduce many issues such as coverage, connectivity, network lifetime, scheduling and data aggregation. In order to prolong the WSN lifetime, energy conservation measures must be considered, scheduling and data aggregation are among the commonly used methods [4].

The energy source of a node is generally considered non-rechargeable. Using energy as a vital resource in WSNs, several MAC protocols have already been proposed to achieve higher energy efficiency during long idle period of the sensors. In WSN, an energy efficient MAC protocol is required for obtaining higher energy efficiency in difficult operating conditions, where node and link failures are common [5]. The MAC protocol manages radio transmissions and receptions on a shared wireless medium. Therefore, MAC has a high effect on network performance and energy consumption. To save the energy of the network, the total number of sensor nodes sending data is to be reduced thereby decreasing the transmission of the redundant data [6].

In [7], an energy-efficient based algorithm has been proposed, in which sensors around the target could control tracking the target statues, according to the target's direction and speed, extend sensors life and reduce the number of working sensors for target tracking system. In addition, a distributed boundary coverage algorithm has been evolved.

An approach based on dynamic sensor collaboration (DSC) for the information-driven has been proposed [8] to determine computation and communication cost using dynamic optimization information utility, by which the sensor network collaboration participants are determined.

A cluster-based distributed tracking scheme (CBDTS) has been discussed in [9]. The extended approach has been introduced by Zhao et al by adding ability to track multiple targets. Moreover, binary sensor network in [10] has tried tracking a moving object and ignoring the object direction, whether it's moving toward or away from the sensor.

The LEACH protocol has been discussed in [11]. This protocol is a remarkable scheme for immovable senor nodes. It joins the based routing of energy-efficient cluster ideas with the selected cluster head of each ones, enables saving communication resources by efficient signal processing distribution.

The rest of paper is organized as follows: Section II presents the proposed algorithm for the target tracking algorithm. Section III contains the simulation results. Finally Section IV concludes the paper.

II. PROPOSED ALOGARITHM

Energy consumption in WSN within the process of target tracking and data collection is the most critical factor. A new and accurate protocol for target tracking with minimum energy consumption and long nodes' life-time is designed in this paper. In the proposed approach, we consider that immovable homogenous nodes are distributed randomly in the field.

The network in cluster-based methods is divided into clusters. Each cluster includes member sensor nodes and a cluster head (CH). The CH is responsible for gathering data from its cluster's members and calculates the present target location and reports it to sink. The energy of the nodes is limited while the sink energy is assumed to be unlimited.

The proposed algorithm assumes that the nodes with the highest energy level in each zone at each interval time will be selected as CH according to LEACH algorithm. When the selections of CH are done, CH broadcasts control packet to its neighbor nodes, which reply with node ID and location of these nodes to the CH.

The sensors that find the target in its vicinity will promote one of them as a leader and transmit distance message to it. After that and by the evaluation of the received messages, the localization will be performed by leader. When the participants are apart, the Dynamic Source Routing (DSR) will be used for multi-hop communication.

The proposed approach uses the Static Cluster and Dynamic Cluster Head mechanism (SCDCH) to collect data from each member in the network, and then send it to the CH. Newton-Gaussian algorithm is used to calculate trajectory in addition to exact estimation method (Trilateration Algorithm). The comparison results demonstrated in the simulation section with the Static Clustering with Static Cluster Head Protocol and the proposed one using MATLAB software proves the enhancements of our proposed Protocol.

A. Network Topology

A network of M nodes is described as an M-vertex undirected graph S = (R = {1 …*M*}), every position *r* of each vertices *j* is defined by set of coordinates r_j = (x_j and y_j respective). A length $d_{j,i}$ between two vertices *j, I* ∈ R is referred to as G, which is the edges matrix. The network topology is translated by a three row matrix, where the positions of nodes are repositioned in columns as shown in Table 1.

Table 1: Topology Matrix

ID	1	2	3	.....	M
$X_{intercept}$	x_1	x_2	x_3	.....	x_M
$Y_{intercept}$	y_1	y_2	y_3	.....	y_M

According to simulation, all nodes will be reachable and fully connected based on multi hoped communication. The range of radio for the nodes should be configured optimally because that the connectivity is based on the radio range. To print out the topology network, the network matrix should be created and the radio range should be known. The network's layout includes vertices and edges between them. If the Euclidean spaces $d_{j,i}$ between the nodes *i, j* are smaller than the communication radius *R* of the considered nodes, the edges or line between the nodes can be printed, and since the wireless links are bidirectional and symmetric so $d_{i,j}=d_{j,i}$.

B. Routing Mechanism

By using the links comparison base, the routing data in (WSN) will be realized. It can compare the adopted links between the sender and the receiver in respect of links quality, the residual energy of nodes pairs or the length. Notwithstanding, the route for data delivery has been selected depending on the path with smallest investigated value. A MATLAB implementation of Dijkstra algorithm is used to discover the optional route between two nodes in the graph data structure. By using grshortpath function, the Dijkstra's algorithm is applied as in grTheory MATLAB toolbox.

The grshortpath function establishes a G matrix of neighbors, the source i and the destination j are the input arguments for the matrix. A dsp matrix with the shortest path between all node pairs in the network will be returned. Also, it returns vector with the shortest path between nodes *i, j*. The G matrix must have an exact form of the correct grShortPath algorithm processing. It contains three columns, where the first two columns contain all the nodes in the network that are neighbors of each other and the third column contains the Euclidean distance in meters between them. The G matrix can be created during the network layout printing as shown in Table II. As what mentioned before, the link lengths between all node pairs are compared with the uniform radio range R. If the condition $d_{i,j}$ <R is accomplished, the link will be assumed.

The G matrix created with these conditions since all required parameters such as IDs of two neighbors and their distance d must be known. Also, the condition of lengths is accomplished .The context of the G matrix is visualized in Table 2**.** G matrix is illustrated in Table 3**.**

Table 2: Pseudo code of layout visualization and E matrix definition.

Algorithm: G=createNbrTable
1: row = 1; *2: for all node pairs* *3: x = abs (x_i - x_j);* *4: y = abs (y_i _yj);* *5: dist = sqrt (x^2 + y^2);* *6: if $dist_{i,j}$< R;* *7: plot ([x_i , x_j],[y_i , yj]); % draw edge* *8: E(row,1) = ID_i;* *9: E(row,2) = ID_j ;* *10: E(row,3) = $dist_{i,j}$;* *11: row++;*

In case of using residual energy or link quality for establishing the route instead of the distance between nodes, we can replace the contained information in the third column of E matrix by the required information. Accordingly, the path between two nodes will be chosen as a path of high quality or with maximum energy.

Table 3: G matrix

ID_i	ID_j	$dist_{i,j}$
1	2	15.65
1	3	9.21
1	8	21.54
2	1	15.65
2	8	11.12
…	…	…

Fig. 1 shows the layout of the network which displays the links between all node pairs in the proposed protocol, a link is assumed when the length of the link between any two nodes is less than or equal to the communication range (R=80m). We deploy 80 nodes randomly in the land scape.

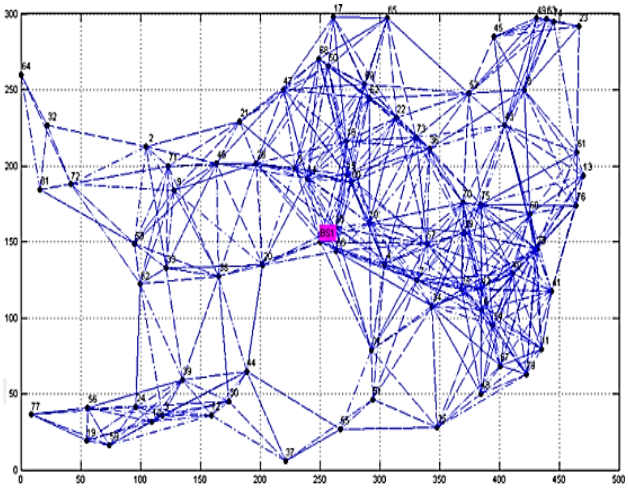


Fig. 1: Network layout with shortest path estimated by the grShortPath function.

C. Energy Model

We notice that, the energy model used in our simulation is the same one used in [12]. In this model, the energy of the sent (ET_x) and the received (ER_x) packets in the network can be described as in the following two equations:

$$E_{Tx}(k,d) = E_{elec} + E_{amp} * d^{a} \quad (1)$$

$$E_{Rx}(k) = E_{elec} \quad (2)$$

Free space model is considered in the short distance, a=2. In the simulation, the electric energy (E_{elec}) is assumed to be 50 nJ/bit and the amplifier energy (E_{amp}) is equal to 10 pJ/bit/m^2. Fig. 2 shows the sensor radio energy model main components.

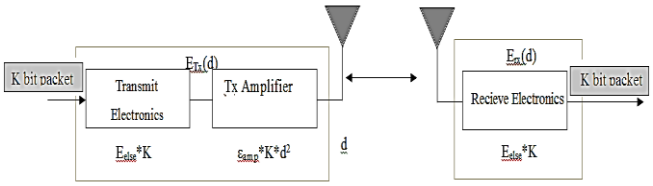


Fig.2: sensor radio energy model [13].

III. RESULTS AND DISCUSSION

In our simulation environment, the landscape has 400 nodes including CHs according to the number of zones. CHs and other nodes are randomly distributed between (x = 0, y = 0) and (x = 500, y=300) with sink at location (250,150). Each node will be initiated with 0.02j. (45-70) meters will be considered the range of communication among all nodes, sensing range is Rs=30 meters.

Target trajectory is selected randomly. Tracking period is set to 1 sec and simulation time is 3600 seconds (60 minutes), the free space model is considered in the short distance a=2 since sending distance between nodes in the cluster-based (CB) architecture that is used in the proposed protocol is assumed to be short. Fig. 3 shows the environment nodes distribution algorithm for 400 nodes.

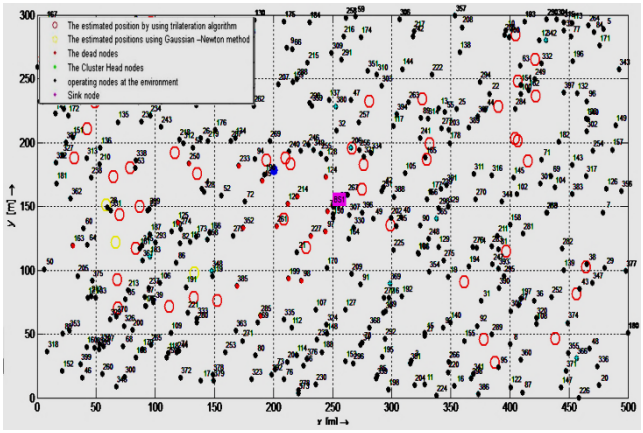


Fig. 3: Environment nodes distribution algorithm for 400 nodes

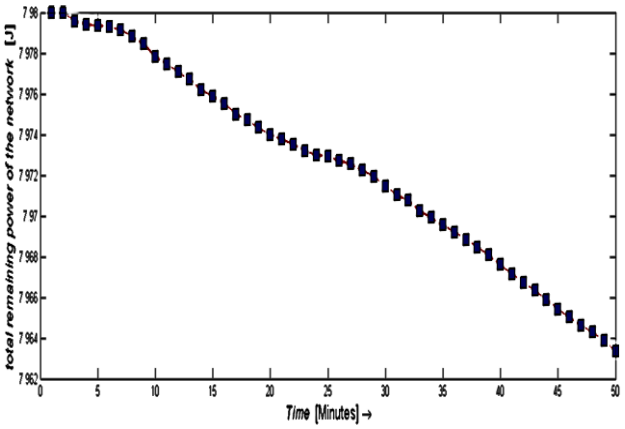


Fig. 4: The energy of the whole network for 400 node environment

Fig. 4 shows the total remaining power of the whole network as a function of simulation time for 400 node environment. From the figure, we observe that the remaining energy is decreased as a function of time. The remaining energy available for whole network at t=50 minutes is equal to7.964 J.

Fig.5 shows the number of dead nodes during the simulation time for 400 nodes. From the figure, it is clear that

after 50 minutes of the simulation, the number of dead nodes reaches 8 nodes for the simulated approach.

The proposed protocol is simulated using MATLAB and compared with well-known protocols, the static clustering approach [7] and the direct communication protocol, with the same simulation conditions. The comparison metrics are; the number of dead nodes, the estimated trajectory of the target, and network life time. The proposed approach shows higher accuracy, more energy saving, larger overall nodes lifetime. The remaining power at any node is inversely proportional to the number of processing messages that sent by this node. The remaining power of the network nodes is increased with the increase of nodes numbers.

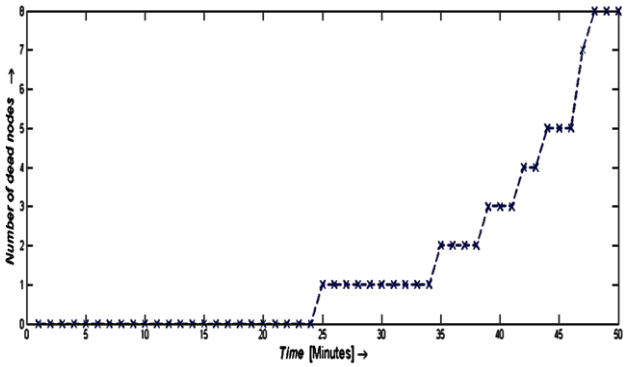


Fig.5: The number of dead nodes during the simulation time for 400 nodes

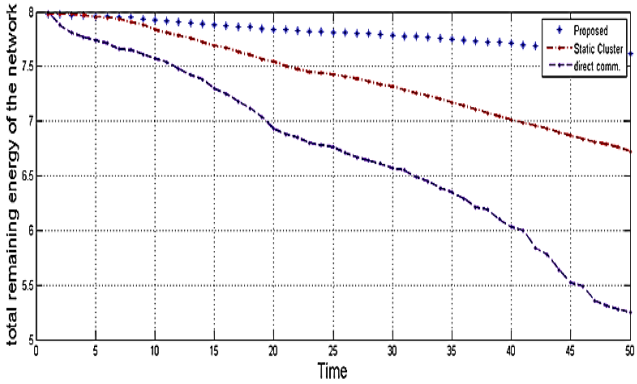


Fig.6: comparison between the total remaining energy for proposed dynamic Cluster Head protocol, the Static Cluster Head Protocol, and Direct communication protocol respectively

Table 4: Number of dead nodes for one round of simulation

Protocol	Number of dead nodes for one round of simulation
Proposed Protocol	17
Static Clustering Protocol	61
Direct Communication Protocol	120

Comparing the number of dead nodes for one round of simulation, Table 4, the proposed protocol has 17 dead nodes for one round of simulation, 61 dead nodes for the Static Clustering protocol, and 120 dead nodes for the direct communication protocol. We can conclude that the proposed protocol achieves significant reduction in energy consumption compared to the Static Cluster and direct communication protocols under the same conditions.

For the seek of comparing the network life time, Fig. 6 shows the total remaining energy of the network as a function of time. From the figure, one can see that the remaining energy within one simulation round is equal to 7.7 J, 6.70 J and 5.4 J for the proposed protocol, static CHs and the direct transmitting respectively. In addition to reducing energy dissipation, LEACH successfully distributes energy-usage among the nodes in the network such that the nodes die randomly and at essentially the same rate.

IV. CONCLUSION

In this paper, a new and accurate protocol for target tracking with minimum energy consumption and long life time nodes is designed. The proposed approach uses the static cluster and dynamic cluster head mechanism (SCDCH) to collect data from each member of the network and then send it to the cluster head (CH). The proposed protocol is simulated using MATLAB and compared with well-known protocols showing very good results in terms of the total remaining energy and number of dead nodes. The estimation error between correct and calculated trajectory using the proposed approach will be discussed and compared with recent published ones as an expansion work to this paper.

REFERENCES

[1] I. F. Akyildiz,W. Su,Y. Sankarasubramaniam, E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, 40 (8):104–112, 2002.

[2] J. Deng, R. Han, and S. Mishra, "INSENS: Intrusion-tolerant routing in wireless sensor networks," *Technical Report CU-CS-939-02, Department of Computer Science,* University of Colorado at Boulder, 2002.

[3] G. Simon, M. Maroti, A. Ledeczi, G. Balogh, B. Kusy, A. Nadas, G. Pap, J.Sallai, K. Frampton, "Sensor network-based countersniper system," *Proc. of the Second International Conference on Embedded Networked Sensor Systems (Sensys)*, Baltimore, MD, 2004.

[4] K. Lorincz, D. Malan, T.R.F. Fulford-Jones, A. Nawoj, A. Clavel, V.Shnayder, G. Mainland, M. Welsh, S. Moulton, "Sensor networks for emergency response: challenges and opportunities," *Pervasive Computing for First Response (Special Issue), IEEE Pervasive Computing*, 2004.

[5] He T., Krishnamurthy S., Stankovic J., Abdelzaher T., Luo L., Storelu R., Yan T., Gu L., Hui J., and Krogh B, "Energy-efficient surveillance system using wireless sensor networks," *Proc. of the 2nd International Conference on Mobile Systems (MobiSYS)*, Boston, USA, 2004 .

[6] F. Sun and M. Shayman, "Prolonging Network Lifetime via Partially Controlled Node Deployment and Adaptive Data Propagation," *Proc. IEEE Intl. Conf. on Information Sciences and Systems (CISS)*, March 2010.

[7] Fatmeh Deldar, Mohammad Yaghmaee, "Designing a Prediction- based Clustering Algorithm for Target Tracking in Wireless Sensor Networks," *IEEE International Conference on Computer Network and Distributed Systems (CNDS)*, Feb. 2011.

[8] Zhao F., Shin J., and Reich J., "Information-driven dynamic sensor collaboration for tracking applications," *IEEE Signal Proces. Mag.*, March 2002.

[9] Liu J., Liu J., Reich J., Cheung P., and Zhao F., "Distributed group management for track initiation and maintenance in target localization applications," *Proc. Int. Workshop on Information Processing in Sensor Networks (IPSN)*, 2003.

[10] Aslam J., Butler Z., Crespi V., Cybenko G., and Rus D., "Tracking a moving object with a binary sensor network," *Proc. ACM Int. Conf. Embedded Networked Sensor Systems (SenSys)*, 2003.

[11] Wendi B. Heinzelman, Anantha P. Chandrakasan, "An Application – Specific Microsensors Networks," *IEEE Transactiona on Wireless Communication*, vol.1, no.4, Oct. 2002.

[12] C. D. Cordeiro, D. P. Agrawal, "AD HOC and SENSOR NETWORK, Theory and Applications," 1st ed. USA: Word Scientific Publishing Co., 2006, ch. 8, sec. 5, pp. 416-419.

[13] S. Amiri, H. Saidi, A. Ghiasian, and M. Hashemi, "Enhanced Lifetime Maximization Algorithm for Wireless Sensor Network," *IEEE International Wireless Communications and Mobile Computing conference (IWCMC),* pp. 302-307, July 2011.

New Efficient On-Demand Multicast Routing Protocol over Mobile Ad hoc Networks

Khalid A. Darabkh and Ola A. Alsukour
The University of Jordan/Computer Engineering Department, Amman, 11942, Jordan
Emails: k.darabkeh@ju.edu.jo and ola.saqour@gmail.com

Abstract— **Due to its simplicity, efficiency, and robustness to mobility, the On-Demand Multicast Routing Protocol (ODMRP) become a standout amongst the most broadly utilized multicast routing protocols in Mobile Ad hoc networks (MANETs). Essentially, the robustness of this protocol is achieved via periodic route refreshing. However, the robustness of ODMRP comes at the expense of incurring a high control overhead to the network. In this paper, a protocol, as an improvement to ODMRP, is proposed where the refresh interval is basically adapted based on the Source Moving Speed (SMS) and number of disconnections reported by multicast members, namely, Improved ODMRP with Slow Start Join Query Refreshment (IODMRP-SSJQR). Since the majority of nodes in MANET rely on batteries, thus the main contribution of this research is to limit the amount of control information that is passed between nodes (i.e., reducing the control overhead over that in ODMRP).**

Index Terms — **On-demand multicast routing protocols, join query refreshment, source moving speed**

I. Introduction

Mobile Ad hoc networks (MANETs) are becoming increasingly popular and going to be an integral part of the future generation of mobile services [1-6]. A MANET is a vigorously reconfigurable wireless network wherein the mobility of nodes results in speedy and eccentric changes in the network topology [7-8]. Each node functions not only as a network host for transmitting and receiving data, but also as a network router for relaying packets to other nodes [9-12]. As the nodes move randomly and make frequent topology changes in MANETs, power resources, bandwidth, and processing abilities become critical [13]. To cope with the frequent changes in topology, multicast algorithms are always in need of improvements [14-15]. Multicast routing can basically be divided into four main procedures which include forwarding mesh creation, links breakage detection, local recovery, as well as join query refreshment [16].

One of the most important mesh-based multicast routing protocols is the On-Demand Multicast Routing Protocol (ODMRP) [16-17]. In this protocol, group membership and multicast routes are established and updated by the source on demand. When a multicast source has packets to send, but no route to the multicast group is available, it broadcasts a *JoinQuery* control packet to the entire network. In point of fact, this *JoinQuery* control packet is periodically broadcasted (i.e., every 3 seconds) to refresh the membership information and update routes. When an intermediate node receives the *JoinQuery* control packet, it stores the upstream node ID and reboradcasts that packet. When the *JoinQuery* packet reaches a multicast receiver, it creates and broadcasts a *JoinReply* control packet to its neighbors. When a node receives a *JoinReply*, it checks if it is situated on the path to the source and thus is a part of the forwarding mesh and broadcasts its *JoinReply* packet. In this way, every *JoinReply* control packet is relayed back to the source using the reverse path traversed by the *JoinQuery* control packet.

Notwithstanding the simplicity, efficiency, and robustness of ODMRP to mobility, it has some impediments. The static Join Query Refresh Interval Time (JQRIT) introduced in this protocol has a crucial role in its robustness via retaining redundant routes. Thus, the JQRIT is of a most extreme imperativeness to the performance of ODMRP. Such refresh interval cannot be chosen arbitrary since it has a terrible impact on the network performance. In ODMRP, the preferred value of this parameter is 3 seconds. This is justified by the need of achieving the highest possible packet delivery ratio. Unfortunately, one important dimension has been neglected which is the control overhead incurred to the network. In fact, the robustness is achieved at the cost of increasing the control overhead. In this article, we propose a protocol called Improved ODMRP with Slow Start Join Query Refreshment (IODMRP-SSJQR). This protocol follow the same forwarding mesh construction procedure employed in ODMRP. It is good to mention that there is no local recovery proposed in ODMRP. However, our proposed protocol proposes not only a new algorithm for local recovery, but also a new dynamic algorithm for join query refreshment that considers the source moving speed (SMS) and number of disconnections received from multicast receivers in its calculations.

The remainder of the paper is organized as follows: The proposed protocol is elucidated in Section II. Section III details the simulation environment. Experimental results along with observations and necessary discussions are presented in Section IV. The paper is summarized in Section V.

II. Proposed Protocol

In our local recovery algorithm, we aimed at reducing the refresh requests flooding by taking advantage of the fact that adjacent nodes have a higher probability to be detached simultaneously from the network within a short time. We employ a flooding aggregation technique and create a special control packet named as *DelegateRefreshRequest* packet. We inserted a new field inside *DelegateRefreshRequest* control packet called *HopCount*. Any node that receives the *DelegateRefreshRequest* packet creates an entry inside a local maintained table called *RefreshRequestTable* and stores the *HopCount* for a specific period (i.e., a minimum join query refresh interval time, 3 seconds). Thereafter, the node increases the *HopCount* field by one and rebroadcasts an updated *DelegateRefreshRequest* packet. When a node fails in local recovery and intends to generate *DelegateRefreshRequest* packet, it uses the *RefreshRequestTable* to detect if any neighboring node (*HopCount* is up to two) does generate a *DelegateRefreshRequest* packet recently. If a neighboring node does, the disconnected node refrains from sending a new *DelegateRefreshRequest* packet. Using this technique makes the suppression more efficient and representative for a group of disconnected neighboring nodes from the forwarding mesh, which will save the network recourses and enhance the overall network performance.

In our SSJQR proposed algorithm, we divided the source node moving speed into two ranges: low mobility speed (i.e., 1 - 30 mps) and high mobility speed (i.e., 31 - 50 mps). This division of the source node moving ranges was constructed based on the number of disconnections obtained through our simulation experiments while simulating the ODMRP. When the source node moves in a low mobility speed, the source node starts broadcasting *JoinQuery* control packet with a minimum refresh interval time (MinJQRIT, 3 secs) and exponentially increases the refresh interval time until two-thirds of the maximum refresh interval time (MaxJQRIT, 30 secs) unless no *DelegateRefreshRequests* are received. If a refreshment time exceeds the two-thirds of the maximum refresh interval with *DelegateRefreshRequests* absence, the increase of the JQRIT becomes linearly increasing by one-third of the previous JQRIT. The selection of the refreshment time behavior before and after the two-thirds of the maximum refresh interval was designed based on experiments to make the refresh interval increase properly with time. However, under source high mobility conditions, the JQRIT decreases proportionately to the difference between the previous and current source node speeds. Whenever the source node receives a *DelegateRefreshRequest* control packet, it decreases the JQRIT to the half of the current defined one (i.e., flight size). Figure 1 explains the whole adaptation of SSJQR algorithm.

III. Simulation Environment

In our simulation experiments, we assume that the network consists of 100 wireless nodes which were randomly distributed over a 1200 m x 800 m area. Addtionally, there is one multicast group, source, and 20 multicast receivers. As employed in [17], no packet or channel error model is considered in our simulations. The mobility model used is our simulations is *Random Waypoint* model where nodes in the simulation area select a certain destination and move towards it at a random speed that is uniformly chosen between the specified minimum speed (i.e., 1 mps) and selected maximum speed (i.e., 50 mps). Moving directions of each node were also chosen randomly and when nodes reached the simulation terrain boundary, they bounced back and continued to move. To demonstrate the effectiveness of our proposed protocol, we simulated it using The QualNet communications simulation platform (QualNet 7.1) on a PC with an Intel Core (TM) i7 CPU working at 2.4 GHz with a 2 MB cache and 8 GB RAM. Data points presented in the simulation results were calculated by taking the average of 20 simulation runs to eliminate the effect of any anomalous individual result.

IV. Results and discussions

Figure 2 shows the number of control packets generated by ODMRP and IODMRP-SSJQR against different node speeds. It is clearly seen that the ODMRP induces nearly 35000 control packets, more than the two times dose of the other protocol. This can be justified as the ODMRP induces significantly more *JoinQuery* wide flooding packets regardless of the mobility condition to maintain a robust forwarding mesh structure and exhibit better reliability. It is good to point out that this huge number of control packets, generated by ODMRP, has a severe impact on the network in term of bandwidth and node battery. However, for efficient results interpretations, Table 1 is included. As shown in table 1, the IODMRP-SSJQR cause an increase in the number of control packets as the network mobility speed increases, which is quite logical and convincing. In fact, this is in view of involving the source node speed factor in its adaptation.

TABLE I. The Number of Control Packets with Different Node Speeds

Maximum Node Speed (m/s)	IODMRP-SSJQR	ODMRP
10	7629	34475
20	8242	34752
30	8692	34927
40	10079	34944
50	11022	35097

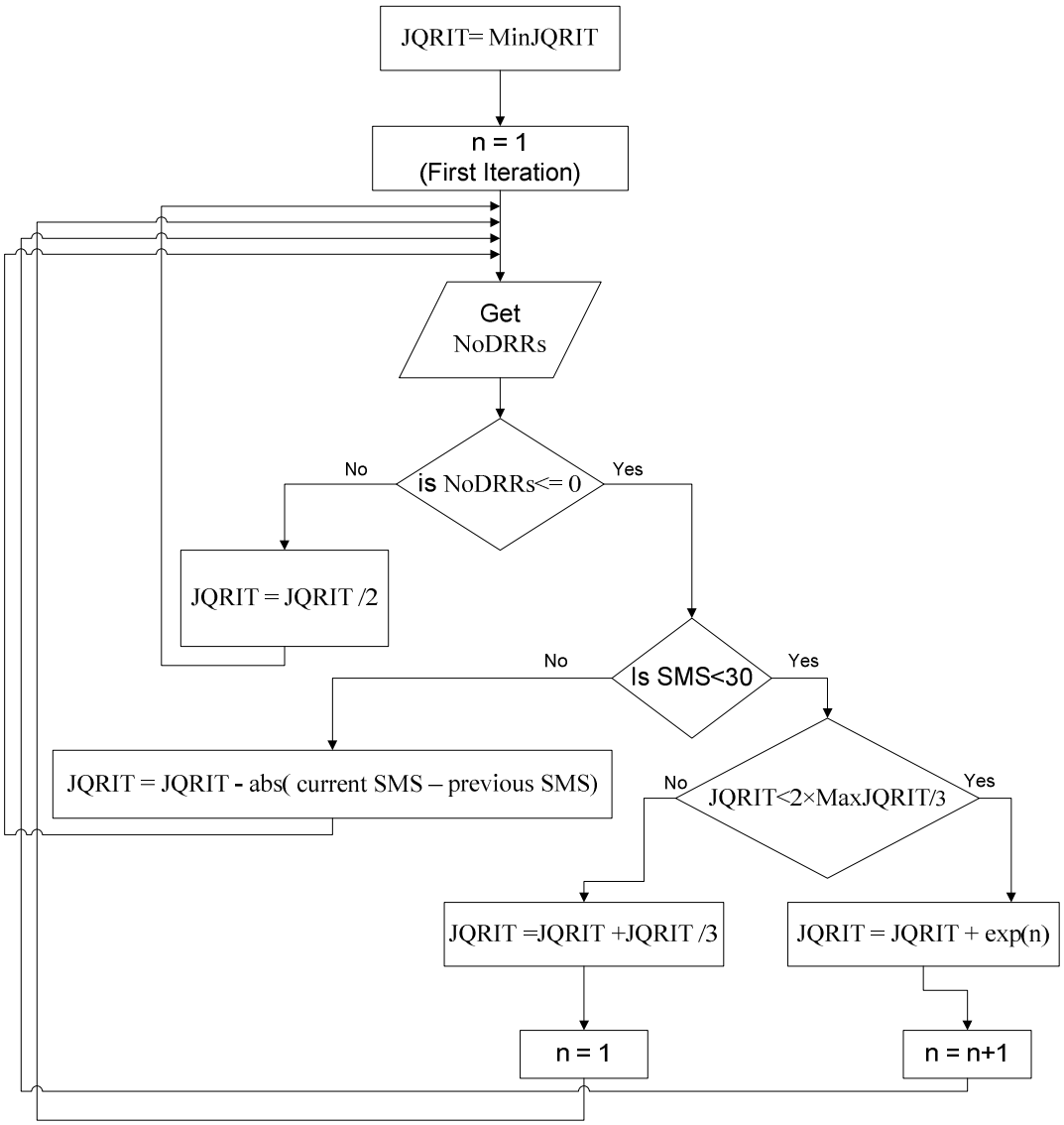


Figure 1. Flow diagram of SSJQR algorithm (where NoDRRs parameter refers to the number of *DelegateRefreshRequests*).

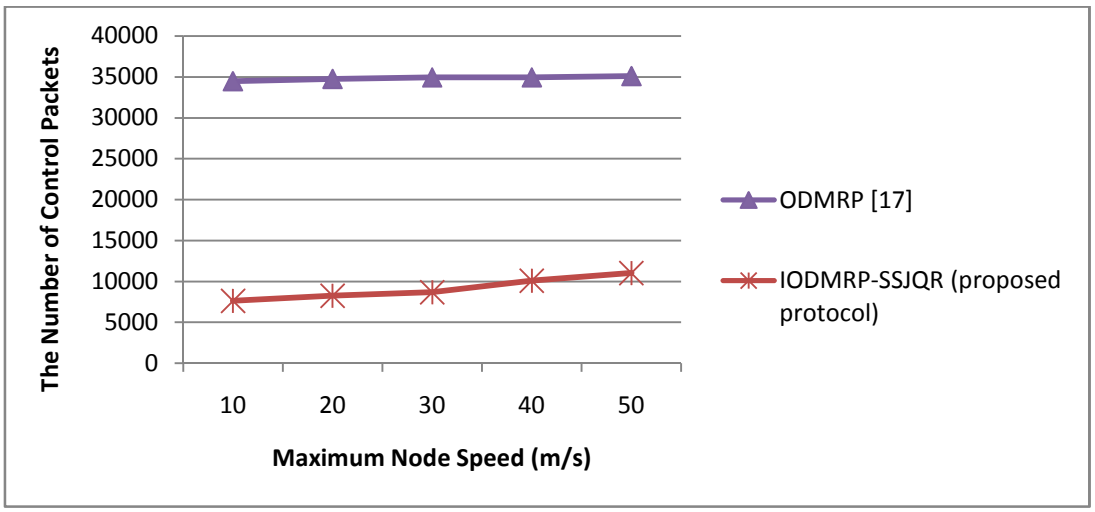


Figure 2. The number of control packets with different node speeds.

V. CONCLUSION

In this article, a new on-demand multicast routing protocol has been proposed as an improvement to ODMRP in Mobile Ad hoc networks, that is, IODMRP-SSJQR. In this protocol, two new algorithms have been proposed which include local recovery and dynamic join query refreshment interval time. Interestingly, the join query refreshment interval time is selected based on the SMS and number of disconnections reported by multicast members. The performance of this proposed protocol has been examined thoroughly utilizing the number of control packets metric. Many simulation experiments have been conducted using QualNet 7.1 and eventually concluded that our proposed protocol achieves an extraordinary improvement over ODMRP with respect to control overhead.

REFERENCES

[1] Raed Al-Zubi, Marwan Krunz, Ghazi Al-Sukkar, Mohammed Hawa, and Khalid A. Darabkh, "Packet Recycling and Delayed ACK for Improving the Performance of TCP over MANETs*," Wireless Personal Communications,* vol. 75, no. 1, pp. 943-963, March 2014.

[2] Mohammad Shurman, Mohammad Alfawares , Mamoun F. Al-Mistarihi, and Khalid A. Darabkh, "A Collaborative Reputation Approach to Avoid Misbehaving Nodes in MANETs," *In Proceedings of the 2014 IEEE International Multi-Conference on Systems, Signals & Devices, Conference on Communication & Signal Processing,* Castelldefels-Barcelona, Spain, pp. 1-4, February 2014.

[3] M. Shurman, M. Al-Mistarihi, and Khalid A. Darabkh, "Merging Dynamic Address Autoconfiguration and Security Key Protocols in Mobile Ad Hoc Networks," *In Proceedings of 36th IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2013),* Opatija, Croatia, pp. 441-445, May 2013.

[4] K. A. Darabkh and R. S. Aygün, "TCP Traffic Control Evaluation and Reduction over Wireless Networks Using Parallel Sequential Decoding Mechanism," *EURASIP Journal on Wireless Communications and Networking*, vol. 2007, Article ID 52492, pp. 1-16, 2007.

[5] K. A. Darabkh, "Evaluation of Channel Adaptive Access Point System with Fano Decoding," *International Journal of Computer Mathematics*, vol. 88, no. 5, pp. 916–937, March 2011.

[6] Khalid A. Darabkh, "Fast and Upper Bounded Fano Decoding Algorithm: Queuing Analysis," *Transactions on Emerging Telecommunications Technologies*, DOI: 10.1002/ett.2929.

[7] Xu Li, Tianjiao Liu, Ying Liu, and Yan Tang, "Optimized multicast routing algorithm based on tree structure in MANETs," *China Communications,* vol.11, no.2, pp.90-99, February 2014.

[8] Yanwei Wang, F.R. Yu, H. Tang, and Minyi Huang, "A Mean Field Game Theoretic Approach for Security Enhancements in Mobile Ad hoc Networks," *IEEE Transactions on Wireless Communications,* vol.13, no.3, pp.1616-1627, March 2014.

[9] Wei Liu and Ming Yu, "AASR: Authenticated Anonymous Secure Routing for MANETs in Adversarial Environments," *IEEE Transactions on Vehicular Technology,* vol.63, no.9, pp.4585-4593, November 2014.

[10] Junfei Xie, Yan Wan, J.H. Kim, Shengli Fu, and K. Namuduri, "A Survey and Analysis of Mobility Models for Airborne Networks," *IEEE Communications Surveys & Tutorials,* vol.16, no.3, pp.1221-1238, Third Quarter 2014.

[11] Wei-Liang Shen, Chung-Shiuan Chen, K.C.-J. Lin, and K.A. Hua, "Autonomous Mobile Mesh Networks," *IEEE Transactions on Mobile Computing,* vol.13, no.2, pp.364-376, February 2014.

[12] Wei Sun, Zheng Yang, Xinglin Zhang, and Yunhao Liu, "Energy-Efficient Neighbor Discovery in Mobile Ad Hoc and Wireless Sensor Networks: A Survey," *IEEE Communications Surveys & Tutorials,* vol.16, no.3, pp.1448-1459, Third Quarter 2014.

[13] A. Clementi, F. Pasquale, and R. Silvestri, "Opportunistic MANETs: Mobility Can Make Up for Low Transmission Power," *IEEE/ACM Transactions on Networking,* vol.21, no.2, pp.610-620, April 2013.

[14] Dinh Danh Le, M. Molnar, and J. Palaysi, "Multicast routing in WDM networks without splitters," *IEEE Communications Magazine,* vol.52, no.7, pp.158-167, July 2014.

[15] C. De Morais Cordeiro, H. Gossain, and D.P. Agrawal, "Multicast over wireless mobile ad hoc networks: present and future directions," *IEEE Network,* vol.17, no.1, pp.52-59, Jan/Feb 2003.

[16] S. Lee, W. Su, and M. Gerla, "On-Demand multicast routing protocol," *In Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '99)*, New Orleans, LA, USA, pp. 1298–1302, September 1999.

[17] S. J. Lee, W. Su, and M. Gerla, "On-demand multicast routing protocol in multihop wireless mobile networks," *Mobile Networks and Applications,* vol. 7, no. 6, pp. 441-453, 2002.

Implementation of a HMIPv6 extension in the INET and xMIPv6 simulation framework with dynamic MAP discovery for OMNeT 4.x

E. Kadušić*, M. Škrbić**, N. Živić***, A. Žgalj* and A. Hantalašević*
*BH Telecom d.d., Sarajevo, Bosnia and Herzegovina
**Elektrotehnički fakultet, Sarajevo, Bosnia and Herzegovina
***Naturwissenschaftlich-Technische Fakultät, Department Elektrotechnik und Informatik, Siegen, Germany
esad.kadusic@gmail.com

Abstract - In order to perform a simulative evaluation of the Hierarchical Mobile IPv6 protocol (HMIPv6) OMNeT++ simulation environment was used. OMNeT++ simulation package – INET framework and xMIPv6 was enhanced with an extension for HMIPv6, which includes Neighbor Discovery protocol, IPv6 router and the implementation of a Mobility Anchor Point (MAP). The simulation model was implemented as defined in RFC 5380 Hierarchical Mobile IPv6 – HMIPv6 Mobility Management. Enhancements are recognized in terms of an upgraded Router advertisement (RA) packet with option data about the MAP. We used dynamic MAP discovery, based on MAP option propagation in Router Advertisements from the MAP to the Mobile Node through certain (configured) router interfaces, within the routers in a network.

I. INTRODUCTION

The increase of mobile users, emergence of different radio access technologies and the increasing importance of IP services over wireless as well as wired networks, raises a number of new challenges. Delay sensitive applications, such as VoIP, have to guarantee IP service continuation during handover. In that sense Mobile IP (MIP) and Mobile IPv6 (MIPv6) were proposed to provide application-transparent IP-based mobility support.

Mobile IP is designed for mobility management of IP networks, but may result with high latency, packet loss and signaling overhead during handoff. Therefore, efficient handover implementation requires advanced mobile mechanisms that enhance Mobile IP, as well as adequate QoS support suited to real-time applications.

Hierarchical Mobile IPv6 (HMIPv6) is proposed to improve the performance of MIPv6, on signaling load and potentially reduce long handover latencies. In order to perform a simulative evaluation of the Hierarchical Mobile IPv6 protocol (HMIPv6) we used the OMNeT++ simulation environment and its communication networks simulation package INET Framework.

With the objective to implement a HMIPv6 simulation model for OMNeT++ 4.x, enhancements to the INET framework were needed, since the Extensible MIPv6 (xMIPv6) simulation model only provides mechanisms for the implementation of various MIPv6 protocol extensions.

We enhanced the INET Framework simulation model with an extended RA packet containing a MAP option, and upgraded the INET framework with dynamic MAP discovery. With this enhancement a valid HMIPv6 simulation model for OMNeT 4.x was established, for further simulation enhancements, in terms of simulation accuracy and validation tests.

II. THEORETICAL BACKROUND

Hierarchical mobility management for Mobile IPv6 is designed to reduce the amount of signaling between the Mobile Node, its Correspondent Nodes, and its Home Agent. HMIPv6 introduces a new entity called the Mobility Anchor Point (MAP), which acts as a local Home Agent (HA) for Mobile Node, to limit the amount of Mobile IPV6 signaling outside the local domain, and improve the performance of Mobile IPv6 in terms of handover speed [1]. When a Mobile Node performs handover between two access routers within the same MAP domain, it will send a BU (Binding Update) only to the MAP, not to its HA [2].

HMIPv6 protocol was developed to separate micro-mobility and macro-mobility, in regard to this RFC 5380 introduces extensions to Mobile IPv6 and IPv6 Neighbour Discovery to allow for local mobility handling. HMIPv6, as an extension of Mobile IP for micro-mobility management, involves three phases:

- MAP Discovery,
- MAP Registration,
- Packet Forwarding.

A. *MAP discovery*

The process of MAP Discovery continues as the Mobile Node moves from one subnet to the next. Every time the Mobile Node detects movement, it will also detect whether it is still in the same MAP domain. The Router Advertisement used to detect movement will also inform the mobile node, through Neighbor Discovery and the MAP option, whether it is still in the same MAP domain. The mobile node will always need to know the original sender of any received packets to determine if route optimization is required [1].

For a HMIPv6-aware implementation the Mobile Node will first register with a MAP by sending it a BU containing its Home Address and on-link address (LCoA). The Home Address used in the BU is the RCoA.

The MAP must store this information in its Binding Cache to be able to forward packets to their final destination when received from the different correspondent nodes or HAs. As the Mobile Node roams within a MAP domain, it will continue to receive the same MAP option included in Router advertisements from its AR. If a change in the advertised MAP's address is received, the Mobile Node must act on the change by sending Binding Updates to its HA and Correspondent nodes (CN) [1].

MAP Discovery is based on propagating the MAP option in Router Advertisements from the MAP to the Mobile Node through configured router interfaces within the routers. Router Advertisements are used for MAP Discovery by introducing a new option.

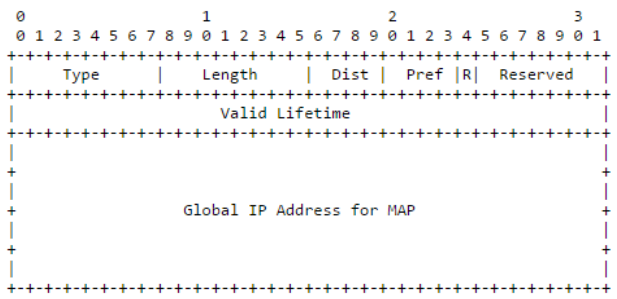


Figure 1. Neighbour Discovery Extension: The MAP Option [1]

The access router is required to send the MAP option (Figure 1. in its Router Advertisements, which are send while informing about its link presence within the IPv6 Neighbor Discovery protocol. The Router Advertisement message format has an ICMP header and message structure, like Figure 2. shows.

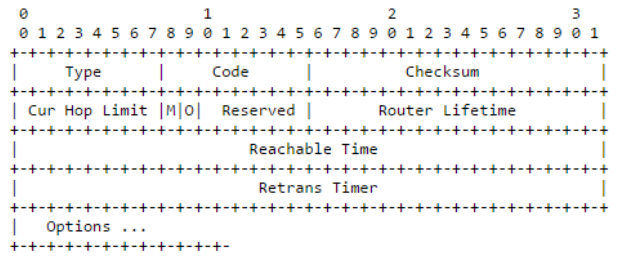


Figure 2. Router Advertisement Message Format [3]

Router Advertisements contain prefixes that are used for determining whether another address share the same link (on-link determination).

The MAP option must be available to all access routers (ARs), to be able to advertise it. The HMIPv6 specification introduces two different methods that access routers can be triggered to advertise a MAP option (MAP discovery):

- Static Configuration – Network operator manually configures all the ARs to send the MAP option, or
- Dynamic MAP discovery – Based on propagating the MAP option in Router Advertisements from the MAP to the Mobile Node through certain (configured) router interfaces within the routers.

We chose the implementation of dynamic MAP discovery.

B. MAP registration

Once the MAP has been discovered and selected, the next phase is MAP registration. MN needs to configure an RCoA on the MAP link, using the prefix advertised in the MAP option, and create a binding between the RCoA and LCoA at the MAP. The MN has to send an LBU message to the newly discovered MAP. The LBU message format is shown in Figure 3.

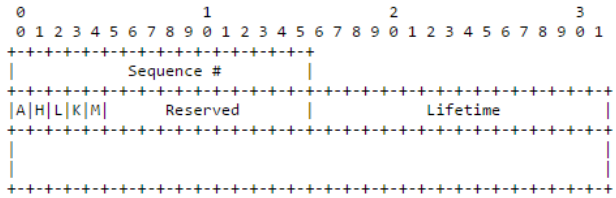


Figure 3. Local Binding Update [1]

A new flag is added: the M flag, which indicates MAP registration. When a Mobile Node registers with the MAP, the M and A flags must be set to distinguish this registration from a BU being sent to the HA or a correspondent node. If flag M is set to 1 it indicates a MAP registration.

The MN uses RCoA as the source IP address in all its packets, but tunnels them to the MAP using LCoA as the source IP address for the tunnel. In the next phase the MAP forwards packets towards the HA and CN's.

C. Packet forwarding

Following a successful registration with the MAP, a bi-directional tunnel between the Mobile Node and the MAP is established. All packets sent by the Mobile Node are tunneled to the MAP, which acts like a HA. The BAck message outer header contains the Mobile Node's LCoA in the source address field and the MAP's address in the destination address field for all outgoing packets. The inner header contains the Mobile Node's RCoA in the source address field and the peer's address in the destination address field.

For incoming packets the tunnel has to check the addresses with the changed source and destination address fields. MAP establishes a binding cache entry for RCoA, which enables packet forwarding of incoming packets to LCoA. For contrary direction a binding entry between the destination address of the tunnel and the interior package is established, to ensure internal packet forwarding to the HA and the CN's.

III. SIMULATION MODEL

To perform a simulative evaluation of the Hierarchical Mobile IPv6 protocol (HMIPv6) OMNeT++ simulation environment was used. OMNeT++ is an extensible, modular, component-based C++ simulation library and framework, primarily for building network simulators [4].

OMNeT++ requires the INET Framework, which is an open-source communication networks simulation package for the OMNeT++ simulation environment. The INET Framework contains models for several wired and wireless networking protocols, including UDP, TCP, SCTP, IP, IPv6, Ethernet, PPP, 802.11, MPLS, OSPF, and many others [5].

Since the earlier released Extensible MIPv6 (xMIPv6) simulation model [6], [7] does not scale well to new developments in the INET framework, enhancements to the INET framework were needed, to enable the implementation of HMIPv6 in the latest versions of OMNeT++ 4.x and INET Framework.

Authors [7] released IPv6SuiteWithINET simulation model to the OMNeT++ community with the base MIPv6 implementation in OMNeT++ 3.2, and assured that HMIPv6 protocol has been implemented, but needed to undergo validation tests, to be released in the coming months.

The available xMIPv6 simulation material was used as a base for our INET simulation model with HMIPv6 extension, which was implemented in accordance to RFC 5380 Hierarchical Mobile IPv6 – HMIPv6 Mobility Management. Several upgrades and modifications to the INET Framework simulation model were needed to create a working simulation model for HMIPv6, using thereby the latest RFC specification, and latest versions of OMNeT++ 4.x and INET Framework.

IV. SIMULATION RESULTS

The process of MAP discovery is needed to inform Mobile Nodes about the presence of the MAP and the distance of the MAP from the Mobile Node, and it continues as the Mobile Node moves from one subnet to the next. Every time the Mobile Node detects movement, it will also detect whether it is still in the same MAP domain.

The INET Framework simulation model was enhanced in a way that the RA packet contains a MAP option (Figure 4. which includes the MAP's global IP address, valid lifetime field which indicates the validity of the MAP's address and consequently the time for which the RCoA is valid, and other information like the distance vector from the MAP to the current access router (measured in hops, but need not be interpreted as the number of hops between MAP and the Mobile Node. Distance is set to 1 if the MAP is on the same link as the Mobile Node), the preference for this particular MAP (decimal value of 15 indicates the highest availability).

We upgraded the INET framework with an implementation of dynamic MAP discovery. In our simulation model the MAP_1 node was configured to act like a MAP on two network interfaces – Firstly as a MAP to the Internet (Internet Gateway node), whose network prefix is used for RCoA calculation, and secondly as a MAP toward to the Mobile Node which visits the MAP domain to forward packets, i.e. bi-directional tunneling, which enables sending of RA packets with MAP option, to routers in the downstream network.

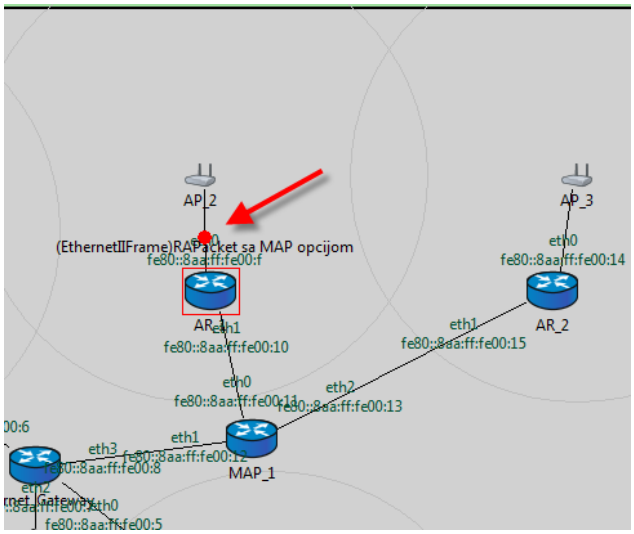


Figure 4. Router Advertisement packet with additional MAP option

Figure 5. shows the sequence diagram for the RA packet with MAP option in the HMIPv6 simulation model. AR_1 and AR_2 Router6 modules receive RA packets with MAP option from the MAP. Router6 modules were enhanced and configured to carry MAP options through interfaces connected with WLAN access points (AP_2 and AP_3), previously increasing Distance field for 1.

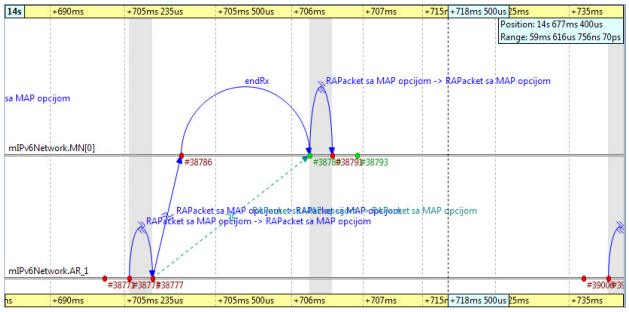


Figure 5. MAP option sequence diagram

At the end of the MAP discovery phase, the HMIPv6-aware Mobile Node MN[0] receives the Router Advertisement with MAP option, obtains the MAP IP address, obtains the Distance field of the MAP form current access router, and verfies if the lifetime is equal zero for the current MAP (a valid lifetime of zero indicates a MAP failure). If it is non-zero then MN[0] has to select a MAP with higher preferences. We enhanced the MN[0] WirelessHost6 module to process RA packets with MAP option.

Once the MN[0] has discovered and chosen MAP_1, it configures RCoA address, using the network prefix, advertised in the MAP option, and creates a binding between RCoA and LCoA at the MAP, assuming MN[0] has already configured a valid LCoA. Our INET framework simulation model was enhanced in the way that MN[0] sends the LBU packet to the newly discovered MAP (Figure 6.

The source IP address is a valid LCoA address, while RCoA is the Home Address in the LBU message. MAP_1

performs DAD (Duplicate Address Detection) procedure for the RCoA and sends BAck, which is identical to the MIPv6 BAck packet.

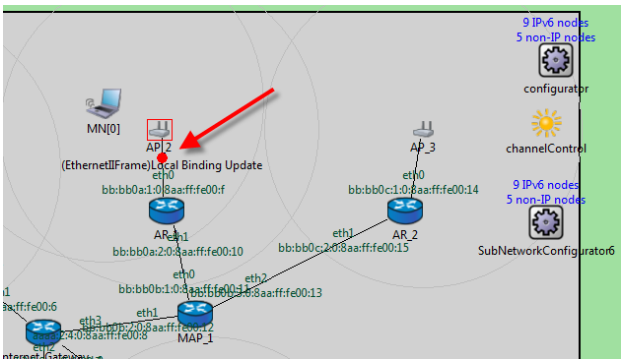


Figure 6. MN[0] sends LBU to discovered MAP

If BAck does not contain type 2 router header with the RCoA, MN[0] will reject BAck, previously controlling the error code in BAck, to check if a new RCoA should be calculated and a new LBU message send. Figure 7. shows the LBU message sequence diagram.

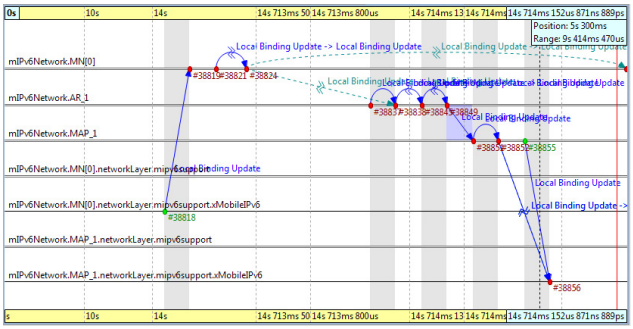


Figure 7. LBU sequence diagram

If a successful binding with the MAP cannot be done, MN[0] will approach the basic MIPv6 and send BU message to the Home Agent (HA) using LCoA as CoA address. If the binding with the MAP, MN[0] will use its RCoA as CoA with Home Agents and Correspondent Nodes. MN[0] has to wait for the BAck, before sending BU messages to Home Agents, or introducing the Return Routability procedure with Correspondent Nodes.

It should be noted that when binding the RCoA with the HA and Correspondent Nodes, the binding lifetime must not be larger, than the Mobile Node's binding lifetime with the MAP, which is received in the Binding Acknowledgement.

Following a successful registration with the MAP, a bi-directional tunnel between the Mobile Node and the MAP is established. All packets sent by the Mobile Node are tunneled to the MAP, which acts like a Home Agent. The BAck message outer header contains the Mobile Node's LCoA in the source address field and the MAP's address in the destination address field for all outgoing packets. The inner header contains the Mobile Node's RCoA in the source address field and the peer's address in the destination address field.

For incoming packets the tunnel has to check the addresses with the changed source and destination address fields. MAP establishes a binding cache entry for RCoA, which enables packet forwarding of incoming packets to LCoA. For contrary direction a binding entry between the destination address of the tunnel and the interior package is established, to ensure internal packet forwarding to the HA and the CN's.

The structure of the simulation model was defined in the NED (Network Description) language, and the enhancement implies a new module, i.e. SubNetworkConfigurator6 (Figure 8. needed to configure an IPv6 subnet in OMNeT++ 4.x. SubNetworkConfigurator6 module configures subnet prefixes for MAP_1, AR_1 and AR_2.

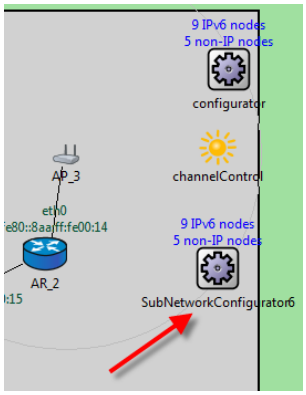


Figure 8. SubNetworkConfigurator6 module

The following table contains all files used in the simulation model, including their role description.

TABLE 1 INET FRAMEWORK FILES ENHANCED WITH HMIPV6 EXTENSION

File	Description
\inet\examples\mobileipv6\ HMIPv6Network.ned	Network topology definition in NED language
\inet\examples\mobileipv6\ omnetpp.ini	Configuration folder for simulation mode control
\inet\src\networklayer\autorouting \ipv6\ SubNetworkConfigurator6.cc	Configures IPv6 adressess and routing tables for hierarhy network topology
\inet\src\networklayer\icmpv6\ IPv6NeighbourDiscovery.cc	Implements Neighbor Discovery protocol for IPv6 (RFC 2461)
\inet\src\networklayer\icmpv6\ IPv6NDMessage_m.cc	HA information, HA Prefix information, MAP Option, etc.
\inet\src\networklayer\xmipv6\ xMIPv6.cc	Implements RFC 3775 Mobility Support for IPv6
\inet\src\nodes\ipv6\ Router6.ned	IPv6 router NED module

V. CONLUSION

We implemented an INET simulation model with HMIPv6 extension in accordance to RFC 5380 Hierarchical Mobile IPv6 – HMIPv6 Mobility Management. In that sense we enhanced the INET Framework simulation model with an extended RA packet containing a MAP option. Further we upgraded the INET framework with an implementation of dynamic MAP discovery. We made it possible for our simulation Mobile Node MN[0] to send the LBU packet to the newly discovered MAP.

Since a valid HMIPv6 simulation model for OMNeT 4.x has been established, further simulation enhancements are enabled in terms of simulation accuracy and validation tests. Further work will be directed to the development of an extended resource management scheme, which takes into account the QoS requirements of handover, such as packet delay, jitter and loss guarantees, and reduces the overall handover latency.

REFERENCES

[1] RFC 5380 Hierarchical Mobile IPv6 (HMIPv6) Mobility Management, 2008.

[2] Sophon Mongkolluksamee, Vasaka Visoottiviseht; "Diffserv Conditionalized Handover for HMIPv6", Feb. 17-20, 2008 ICACT 2008.

[3] RFC 4861 Neighbor Discovery for IP version 6 (IPv6), 2007.

[4] OMNeT++ Community, http://www.omnetpp.org/ [05.01.2015.].

[5] INET framework, http://inet.omnetpp.org/ [05.01.2015.].

[6] Project xMIPv6: Mobility Management Engine for the NGN, http://www.kn.e-technik.tu-dortmund.de/de/forschung/ausstattung/xmipv6.html [12.12.2014.].

[7] Faqir Zarrar Yousaf, Christian Bauer, and Christian Wietfeld. An accurate and extensible mobile IPv6 (xMIPV6) simulation model for OMNeT++. Proc. of the 1st Int. Conf. on Simulation tools and techniques for communications, networks and systems & workshops (Simutools '08). ICST Brussels, Belgium, Article no. 88, 8 pp.

NEW METHODS & SOLUTIONS

Matrix inversion in the design of active queue management mechanisms

Pawel Mrozowski and Andrzej Chydzinski
Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland
andrzej.chydzinski@polsl.pl

***Abstract*—A router exploiting the active queue management can drop arriving packets even if its buffer is not full. This is meant to notify the TCP senders to reduce their sending rates. One type of the active queue management is based on dropping arriving packets with probability that is a function of the queue size observed upon packet arrival. Although the queuing models of algorithms of this type have been developed, they present some numerical difficulties that have to be overcome in order to be practically usable, e.g. they require inverting large matrices. In this paper we compare several inversion methods with respect to their applicability in the parameterization of the dropping mechanism of the active queue management. In addition to their implementations on a CPU, we also check their performance when implemented on a graphics processor unit.**

***Index Terms*—active queue management, Internet routers, dropping function, numerical calculations, matrix inversion, graphics processor unit**

I. Introduction

The idea of the active queue management (AQM) in Internet routers is that arriving packets can sometimes be dropped even if the router's buffer is not full (see e.g. [1], [2], [3], [4], [5], [6]). This action is supposed to notify the TCP senders to reduce their sending rates, before the actual congestion happens at the bottleneck link. It has several positive effects discussed in the literature. Two the most important benefits are reduction of queue sizes and queueing delays as well as reduction of the effect of global synchronization of TCP flows.

An important type of the active queue management is based on dropping the arriving packets with probability that is a non-decreasing function of the queue size observed upon packet arrival. In the well-known RED (Random Early Detection) algorithm, [1], this function is linear. It can have other forms as well, e.g. exponential or doubly linear.

The queuing models of such algorithms have been solved in [7], [8], [9], using transform techniques[1]. In particular, the formulas for the most important performance characteristics, including the average queue size and the overall packet drop ratio, were provided. These formulas, however, present some numerical difficulties that have to be overcome in order to be practically usable. Especially demanding, but also the most practically useful, are the models which are able to mimic the self-similar structure of the network traffic. Their usefulness is connected with the fact that since the famous paper by Leland et. al., [12], we have known that network traffic posses some sophisticated statistical characteristics (e.g. self similarity), which have to be taken into account in models of queues in routers. Otherwise, all the obtained performance characteristics are optimistically underestimated. On the other hand, the models mimicking self-similarity are computationally very demanding. In particular, they require inverting huge matrices.

Detailed formulas will be provided in the next section, but just to get a picture of these demands, consider a link of capacity 1Gb/s and the average packet size on this link of 1000 bytes. The commonly used rule of thumb for dimensioning the routers buffers (e.g. [13]) is the bandwidth-delay product for 100ms or 200ms of buffering delay. Therefore, in our example we obtain the buffer for 12500 packets. On the other hand, a fine mimicking of the self-similar properties of the traffic may require matrices of sizes as big as 56×56 (see e.g. [14]). Therefore, as follows from formulas (1), (2) and (3) of Section II, we have to invert a matrix with $(56 \times 12501)^2 \approx 49 \times 10^9$ elements. Assuming a floating-point number format that occupies 8 bytes, the memory requirement for such matrix would be almost 4 tera bytes. Inverting matrices of this size is not trivial.

In practice the problem is a bit smaller due to the following two factors. Firstly, a more rough parameterization of the traffic can be used, for instance with 5×5 matrices (see e.g. [10], [15]), or even 3×3 (see e.g. [16]). Secondly, the application of the active queue management allows for a significant reduction of the buffer sizes. Therefore, instead of 12500, the buffer for 1000-2000 would be large enough. All off these lead to matrices with hundreds of milion elements, which still may take a lot of time to invert, when approached in a standard way.

Therefore, in this paper we compare several inversion methods with respect to their applicability in the parameterization of the dropping mechanisms in the active queue management. We try to speed up the computations in two ways: by using various algorithms for matrix inversion and multiplication and by implementing them on a GPU and CPU with multiple cores. Apparently, for a two-hundred-million-element matrix these operations can reduce the computation time by at least three orders of magnitude.

The rest of the paper is organized as follows. In Section II, the active queue management model and the most important formulas are presented. In Section III, the inversion algorithms are recalled. Then, in Section IV, the modifications and imple-

[1]For other queueing models solved in this way see e.g. [10], [11].

mentations on GPUs and multiple core CPUs are discussed. Section V presents the results of the tests performed using eleven different implementations of the inversion algorithms. In Section VI remarks concluding the paper are presented.

II. Queuing model

We deal with a single-server queue whose arrival process is the Markov-modulated Poisson process parameterized by the intensity matrix of the modulating chain, Q, of size $m \times m$, and the vector of corresponding intensities of Poisson arrivals, $[\lambda_1, \ldots, \lambda_m]$. Let $\Lambda = diag[\lambda_1, \ldots, \lambda_m]$.

The service time is of general type, described by a distribution function $F(t)$. The total number of packets present in the system is limited by the buffer size, b. Packets arriving when the buffer is full are dropped and lost.

Moreover, any arriving packet can be dropped with probability $d(n)$, where n denotes the queue length at the time of arrival of this packet. (This is the active queue management mechanism). Therefore, the dropping probability is a function of the queue length. The finite buffer means that there is $d(n) = 1$ for $n \geq b$.

Let I denote the unit matrix of size $m \times m$, $\mathbf{0}$ denote an $m \times m$ matrix of zeros, $\mathbf{1}$ denote the column vector of 1's, and δ_{ij} is 1 if $i = j$ and 0 otherwise. In [9] we proved that the Laplace transform of the queue length distribution at time t has the following form:

$$\phi(s,l) = G^{-1}(s)R(s,l), \qquad 0 \leq l \leq b, \tag{1}$$

with

$$G(s) = [G_{ij}(s)]_{i=0\ldots b, j=0\ldots,b}\,, \tag{2}$$

and

$$G_{ij}(s) = \begin{cases} A_{i,j-i+1}(s), & \text{if } 1 \leq i \leq b-2, i < j \leq b-1, \\ A_{i,1}(s) - I, & \text{if } i = j, 1 \leq i \leq b-1, \\ A_{i,0}(s), & \text{if } i = j+1, \\ U_0(s) - I, & \text{if } i = 0, j = 0, \\ V_0(s), & \text{if } i = 0, j = 1, \\ -I, & \text{if } i = b, j = b, \\ \mathbf{0}, & \text{otherwise,} \end{cases} \tag{3}$$

where

$$R(s,l) = [R_0(s,l), \ldots, R_b(s,l)]^T\,, \tag{4}$$

$$R_i(s,l) = \begin{cases} -\delta_{0l} z(s), & \text{if } i = 0, \\ -r_i(s,l), & \text{if } 1 \leq i \leq b, \end{cases}$$

$$A_{n,k}(s) = [a_{n,k,i,j}(s)]_{i,j}\,, \qquad 1 \leq i,j \leq m,$$

$$a_{n,k,i,j}(s) = \int_0^\infty e^{-st} H_{n,k,i,j}(t) dF(t),$$

$$U_n(s) = \left[\frac{(\lambda_i - Q_{ii})p_{ij} + d(n)\Lambda_{ij}}{s + \lambda_i - Q_{ii}}\right]_{i,j}, \qquad 1 \leq i,j \leq m, \tag{5}$$

$$p_{ij} = \begin{cases} 0 & \text{if } i = j, \\ \frac{Q_{ij}}{\lambda_i - Q_{ii}} & \text{if } i \neq j, \end{cases}$$

$$V_n(s) = \left[\frac{(1 - d(n))\Lambda_{ij}}{s + \lambda_i - Q_{ii}}\right]_{i,j}, \qquad 1 \leq i,j \leq m, \tag{6}$$

$$z(s) = \left[\frac{1}{s + \lambda_1 - Q_{11}}, \ \ldots\ , \frac{1}{s + \lambda_m - Q_{mm}}\right]^T, \tag{7}$$

$$r_n(s,l) = \begin{cases} \mathbf{0} \cdot \mathbf{1}, & \text{if } l < n, \\ \overline{D}_{n,l-n}(s) \cdot \mathbf{1}, & \text{if } n \leq l \leq b, \end{cases} \tag{8}$$

$$\overline{D}_{n,k}(s) = [d_{n,k,i,j}(s)]_{i,j}\,, \qquad 1 \leq i,j \leq m, \tag{9}$$

$$d_{n,k,i,j}(s) = \int_0^\infty e^{-st} H_{n,k,i,j}(t)(1 - F(t))dt,$$

and $H_{n,k,i,j}(u)$ denoting the counting function of the arrival process filtered by the dropping mechanism.

One of the main challenges when using formula (1) in practice is inversion of matrix $G(s)$. As $A_{i,j}(s)$, $U_0(s)$, $V_0(s)$, I and $\mathbf{0}$ are square matrices of size $m \times m$, from (2) and (3) it follows that matrix $G(s)$ is of size $(b+1)m \times (b+1)m$. Therefore, as discussed in Introduction, it can be huge even if the buffer size is moderate.

III. Inversion algorithms

In this section we recall briefly a few chosen inversion methods. To allow calculations for large matrices, the researchers try to reduce the computational complexity $O(\omega)$ by decreasing ω bound, [17], [18], [19]. Other ideas exploit more computing power, by using distributed programming and parallel programming. The latter approach focuses often on cheap computing power of general-purpose graphics processing units (GP-GPUs, see e.g. [20], [21]). Finally, there is a research direction on inverting sparse matrices, but this is beyond the scope of this paper - the matrices are not sparse herein.

A. Gaussian elimination

Gaussian elimination is the most popular method for finding the inverse of a square matrix. The time complexity of Gaussian elimination is $O(n^3)$. In order to calculate A^{-1}, the following elementary operations on matrix rows must be done.

1. For each row $k = 1, \ldots, n$ calculate:
   $a_{k,j}^{k+1} = \frac{a_{k,j}^k}{a_{k,k}^k}$ for $j = 1, \ldots, n$ and $a_{k,k} \neq 1$.
2. If $a_{k,k} = 1$ then:
   $a_{i,j}^{k+1} = a_{i,j}^k - a_{i,k}^k a_{k,j}^{k+1}$ for $i, j = 1, \ldots, n$ and $i \neq k$.
3. Extract inverted matrix:
   $[A|I] = \left[I|A^{-1}\right]$.

B. LU decomposition

The method was invented by A. Turing and it exploits the factorization of matrix A:

$$A = LU, \tag{10}$$

where L is a lower triangular matrix, while U is an upper triangular matrix. Two most popular decomposition algorithms are by Doolittle and Crout. The first one gives the unit lower triangular matrix, the second gives the unit upper triangular matrix.

1) LU factorization - Doolittle: The algorithm is derived from matrix multiplication:

$$a_{i,j} = \sum_{k=1}^{n} l_{i,k} \cdot u_{k,j}.$$

As L and U are triangular, we consider two cases. For $i \leq j$ we have:

$$u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k} \cdot u_{k,j}. \tag{11}$$

For $i > j$:

$$l_{i,j} = \frac{1}{u_{j,j}} \left(a_{i,j} - \sum_{k=1}^{j-1} l_{i,k} \cdot u_{k,j} \right). \tag{12}$$

The steps of the algorithm are:

1. Create matrices L, U and fill them with zeros, the diagonal in L fill with ones.
2. For $i = 1, 2, \ldots, j$ calculate $u_{i,j}$ using (11).
3. For $i = j+1, j+2, \ldots, n$ calculate $l_{i,j}$ using (12).

Finally, obtaining A^{-1} is done by forward or backward substitution on the columns of identity matrix.

The LU method is easy to implement and is relatively fast. It is also worth mentioning that there are modifications of the LU decomposition, aiming at error minimization, called "partial pivoting". More on this subject can be found in [22], [23].

C. QR decomposition

This factorization of the matrix is done as follows

$$A = QR, \tag{13}$$

where Q is orthogonal, while R is upper triangular. There are at least three different QR decomposition methods: Gram-Schmidt (orthogonalization), Givens rotations and Householder reflections, [24].

The algorithm consists of the following steps.

1. QR decomposition of A.
2. Inversion of the upper triangular factor R.
3. Multiplication according to the formula:

$$A^{-1} = R^{-1} Q^T. \tag{14}$$

D. Strassen method

Although the original paper [25] was devoted to matrix multiplication in time $O(n^{2.807})$, the author showed also that the algorithm breaks down the inversion of matrices into several multiplication operations and that the problem has in fact the same time computational complexity. To apply this method, we have to divide A into four submatrices of size $k = n/2$:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{15}$$

The inverted matrix has a form of block matrix:

$$A^{-1} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \tag{16}$$

$$= \begin{bmatrix} a_{11}^{-1} + a_{11}^{-1} \cdot a_{12} \cdot S^{-1} \cdot a_{21} \cdot a_{11}^{-1} & -a_{11}^{-1} \cdot a_{12} \cdot S^{-1} \\ -S^{-1} \cdot a_{21} \cdot a_{11}^{-1} & S^{-1} \end{bmatrix} \tag{17}$$

where $S = a_{22} - a_{21} \cdot a_{11}^{-1} \cdot a_{12}$ is called Schur complement of a_{11} in matrix A. To speedup multiplications, temporary matrices are denoted by R_i, where $i = 1, \ldots, 7$. We have:

$$\begin{array}{ll} 1: R_1 = a_{11}^{-1}, & 2: R_2 = a_{21} \cdot R_1, \\ 3: R_3 = R_1 \cdot a_{12}, & 4: R_4 = a_{21} \cdot R_3, \\ 5: R_5 = R_4 - a_{22}, & 6: R_6 = R_5^{-1}, \\ 7: c_{12} = R_3 \cdot R_6, & 8: c_{21} = R_6 \cdot R_2, \\ 9: R_7 = R_3 \cdot c_{21}, & 10: c_{11} = R_1 - R_7, \\ 11: c_{22} = -R_6. & \end{array}$$

Finally we must rebuild the inverted matrix according to formula 16.

IV. Modifications

A. CPU parallelization

Analyzing the algorithm described in the previous subsection, we can see that some of its steps can be parallelized, since there are no internal dependencies between them. In particular, the operations can be divided into seven groups, I-VII, which are performed consecutively, while operations inside each group can be done in parallel way:

$$\begin{array}{ll} I: 1; & II: 2, 3; \\ III: 4; & IV: 5; \\ V: 6; & VI: 7, 11, 8; \\ VII: 9, 10. & \end{array}$$

B. Modified multiplication

In the Strassen method, the crucial role is played by multiplication operation, which influences severely the overall method complexity. Therefore we will discuss now multiplication methods, their improvements and GPU implementations.

1) Classic multiplication: If A is a $N \times M$ matrix and B is a $M \times P$ matrix, then matrix product AB is a matrix composed of the following elements:

$$(AB)_{ij} = \sum_{k=1}^{M} A_{ik} B_{kj} \quad for \;\; 1 \leq i \leq N \;\; and \;\; 1 \leq j \leq P. \tag{18}$$

The naive implementation of this formula may have the form:

```
1. For i = 1 to N
2.         For j = 1 to P
3.                 For k = 1 to M
4.                   AB(i,j) = AB(i,j) + A(i,k)B(k,j)
```

Surprisingly, the following minor modification of the code:

```
1. For i = 1 to N
2.         For k = 1 to M
3.                 For j = 1 to P
4.                   AB(i,j) = AB(i,j) + A(i,k)B(k,j)
```

may result in even 10 times shorter computations time. This is due the fact that compilers can optimize the latter code by creating temporary variable where the values $A[i][k]$ are

stored. The last nested loop requires fewer lookups into the arrays.

The classic implementation is available on GPUs. Namely, it is offered by function $cublasDgemm$ Level-3 Basic Linear Algebra Subprograms from [26], which was used in our experiments.

2) Strassen multiplication: In order to multiply two matrices A and B by the means of Strassen method, we use following decomposition of A and B into four sub-matrices. The multiplication is carried out by recursive partitioning into smaller blocks in the following way:

$$AB = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}. \tag{19}$$

Firstly, we compute:

$$\begin{aligned} 1 : M_1 &= (a_{11} + a_{22})(b_{11} + a_{22}), \\ 2 : M_2 &= (a_{21} + a_{22}) b_{11}, \\ 3 : M_3 &= a_{11}(b_{12} - b_{22}), \\ 4 : M_4 &= a_{22}(b_{21} - b_{11}), \\ 5 : M_5 &= (a_{11} + a_{12}) b_{22}, \\ 6 : M_6 &= (a_{21} + a_{11})(b_{11} + a_{12}), \\ 7 : M_7 &= (a_{12} + a_{22})(b_{21} + a_{22}). \end{aligned}$$

Then we obtain:

$$AB = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix},$$

where

$$\begin{aligned} c_{11} &= M_1 + M_4 - M_5 + M_7, \\ c_{12} &= M_3 + M_5, \\ c_{21} &= M_2 + M_4, \\ c_{22} &= M_1 - M_2 + M_3 + M_6. \end{aligned}$$

3) Coppersmith-Winograd multiplication: The algorithm described in [17] is the one with the lowest time complexity in this paper, which is $O(n^{2.376})$. There are numerous modifications in the literature, resulting in slightly better performance than the original version, [18], [19].

In the algorithm, matrices A and B are divided into four sub-matrices the size of $k = n/2$, as in (19). Then, the following temporary matrices are calculated:

$$\begin{aligned} &1 : S_1 = a_{21} + a_{22}, && 2 : S_2 = S_1 - a_{11}, \\ &3 : S_3 = a_{11} - a_{21}, && 4 : S_4 = a_{12} - S_2, \\ &5 : T_1 = b_{12} - b_{11}, && 6 : T_2 = b_{22} - T_1, \\ &7 : T_3 = b_{22} - b_{12}, && 8 : T_{43} = b_{21} - T_2, \\ &9 : P_1 = a_{11} \cdot b_{11}, && 10 : P_2 = a_{12} \cdot b_{21}, \\ &11 : P_3 = S_1 \cdot T_1, && 12 : P_4 = S_2 \cdot T_2, \\ &13 : P_5 = S_3 \cdot T_3, && 14 : P_6 = S_4 \cdot T_4, \\ &15 : P_7 = a_{22} \cdot T_4, && 16 : U_1 = P_1 + P_2, \\ &17 : U_2 = P_1 + P_4, && 18 : U_3 = U_2 + P_5, \\ &19 : U_4 = U_3 + P_7, && 20 : U_5 = U_3 + P_3, \\ &21 : U_6 = U_2 + P_3, && 22 : U_7 = U_6 + P_6. \end{aligned}$$

Finally, we obtain:

$$\begin{aligned} &23 : c_{11} = U_1, && 24 : c_{12} = U_7, \\ &25 : c_{21} = U_4, && 26 : c_{22} = U_5. \end{aligned}$$

The algorithm requires 7 multiplications and 15 additions/substractions.

C. GPU parallelization

The idea behind GPU parallelization lies within the parallel execution of the code using multiple threads. The main difference between CPUs and GPUs is that the latter has a lot more ALUs (arithmetic logic units) performing computations. The structure of a program for GPU consists of one kernel at a time divided into blocks which use independent groups of ALUs. Each block is comprised of threads. The threads within a block typically work together to compute a value and they share the same area of memory.

The main question is how to parallelize the code. The answer is not simple and depends on the algorithm itself. Good candidates are simple, naive algorithms, which perform many independent operations. In our calculations we benefited from large speedup of the following simple algorithms rewritten into CUDA (Compute Unified Device Architecture) [27] kernels: matrix addition, matrix subtraction, matrix multiplication. It is also possible to extend the idea of basic kernels by combining operations, e.g. multiplication with addition, etc.

Currently, a lot of research is carried out to answer the following questions (see e.g. [28]): How many threads per block should be used? How many blocks per grid should be used? The answers are not discussed herein, but they can potentially give a significant performance boost.

As always, there are some drawbacks. The first one is the amount of memory available for a GPU. Today, the most powerful graphical accelerators have 24GB of memory (12GB per GPU). As argued in Introduction, problems can sometimes be much bigger than that. One of the possible solutions is data partitioning. The second problem is that not all GPUs perform well with double-precision values (that is why the most of comparisons in the literature are done on single-precision numbers). In practice, we need to use high-precision numbers quite often. Another problem is relatively high cost (milliseconds) of transferring the data between host and device. Finally, it has to be mentioned that the GPU performs at top speed only when it is busy. Single or occasional jobs scheduled on a GPU take, very often, more time on the GPU, than CPU. The vendors try to overcome this problem by means of following technologies/frameworks like CrystalGPU [30], Grophecy [31] or virtualization, e.g. Pegasus project [32].

V. Test results

We compared the performance of the described above algorithms and their modifications using matrices up to 225 million of elements. Namely, the performance of the following eleven implementations were compared:

- Gaussian elimination on one CPU core,
- LU decomposition on one CPU core,
- QR decomposition on one CPU core,
- LU decomposition on CUDA,
- LU decomposition on 12 CPU cores,
- QR decomposition on 12 CPU cores,
- Strassen method with classic multiplication on 12 CPU cores,

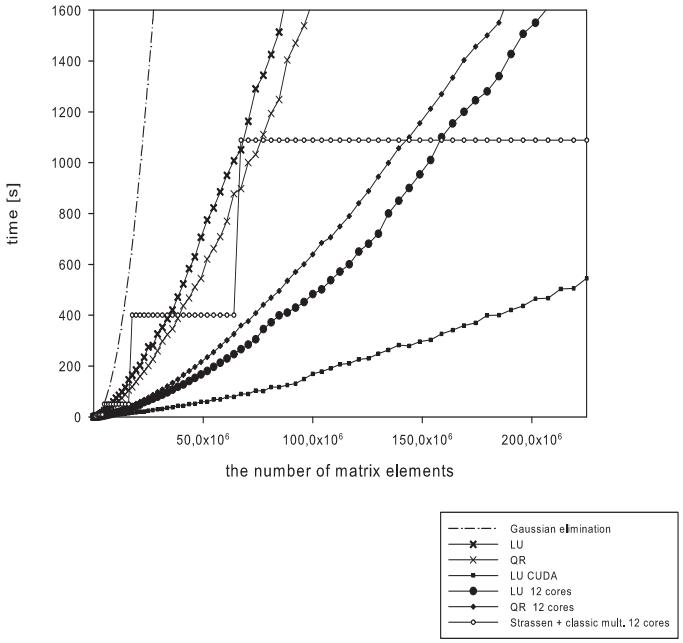


Fig. 1. Inversion time – comparison of various methods.

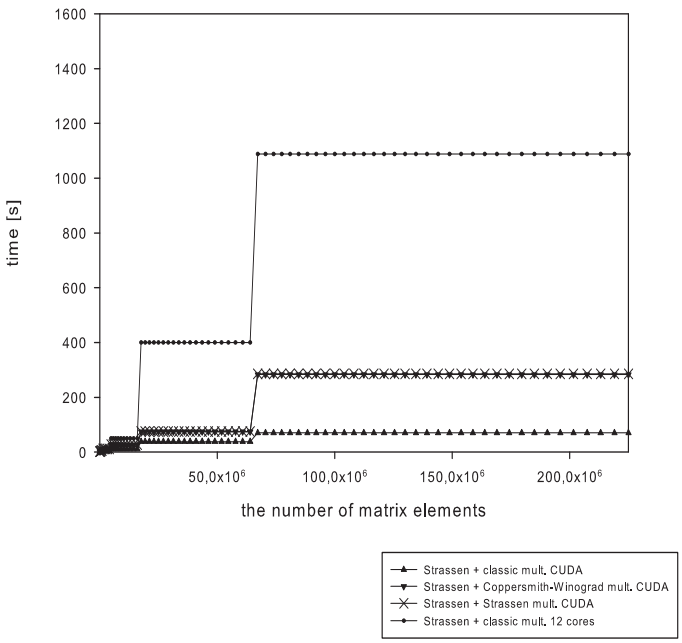


Fig. 2. Inversion time – Strassen method variants.

- Strassen method with classic multiplication on CUDA,
- Strassen method with Coppersmith-Winograd multiplication on CUDA,
- Strassen method with Strassen multiplication on CUDA,
- Gaussian elimination on CUDA.

The implementations were done in Java mostly by authors. As for the LU and QR algorithms, their single-core versions from the library JAMA were used, whilst the multi-core versions – from the library Ojalgo. The Gaussian elimination from the library Apache was exploited.

In the experiments, the inversion times for dense matrices of type (2), (3) in function of their number of elements were checked. For experimental purposes, we used several different workstations and different types of the GPU. The results shown herein were obtained on the workstation with two CPUs of Intel Xeon E5-2630 v2 type, 2.6GHz, 6 cores each, and the GPU of Nvidia Quadro 4000 type, with 3GB of memory and double-precision numbers.

The results are presented in Figs. 1, 2 and 3. In particular, in Fig. 1, the inversion times for various methods and large matrices are depicted. In Fig. 2, the inversion times for different implementations of Strassen method are shown. In Fig. 3, the inversion times for smaller matrices are presented.

As we can see, the Strassen method was the fastest among those we tested. Its version with classic multiplication implemented on CUDA provided the shortest computation time of 70 seconds for a matrix with 225×10^6 elements (see Fig. 2). Surprisingly, the Strassen method with Coppersmith-Winograd multiplication performed a bit worse. Most likely, it requires even larger matrices to fully benefit from its theoretical $O(n^{2.376})$ complexity.

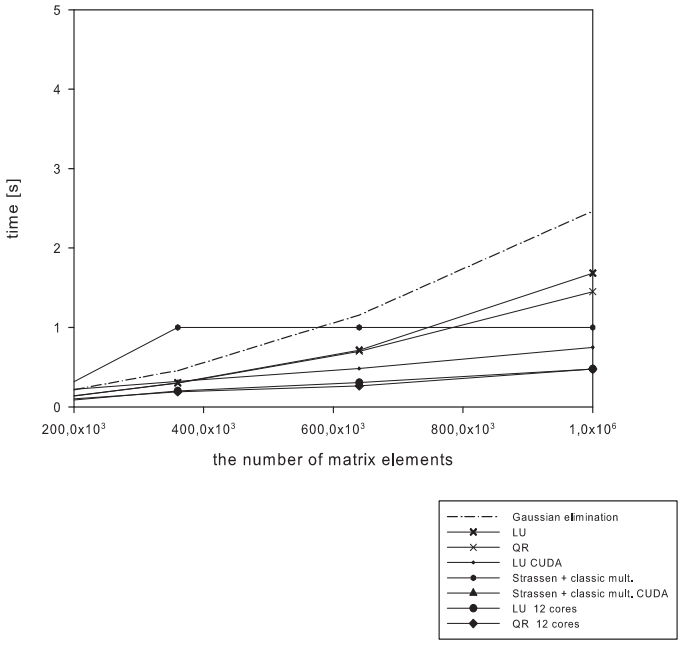


Fig. 3. Inversion time – moderate number of elements.

From Fig. 2 it follows also that exploiting the GPU is very beneficial and outperforms the multicore implementation.

The main disadvantage of the Strassen method is that the matrix size must be the power of 2. Otherwise the matrix must be padded with zeros (see [29]) and put in the left upper corner.

This explains the constant values on Strassen's method plots. This also increases the sizes of matrices to multiply, add and subtract. In the worst case scenario, we perform computations on matrices with almost doubled size we need.

As expected, the slowest method was the Gaussian elimination. The largest matrix we were able to compute using Gaussian elimination consisted of 169 million elements and it took more than 15 hours to finish computations.

For matrices of moderate sizes (up to 10^6 elements), other than Strassen methods can be recommended, e.g. the parallel version of QR algorithm (see Fig. 3). It is also worth underlining the strange behavior of GPU for moderate matrices. As wee can see, the computation time may decrease slightly between $4 \cdot 10^5$ elements and 10^6 elements when CUDA is used. This is the well-known anomaly related to "warm-up" period of the GPU and host-device memory transfer penalty.

The Java programming language turned out to be mature and efficient environment with good libraries for linear algebra. It offers a fully portable environment so we are not restricted by certain operation system and/or CPU architecture. The libraries that enabled us to use CUDA technology within Java are somewhat limited. In our opinion they do not use the full potential of CUDA technology, but they are easy to use and can be used in short time by the developers to speed up calculations. More on Java and GPU applicability to solve numerical problems can be found in [33].

VI. Conclusions

We compared several matrix inversion methods and their implementations for multi-core CPUs and GPUs with respect to their applicability in the design of dropping mechanisms of the active queue management. We demonstrated that the computation time can be reduced by at least three orders of magnitude when a properly chosen algorithm is used and implemented on a GPU. Of course, there is no need to equip the router with a GPU, because all the necessary calculations can be carried out by an external workstation, while the resulting dropping function can be uploaded to the router as a simple configuration file.

Acknowledgement: The infrastructure was supported in part by POIG.02.03.01-24-099/13 grant: GeCONiI–Upper Silesian Center for Computational Science and Engineering.

References

[1] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance, IEEE/ACM Transactions on Networking, vol. 1, pp. 397–413, (1993).

[2] W. Feng, K. G. Shin, D. D. Kandlur and D. Saha, The BLUE active queue management algorithms. IEEE/ACM Transactions on Networking, pp. 513 – 528, Vol. 10, Iss. 4, (2002).

[3] S. S. Kunniyur and R. Srikant, An adaptive virtual queue (AVQ) algorithm for active queue management. IEEE/ACM Transactions on Networking, pp. 286–299, Vol. 12, Iss. 2, (2004).

[4] Z. Na, Q. Guo, Z. Gao, J. Zhen and C. Wang C, A novel adaptive traffic prediction AQM algorithm Journal, Telecommunication Systems, Volume 49, Issue 1, pp 149-160, (2012).

[5] M. H. Suzer, K.-D. Kang and C. Basaran, Active queue management via event-driven feedback control, Computer Communications, vol. 35, no. 4, pp. 517-529, Feb. (2012).

[6] N. Farzaneh, R. Monsefi, M. H. Yaghmaee and A. H. Mohajerzadeh, A novel congestion control protocol with AQM support for IP-based networks, Telecommunication Systems, Volume 52, Issue 1, pp 229-244, (2013).

[7] A. Chydzinski, L. Chrost, Analysis of AQM queus with queue sitze packed based dropping, Int. J. Appl. Math. Comput. Sci., Vol. 21, No. 3, pp. 567–577, (2011).

[8] W. Kempa, A Direct Approach to Transient Queue-Size Distribution in a Finite-Buffer Queue with AQM, Appl. Math. Inf. Sci. 7, No. 3, pp. 909–915 (2013).

[9] P. Mrozowski, A. Chydzinski. Queues with dropping functions and autocorrelated arrivals. Submitted for publication.

[10] A. Chydzinski. Transient Analysis of the MMPP/G/1/K Queue. Telecommunication Systems, vol. 32, n. 4, pp. 247–262, (2006).

[11] A. Chydzinski. Duration of the buffer overflow period in a batch arrival queue. Performance Evaluation, vol. 63:4-5, pp. 493–508, (2006).

[12] W. Leland, M. Taqqu, W. Willinger and D. Wilson. On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Transactions on Networking 2(1): pp. 1–15, (1994).

[13] A. Dhamdhere, H. Jiang, and C. Dovrolis, Buffer sizing for congested internet links, INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, vol. 2, pp. 1072–1083 vol. 2, (2005).

[14] P. Salvador R. Valadas and A. Pacheco. Multiscale Fitting Procedure Using Markov Modulated Poisson Processes. Telecommunication Systems 23:1-2, pp. 123–148, (2003).

[15] A. Chydzinski, R. Wojcicki and G. Hryn. On the Number of Losses in an MMPP Queue. Lecture Notes in Computer Science 4712, pp. 38–48, (2007).

[16] A. Chydzinski. Queue Size in a BMAP Queue with Finite Buffer. Lecture Notes in Computer Science 4003, pp. 200–210, (2006).

[17] D. Coppersmith, S. Winograd, Matrix multiplication via Arithmetic Progressions, Proc. 19th ACM Symposium on Theory of Computing, pp. 1–6, (1987).

[18] F. Le Gall, Powers of Tensors and Fast Matrix Multiplication, arXiv:1401.7714 [cs.DS], Jan. (2014).

[19] V.Vassilevska Williams, Multiplying matrices in $O(n^{2.373})$ time, Stanford University, (2014).

[20] J. Li, S. Ranka, S. Sahni. Strassen's Matrix Multiplication on GPUs. IEEE 17th International Conference on Parallel and Distributed Systems, pp.157-164, (2011).

[21] G. Sharma, A. Agarwala, B. Bhattacharya. A fast parallel Gauss Jordan algorithm for matrix inversion using CUDA, Computers & Structures, Vol. 128, pp. 31-37, (2013).

[22] J. W. Demmel. Applied Numerical Linear Algebra. SIAM, Philadelphia, (1997).

[23] G. Golub and C. Van Loan. Matrix Computations. Johns Hopkins University Press, Baltimore, MD, 3rd edition, (1996).

[24] A. Bjorck. Numerics of Gram-Schmidt Orthogonalization. Linear Algebra and Its Applications 198:297, p.316, (1994).

[25] V. Strassen. Gaussian elimination is not optimal. Numerische Mathematik 14, no. 3, 354–356, (1969).

[26] http://docs.nvidia.com/cuda/cublas

[27] http://www.nvidia.com

[28] H. Wong, M. M. Papadopoulou, M. Sadooghi-Alvandi, A. Moshovos, Demystifying GPU Microarchitecture through Microbenchmarking, Proc. Performance Analysis of Systems & Software (ISPASS), pp. 235 – 246, (2010).

[29] S. Huss-Lederman et. al., Implementation of Strassen's algorithm for matrix multiplication, Proc. ACM/IEEE Conference on Supercomputing, pp. 1-25, (1996).

[30] A. Gharaibeh, S. Al-Kiswany, S. Gopalakrishnan, and M. Ripeanu, A GPU accelerated storage system, Proc. ACM Int. Symposium on High Performance Distributed Computing, Chicago, (2010).

[31] J. Meng et. al. GROPHECY: GPU performance projection from CPU code skeletons, Proc. Int. Conference for High Performance Computing, Networking, Storage and Analysis, Seattle, pp. 1-11, (2011).

[32] V. Gupta et. al., Pegasus: coordinated scheduling for virtualized accelerator-based systems, Proc. of the 2011 USENIX conference on USENIX annual technical conference, Portland, (2011).

[33] J. Docampo et. al. Evaluation of Java for General Purpose GPU Computing, Proc. Advanced Information Networking and Applications Workshops, pp. 1398–1404, Barcelona, (2013).

An Overview of Various Types of Waveguide Grating Based Demultiplexors in WDM Systems

J. Tóth*, Ľ. Ovseník * and J. Turán*
* Technical University of Košice/Department of Electronics and Multimedia Communications, Košice, Slovakia
jan.toth@tuke.sk, lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstract - Fiber optic communication has been developed and studied significantly for several decades. Firstly, there were problems related to fiber attenuation and a choice of some suitable materials. Nowadays there is a great demand on increasing data rate and bandwidth. Despite some persisting optical fiber's weaknesses, this technology seems to be a leading one in today's telecommunication systems. The cost of this technology is perhaps one of the most limiting factors. This problem could be possibly solved by substituting and surrounding of the existing network topologies with WDM systems. Not a massive production of WDM devices makes WDM systems still quite expensive. Besides of optical sources, fibers and optical amplifiers there is the main WDM component called a multiplexor or demultiplexer respectively. This paper will provide a basic overview of some Arrayed Waveguide Grating (AWG) based demultiplexers. AWG demultiplexer is a passive WDM element which is able to combine and separate a broadband high frequency optical signal over large distances.

I. INTRODUCTION

Growing demands on increasing data bit rate and bandwidth have inspired and motivated researchers to come up with new telecommunication concepts. Wavelength Division Multiplexing (WDM) systems have made a great revolution in communication technology. The first WDM prototype came up in 1970. Laboratory realization of the first real WDM architecture was presented 8 years later. WDM architecture is a generalization of orthogonal OFDM transmission system. As any other conventional communication system, a basic WDM link consists of a transmitter, a modulation element a multiplexor/demultiplexer, an amplifier (usually in between MUX and DEMUX), a demodulation circuit and a receiver. There have been a few main moments in history which made WDM systems happened. In 1960 there was a first laser invented [1]. Six years later Kao and Hockham worked with a concept of a dielectric optical fiber – a potential medium for data transmission [2]. The achievement of low-loss fiber (~20dB/km) was introduced in 1970 [3]. Later on, in 1978 the first a low-loss single mode fiber was introduced. Its attenuation was only about 0,2 dB/km which was a big step forward. In the same year the first EDFA (Erbium-doped Fiber Amplifier) was developed [4]. Finally the first commercially used WDM system was available in 1995. All mentioned developments were accomplished by gradual improving of a few fundamental WDM components like laser diodes and many kinds of single mode fibers. First high speed single channel transmission link was realized at 1300 nm wavelength (SMF transmission). This is the region of the lowest chromatic dispersion and attenuation. Fiber attenuation has the lowest value around 1550 nm wavelength. It's basically a material characteristic and cannot be shifted and influenced (for Si optical fibers). In order to maximize the transmission length and a data bit rate there was a high effort to produce an optical fiber with a low chromatic dispersion in different wavelength regions. Chromatic dispersion consists of material and waveguide (a fiber property) dispersion which makes this parameter partly tunable. There were lots of attempts to design an optical fiber with so called shifted (zero – level) chromatic dispersion. This was successfully achieved by introducing of DSF (Dispersion Shifted Fiber) around 1990. It didn't take a long time facing some new challenges with the invention of WDM systems. More channels (carriers) and longer geographical distances were asking for higher input power level at the transmitter end. Fairly new kinds of problems – nonlinear effects – showed up which had not been taken under the scope before. Despite the fact that there had been some work made on fiber nonlinearities, one of nonlinear effects – FWM (Four Wave Mixing) – started causing serious difficulties in terms of data bit rate in WDM systems. FWM effect appears when mixing neighboring signals are matching in phase. This was happening as a consequence of WDM operation closely to zero chromatic dispersion level (~1550 nm) and seriously limited WDM transmission length. This problem led to the development of new SMFs. This new fibers are called NZ – DSFs (Nonzero – DSFs) which allow a certain amount of a chromatic dispersion [5]. NZ – DSFs were a reasonable compromise between negative (linear and nonlinear) effects and reach in WDM systems. In fact de/multiplexer is a fundamental WDM element which combines and separates particular wavelengths carrying the actual information. It turns out that waveguide based de/multiplexors (so called AWG DE/MUX, from now only AWG demultiplexers) provide the best WDM performance. Nowadays, there are several AWG concepts challenging themselves. There are several kinds of grating multiplexers based like arrayed waveguide gratings, concave gratings as well as their various modifications [6]. There is a high demand to develop a device which would be able to carry as many channels as possible with satisfactory low crosstalk, high transmission, small physical proportions and low manufacturing costs.

II. PRINCIPLE OF ARRAYED WAVEGUIDE GRATING DEMULTIPLEXER

The principle of AWG demultiplexer in based on a Mach – Zehnder interferometer (MZI). AWG demultiplexer consists of two multi – port couplers interconnected by a waveguide array. An optical signal propagating through waveguide array is gradually shifted in phase (in a specific phase value). This device has a few practical implementations like reconfigurable AWG routers and switches [7].

A. AWG demultiplexer

Contrary to MZI, the AWG demultiplexer experiences lower losses, cheaper manufacturing and higher optical compatibility with circuits. Both couplers and waveguide array are integrated on the same substrate. A substrate is usually made of silicon and waveguide array require some additional substances ($GeSiO_2$, SiO_2 – Ta_2O_5). AWG demultiplexer is specifically designed for particular channel spacing defined for ITU CWDM, DWDM standards [8]. Temperature fluctuations also have considerable impact on this device that's why the temperature management has to be carefully controlled. An input signal in figure (Figure 1.) is getting through so called FPR1 (Free Propagation Region) and propagates towards the waveguide array.

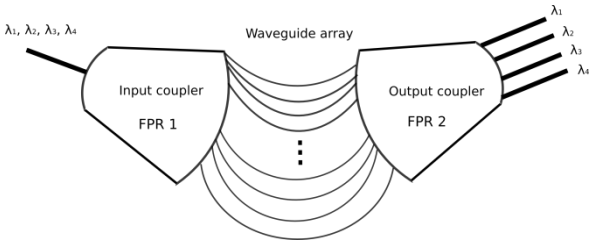


Figure 1. A basic scheme of Arrayed Waveguide Grating (AWG) demultiplexer with its two free propagation regions (FPRx).

This signal is coupled into the each waveguide from the waveguide array [9]. A physical length of each particular waveguide gets longer in a specific value ΔL from a previous one. It turns out that the particular waveguide phase shift will be different at so called grating circle of second FPR2 as shown in figure (Figure 2.).

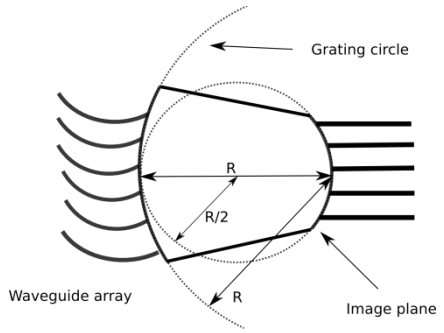


Figure 2. AWG demultiplexer's FPR2 with its image and grating circle.

B. Design of AWG demultiplexer

In order to achieve constructive interference at the image plane as shown in figure (Figure 2.), the FPR2 has to be specifically designed. Transparent grating is placed along grating circle. From the wave theory of light it turns out that every slit serves as a secondary source of light. Thanks to the phase mismatch (caused by ΔL) among waveguides in waveguide array there will be light diffraction happening. This is observed on image plane of FPR2. Optical signal has ultra-broad bandwidth in comparison to high-frequency links and particular informational channels are represented by certain carrier frequencies (wavelengths). The actual components of optical signal (λ_1, λ_2, λ_2…) are diffracted at grating circle, then travelling along FPR2 and finally spatially separated at image plane of FPR2 (Figure 2.). That's why constructive interference is desired in order to spatially isolate particular frequency components (channels) of the optical signal. This can be accomplished by following of the Rowland circle architecture applied for FPRs [10].

C. Free Propagation Region (FPR) Rowland geometry

Optical signal (light) passes through two FPRs of AWG demultiplexer and gets convergent. Diffraction gratings are located along grating circles (of both FPRs) and provide optical signal spatial separation. So called Rowland geometry satisfies the condition of constructive interference. Constructive interference occurs when various optical signal components from waveguide array are matching in phase at the certain spatial points of the FPR2 image plane. Optical fibers are then joined at these particular points [10]. The actual Rowland geometry looks like shown in figure (Figure 3.).

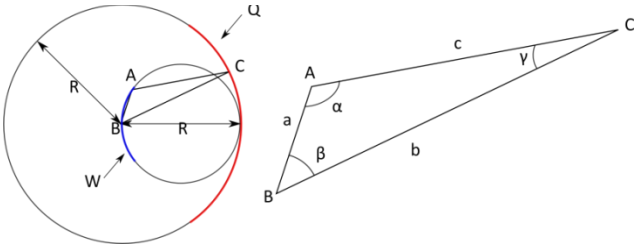


Figure 3. Free Propagation Region (FPR) of AWG demultiplexer – Rowland spatial geometry.

A fundamental condition for constructive interference is that a total phase shift has to be $n2\pi$ (n – integer). That is why it is very important to find a condition which will satisfy this requirement. In other words there is a need to know a distance c of triangle shown in figure (Figure 3.). This distance is needed to calculate a phase shift in FPRs for a particular wavelength. A total phase shift for FPR $\Delta\Phi_2$ can be expressed as shown in the following equation

$$\Delta\Phi_2 = \frac{2\pi}{\lambda} c\,, \tag{1}$$

where λ refers to an actual wavelength component. It is possible to calculate a c parameter from cosine formula

$$c = a^2 + b^2 - 2ab\cos(\gamma) \tag{2}$$

It is obvious that parameter $b=R$ as shown in figure (Figure 3.). On the other hand a parameter a can be evaluated from figure (Figure 4.).

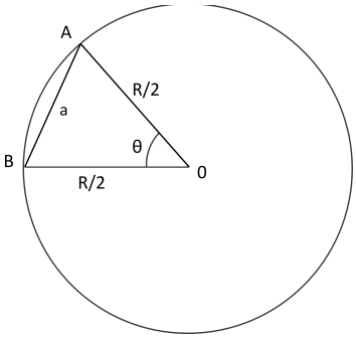


Figure 4. Parameter *a* calculation.

Parameter *a* equals to a length |AB| as it can be seen in both figures (Figure 3.), (Figure 4.).

$$a = R\sin\frac{\vartheta}{2} \tag{3}$$

Furthermore, a parameter *c* can be evaluated by using cosine formula from equation (2) and substituting *a* from (3).

$$c^2 = R^2\sin^2\left(\frac{\theta}{2}\right) + R^2 - R\sin\left(\frac{\theta}{2}\right)R\cos(\gamma) \tag{4}$$

$$c = R\sqrt{\left(1+\sin^2\left(\frac{\theta}{2}\right) - 2\sin\left(\frac{\theta}{2}\right)\cos(\gamma)\right)} \tag{5}$$

A final form for a total phase shift within a FPR can be expressed as following

$$\Delta\Phi_2 = \frac{2\pi}{\lambda}R\sqrt{\left(1+\sin^2\left(\frac{\theta}{2}\right) - 2\sin\left(\frac{\theta}{2}\right)\cos(\gamma)\right)} \tag{6}$$

D. A total phase shift of AWG demultiplexer

In order to work out a total phase shift for AWG demultiplexer it's important to consider a phase shift of the actual waveguide array $\Delta\Phi_1$ [11]. Particular waveguide's lengths differ from each other in a specific value ΔL

$$\Delta L = m\frac{\lambda_c}{n_{eff}}, \tag{7}$$

where *m* is an order of the actual waveguide, λ_c refers to a central demultiplexer wavelength and n_{eff} is an effective refractive index. It turns out that gradual ΔL increasing will cause a various phase difference $\Delta\Phi_1$

$$\Delta\Phi_1 = \beta\Delta L = 2\pi\Delta L\frac{n_{eff}}{\lambda}, \tag{8}$$

where β is a propagation constant. An optical signal is propagating through FPRs and waveguide array. As it has been already mentioned, it's needed to make sure that a total phase shift equals to multiple of 2π (9) as shown in figure (Figure 5.).

$$\Delta\Phi_1 + 2\Delta\Phi_2 = 2\pi \tag{9}$$

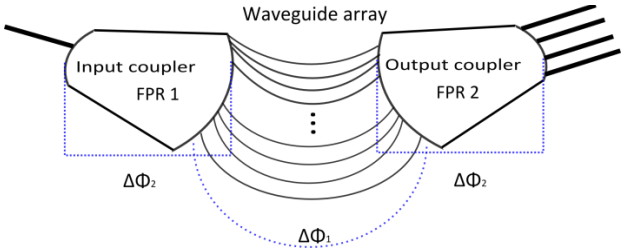


Figure 5. A total AWG demultiplexor's phase shift.

Considering previously mentioned equations, it's possible to define a final condition for AWG demultiplexer phase shift as shown in a following

$$2\pi\Delta L\frac{n_{eff}}{\lambda} + \frac{4\pi}{\lambda}R\sqrt{\left(1+\sin^2\left(\frac{\theta}{2}\right) - 2\sin\left(\frac{\theta}{2}\right)\cos(\gamma)\right)} = n2\pi \tag{10}$$

E. Planar Concave Gratings

PCG (Planar Concave Grating) demultiplexer is one of the competitive technologies besides AWG. In past, there were some technological problems to manufacture a PCG demultiplexer, which made it less popular. A recent progress in material research and more precise manufacturing technologies made PCG components usable in WDM systems. PCG demultiplexer has basically the same function as any other demultiplexer – to separate frequency components. There are two main differences in comparison to AWG systems [12]. PCG demultiplexer has only one FPR and its construction is constrained by Rowland geometry as shown in figure (Figure. 6). Contrary to AWG demultiplexer, an AWG doesn't have any waveguide array. An optical signal is separated on blazed nontransparent gratings.

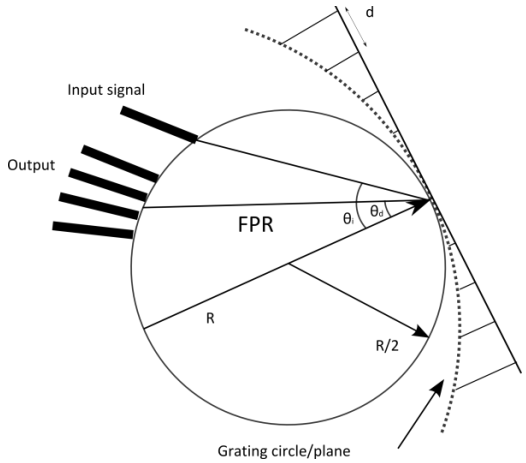


Figure 6. Planar Concave Grating demultiplexer.

A concave grating has several functions. It reflects and focuses light as shown in figure (Figure. 6) as well as separates different light components (diffraction of light). PCG devices have also satisfy grating formula

$$d\left(\sin\theta_i + \sin\theta_d\right) = m\frac{\lambda}{n_{eff}}, \tag{11}$$

where θ_i and θ_d are angles on the Rowland circle as shown in figure (Figure 6.). Parameter m refers to diffraction order and d is so called grating period [13].

F. *Hybrid demultiplexers*

In order to use an optical signal spectrum as efficiently as possible, there is one more interesting concept for WDM demultiplexer. Sometimes it's also called 2D demultiplexor because it combines AWG demultiplexor with a concave grating demultiplexer [15].

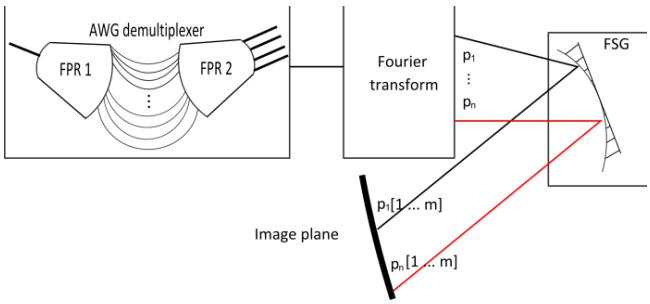


Figure 7. 2D hybrid AWG – Free Space Grating demultiplexer.

An optical signal is coming through the AWG demultiplexor which separates wavelengths along a vertical line (1D separating). Every AWG demultiplexer is capable to distinguish a finite number of diffraction orders *p*. Each filtrated channel has its bandwidth. It is 0.8nm for a DWDM standard which is about 100GHz of bandwidth. Such a range is still quite wide. There have been many experiments made lately to investigate the possibilities of using narrower bands [16]. These bands should satisfactory serve as fully reliable transmission channels for WDM systems in future. There are still problems with crosstalk's and proper detection and separation of particular channels. However this technology seems to be promising. The second part of such a demultiplexer is made of so called FSG (Free Space Grating). This component will divide each particular diffraction order p_i to the finite number of channels m_i. New channel focusing plane (perpendicular to the AWG demultiplexer output line) happens to appear. Finally, there is 2D matrix coming up as shown in figure (Figure 7.).

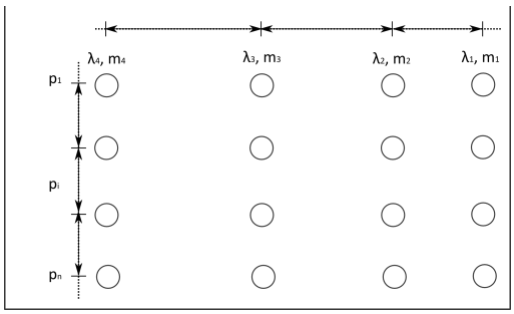


Figure 8. 2D hybrid demultiplexer's channel matrix.

The finite number of satisfactory filtrated channels (as well as channel spacing) is then specified and given by an actual AWG demultiplexor, laser's bandwidth and FSG demultiplexor.

III. Experiments And Results

We will picture a basic DWDM link with AWG demultiplexor as a fundamental element in this section. We decided to use a 16 channel ultra – dense DWDM (channel spacing 0.25 nm) system with EDFA amplifier on 200 km transmission link as shown in figure (Figure 9.).

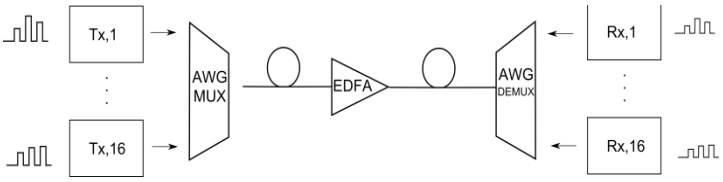


Figure 9. 16 channel DWDM transmission link with EDFA amplifier.

The most important WDM link's evaluative and qualitative indicator is BER as a function of Q – factor. There are several comparisons displayed in figure (Figure 10.)

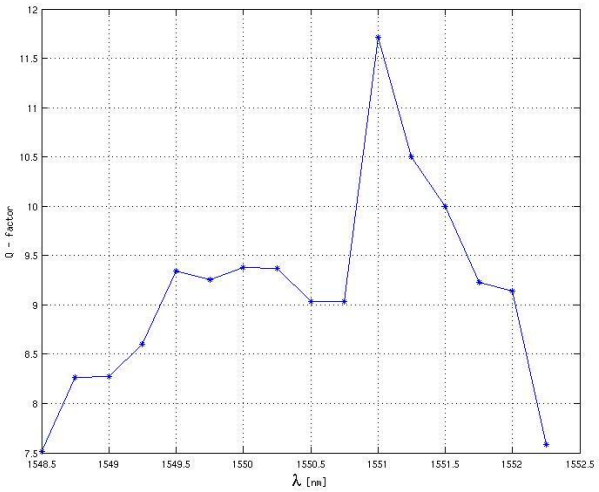


Figure 10. Q – factor as a function of wavelength

Figure (Figure 10.) shows the dependence of Q - factor as a function of a particular wavelength. The second chart (Figure 11.) refers to the BER evaluation as a function of wavelength.

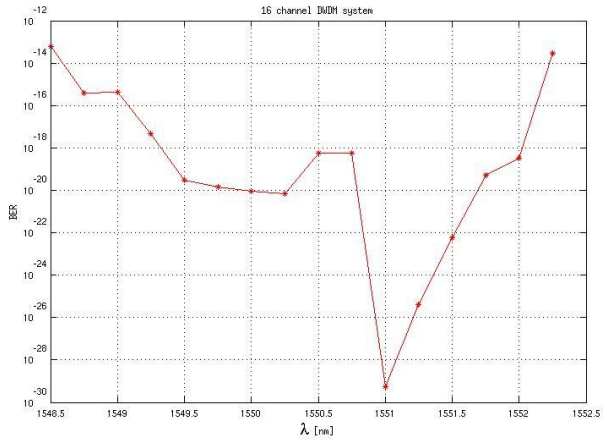


Figure 11. BER for particular channels as a function of wavelength.

The comparison in the figure (Figure 12.) shows an ideal BER curve as a function of Q – factor with simulated values (red color).

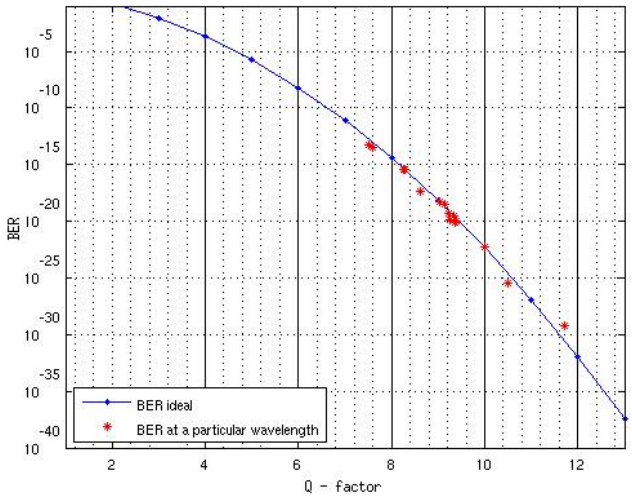


Figure 12. Comparison of BER and Q – factor for an ideal and experimentally simulated values.

IV. CONLUSION

We provided a fundamental WDM demultiplexors overview in this paper. As the most commonly used demultiplexor nowadays is AWG demultiplexor. However, there are some other options available. Hybrid demultiplexor seems to be the most promising future technology. Our simulations have pointed out at the ultra – dense DWD link with 16 channels in the experimental section. Chosen channel spacing was picked to investigate possibilities of using some alternative WDM setup in between two standardized regions for DWDM systems. Results show that these alternative spacing seems to be also appropriate for use in real operation.

ACKNOWLEDGMENT

This work was supported by Cultural and Educational Grant Agency (KEGA) of the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project no. „006TUKE-4/2014 - The Use of TUKE PON Experimental Model in Teaching. This work is also supported by project no. 059TUKE-4/2015: Implementation of the FSO long-term experimental measurement results into the educational process in optoelectronic sensor and communication systems field. This work is also the result of the contract no. "APVV-0025-12 - Mitigation of Stochastic Effects in High-Bitrate All Optical Networks" supported by the Slovak Research and Development Agency.

REFERENCES

[1] T.H. Maiman, "Stimulated optical radiation in ruby," vol. 187, pp. 493–494, 1960.

[2] K.C. Kao, G.A. Hockham, "Dielectric-fiber surface waveguide for optical frequencies, " IEEE, vol. 113, pp. 1151, 1966.

[3] D.B. Keck, "On the ultimate lower limit of attenuation in glass optical waveguides," Appl. Phys. Lett., vol. 22, 1973.

[4] R.J. Mears, "Low threshold tunable CW and Q-switched fiber laser operating at 1.55 μm," Electron. Lett., vol. 22, pp. 1026-1028, 1986.

[5] G.P. Agrawal, "Fiber-Optic Communication Systems." John Wiley & Sons. Inc..Publication. The Boulevard. Langford Lane. Kidlington. Oxford OX5 1GB. UK. ISBN 978-0-12397-023-7. 2013.

[6] H. Abramczyk, "Dispersion Phenomena in Optical Fibers" <http://www.mitr.p.lodz.pl/evu/lectures/Abramczyk3.pdf>, 2008.

[7] G.T. Reed, A.P. Knights, "Silicon Photonics An Introduction," The Atrium, Southers Gate, Chichester, West Susex, England, 2004.

[8] Telecommunication standardization sector of itu. "Spectral grids for WDM applications: DWDM frequency grid.", Recommendation ITUT G.694.1,<https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-G.694.1-201202-I!!PDF-E&type=items>.

[9] S. Pathak, W. Bogaerts, E. Lambert, P. Dumon, D. Van Thourhout "Compact SOI-Based AWG With Flattened Spectral Response Using a MMI," IEEE, 2012< http://www.photonics.intec.ugent.be/download/pub_ 2774.pdf>.

[10] D. Meena, O.S. Sunishkumar,D.C. Pande, "A Geometrical Model for Arrayed Waveguide Grating based Optical Multiplexer/Demultiplexer," Progress In Electromagnetics Research M, Vol. 35, 2014 [cit. 2014-12-05]. <http://www.jpier.org/PIERM/pierm35/10.14011906.pdf>

[11] M.K. Smit, C. Dam, "PHASAR-Based WDM-Devices: Principles, Design and Applications," IEEE Journal of Selected Topics in Quantum Electronics, Vol. 2, No. 2, June, 1996 <http://repository.tudelft.nl/assets/ uuid:37a5ebe1-766e-42cf-9511-573882edc4b8/ieee_smit_1996.pdf>.

[12] Muneeb, M.; Chen, X.; Ryckeboer, E.; Malik, A.; Mashanovich, G. Z.; Roelkens, G. Silicon-On-Insulator Mid-Infrared Planar Concave Grating Based (De)Multiplexer [online.]. IEEE, 2013 [cit. 2014-08-24]. Dostupné na internete: < http://www.photonics.intec.u gent.be/download/pub_3214.pdf>.

[13] Brouckaert, J.; Bogaerts, W.; Dumon, P.; Thourhout, D.; Baets, R. Planar Concave Grating Demultiplexer Fabricated on a Nanophotonic Silicon-on-Insulator Platform [online]. Journal of Lightwave Technology, vol. 25, no. 5, May, 2007 [cit. 2014-12-05].
Dostupné na internete: <http://www.photonics.intec.ugent.be/download/pub_2111.pdf>

[14] J. Pello, M. Muneeb, S. Keyvaninia, "Planar Concave Grating Demultiplexers on an InP-Membrane-on-Silicon Photonic Platform," IEEE Photonics Technology Letters, Vol. 25, No. 20, October, 2013<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6589120>.

[15] T.K. Chan,J. Karp,R. Jiang, N. Alic, CH.F. Marki, J.E. Ford "1092 Channel 2-D Array Demultiplexer for Ultralarge Data Bandwidth," Journal of Lightwave Technology, Vol. 25, No. 3, March, 2007 <http://psilab.ucsd.edu/publications/42.%20JLT%20wavelength%20demux.pdf>.

A Powerful Early-based Video Discarding Policy

Khalid A. Darabkh[1], Abeer M. Awad[1], and Ala' F. Khalifeh[2]
[1]Computer Engineering Department, The University of Jordan, Amman 11942, Jordan
[2]Communication Engineering Department, German Jordanian University, Amman 11180, Jordan
Emails: {{k.darabkeh, abeer.awad}@ju, ala.khalifeh@gju}.edu.jo

***Abstract*— This paper proposes an improved version of early video frame discarding policy that adopts a threshold at the buffers of network elements to control the ingoing video traffic. This improvement is performed through taking the dependencies between video types into consideration to make decision whether to accept or discard any stream of video frame packets. Besides, two thresholds are employed to maintain the buffers of network elements. Furthermore, any frames where their decoding possibilities at the receiver side are almost negligible are prevented from being queued at the buffers of network elements. However, our proposed policy is evaluated in terms of the rate of non-decodable frames and peak signal-to-noise ratio (PSNR). Our proposed policy behaves very well and achieves a remarkable enhancement over the conventional early video frame discarding policy.**

***Index Terms*— Video discarding policies; video traffic; group of pictures; rate of non-decodable frames**

I. INTRODUCTION

In recent years, there has been a tremendous increase in demand for video streaming over wired and wireless networks [1-3]. Applications of digital video streaming include learning via distance [4], video teleconferencing [5], interactive digital TV [6], remote diagnosis/surgery, distance/remote sensing, and process monitoring [7]. Moving Picture Experts Group (MPEG) is one type of video compression techniques that was formed by ISO [8]. New advanced video coding (AVC) schemes appear like MPEG-4 AVC (H.264) that is used for compressing streaming media (audio or video) and distinguished over previous schemes (e.g., MPEG or MPEG-2) as of being capable of affording a lower data rate with comparable picture quality or a comparable data rate with considerably higher picture quality [9].

The compressed MPEG video data has three video frames: I-, P-, and B-frames whereas their repeated pattern is called the Group of Pictures (GOPs) [10, 11]. I-frame is considered an independent and reference frame as of being decoded at the receiver side independently of other P- and B-frames. P-frame is also considered a reference frame bearing in mind that in the decoding process, it depends only on the previous I- or P-frames. B-frames are ordinarily decoded based on the former or posterior I- or P-frames. Additionally, it is decoded using the interpolation of the former and posterior I- or P-frames. In light of previous discussions concerning video frames, it is good to stress on the fact that losing I-frames is more disadvantageous than losing P-frames though both are considered reference frames [12]. In other words, discarding the initial I-frame leads to having all subsequent frames of the same GOP useless. Analogously, losing P-frames may have severe consequences on the subsequent P- and B-frames. On the contrary, losing B-frames may not have any impact of the succeeding frames. Due to carrying the frame header information, losing the head-of frame packet is as worse as losing the whole frame [13].

In this paper, we propose a powerful video frame discarding policy called eclectic early frame discarding-with dependency (EEFD-wD). This policy is considered a novel extension to early frame discarding-without dependency (EFD-woD) policy proposed in [14]. Complete details about our proposed policy and EFD-woD policy will be provided shortly. However, our policy is evaluated based on the following performance metrics: the rate of non-decodable frames and peak signal-to-noise ratio (PSNR). Our proposed policy not only outperforms EFD-woD policy, but also achieves an extraordinary improvement.

This paper is organized as follows. The following section provides details about related work. Section III illustrates our proposed discarding polices. The simulation environment, results, and discussions are discussed in Section IV. Section V concludes our paper.

II. RELATED WORK

There are works in the literature focused on efficient message or email files discarding for TCP/IP-based transmissions utilizing discrete-time [15-16] and continuous-time [17-23] queuing systems. Additionally, there is a work proposed in [24] which predominantly involves two partial frame discarding policies for UDP-based transmissions including frame dependencies and is based on partial frame discarding policy proposed in [14] in which the frame dependencies were excluded.

It is leading light that the most relevant continuous-time-based work to our proposed discarding policies is [14] in which early frame discarding without dependency (EFD-woD) policy has been extensively evaluated in terms of frameput and average buffer contents considering the case where video frames are inputs to the buffers of network elements. The network and buffer model used in this policy is based on M/M/1/N system model in which *M*, 1, and *N* refer to the memoryless property, number of servers used, and number of customers that can be accommodated, respectively. The first *M* refers to the situation where packets that belong to a certain video frame arrive according to a Poisson process with arrival rate λ while the second *M* refers to case where packets at the

head of the buffer are served according to a Poisson process with a service rate μ (the service time follows exponential distribution). The system load (utilization factor) is expressed as ρ (i.e., $\rho=\lambda/\mu$). The length of a frame (in terms of packets) is assumed to be geometrically distributed where a packet belongs to the same frame with probability p, which is equal to $(1-q)$, and is discarded with the same probability. The first packet of a frame is called as a head-of-frame packet and arrives with probability q. In EFD-woD policy, a packet is queued with respect to a threshold K $(0 \le K \le N)$ that is employed to maintain the buffer occupancy level. In different words, once the buffer occupancy reaches at or above that threshold, no packets of a new frame are accepted. On the contrary, the subsequent packets, which belong to a frame has been partially accepted, are accepted.

III. Proposed Video Discarding Policy

Eclectic early frame discarding-with dependency policy (EEFD-wD) policy is basically based on EFD-woD policy proposed in [14]. Unlike EFD-woD policy, the EEFD-wD policy takes the dependencies between video types into consideration to make decision whether to accept or discard any stream of packets. This is accomplished by preventing the frames where the probability to be decoded at the receiver side is too low, as of losing their reference frames, from being buffered and served. Dissimilar to EFD-woD policy, the EEFD-wD policy employs two thresholds, K_L and K_H to maintain the buffer occupancy. Besides, the buffer records the frame type for each arriving packet. Interestingly, if packets belong to B-frame arrive at buffer occupancy lower than or equal K_L, then these packets will be unconditionally accepted. Once the buffer occupancy rises above K_L and stays below or reaches K_H, then arriving packets that belong to partially accepted B-frame will be accepted with probability (1-P_d). It is worth noting that a head-of-B-frame packet will not be accepted during this time interval. Furthermore, arriving B-frame packets at buffer occupancy lies above K_H will be completely rejected. The dropping function of B-frame packets is proposed as:

$$P_d = \begin{cases} 0, & buffer\ occupancy \le K_L \\ e^{\frac{\gamma(buffer\ occupancy - K_H)}{(buffer\ occupancy - K_L)}}, & K_L < buffer\ occupancy \le K_H \\ 1, & buffer\ occupancy > K_H \end{cases} \quad (1)$$

where γ is the exponential factor. The idea of proposing the exponential function shown above is to satisfy the shape shown in Fig. 1. As seen from this figure, the probability of accepting subsequent B-frame packets (i.e., partially accepted B-frame) is high as long as the buffer occupancy stays closer to K_L. Once the buffer occupancy gets higher and higher (i.e., comes closer to K_H) and for the sake of giving other reference frames the opportunity to get totally queued, the probability of buffering successive B-frame packets gets lower and lower.

Due to their importance, the way we deal with P- or I-frames is different. In other words, after conducting many experiments seeking for optimality, we propose that the I- and P-frame packets are accepted at any buffer occupancy level, whether the arrivals represent a head-of-frame packet or packets belong to already accepted frame. Furthermore, if I- and P-frame packets arrive when the buffer occupancy is at or above K_L where the empty space left is not enough, then eclectic mode will be incorporated. In this mode, which represents preventive expedient, a massive buffer search for B-frame packets will occur in which those packets that have the least average waiting time and their count suits or complies with the number of arriving I- or P-frame packets are selected to be substituted with reference frame packets. When there are no possible ways to accept these packets, either by not finding enough B-frame packets or originally not having B-frame packets, the arrivals will be discarded. Consequently, the subsequent packets of that frame will be also rejected. Not only to this extent, but also the rest frames of the current GOP will be disposed of. As an efficient reaction to this situation (i.e., best buffering management), the EEFD-wD policy will eject those buffered packets of that impaired frame, knowing that this frame will not be decoded at the receiver side. Actually, this is also applied for impaired B-frames.

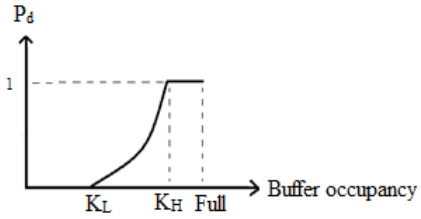


Fig. 1. Drop function of EEFD-wD policy.

IV. Simulation Results and Discussions

This section demonstrates the setup made for simulating our proposed discarding policy along with its obtained results. Subsection *A* discusses the simulation environment. Subsection *B* illustrates the performance metrics used. Finally, the results and discussions are presented in Subsection *C*.

A. Simulation environment

Many experiments were conducted to simulate our proposed discarding policy using Matlab 2010R on a Windows 7 and PC with an Intel Quad Core CPU working at 2.3 GHz with a 2MB cache and 2 GB RAM. These experiments are repeated 100 times and require 15 days to end. However, several benchmark raw videos are used in simulations which include *mobile_sif_mono*, *paris_cif*, *hall_monitor_cif*, *bridge_close_cif*, and *garden_sif* and are available in [25]. The obtained results considering all aforementioned raw videos were relatively similar. Hence, we consider in our simulation results the raw video (*paris_cif*) that has duration of 35.52 seconds. Fast forward MPEG encoder is used to compress the adopted raw video and then encode the generated H.264 video file to eventually extract the I-, P-, and B-frames, which are 1066 frames, utilizing a picturing rate of 30fps. In our proposed policies, the existence of B frames is very helpful. Thus, the GOP pattern used is IBP.

Video frames have different sizes compared to each other. I- frames are the largest frames while P-frames are larger than B-frames [26]. These frames must be fragmented into small packets before they are transmitted to meet the maximum transmission unit constraints. Typically, the mean frame sizes of I-, P-, and B-frames are chosen based on the bit rate used.

To have different load utilizations, we make the service rate stick to 3 (packets/sec) and arrival rate range from 2.5 to 20 (packets/sec). These values can be adjusted to keep the values of load utilization factors. The buffer is assumed to accommodate up to 20 packets as employed in [14]. As a summary of performing many experiments, seeking for those values of K_L, K_H, and γ that maximize the performance of our proposed discarding policy, the chosen values are $N/2$, $1.6xK_L$, $(K_H \geq 0.75\,N)$, 0.1, respectively.

B. Performance Metrics

To assess the performance of our proposed discarding policy, two performance metrics are used which include the rate of non-decodable frames and PSNR. The rate of non-decodable frames represents the ratio of accumulative number of frames that are served but will not be decoded at the receiver side to the number of total frames. The PSNR is one of the best objective ways and widely used to measure the quality of videos. PSNR is measured on a logarithmic scale based on the mean squared error (MSE) between the original video and an impaired one. Higher PSNR values usually refer to lower video impairments [27-28]. It can be found as follows [27]:

$$PSNR = 10\log_{10}\frac{(2^n-1)}{MSE} \qquad (2)$$

$$MSE = \frac{1}{H*W*T}\sum_{t=1}^{T}\sum_{h=1}^{H}\sum_{w=1}^{W}[f(t,h,w)-\tilde{f}(t,h,w)]^2 \qquad (3)$$

where H and W are the horizontal and vertical dimension in pixels of the video frame, respectively. T is the number of frames in the video, n is the number of bits per pixel while f and $\tilde{f}$ are the original and impaired video frames, respectively.

C. Results

Fig. 2 depicts the results of the rate of non decodable frames versus arrival rate for a buffer size of 20. The GOP length and pattern used are IBP and 15, receptively. Additionally, the bit rate used is 200kbps which implies that the mean frame sizes for I-, P-, and B-frames are 20, 3, and 1 packets, respectively. However, the negative effect of increasing the arrival rate on all policies is clearly observed. Unlike the EEFD-wD, the EFD-woD policy achieves higher rates of non-decodable frames. This is expected since frame acceptance is based on only buffer availability. To clearly distinguish between EEFD-wD and EFD-woD policies, they have been subplotted separately.

Fig. 3 delineates the PSNR versus arrival rate considering a buffer size of 20. The parameters related to GOP and bit rate are the same as previously mentioned. Obviously, the improvement of PSNR values as a result of employing EEFD-wD policy is extremely impressive. As a matter of fact, the policy that achieves lower rates of non-decodable frames should have higher values of PSNR. Genuinely, increasing the buffer size will improve the PSNR results for all policies. However, the decrease in PSNR values as the traffic load increases is also noticed for all policies.

V. CONCLUSIONS

This paper proposes an efficient video frame discarding policy (EEFD-wD) to control and maintain the buffers of wired/wireless hops under heavy loads of video stream. the proposed policy is based on EFD-woD policy proposed in [14] in which major amendments are made, out of which incorporating the dependencies existing among video frames, operating two buffer thresholds, and preventing any frames where their decoding possibilities at the receiver side are almost negligible from being queued at the buffers of network elements. In EEFD-wD policy and as soon as the buffer occupancy gets located between buffer thresholds, B-frame packets are accepted according to an exponential function which helps to limit the maximum number of B-frame packets that could exist in the buffer without degrading the overall video quality. Under unexpected video traffic and on account of accepting reference frames (I- or P-frames), those buffered B-frames are obliged to expel. The performance metrics used to evaluate the proposed discarding policies include the rate of non-decodable frames and PSNR. The effectiveness of our results is verified by comparing with EFD policy.

REFERENCES

[1] Khalid A. Darabkh, Abeer M. Awad, and Ala' F. Khalifeh, "Efficient PFD-Based Networking and Buffering Models for Improving Video Quality over Congested Links," *Wireless Personal Communications*, vol. 79, no. 1, pp. 293-320, November 2014.

[2] K. A. Darabkh and R. S. Aygün, "TCP Traffic Control Evaluation and Reduction over Wireless Networks Using Parallel Sequential Decoding Mechanism," *EURASIP Journal on Wireless Communications and Networking*, vol. 2007, Article ID 52492, pp.1-16, November 2007.

[3] Khalid A. Darabkh, "Fast and Upper Bounded Fano Decoding Algorithm: Queuing Analysis," *Transactions on Emerging Telecommunications Technologies*, DOI: 10.1002/ett.2929.

[4] Taralynn Hartsell and Steve Chi-Yin Yuen, "Video Streaming in Online Learning," *AACE Journal*, vol. 14, no. 1, pp. 31-43, January 2006.

[5] Susanne Naegele-Jackson, Peter Holleczek, Thomas Rabenstein, Juergen Maiss, and M. Sackmann, "Influence of Compression and Network Impairments on the Picture Quality of Video Transmissions in Tele-Medicine. *In Proceedings of the 35th Annual Hawaii International Conference on System Sciences,* Hawaii, pp. 2060-2068, January 2002.

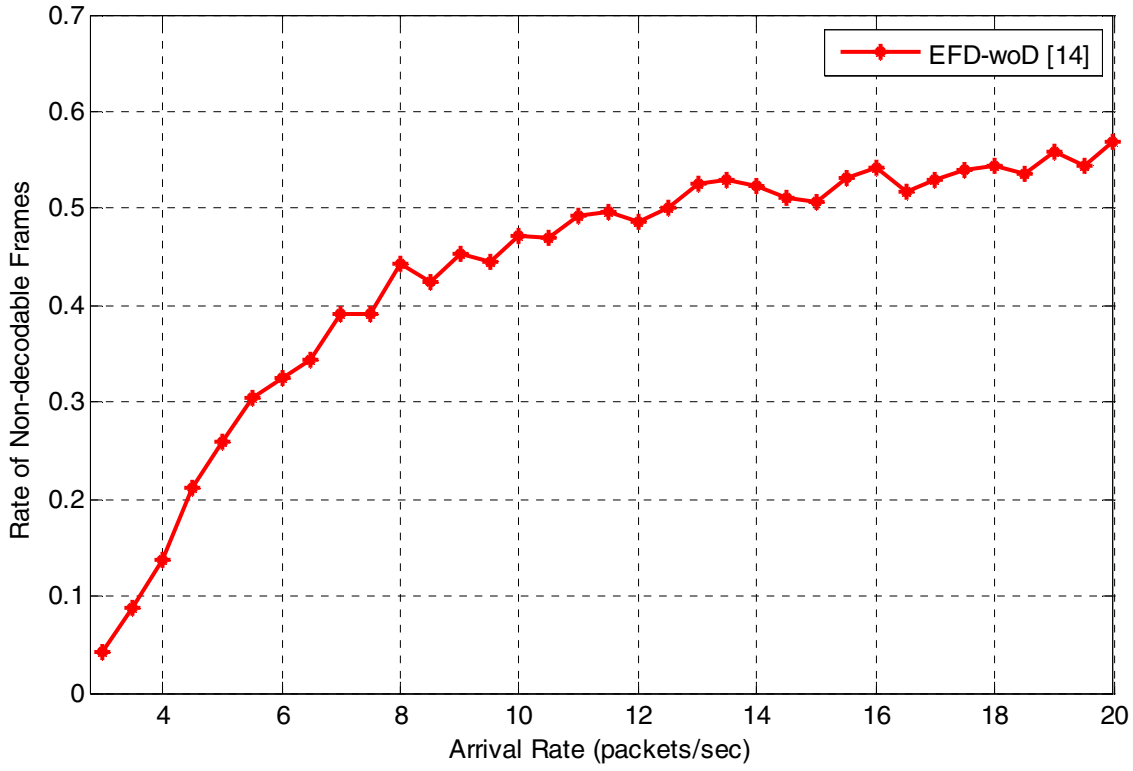
[6] The six best digital media-streaming devices, http://www.pcadvisor.co.uk/news/network-wifi/10736/the-six-best-digital-media-streaming-devices/

[7] Toufik Ahmed, Ahmed Mehaoua, Raouf Boutaba, and Youssef Iraqi, "Adaptive Packet Video Streaming Over IP Networks: A Cross-Layer Approach," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 385-401, February 2005.

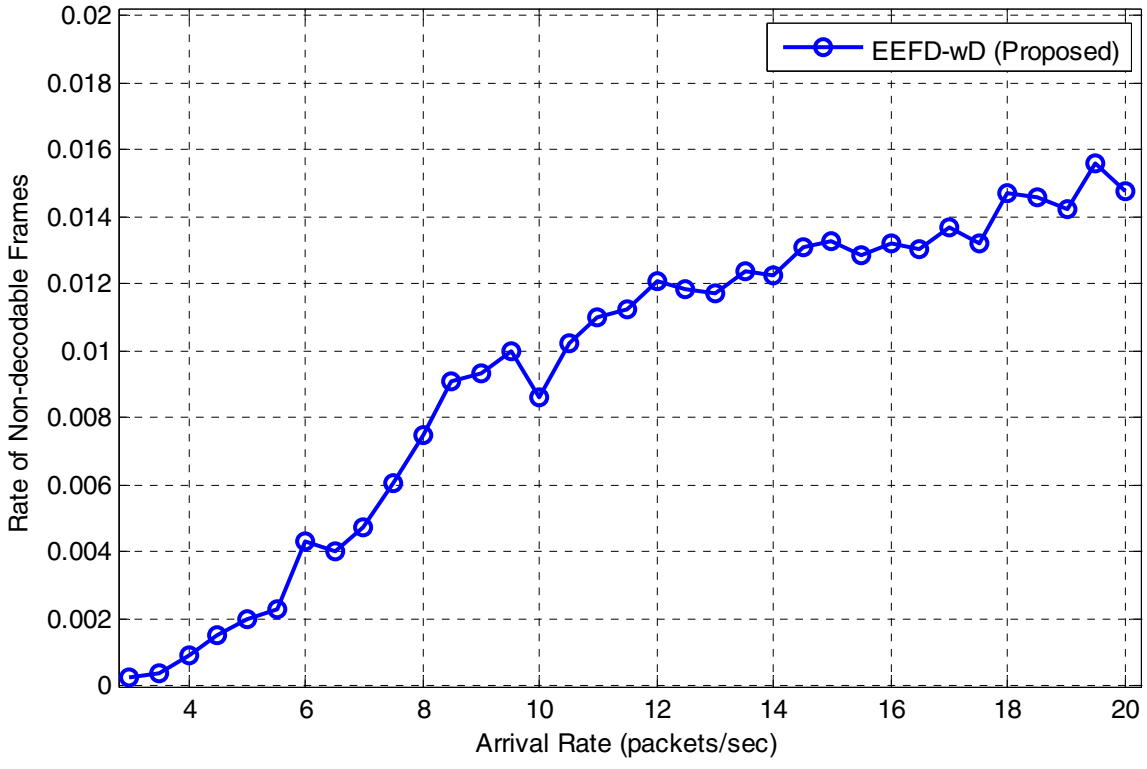
[8] Didier Le Gall, "MPEG: A video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, no. 4, pp. 46-58, April 1991.

[9] Richardson Iain, *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*, John Wiley & Sons, February 2004.

[10] Susie J Wee, "Manipulating temporal dependencies in compressed video data with applications to compressed-domain processing of MPEG video," *In Proceeding of 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, vol. 6, pp. 3129-3132, March 1999.



(a)



(b)

Fig. 2. The rate of non-decodable frames versus arrival rate for a buffer size of 20 considering (a) EFD-woD policy (b) EEFD-wD policy.

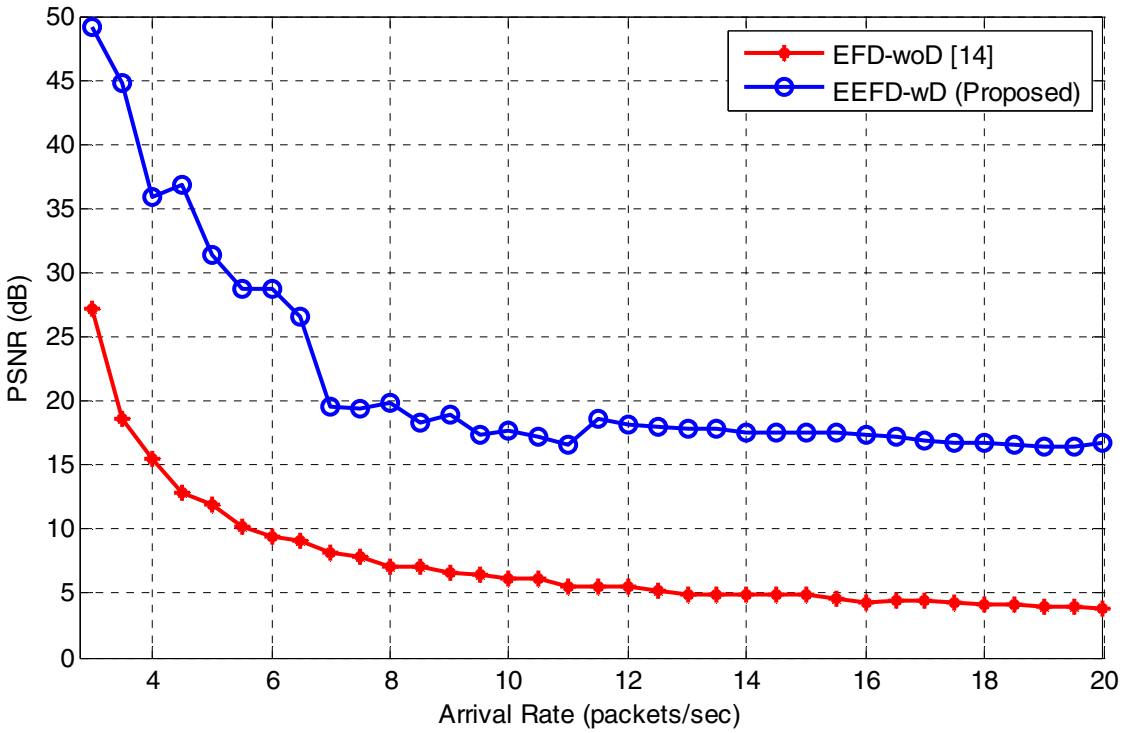


Fig. 3. PSNR versus arrival rate for different policies and a buffer size of 20.

[11] Shi Yun and Huifang Sun, *Image and Video Compression for Multimedia Engineering: Fundamentals, Algorithms and Standards*, CRC Press, March 2008.

[12] G. Ravindra, N. Balakrishnan, and K. Ramakrishnan, "Active Router Approach for Selective Packet Discard of Streamed MPEG video under Low Bandwidth Conditions," *In Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2000)*, New York, NY, pp.739-742, August 2000.

[13] Mario Marques da Silva, *Multimedia Communications and Networking*, CRC Press, March 2012.

[14] K. Darabkh and R. Aygun, "Improving UDP Performance using Intermediate QoD-Aware Hop System for Wired/Wireless Multimedia Communication Systems," *International Journal of Network Management*, vol. 21, no. 5, pp. 432-454, September 2011.

[15] Wen-Hui Zhou and Ai-Hu Wang, "Discrete-time queue with Bernoulli bursty source arrival and generally distributed service times," *Applied Mathematical Modeling*, vol. 32, no. 11, pp. 2233-2240, November 2008.

[16] Wen-Hui Zhou and Yong-Jun Li, "A discrete-time queueing model of EMD policy in high-speed networks," *Applied Mathematics and Computation,* vol. 181, no. 1, pp. 543-551, October 2006.

[17] Yael Lapid, Raphael Rom, and Moshe Sidi, "Analysis of Discarding Policies in High-Speed Networks," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 764–777, June 1998.

[18] K. Kawahara, K. Kitajima, T. Takine, and Y. Oie, "Packet Loss Performance of Selective Cell Discard Schemes in ATM Switches," *IEEE Journal of Selected Areas in Communications*, vol. 15, no. 5, pp.903–913, June 1997.

[19] Parijat Dube and Eitan Altman, "Queueing and Fluid Analysis of Partial Message Discard Policy," *In Proceedings of 9th IFIP Working Conference on Performance Modeling and Evaluation of ATM and IP Networks*, Budapest, Hungary, June 2001.

[20] Parijat Dube and Eitan Altman, "Queueing analysis of early message discard policy", *In Proceedings of IEEE International Conference on Communications (ICC 2002)*, vol. 4, pp. 2426–2430, May 2002.

[21] Parijat Dube and Eitan Altman, "On Fluid Analysis of Queues with Selective Message Discarding Policies," *In Proceedings of 38th Annual Allerton Conference on Communication, Control and Computing*, Monticelio, Illinois, USA, October 2000.

[22] Parijat Dube and Eitan Altman, "Fluid Analysis of Early Message Discard Policy Under Heavy Traffic," *In Proceedings of INFOCOM 2002*, New York City, USA, pp.1307-1315, June 2002.

[23] Parijat Dube and Eitan Altman, "Goodput Analysis of a Fluid Queue with Selective Burst Discarding and a Responsive Bursty Source", *In Proceedings of INFOCOM 2003*, San Francisco, CA, USA, pp. 166-176, April 2003.

[24] Khalid A. Darabkh, Abeer M. Awad, and Ala' F. Khalifeh, "Intelligent and Selective Video Frames Discarding Policies for Wireless Networks," *In Proceedings of the 2013 IEEE International Symposium on Multimedia (ISM 2013)*, Anaheim, California, USA, pp. 297-300, December 2013.

[25] Video Test Media, http://media.xiph.org/video/derf/

[26] Damir Isovic, Gerhard Fohler, and Liesbeth Steffens, "Some Misconceptions about Temporal Constraints of MPEG-2 Video Decoding," *In Proceedings of the WiP of the 23rd IEEE International Real-Time Systems Symposium (RTSS 2003*), Austin, Texas, 2002.

[27] A. Vishwanath, P. Dutta, M. Chetlu, P. Gupta, S. Kalyanaraman, and A. Ghosh, "Perspectives on quality of experience for video streaming over WIMAX," *ACM SIGMOBILE Mobile Computing and Communications Review,* vol. 13 no. 4, pp.15-25, October 2009.

[28] D. Gangadharan, L. Phan, S. Chakraborty, R. Zimmermann, and I. Lee, "Video Quality Driven Buffer Sizing via Frame Drops," *In Proceedings of 17th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'11)*, Toyama, Japan, pp. 319–328, August 2011.

Mobile node localization using infrared angle of arrival sensor

D. Frantal, D. Arbula
Faculty of Engineering Rijeka, Computer science department, Rijeka, Croatia
dfrantal@riteh.hr, damir.arbula@riteh.hr,

Abstract - This paper presents mobile node localization algorithm for environments in which GPS or other positioning systems are not applicable or available. Existing localization systems are using propagation characteristics such as attenuation or speed of radio or ultrasound signals or timings of controlled laser radiation to estimate node azimuth and/or distance to nodes with known position (anchor nodes). Presented algorithm is using azimuth estimation based on measured angle of incidence of controlled infrared radiation from anchor nodes. Angle of arrival estimation and localization algorithms are implemented on low cost, small footprint and low power wireless nodes based on 8-bit microcontroller. Experimental evaluation results are promising and with further sensor optimization can guarantee high precision of estimated locations.

I. INTRODUCTION

Solution to mobile node localization problem in environments where general purpose localization methods are unavailable is presented in this paper. Method is implemented on low cost and low power wireless sensor network nodes [1]. Each node consists of microcontroller, radio module, and angle of arrival sensor prototype (Figure 1).

Node with unknown location receives true coordinates of limited set of nodes with known position (anchor nodes). Subsequently, using angle of arrival sensor it measures angle of incidence of infrared radiation emitted by anchor nodes and estimates their azimuths. With the data obtained from at least 3 anchor nodes, mobile node is able to calculate its position.

Existing solutions are based on radio interference measurements and require more expensive hardware with three [2] or two [3] antennas. There are also solutions that use rotating lasers [4] but they come with significant setup overhead.

A. Platform

Wireless sensor node platform used is commercially available under name JeeNode. This low cost platform consists of an 8-bit AVR RISC Atmel ATmega328P microcontroller, RFM12B radio module and 4 input/output ports. Communication with PC is possible through serial UART port. Each of 4 available ports provides digital and analog input/output, power lines, interrupt line and ground. The whole platform has small footprint and low power usage.

The work presented in this paper is supported by the University of Rijeka research grant Grant 13.09.2.2.16.

II. HARDWARE

A. Angle of arrival sensor

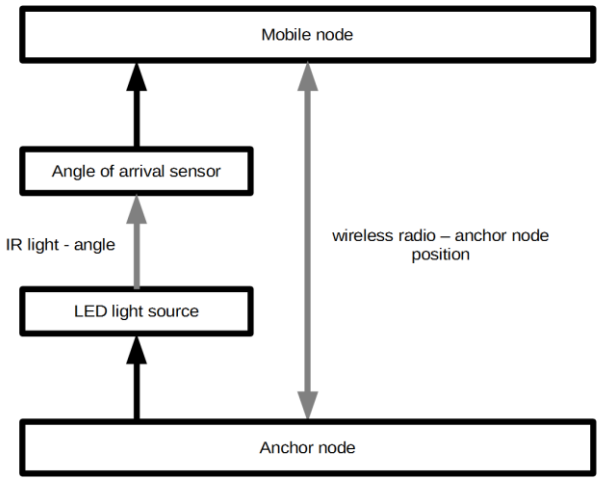


Figure 1. Conceptual scheme

Angle of arrival (AoA) sensor consists of 12 infrared phototransistors (Figure 4) [5] evenly distributed along the circular rim of the sensor. Infrared radiation is chosen to minimize expected interferences with visible light.

Each phototransistor emitter is connected to multiplexer input and with resistor in series to ground. Voltage is fed to phototransistor collector. Multiplexer channel selection is controlled by dedicated slave microcontroller and this way each voltage drop can be measured using 10 bit A/D converter on microcontroller's analog input (Figure 2).

B. IR radiation source

IR radiation source consists of 15 evenly circularly distributed infrared LE diodes [6] that can be switched on and off by microcontroller. LEDs are triggered (for short duration) by microcontroller digital output pin (DOP) that is connected to MOSFET because digital output pin can only provide 40 mA of current which is not enough for powering all LEDs. LEDs are connected directly to battery power source that provides 5 V (Figure 3).

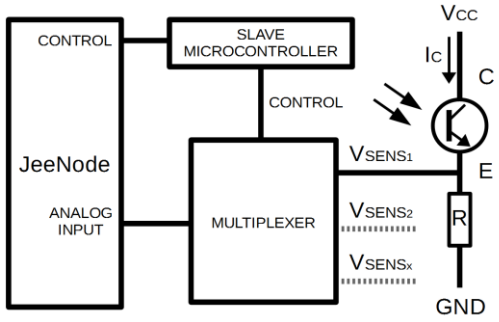


Figure 2. Phototransistor circuit collector current I_C (max 0.31 mA) is dependent on IR radiation intensity, V_{SENS} (max. 3.1 V) is voltage detected by microcontroller, pin is connected to multiplexer input.

III. NODE LOCALIZATION

A. Localization method

In the current setup, mobile node is the node with unknown position, likewise nodes with known coordinates or anchor nodes are stationary [7]. Anchor nodes are equipped with omnidirectional IR radiation source and mobile node is equipped with IR AoA sensor. If the mobile node is in the communication and AoA sensor range of at least 3 anchor nodes, then it is able to localize itself. In this paper only 2D localization is considered.

B. Localization procedure

Based on the stated assumptions, procedure for node localization is as follows:

1. Mobile node broadcasts request for nearby node IDs. All nearby anchor nodes in range send their ID with different delay (ID based) to mobile node.
2. Mobile node measures ambient IR radiation, that is, IR radiation in the environment with all anchor nodes having their IR diodes turned off.
3. Mobile node sends request to the first (by ID) nearby anchor node in range for position and IR radiation.
4. Anchor node switches its IR LEDs on, transmits location and lights on confirmation.
5. Mobile node using AoA sensor measures IR radiation intensity from different directions. To be able to calculate intensity that originates from anchor node IR LED, ambient IR radiation that was measured in step 1 is subtracted from values obtained in this step.
6. Mobile node estimates azimuth of the anchor node using measured intensity values (from each phototransistor, thus in each direction) of its IR radiation and sends confirmation that the location is received and angle of light arrival is measured. Estimated azimuth and anchor location are stored in memory.

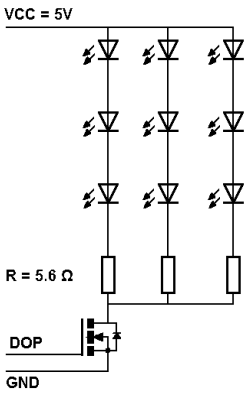


Figure 3. Radiation source scheme, total current drawn from voltage source by LEDs is 565 mA.

7. Localized node turns off LED light and transmits light off confirmation
8. After light off confirmation is received the steps 2. through 7. are repeated for each anchor node in the range.
9. After all anchors in the range have sent their locations and all azimuths are estimated mobile node selects three anchors with minimum geometric dilution of precision (GDOP). Localization precision depends on the relative locations of anchor nodes and this selection is mandatory to select most appropriate anchors triplet.
10. Location is estimated using selected anchors.

C. Azimuth estimation

Azimuth of the anchor node is estimated using measured voltage on each photodetector. Photodetectors on sensor are distributed evenly in order to cover fields which cover equal azimuth of circle. (Figure 4)

Radiation intensity is represented by voltage drop on a resistor in series with photodetector (phototransistor) and equals a numerical value.

Two phototransistors with maximum measured intensity are selected and their values V_1 and V_2 are used in estimation using the following expression

$$\beta = \frac{V_1 \times \alpha_1 + V_2 \times \alpha_2}{V_1 + V_2} \quad (1)$$

In which β is the angle of arrival, α_1 orientation of the phototransistor with highest voltage and α_2 orientation of the phototransistor with second highest voltage value.

D. Location estimation

Location estimation method is based on assumption that there are three anchor nodes in range (q_1, q_2, q_3) with known position (x_j, y_j) and unlocalized node p with unknown position and orientation (x, y, θ) (Figure 5) [8].

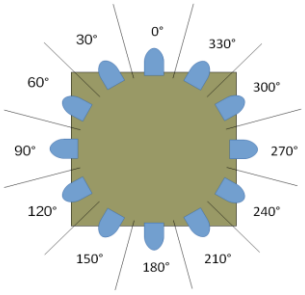


Figure 4. AOA sensor distribution of fields for sensor with 12 photodetectors

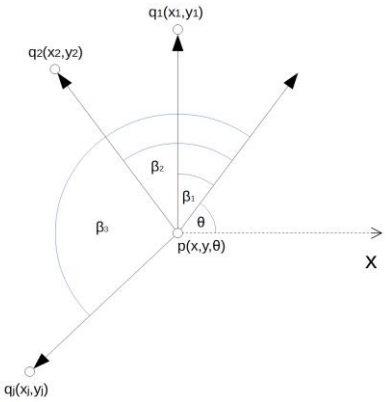


Figure 5. Nodes in 2D environment, q_1 q_2 and q_3 are selected anchors with known position and p is node with unknown position and orientation.

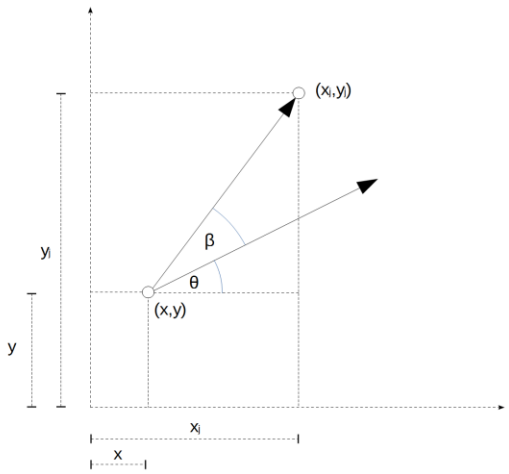


Figure 6. Nodes relation

$$q_j = \left[x_j + y_j\right], j = 1,2,3 \tag{2}$$

$$p = \left[x, y, \theta\right] \tag{3}$$

Anchor j azimuth β_j can be expressed using:

$$\tan\left(\theta + \beta_j\right) = \frac{y_j - y}{x_j - x}, j = 1,2,3 \tag{4}$$

where (x_j, y_j) represent j-th anchor node location, (x, y) mobile node location, θ mobile node orientation, and β is the angle between mobile node and anchor node as shown on Figure 6.

Equation (4) can be written as:

$$(x_j - x)\tan\left(\theta + \beta_j\right) = y_j - y, j = 1,2,3 \tag{5}$$

and by using trigonometric addition equations (5) it can be transformed to:

$$\frac{\tan\theta + \tan\beta_j}{1 - \tan\theta\tan\beta_j}\left(x_j - x\right) = y_j - y, j = 1,2,3 \tag{6}$$

$$\left(\tan\theta + \tan\beta_j\right)\left(x_j - x\right) = \left(y_j - y\right)\left(1 - \tan\theta\tan\beta_j\right) \tag{7}$$

After (7) multiplication and grouping:

$$\begin{aligned}&\tan\beta_j\left(x + y\tan\theta\right) - \left(y - x\tan\theta\right) - \\ &- \left(x_j + y_j\tan\beta_j\right)\tan\theta = x_j\tan\beta_j - y_j\end{aligned} \tag{8}$$

To simplify (8) can be multiplied with $\cos\beta_j$:

$$\begin{aligned}&\sin\beta_j\left(x + y\tan\theta\right) - \cos\beta_j\left(y - x\tan\theta\right) - \\ &- \left(x_j\cos\beta_j + y_j\sin\beta_j\right)\tan\theta = \\ &= x_j\sin\beta_j - y_j\cos\beta_j\end{aligned} \tag{9}$$

One can introduce following auxiliary variables:

$$\begin{aligned}u_1 &= x + y\tan\theta \\ u_2 &= y - x\tan\theta \\ u_3 &= \tan\theta\end{aligned} \tag{10}$$

So we can write (10) as:

$$\begin{aligned}&\left(\sin\beta_j\right)u_1 - \left(\cos\beta_j\right)u_2 - \left(x_j\cos\beta_j + y_j\sin\beta_j\right)u_3 = \\ &= x_j\sin\beta_j - y_j\cos\beta_j\end{aligned} \tag{11}$$

Equation (11) can be used for three nodes with known location to get three linear equations with three unknown variables in matrices:

$$A = \begin{bmatrix}\sin\beta_1 & -\cos\beta_1 & -x_1\cos\beta_1 & -y_1\sin\beta_1 \\ \sin\beta_2 & -\cos\beta_2 & -x_2\cos\beta_2 & -y_2\sin\beta_2 \\ \sin\beta_3 & -\cos\beta_3 & -x_3\cos\beta_3 & -y_3\sin\beta_3\end{bmatrix} \tag{12}$$

$$b = \begin{bmatrix}-x_1\cos\beta_1 & -y_1\sin\beta_1 \\ -x_2\cos\beta_2 & -y_2\sin\beta_2 \\ -x_3\cos\beta_3 & -y_3\sin\beta_3\end{bmatrix} \tag{13}$$

$$U = \begin{bmatrix}u_1 \\ u_2 \\ u_3\end{bmatrix} \tag{14}$$

Then we can write:

$$AU = b \tag{15}$$

$$U = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \left(A^T - A\right)^{-1} A^T b \qquad (16)$$

From (10) one can substitute:

$$\begin{bmatrix} x \\ y \\ \theta \end{bmatrix} = \begin{bmatrix} \frac{u_1 - u_3 u_2}{1 + u_3^2} \\ \frac{u_2 - u_1 u_3}{1 + u_3^2} \\ \tan^{-1} u_3 \end{bmatrix} \qquad (17)$$

IV. EXPERIMENTS

Experiments were carried out to evaluate performance and precision of the angle of arrival sensor and mobile node location estimation.

A. Azimuth estimation accuracy

Accuracy was measured in three different luminance conditions with different light sources:

- Complete darkness (0 lux)
- Incandescent light bulb (90 lux)
- Sunlight (1280 lux)

TABLE I. MEASURMENTS IN DARKNESS

Luminance:	0 lux
Distance between LED light source - sensor	**Angle measurement root mean squared error**
180 cm (edge of detection range):	30°
90 cm (half range distance)	3.58°
2.5 cm (source and sensor touching)	7.15°

The data in Table I. represents angle measurement root mean squared error at different distances between the source and the sensor. Distances that were used are:

- Edge of IR detection range – when angular movement of light source can be barely detected by sensor
- Half range distance
- Source and sensor touching – source IR LED and sensor photodetector are touching

Figure 7 shows measured angle deviations from real angle values in range from 76° to 105° in complete darkness for 90 cm distance (half range) from light source to sensor.

Figure 7. Estimated azimuth deviation from real values in darkness

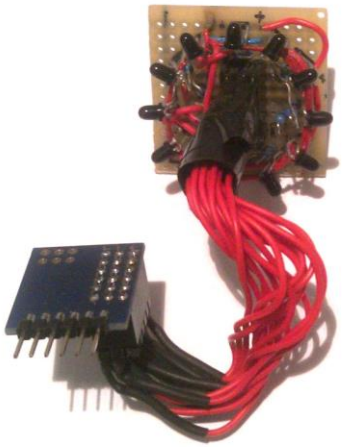


Figure 8. AoA sensor prototype

The best results were observed in complete darkness because there are no infrared components in light to interfere with the sensor operation and best accuracy is observed at half range distance. It is important to notice that observed results can be greatly improved by developing AoA sensor with smaller tolerances than prototype version (Figure 8) and by employing better filtering of surrounding noise (IR remote controllers, pulsating lights…)

B. Location estimation

Location estimation was carried using 3 anchor nodes at fixed coordinates (100,120; 10,100; 50, 20) and mobile node at different locations in range. True and estimated locations of the mobile node are presented in Table II.

TABLE II. LOCATION ESTIMATION RESULTS

True position x, y[cm]	Estimated position x, y [cm]	Deviation x, y [±cm]	Error distance [cm]
0,0	3,6	+3,+6	6,71
50,50	44,56	-6,+6	8,49
80,80	74,82	+4,+2	6,32
100,0	95,-8	-5,-8	9,43
0,120	-6,131	-6,+11	12,53
50,100	42,98	-8,-2	8,25
100,50	86,47	-14,-3	14,32
Error distance root mean square:			9,83

Position estimation is quite satisfactory if we take into account sensor price and still prototype development phase.

V. CONCLUSION

Although current platform and methods for azimuth and location estimation are still under development these initial results are very encouraging. Location estimation precision is high enough to be used for example in indoor mobile robot localization or assets tracking.

Future work includes integrated circuit development with both AoA sensor and IR LEDs using surface mount technology. Benefits would be twofold: precisely positioned phototransistors would allow better azimuth estimation accuracy and it would make platform more compact. As further optimizations each photodetector could be calibrated in order to compensate for systematic errors caused by imprecise phototransistors placements.

Another improvement could be more sophisticated algorithm for azimuth estimation that takes in account more than two phototransistors. Also, AoA range can be increased by driving IR LEDs with higher current for shorter period of time.

REFERENCES

[1] Holger Karl, Andreas Willig: „Protocols and Architectures for Wireless Sensor Networks“, John Wiles & Sons, 2005

[2] Amundson, I., Sallai, J., Koutsoukos, X., Ledeczi, A., Maroti, M., “RF angle of arrival-based node localisation”, Int. J. Sen. Netw., Vol. 9, No. 3/4, page. 209–224, May 2011

[3] Jiang, J.-R., Lin, C.-M., Lin, F.-Y., Huang, S.-T., “ALRD: AoA localization with RSSI differences of directional antennas for wireless sensor networks”, International Journal of Distributed Sensor Networks, Vol. 2013, Mar. 2013

[4] Nasipuri, A., el Najjar, R., “Experimental evaluation of an angle based indoor localization system”, in 2006 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, page. 209–224, Apr. 2006x

[5] Siemens, “SFH309FA, Silicon NPN Phototransistor, technical data sheet”

[6] Vishay, “TSUS5202, Infrared Emitting Diode, technical data sheet”

[7] Arbula Damir, Distributed algorithm for anchor-free network localization using angle of arrival, IEEE International Symposium on Industrial Electronics, 2008. ISIE 2008

[8] Huajie Shao, Xiaoping Zhang, Zhi Wang, Zhibo Wang: „Efficient Closed-Form Algorithms for AoA Based Self-Localization of Sensor Nodes Using Auxiliary Variables“, IEEE Transactions on Signal Processing, 2014

A Profile- and Community-Driven Book Recommender System

I. Petrović, P. Perković and I. Štajduhar
University of Rijeka, Faculty of Engineering, Department of Computer Engineering, Rijeka, Croatia
iva.petrovic5@gmail.com

Abstract - Book lovers often struggle to find something new to read. The choice of possible books to read can be overwhelming and confusing, making this subjective decision-making process quite difficult. Searching the Internet we often stumble upon opinions and book ratings of people who are strangers to us. We do not know what their favorite categories, authors, publishers and taste in books are, nor whether we should trust them. As a possible solution to the above mentioned problems, we present *Pickbooks* - a Web application that is partly a social network and partly a book database. It integrates the social aspects of today's popular social networks into a book recommender system, and not the other way around. It features book recommendation based on previously read books or personal preferences. It allows users to message and follow people they like, see their updates, reviews and get recommendations from them, and vice versa. The users are encouraged to rate a book they have read before, as those ratings affect the global lists of top rated and currently popular books. There are other features which are in the works and will allow us to further improve the presented application and attract more audience.

I. INTRODUCTION

Many book fans often find themselves in a situation where they do not know what to read next. Recommender systems are there to help them in these situations. Consider, for example, the *Internet Movie Database* (IMDb) website (http://www.imdb.com/). It is, most certainly, today's most popular movie database website. What if we could create something similar to that, but for book recommendations, i.e. a website that would incorporate those features and adapt them to books? Motivated and inspired by this idea, *Pickbooks* emerged.

A lot of websites give their users lists of top rated books, but that is no guarantee the users will like them. The main reason for this is the user's possible (and quite probable) disagreement with the interests and tastes of that author's list. One viable solution to this problem is to develop a recommender system embedded with some social network characteristics, enabling us to connect people with mutual interests. Broadening this vision of the recommender system by using these features provides the potential to bring books closer to the user.

In short, *Pickbooks* is a Web application, which fuses a book database, a social network and a book recommender system, into one product, available to users free of charge. Imagine binding a *Twitter*-like application (https://twitter.com/) [1] with a book recommender system, and you would understand the proposed idea.

The paper is structured in the following way. In the following section several popular and comparable related recommender systems are presented and the advantages of the proposed system over its direct competitor, *Goodreads*, are described. In section III, several important approaches to recommender systems are explained, introducing additionally some of the more popular consumer services as their respective representatives. The details concerning the proposed application and its' recommender system are outlined in section IV. In section V, final remarks are given together with an insight into the more probable future development plans for *Pickbooks*.

II. RELATED WORK

Browsing the Internet, one can find a lot of different types of recommender systems. Some of the most popular recommender systems, comparable to *Pickbooks* in several ways, are:

- *Last.fm* (http://www.last.fm/) is a music website that uses a recommender system called *Audioscrobbler*. Its goal is to create a detailed user profile based on the user's musical tastes. In this case, the profile is based on recording details of the tracks the user listens to.
- *Netflix* (*https://www.netflix.com/*) provides users with the ability to stream online videos on demand. Today, *Netflix* uses one of the best available recommendation algorithms. An extensive personalized video-recommendation system is based on ratings and reviews by its users.
- *Foursquare* (https://foursquare.com/) is a local search, recommendation system that learns what you like and leads you to places you love. Their recommendation system records places visited by the user, the things they like (reported to the application), and other users, whose advice they trust.

The recommendation system introduced in this paper is very similar to the recommender systems mentioned above. All three systems create a detailed profile based on a user's tastes. In the case of *Pickbooks*, this applies to book tastes, e.g. what book categories the user likes to

The work presented in this paper is partially supported by the University of Rijeka research grant 13.09.2.2.16.

read or what authors he or she prefers. Just like *Netflix*, the introduced recommender system is based on user's book ratings.

Pickbooks uses the best features of the most famous recommender system algorithms, combining them into one system that strives to give its users best book recommendations. What is new in this recommendation system? The goal is not only to collect data about users. It is also to provide users with the opportunity to give recommendations to their friends.

The main competitor of *Pickbooks* is *Goodreads* (*http://www.goodreads.com/*). It uses a recommendation algorithm to analyze which books people might like, based on books they have enjoyed in the past and books that people with similar tastes have liked. In order to get a book recommendation from this system, the user has to assign a rating to 20 books, at least.

For the authors of this paper, contemplating this from the perspective of *Goodreads* users, it was a big disadvantage. Sometimes rating books can be exhausting, especially if one hasn't read the books recently and cannot remember which 20 books to rate. It would be better to enable users to get a book recommendation immediately after registering. The proposed recommender system gives a book recommendation at the exact same moment the user starts searching for books or authors. Users are also given a list of top rated and trending books.

As was already pointed out, the idea was not only to create a recommender system, but also to interweave it with social network aspects. *Goodreads* implements a similar idea, i.e. it also enables their users to find friends with similar interests and to engage in conversations with them. *Pickbooks* also consists of a friend recommendation algorithm that monitors users' activities, what books they have read and how they have rated them, and based on that information it recommends cliques of users sharing the same interests. Another factor for the algorithm are the users' lists of friends. Using a user's list of friends, friends of his or hers friends are recommended. Most of social networks work on that principle, e.g. *Facebook* (https://www.facebook.com/) [2].

There are some features *Goodreads* is missing, which are implemented in *Pickbooks*:

- Generic representation of the system's data flow;
- Book recommendation when searching for book/author (e.g. based on book information the user is searching for, the system recommends similar books from the same category or books from the same author);
- Friend recommendations based on same interests or on mutual friends;
- Search for friends based on their name, e-mail address, city they come from or mutual friends;
- Easy way to recommend books to friends;
- Posting statuses about the books the user read/likes;
- Recommendation of top rated books from all categories.

III. RECOMMENDER SYSTEM

According to [3], recommender systems represent a powerful method for enabling users to filter through large information and product spaces, and, thus, pose an important part of the information and e-commerce ecosystem. The goal of a recommender system is to generate meaningful recommendations to a collection of users, for items or products that might interest them. There are several popular approaches to creating a recommendation algorithm: collaborative filtering, content-based filtering and a hybrid approach. They are described next.

A. *Collaborative filtering*

Collaborative filtering (CF) is a popular recommendation algorithm that bases its predictions and recommendations on the ratings or behavior of other users in the system. This type of approach collects and analyses a large amount of information on users' behavior, activities and personal preferences, and thus predicts what the users will like, based on their similarities to other users using the same service. The key advantage is that it does not rely on machine analyzable content and, as a result, it can generate recommendations without requiring an understanding of the item being recommended [4,5].

CF operates on the assumption that people will like similar things, which they liked in the past, and so can be slow to adapt at sudden changes in user preferences and other related information. The quality of recommendations are solely based on how much information a user has shared with the system, thus more information means better recommendations. Often, the algorithms operate on a user database that counts millions of entries and, as such, needs a large amount of computational power to calculate recommendations [6].

Last.fm, *Facebook*, *MySpace*, *LinkedIn* (https://www.linkedin.com/) and other social network services use collaborative filtering to recommend music, new friends, social connections, etc.

B. *Content-based filtering*

The second popular approach to creating a recommendation system is content-based filtering (CBF). In CBF, all items have a description and all users have a profile of their preferences (user history, ratings, personal information, etc.). The algorithm then matches the keywords from these two sources and generates a list of recommendations. The algorithm sorts and recommends items, based on their weight of importance.

One of the key challenges with this type of approach is concerned with enabling the system to use collected data from one source and apply the filtering to other content sources. For example, if a user likes a book category by the name of "Romance", it would be useful to be able to recommend movies, videos and news articles that contain that exact keyword [6].

CBF filtering techniques suffer from problems like limited content analysis, overspecialization and new users. Limited content analysis refers to the problem when a set of attributes, attached to every item in the collection, is rather small. No suitable suggestions can exist if the analyzed content does not contain enough information to differentiate items the user likes from items the user does not like. Overspecialization occurs when the system judges individual items based on a limited number of indicators. This results in the system recommending items that are either too alike to what the user has seen in the past or items that are content-identical to other recommendations shown at the same time. New users present a problem, because the recommendation algorithm has yet to receive profile information from the user to be able to create a recommendation based on criteria [7].

Examples of services that use this approach are: *Pandora Radio* (http://www.pandora.com/), *IMDB*, *8Tracks* (http://8tracks.com/), *Foursquare*, etc.

C. *Hybrid recommender system*

Hybrid recommender system (HRS), as the name clearly suggests, is a hybrid of the two aforementioned approaches to recommendation algorithms [8]. In recent history, researchers speculated that this kind of approach, combining collaborative and content-based filtering, may yield better and more accurate results, at least in some cases. Using the hybrid method, problems inherent to pure implementations (CF, CBF), like cold start and sparsity, can be overcome [9,10].

Netflix is one of the most popular services, which combines both collaborative and content-based filtering for recommending movies. Table I depicts the links between several popular Web services and their respective recommender system types.

IV. APPLICATION DESCRIPTION

Pickbooks is a Web application that allows people to discover new books, and share their thoughts about the ones they have already read. The application is powered by our own recommendation algorithm, which uses the hybrid approach for generating recommendations. Book recommendations are generated by leveraging user's personal preferences and connections with their friends. Previous searches, book ratings, the same book category and author preference with friends and friends of friends, book popularity - all are taken into account when learning about a user's reading preferences and habits.

TABLE I. RECOMMENDER SYSTEM TYPE USED BY A PARTICULAR WEB SERVICE

Web service name	CF	CBF	HRS
Last.fm	X		
Netflix			X
8Tracks		X	
IMDB		X	
Facebook	X		
Foursquare		X	
Pandora Radio		X	
MySpace	X		
LinkedIn	X		

Upon arriving on the landing page (Fig. 1), before starting using our recommender application, it is mandatory to give some information about oneself. The registration process requires the user to choose a unique username, password and enter an email address for further authentication. After a successful registration, the user is immediately able to search through the system's database, connect with his or her friends, meet other book lovers, rate his or her favorite books, keep a list of what to read next, etc. The system's topology is illustrated in Fig. 2. A more thorough analysis of its architecture is described next.

Our system can be roughly divided into five main parts:

- Home page,
- Profile page,
- Library,
- Friends, and
- Chat.

Home page (Fig. 3) is the main page of our Web application, where users are encouraged to search for recommendations for top rated and currently popular

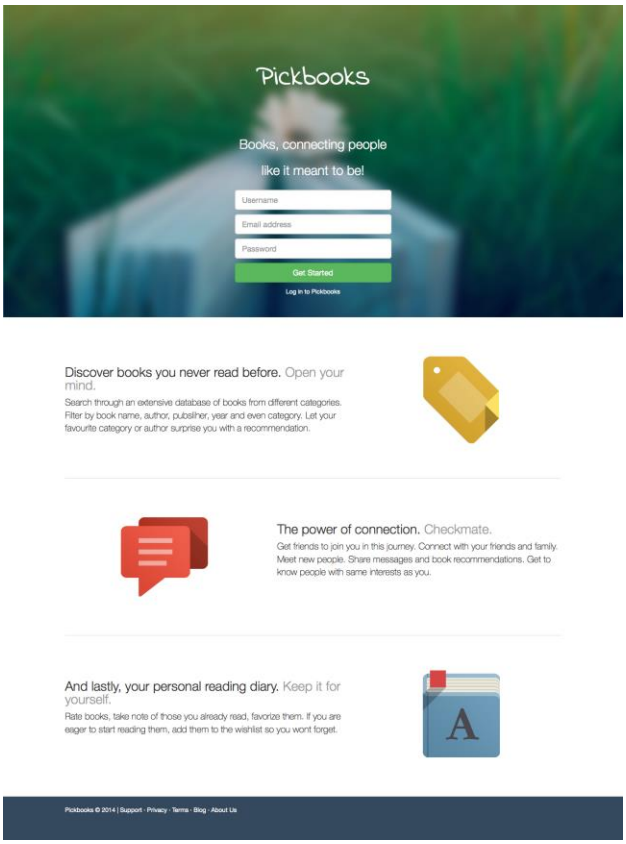


Figure 1. Landing page

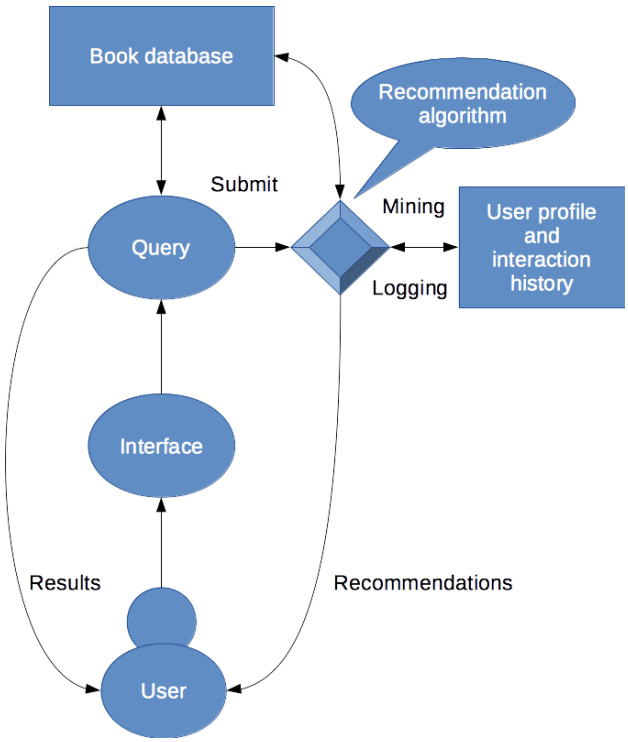


Figure 2. Generic representation of the system's data flow

books, among other things. An important part of this page is the news feed. There, users can find out who added them as a friend, and see what books their friends liked recently. Additionally, users are also able to share their opinions on interesting books they read with their friends through the option of posting statuses. They can also search for other users, sharing the same interests, or friends of friends, add them as their friends or send them messages. Home page combines basic recommendations and social network aspects, and puts them together in one place.

The profile page contains the public profile of the user. A user profile is personally curated by the owner, and contains the user's personal information (profile picture, full name, location, gender, biography and website), books he or she has read, his or her favorite books, books they would like to read next and personal statuses. Additionally, a list of people who are being followed by the user, and a list of people following the same user, are also available. All of the mentioned information can be explicitly hidden from the public, if the user chooses so, and can be adjusted using the security options.

Library page (Fig. 4) basically displays a book database. Searching for new books can sometimes be really exhausting. Our system enables filtering books by their name, author, publisher, publication year or category. To help the user even more in deciding what to read, the system provides a short book description, so he or she can have a quick look at the book content. Whilst reading the short description and basic information about a book, the user is also given a list of books from the same author and/or category.

Additionally, the user can recommend books to his or her friends. They will receive a notification about the recommendation shortly after it has been sent. Most of the book fans have their favorite authors, and want to read all of his or her books. Often they do not know which books they have read, and which they have not read yet. That is why it is useful to have a list containing all the books one has read. Our system allows the creation of lists like that one, and even more. The user is able to make his own wish list, a list of favorite books (once read), and even rate them.

The next part is the friends page. As it was already stated, our idea was to entwine a recommender system with a social network. Therefore, this approach required the introduction of a user's friends. This page allows the user to search for friends, or to find people sharing the same interests. There is an option to search and filter friends by name, email, location or mutual friends. Also, there is a generated list of people the user may know. When users find their friends, they can start following them, gaining the ability to receive information about their activities in the news feed on their home page.

The chat hub is the last main part of the Web application. On this page, the user can exchange messages in real time with his or hers friends, anytime he or she wants. If they are offline in that exact moment, that poses no real problem. They will receive a message when they come back online.

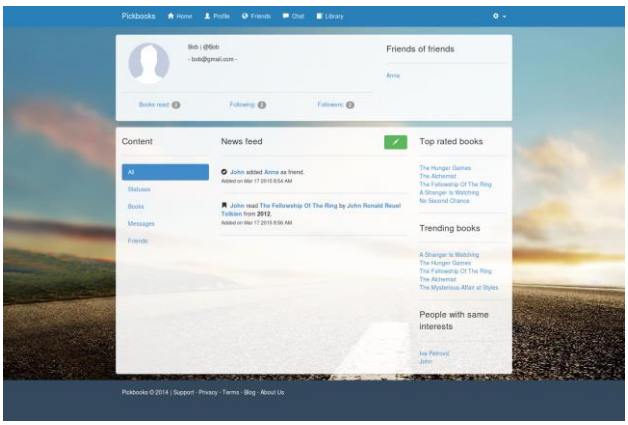


Figure 3. Home page

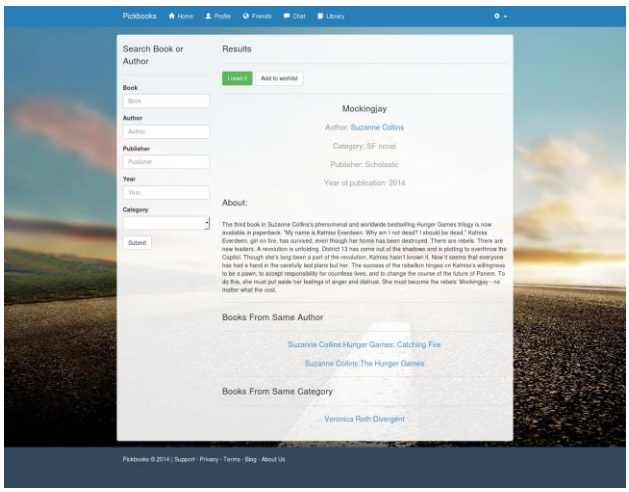


Figure 4. Library page

Pickbooks was developed using only free and open source Web technologies. For the frontend design, we used Bootstrap framework [11] and other custom HTML (Hypertext Markup Language), CSS (Cascade Style Sheet) and JS (JavaScript) scripts. The system's backend consists of both PHP (PHP Hypertext Preprocessor) and MySQL scripts, working together to enable users to interact with the application.

V. CONCLUSION AND FUTURE WORK

In this paper, a Web application for book lovers is presented to discover books and connect with friends. Although the expectations were met and the application has successfully evolved through development, there is still some work that needs to be done.

The idea for this application arose from the disappointment, evoked by the lack of good internet services regarding books. We found that Amazon's *Goodreads* has the biggest market share and that there are no other significant competitors in the market. This is the main reason this application was developed, thus fulfilling our vision regarding the approach to this kind of social experience and making it a reality.

We have a few features in store to improve the overall experience of *Pickbooks*, which will be added at a later date:

- Book lists: Every user will be able to create personal book compilations, like top lists, for every kind of occasion. The lists can then be shared with the public, for others to see, or just with friends. We believe this will allow users to find themselves a focused group of recommendations, created exclusively by other users, ones that will be very hard to reproduce by the algorithms or otherwise.
- Community driven database expansion: We will enable users to submit proposals for adding books to the library. They will be able to provide information about books and their authors. After the submission of data, administrators will confirm or reject books based on information authenticity.
- Administration page: In future we are going to develop an administration page where administrators will have the ability to confirm or reject user's book submissions. Also, another very important role of administrators will be to moderate the activity of the overall community, like deleting inappropriate content.
- Mood centric recommendations: A user's search for a book by categories can result in recommendations that don't necessarily suit his or her mood. We would like to extend our recommendation system, by adding an option of recommending books based on user's mood. This feature will be accessible only on demand. Getting book feedback from other users and analyzing keywords will allow us to get an estimate of a book's psychological effect.
- Online book purchases from retail stores: We would like to offer our users a quick and easy way to find a book and buy it online from different retail stores, available immediately.
- Further security and user privacy improvements: We know that user's personal information are indeed personal, therefore we would like to further improve security and privacy aspects of *Pickbooks*, giving users more control over these settings.
- As always, we strive to improve the accuracy and authenticity of book recommendations. We will take a look at our recommendation algorithms, see where we can improve, and try to fine-tune those areas.

REFERENCES

[1] A. Java, X. Song, T. Finin and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, ACM, 2007, pp. 56-65.

[2] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of Facebook "friends:" Social capital and college students' use of online social network sites," Journal of Computer-Mediated Communication, vol. 12, no. 4, 2007, pp. 1143-1168.

[3] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, "Collaborative Filtering Recommender Systems," Foundations and Trends in Human-Computer Interaction, vol. 4, no. 2, 2010, pp. 81-173.

[4] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, 1998.

[5] X. Su, and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," Advances in artificial intelligence, vol. 4, 2009.

[6] J. Leskovec, "Recommender Systems: Content-based Systems & Collaborative Filtering", Stanford University.

[7] G. Adomavicius, and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, 2005, pp. 734-749.

[8] J. Basilico, and T. Hofmann, "Unifying collaborative and content-based filtering," in Proceedings of the 21-st international conference on Machine learning, ACM, 2004.

[9] R. Burke, "Hybrid recommender systems: Survey and experiments," User modeling and user-adapted interaction, vol. 12, no. 4, 2002, pp. 331-370.

[10] G. Adomavicius, and A. Tuzhilin, "Context-aware recommender systems," Recommender systems handbook, Springer US, 2011, pp. 217-253.

[11] S. Santo, "Best way to customize Bootstrap CSS style," Bootstrap framework, 2014.

Extending LARIS with Session Types: A Case Study in the Railway Interlocking

Tibor Kiss, Florin Craciun, and Bazil Parv
Department of Computer Science, Faculty of Mathematics and Computer Science, Babes-Bolyai University, Cluj-Napoca
Email: {kisst, craciunf, bparv}@cs.ubblcuj.ro

***Abstract*—Choosing the DSL LARIS (LAnguage for Railway Interlocking Specifications) to develop control software for distributed railway interlocking systems can give us a lot of advantage over the general purpose languages, that have been proven in the industry over a long time. We present an extension of the LARIS language with session types to model the communication between railway entities. Our approach improves the safety and reliability of the system and also reduces the verification time by enabling the verification of deadlocks and livelocks of these systems. We evaluate the feasibility of our approach on a case study in railway interlocking.**

I. Introduction

Formal and automated verification of code correctness is a key component in the development of safety critical systems. Applying this principle in the software development of geographic railway interlocking systems (which must be concurrent and distributed) is more challenging, from several points of view. First, it is very difficult to specify global constraints in such systems. Secondly, assuming that these previously mentioned constrains are specified, it would be very challenging to verify them, because of the abstract and incomplete nature of the software specification.

Analyzing the current industrial practices, along with the state of the art programming languages and tools to develop distributed railway interlocking software, we concluded the following: Without a way to routinely and reliably build concurrent and distributed systems, following some correct-by-construction principles, the technical progress of these systems will draw to a close. Furthermore the price of such systems considering the development and maintenance costs, remains high. On the other hand, trying to follow the functional development principle [1], [2] and to explore the state space of such a system, it seems to be also generally impossible [2]. The verification leads approximately always to state explosion, therefore a new methodology to verify interlocking software correctness is highly welcome.

We have analyzed several geographic interlocking languages and methods [3], [4], and we concluded to choose LARIS [3] for our research. LARIS is a textual typed language for specifying geographic interlocking control systems.

The goal of our research is to extend the LARIS in order to enable a larger number of system requirements specification and formal verification on global level without compromising the current development process.

We propose the adoption of a new state of the art global specification principle named Session types in LARIS in order to enable static verification of global properties in the source code. This approach is in correlation with previous theoretical research results [5],[6].

The goal of this paper is to extend LARIS in order to ensure the correctness of its source codes against a local protocol specifications. These local protocol specifications are projected from a global protocol specifications developed in "Scribble" which is a "Session typed" protocol specification language [7].

Our contribution is twofold:

- We integrated Scribble into LARIS to enable the session typed protocols specification and verification
- We translated the "Scribble" local protocol specifications according to [5], into some weakest-preconditions and strongest-postconditions plus some static verification rules which can be statically verified with existing tools such as: [8], [9], [10], [11], [12], [13].

The rest of this paper is organized as follows: Sec. 2 introduces the DLS LARIS including the related terminologies and principles of railway interlocking systems, and the basic concept of "Session Types". Sec. 3 presents our LARIS source code verification principles to enforce specification of a local protocol. Sec. 4 illustrates our approach by modelling and verifying a combined scenario of route verification in the domain of the geographical railway interlocking systems; Finally, Sec. 5 draws the conclusion and presents the future works.

II. LARIS and Session types

A. Railway control

The management of a railway system consists of the following three essential layers: logistic,interlocking and infrastructure layer as in Figure 1. The control instructions for the railway yard has to be formulated at the logistic level. Next, it has to be guaranteed that the execution of control instructions does not lead to collisions and derailments. This is realized by means of an interlocking layer, which is the most important layer from safety perspective. This layer is located between the infrastructure and logistic level, and serves as a local solution for a specific railway yard, whereby the logic could be designed to cope with the peculiarities of the railway yard in a way to force the embodiment of safety principles and basic rules, according to which a train moves through a railway yard. On the top these, control instructions have to be passed on to the infrastructure elements like signals, points, tracks

etc.,These elements must change their state according to the received control instruction.

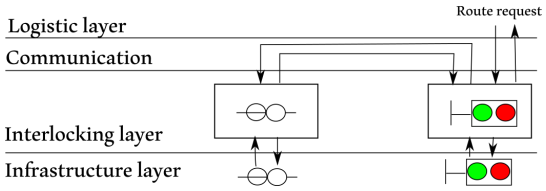


Fig. 1: Illustration of inter-object communication

B. LARIS

LARIS is a textual domain specific language for specifying geographical interlocking software. This language is a textual representation of the EURIS (EUropean Railway Interlocking Specification), a graphical specification language for railway interlocking systems. EURIS was inspired by the Prozes Ablauf Plane[2] being used at Siemens and it was used to specify the control of several railway yards. Hence, EURIS is a well-tested language, it became clear that it was appropriate for its intended use.

LARIS has an object-oriented architecture, consisting of a collection of generic components, representing the elements of the infrastructure such as signals and points. The components, which collectively make up the interlocking logic, communicate with each other by means of data-structures called telegrams. The components can also communicate with the logistic and infrastructure layers. These two layers can be considered the environment of the interlocking control system.

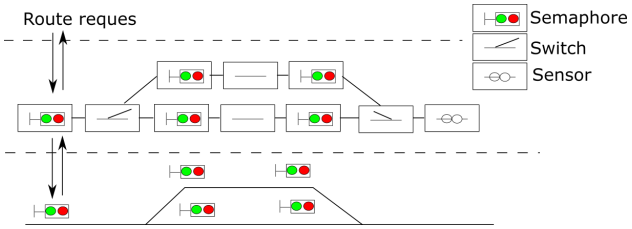


Fig. 2: Illustration of the layered interlocking architecture

The main advantage of LARIS over a functional specification method, is the possibility to reuse the previously defined components. The central part of this approach is the behavior specification of the components, which consists of handling incoming telegrams. When all types of components have been specified in full detail, the specification of a particular railway yard is constructed by simply connecting its separate building blocks in the appropriate manner (like in Figure 2).

Like in EURIS, LARIS uses LSC (Logic and Sequence Chart) for component specification. A typical LARIS LSC is given in [Listing 1.].

Listing 1 : LARIS LSC Example

```
1  LSC name(Ea,Eb:Component;Pa,Pb:Port;ID:Int) =
2    vars  V1,B2,V3:Bool;
3
4    initial skip
5
6    mes Pa?A06(RL:Int) =
7      RL:=RL+TSL;
8      if  TSU
9        then PER:=false;
10       else TRL:=true;
11     Eb |> Pb!A06(RL);
```

Explaining Listing 1, the LSC keyword is followed by a unique name and a list of initialization parameters. These parameters are given by the rail yard developer, who uses the static specification language of LARIS to specify these parameters at the LSC instantiation.

The message listening statement [Listing 1 Line 6] starts with the ”mes” keyword followed by a port identifier (in our case Pa) separated by a question mark from the telegram identifier(in our case A06) and a list of variables. From a semantic point of view, this can be interpreted as follows: if we receive a telegram A06 from the interface Pa and A06 contains the following variables, then the telegram must be processed with the following lines of code (see example [Listing 1 Lines 6:11]). The second statement is the message transmission [Listing 1 Line 11]. In this line of code the message A06 is sent out on the interface Pb to the component Eb, and subsequently the component Eb must have a definition to handle the message properly. A more detailed explanation can be found in [3]. The first advantage of LARIS over EURIS is that is well specified [3]. Another advantage is that of being a typed language: every variable used in LARIS has a certain type, so all the methods and operations must fulfill this type discipline.

C. Session types

In spite of its many benefits, as mentioned above, LARIS also has its disadvantages. Though it is true that the components can be reused anytime, there are times when these systems can have some serious safety or liveness issues. The source of these issues have been identified as lack of global specifications. To avoid this problem we suggest ”Scribble” a global protocol specification language. Scribble has a theoretical foundation, based on the π -calculus and Session Types, to ensure that protocols described using the language are sound, and do not suffer from deadlocks or livelocks.

Session types are a type formalism used to model structured communication-based programming for distributed systems [14]. In particular, binary session types, describe the communication between exactly two participants in such scenarios [15]. Duality is a fundamental notion of session types, as it describes compatible behaviors between communicating parties. When session types are added to a standard communication channel, it must statically enforce that client-server communication proceeds according to a previously defined choreography. The

object and benefits of this encoding are mainly in expressiveness and re-usability in larger system settings.

III. Verification principles

The central idea of session types is to parameterize a channel with some type that represents a protocol, which the type system then enforces [16]. In LARIS, there is no mechanism to encode a protocol. We may enforce a protocol using some special interface constraints, named local protocol type constraints. For the generation of local protocol type constraints, we decided to use the Scribble multiparty protocol specification language. Scribble is a platform-independent description language for the specification of asynchronous, multiparty message passing protocols [17], [18], [19]. Scribble may be used to specify protocols from both the global (neutral) and the local perspective of a particular participant (abstract as a role); at heart, the Scribble language is an engineering incarnation of the notation for global and local types in formal MPST (multiparty session types) systems and their correctness conditions [27]. The Scribble Conversation API provides the local communication operations the possibility to implement the endpoint programs for each role natively in various mainstream languages. The current version of Scribble supports Java [7] and Python [20] Conversation APIs [27].

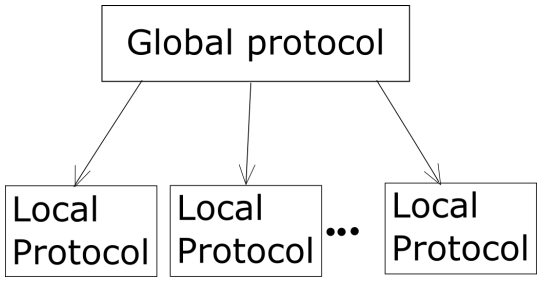


Fig. 3: Local protocol projection

Our development work-flow starts from the explicit specification of the global protocols in Scribble, similarly to existing informal approaches, such as sequence diagrams and communication diagrams. We modified Scribble in a way that it enables the specification of telegrams.

Intuitively, a choreography is well-typed, provided that shared channels are used and processes play roles according to the local protocols [Table II], and session channels are used according to global session protocols [Table I].

global G ::=	$A->B:\langle U\rangle;G$	(communication)
	$A->B:\{l_i:G_i\}_{i\in I}$	(branching)
	end	end
	t	(call)
U::=	S — G@A	
S::=	bool — int — string — ..	

TABLE I: Local protocol specification

Therefore, after the specification of the global protocols, the next step in the Scribble framework (Figure 3) is the projection of local protocols, for each role, from the global protocols. Having this specification, we must enforce it as a type discipline in LARIS, in order to be able to verify the specification correctness.

Local T ::=	$K!\langle U\rangle;T$	send
	$K?\langle U\rangle;T$	receive
	$K\oplus\{l_i:T_i\}_{i\in I}$	selection
	$K\&\{l_i:T_i\}_{i\in I}$	branching
	$\mu t.T\|t\|end$	end

TABLE II: Local protocol specification

To enforce this discipline, we introduce a new interface type constraint, named session type interface, which enables the identification of the roles in order to be able to verify the type discipline correctness.

One of the most challenging parts of this work, was to find an appropriate verification method to check the local protocol specification. We intend to verify only the safety aspects of this specification. In order to enable the safety verification of these aspects in LARIS, some well known formalisms (like: model checking, deductive verification (Hoare Logic, Separation Logic, etc.) etc.) must be adopted or a new formalism must be invented. We were inspired by the work of T.F.Melham [21], Aaron Turon and Mitchell Wand [22] and Peter OHearn and Tony Hoare [23] to transform π -calculus in a specification language with pre- and post-conditions. Intuitively, this transformation must be a one-to-one projection mechanism from the local protocol specification (from the Table [II]) into a specification method in LARIS, otherwise the correctness of the verification can not be assumed. We have identified some translation principles, which enable the total translation of the protocol specification into static verification rules, consisting of existential verification rules plus static specifications in form of Hoare logic pre- and post-condition. A Hoare Triple is of the form $\{P\}\ S\ \{Q\}$, where P is the precondition, Q is the post-condition, and S is the statement(s) that implement the function. These pre- and post-conditions could be verified subsequently to prove the type discipline correctness.

In the next section, we present the transformation.

π -calculus	Static specifications
$K!\langle U\rangle;T$	$\dfrac{x\in\mathbb{N}}{\{K!\langle U\rangle(x)\wedge Q\}K!\langle U\rangle\{K!\langle U\rangle(x+1)\wedge Q\}}$
$K?\langle U\rangle;T$	$\exists mesK?U();mesK?U()\equiv K?\langle U\rangle$
$K\oplus\{l_i:T_i\}_{i\in I}$	$\forall l_i:T_i;mesK?l_i()\equiv K?\langle l_i\rangle$
$K\&\{l_i:T_i\}_{i\in I}$	$\dfrac{\{l_i:T_i\}_{i\in I};\{x_i\in\mathbb{N}\}_{i\in I}}{\{K!\langle l_i\rangle(x_i)\wedge Q\}K!\langle l_i\rangle\{K!\langle l_i\rangle(x_i+1)\wedge Q\}}$

TABLE III: Local protocol translation rules

The specification of the ”receive” statement from [Table II]($K?\langle U\rangle;T$) is translated into this ”*mes* K?U() =” LARIS specification [Table III. In order to enable the static verification of this property, it is necessary to replace K with the

corresponding network interface from the protocol interface specification.

The ”branching” specification from the [Table II] can be reduced to a set of ”receive” specifications [Table III], because of the LARIS asynchronous message listening mechanism. Therefore, the verification consists of an iterative check of these listening statements. The ”end” keyword from the Table II plays no role in our definition because the persistent nature of the channel.

The remaining specifications translation from Table III needs a more sophisticated approach. Because of the asynchronous handling of the messages in LARIS, we can split a local specification into a list of ”receive” operations Table II followed by a list of operations of type ”send” Table II and ”selection” Table II. Our first goal is to avoid the unintended message transmission, the second, to ensure that the intended message transmissions are triggered once and only once, and third to ensure the message transmission order. The message transmission operations can be modeled with integer variables according to [21], therefore this specification can be partially translated into pairs of preconditions and post-conditions on these integer variables.

Let us consider the model of a message transmission an integer variable incrementation. Having this model, we can substitute the transmission operations with the corresponding models. An undefined message transmission on this interface is considered an error. The pre- and post-conditions of the message handler consists of assuming that the initial value of the model is zero and ensuring that the final state of the value is one. The order of the messages is ensured by the following assertion: Considering two messages m_i and m_j, and considering the model of m_i and m_j, $mod(m_i)$ and $mod(m_j)$ then:

$(K!\langle m_i\rangle;T).....(K!\langle m_j\rangle;T) \Longrightarrow m_i \geq m_j$

For the verification of the ”branching” statement from Table II we assume that all the variables, which model a message transmission, are zero. We must ensure the sum of these models will always be one. Although, in case of a protocol invocation, the problem seems to become more complicated, but actually it is simple and can be solved by replacing the sub-protocol body in the protocol.

IV. An Example

In this section we study the applicability of session types in industrial environments such as the Geographical Railway Interlocking Systems development. From the fundamental point of view, we demonstrate the correctness of our work using subsection III. We also illustrate the use of session types with an application, a subproblem from a route allocation problem.

According to Banci [2], [24] if a typical railway interlocking system receives a route request, the answer to this request has to go through a series of checks:

1) a route can not be locked if a track section element is occupied;
2) a route is available if no part of the route is already locked;
3) if the route is available it will be locked;
4) locking a route implies that all its track elements become locked;

Verification requirements from the previous mentioned steps, can be triggered also by the logical layer to verify the state of route elements, in order to prevent the triggering of an impossible route allocation. To prove our verification theory, we have chosen the implementation of the second item from the previous list. The verification process starts by sending a route verification request from the logistic layer to the interlocking layer. This request must be sent to one of the route end elements which must be a light signal. Depending on the light signal internal state, two scenarios are possible: In the first scenario the signal will answer with a fail message to the logical layer. In the second scenario the request must be delegated toward to the next route element in order to verify its state. This element must be the left or the right neighbor of the light signal.

```
1  module ro.ubbcluj.irg.RouteVerification;
2
3  type <telegram> "Telegrams.xts" from
4    "A06" as Route;
5  type <telegram> "Telegrams.xts" from
6    "A11" as Ok;
7  type <telegram> "Telegrams.xts" from
8    "A16" as SegmentError;
9
10 global protocol RouteVerification
11 (role LogisticLayer,role Semaphore,
12   role LeftElement,role RightElement ) {
13
14   verify(Route)
15     from LogisticLayer to Semaphore;
16
17   choice at Semaphore {
18     do RouteElementVerification(
19       Semaphore as SourceElement,
20       LeftElement as DestinationElement);
21   } or {
22     do RouteElementVerification(
23       Semaphore as SourceElement,
24       RightElement as DestinationElement);
25   } or {}
26
27   choice at Semaphore {
28     free(Ok) from Semaphore to LogisticLayer;
29   } or {
30     allocated(SegmentError)
31       from Semaphore to LogisticLayer;
32   }
33 }
```

Listing 2 : Scribble: Global route verification protocol

We consider the previous specification as a requirement. Before starting the development of a railway interlocking software requirement, it is recommended to clearly identify the involved railway elements, the role what each element plays and to specify a consistent communication protocol. These are done by specifying the choreography of communication between the railway track elements in order to have a proper communication mechanism to address the involved track elements without communication deadlock or live-lock. In our case these specifications are done in Scribble protocol

specification language Listing 2, 3.

As can be seen in Listing 2 the specification starts with the module specification. In the next step the necessary telegrams are imported (the telegrams are globally specified package types used by LARIS to transmit data). The roles are enumerated in the global protocol header. The "LeftElement" role means the element located before the semaphore in a railway yard, and the "RightElement" the element located after the semaphore in a railway yard.

The global protocol specification starts with a request transmission from the logistic layer to a semaphore. In order to answer the request, the semaphore must decide if the elements are allocated or not. Therefore the semaphore checks its internal state, and according to them it sends an answer to the logistic layer, or sends another request to its left or right neighbor according to the received route, as parameter from the logical layer, to check the neighbor's state. The neighbor must behave in the same way as the semaphore did, it must check its internal state and according to them must answer with an allocated message to the semaphore or must transmit the message forward until the package reaches the end node. Depending on its internal state, the end node must **send** an allocated or a **free** message. The messages must be propagated back to the semaphore, in order to enable the semaphore to answer the logistic layer with an appropriate response.

```
1  module ro.ubbcluj.irg.RouteElementVerification;
2
3  type <telegram> "Telegrams.xts" from
4    "A23" as Route;
5  type <telegram> "Telegrams.xts" from
6    "A24" as Free;
7  type <telegram> "Telegrams.xts" from
8    "A25" as Error;
9
10 global protocol RouteElementVerification
11   (role SourceElement,role DestinationElement){
12
13   verify(Route)
14     from SourceElement to DestinationElement;
15
16   choice at DestinationElement{
17     free(Free)
18     from DestinationElement to SourceElement;
19   } or {
20     allocated(Error)
21     from DestinationElement to SourceElement;
22   }
23 }
```

Listing 3 : Scribble: Global element verification protocol

Having the protocol specification from Listing 2 and 3, we can start the semaphore implementation. Our extension of Laris enables the verification of the local protocol of the components. We followed the same principle as in [25],[26], and we have connected the protocol specification with LARIS with a new type of interface. The syntax of the specification can be found in Listing 5.

Hereafter, a short presentation of the syntax will be given. The syntax of the interface consists of a new expression "implement role" followed by a list of interfaces. Each interface definition has the following form: the global protocol name followed by a role identifier, which enables the identification of the local protocol [Listing 5 Lines 3:9], followed by a list of role associations, where each role is associated with the corresponding interface and component.

```
1  module ro.ubbcluj.irg
2    .RouteVerification_Semaphore_Ext;
3  type <telegram> "Telegrams.xts" from
4    "A06" as Route;
5  type <telegram> "Telegrams.xts" from
6    "A11" as Ok;
7  type <telegram> "Telegrams.xts" from
8    "A16" as SegmentError;
9  type <telegram> "Telegrams.xts" from
10   "A23" as REV_Route;
11 type <telegram> "Telegrams.xts" from
12   "A24" as REV_Free;
13 type <telegram> "Telegrams.xts" from
14   "A25" as REV_Error;
15
16 local protocol RouteVerification
17 (role LogisticLayer,
18   role LeftElement,role RightElement ) {
19
20   verify(Route) to Semaphore;
21
22   choice at Semaphore {
23     REV_verify(REV_Route) to LeftElement;
24
25     choice at LeftElement{
26       REV_free(REV_Free) from LeftElement;
27     } or {
28       REV_allocated(REV_Error) from LeftElement;
29     }
30
31   } or {
32     REV_verify(REV_Route)to RightElement;
33
34     choice at RightElement{
35       REV_free(REV_Free) from RightElement;
36     } or {
37       REV_allocated(REV_Error) from RightElement;
38     }
39   } or {}
40
41   choice at Semaphore {
42     free(Ok) to LogisticLayer;
43   } or {
44     allocated(SegmentError)
45       to LogisticLayer;
46   }
47 }
```

Listing 4 : Scribble: Local route verification protocol

The rest of the code remains unchanged. We removed some aspects (like the enumeration definitions or the "DestinationElement" role definition and implementation) from the code Listing 5, to make it compact and understandable.

In the remaining part of this section the verification process will be illustrated.

Having the local protocol discipline on our source code Listing 5, we must check if the discipline is respected or not.

The "receive" and "selection" properties from Table II are easy to check. This can be done simply by iterating over the AST (abstract syntax tree) of the LARIS language.

```
LSC Semaphore(EA,EB:Component;
        Pa,Pb:Port;TSL:Int)
implement role RouteVerifcation.Semaphore(
 LogisticLayer=[Log,log],leftElement=[EA,Pa],
 rightElement=[EB,Pb]),
RouteElementVerification.SourceElement(
  DestinationElement=[EA,Pa]),
RouteElementVerification.SourceElement(
  DestinationElement=[Eb,Pb]) =

  initial skip;

  mes log? A06 (RL:Bool[16]) =
    if STATE == ALLOCATED
      then Log |> log ! A16(self,ALLOCATED)
      else
        if RL[0]== true
          then EA |> Pa! A22(RL:Bool[16],1)
          else EB |> Pb! A22(RL:Bool[16],1);

  mes Pa? A24 (RL:Bool[16],size:Int,pos:Int) =
   Log |> Log ! A11();

  mes Pa? A25 (item:Id,state:State,RL:Bool[16],
    size:int,pos:int) =

    Log |> Log ! A16(item,state);

  mes Pb? A24 (RL:Bool[16],size:Int,pos:Int) =
    Log |> log ! A11();

  mes Pb? A25 (item:Id,state:State,RL:Bool[16],
    size:Int,pos:Int) =

    Log |> log ! A16(item,state);
```

Listing 5 : LARIS semaphore partial implementation

```
assume (var1 == 0 && var2 == 0
    && var3 == 0 && var4 ==0)
ensure (var1 + var2 + var3 + var3 == 1)
if STATE == ALLOCATED
      then
        var1++;
      else
        if RL[0]== true
          then var2++
          else var3++;
```

Listing 6 : LARIS send out verification model

To prove the ”send” and ”branching” property, we will apply the modeling and verification principle from the previous section. We have chosen to demonstrate the local protocol correctness from Listing 4. According to this specification, the message handler ”mes log? A06 (RL:Bool[16]) =” from Listing 5 must emit one of the following telegrams: A11 on the ”log” interface, A16 on the ”log” interface, A22 on the ”Pa” interface or A22 on the ”Pb” interface. These messages must be sent out once and only once. Following the modeling and verification rules from the previous chapter we replace ”Log | > log! A11” with ”var1++”,”Log | > log! A16” with ”var2++”, ”EA | > Pa! A22” with ”var3++” and ”EB | > Pb! A22” with ”var4++”. According to our specification we assume that each variable is zero and we ensure that the sum of these values will be one. The last step is straightforward, because there are a lot of static program verifiers [8], [9], [10], [11], [12], [13], which are able to prove the correctness of such a code with regards to the formal specification. This proves the soundness of our theory and its applicability in the practice.

V. Conclusion

In this paper we demonstrated that session types can be combined and used with a domain specific programming language like LARIS. Our work was inspired by the ”Scribble protocol language” combined with Java and Python [27],[28],[17], and also by T. F. Melham [21], Aaron Turon and Mitchell Wand [22] and Tony Hoare and Peter O’Hearn [23] works to translate the π -calculus into Hoare logic or Separation Logic specifications. Promising results have been reported in adopting session types in other programming languages, like Haskell [25],C# [29] and recently the Chorlang [30], but their approach is based on embedding session types into the programming language. We observed that combining ”Scribble” with LARIS can provide a lot of benefits in the verification and validation of a geographically distributed interlocking system. A major benefit of this combination is the local verification of global protocols. We modeled the message sending with an integer in order to enable the verification of the local protocol in LARIS. We encoded the local protocol specification into static verification rules, in addition with specifications in form of pre- and post-conditions to enable the verification of the generated model in order to ensure the safety properties correctness of this typing discipline. We validated our method on a railway interlocking problem implementation in Scribble and LARIS.

We consider that extending the protocol language ”Scribble” with a specification language like in [28] and [5] can open currently unexpected opportunities in the local verification of global properties. We expect that this specification and verification approach allows the framing of newly emerging issues in the world of communication centered programming and micro services, into the mainstream domains of type theories and systems.

References

[1] A. Bonacchi and A. Fantechi, “On the validation of an interlocking system by model-checking,” in *Formal Methods for Industrial Critical Systems*, ser. Lecture Notes in Computer Science, F. Lang and F. Flammini, Eds. Springer International Publishing, 2014, vol. 8718, pp. 94–108. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10702-8_7

[2] M. Banci and A. Fantechi, “Geographical versus functional modelling by statecharts of interlocking systems,” *Electron. Notes Theor. Comput. Sci.*, vol. 133, pp. 3–19, May 2005. [Online]. Available: http://dx.doi.org/10.1016/j.entcs.2004.08.055

[3] W. Fokkink, J. F. Groote, M. Hollenberg, and B. van Vlijmen, *LARIS 1.0, LAnguage for Railway Interlocking Specifications*, ser. Lecture Notes in Computer Science. New York NY: Springer Verlag, 1999, vol. 4204.

[4] F. van Dijk, W. Fokkink, G. Kolk, P. van de Ven, and B. van Vlijmen, *Mathematical Theory of Efficient Piracy*, ser. Lecture Notes in Computer Science. New York NY: Springer Verlag, 1998, vol. 3204.

[5] M. F. Jensen. (2014) Formal verification of distributed programs using session types and coq. [Online]. Available: http://sevengoslings.net/~fangel/thesis.pdf

[6] S. M. Loos, D. Renshaw, and A. Platzer, "Formal verification of distributed aircraft controllers," in *Proceedings of the 16th International Conference on Hybrid Systems: Computation and Control*, ser. HSCC '13. New York, NY, USA: ACM, 2013, pp. 125–130. [Online]. Available: http://doi.acm.org/10.1145/2461328.2461350

[7] Scribble. (2015) Scribble homepag. [Online]. Available: http://www.scribble.org

[8] E. P. David Cok, Joe Kiniry. (2015) Esc/java2. [Online]. Available: http://www.cs.ru.nl/E.Poll/talks/jml_tutorial/2_tool.pdf

[9] D. R. Cok, J. R. Kiniry, and D. Cochran, "Esc/java2 implementation notes," 2008.

[10] EVEREST. (2015) Jack: Java applet correctness kit. [Online]. Available: http://www-sop.inria.fr/everest/soft/Jack/jack.html

[11] W. Ahrendt. (2015) The key tool. [Online]. Available: http://www.key-project.org/

[12] Y. Moy. (2015) Krakatoa verification tool for java. [Online]. Available: http://krakatoa.lri.fr

[13] ——. (2015) Frama-c jessie plug-in. [Online]. Available: http://frama-c.com/jessie.html

[14] O. Dardha, E. Giachino, and D. Sangiorgi, "Session types revisited," in *Proceedings of the 14th Symposium on Principles and Practice of Declarative Programming*, ser. PPDP '12. New York, NY, USA: ACM, 2012, pp. 139–150. [Online]. Available: http://doi.acm.org/10.1145/2370776.2370794

[15] G. Castagna, M. Dezani-Ciancaglini, E. Giachino, and L. Padovani, "Foundations of session types," in *Proceedings of the 11th ACM SIGPLAN Conference on Principles and Practice of Declarative Programming*, ser. PPDP '09. New York, NY, USA: ACM, 2009, pp. 219–230. [Online]. Available: http://doi.acm.org/10.1145/1599410.1599437

[16] S. J. Gay and M. Hole, "Types and subtypes for client-server interactions," in *Proceedings of the 8th European Symposium on Programming Languages and Systems*, ser. ESOP '99. London, UK, UK: Springer-Verlag, 1999, pp. 74–90. [Online]. Available: http://dl.acm.org/citation.cfm?id=645393.756448

[17] K. Honda, R. Hu, R. Neykova, T.-C. Chen, R. Demangeon, P.-M. Denilou, and N. Yoshida, "Structuring communication with session types," in *Concurrent Objects and Beyond*, ser. Lecture Notes in Computer Science, G. Agha, A. Igarashi, N. Kobayashi, H. Masuhara, S. Matsuoka, E. Shibayama, and K. Taura, Eds. Springer Berlin Heidelberg, 2014, vol. 8665, pp. 105–127. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-44471-9_5

[18] K. Honda, A. Mukhamedov, G. Brown, T.-C. Chen, and N. Yoshida, "Scribbling interactions with a formal foundation," in *Distributed Computing and Internet Technology*, ser. Lecture Notes in Computer Science, R. Natarajan and A. Ojo, Eds. Springer Berlin Heidelberg, 2011, vol. 6536, pp. 55–75. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-19056-8_4

[19] S. Tea. (2015) . scribble project github homepag. [Online]. Available: http://www.scribble.org

[20] R. Hu, R. Neykova, N. Yoshida, and R. Demangeon, "Practical interruptible conversations: Distributed dynamic verication with session types and python," *a*, vol. 8174, pp. 148–130, 2013.

[21] T. F. Melham, "A mechanized theory of the pi-calculus in hol," NORDIC JOURNAL OF COMPUTING, Tech. Rep., 1992.

[22] A. Turon and M. Wand, "A resource analysis of the -calculus," *Electronic Notes in Theoretical Computer Science*, vol. 276, no. 0, pp. 313 – 334, 2011, twenty-seventh Conference on the Mathematical Foundations of Programming Semantics (MFPS XXVII). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1571066111001198

[23] T. Hoare and P. O'Hearn, "Separation logic semantics for communicating processes," *Electronic Notes in Theoretical Computer Science*, vol. 212, no. 0, pp. 3 – 25, 2008, proceedings of the First International Conference on Foundations of Informatics, Computing and Software (FICS 2008). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S157106610800265X

[24] M. Banci and A. Fantechi, "Some experiences on formal specification of railway interlocking systems using statecharts," *Electron. Notes Theor. Comput. Sci.*, 2005.

[25] R. Pucella and J. A. Tov, "Haskell session types with (almost) no class," *SIGPLAN Not.*, vol. 44, no. 2, pp. 25–36, Sep. 2008. [Online]. Available: http://doi.acm.org/10.1145/1543134.1411290

[26] A. Vallecillo, V. T. Vasconcelos, and A. Ravara, "Typing the behavior of objects and components using session types," 2003.

[27] N. Yoshida, R. Hu, R. Neykova, and N. Ng, "The scribble protocol language," in *Trustworthy Global Computing*, ser. Lecture Notes in Computer Science, M. Abadi and A. Lluch Lafuente, Eds. Springer International Publishing, 2014, pp. 22–41. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-05119-2_3

[28] R. Neykova, N. Yoshida, and R. Hu, "Spy: Local verification of global protocols," in *Runtime Verification*, ser. Lecture Notes in Computer Science, A. Legay and S. Bensalem, Eds. Springer Berlin Heidelberg, 2013, vol. 8174, pp. 358–363. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40787-1_25

[29] M. Fähndrich, M. Aiken, C. Hawblitzel, O. Hodson, G. Hunt, J. R. Larus, and S. Levi, "Language support for fast and reliable message-based communication in singularity os," *SIGOPS Oper. Syst. Rev.*, vol. 40, no. 4, pp. 177–190, Apr. 2006. [Online]. Available: http://doi.acm.org/10.1145/1218063.1217953

[30] F. Montesi. (2015) Choreography programming language. [Online]. Available: http://www.chor-lang.org/

Monitoring Cerebral Processing of Gustatory Stimulation and Perception Using Emotiv Epoc

Guruprasad Madhale Jadav, Miroslav Vrankić and Saša Vlahinić
Tehnički Fakultet Sveučilište u Rijeci, Rijeka, Croatia
jguruprasad@uniri.hr

Abstract - The affective and cognitive states of the user can be effectively monitored by electroencephalography (EEG). The human taste stimulation and perception produces changes in the EEG signals of the brain at several subdivisions of insula, orbitofrontal cortex and anterior cingulate cortex. In this work we propose a method which can be used to monitor these signals elicited by taste stimulation and perception using a simple gaming EEG device. Emotion elicited by taste recall is stronger than stimulus elicited emotion will be re-established but with a gaming EEG device. The results are presented for a number of subjects and obtained using the research edition of the gaming EEG headset.

I. INTRODUCTION

Neuro mechanism for taste preferences plays a very vital role in our lives. The gustatory stimulation and perception has not been the field in which extensive research has taken place, and the early results are disputed [1].

The acidic and alkaline forming effects of food and liquids that we ingest in our body can disrupt the pH level of some fluids in our body, and our oral fluid is one of them [2]. Our oral hygiene largely depends on the pH of the oral fluid, which directly gets affected by the food or liquids we chew or drink, also the effect of change in pH is clearly sensed by the gustatory organ which in turn causes change in EEG signals, these signals can be recorded at different frontal and temporal locations at several subdivisions of insula, orbitofrontal cortex and anterior cingulate cortex [1].

The research edition of the Emotiv's gaming EEG headset also called as Epoc headset is a consumer brain-computer interface device, which communicates with the computing environment wirelessly. This headset observes temporal pattern of the EEG signals produced by the human brain. A simple to setup device like this can be used to assign control tasks to a particular taste perception by monitoring the changes in Event Related Potentials (ERP). The Emotiv's Epoc headset does not need an experimental laboratory setup where pre and post experimental difficulties like electrode placement and rinsing of subjects hair to remove electrode gel becomes mandatory. It needs minimal adjustment of electrodes and uses cotton felt pads which are to be pre soaked in saline solution to make good contact with the scalp (Bobrov et al., 2011; Stytsenko et al., 2011). The EEG activity for the affective and cognitive states of the user elicited by taste stimulation and perception respectively can be observed at the electrode locations AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7 and T8 according to 10-20 system of electrode placement [3].

The stimulus related EEG can be monitored without the need of active attention of the subject where as taste perception needs active cognitive attention of the subject. In this work we show that taste stimulus elicited EEG signals can be observed at the channel locations AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7 and T8 for Emotiv's Epoc headset, by using time-frequency representation for stimuli of different pH values. We try to determine if emotion elicited by taste recall is stronger than stimulus elicited emotion. We also tried to evaluate if there is a clear difference in the frequency pattern for stimuli of different pH levels. Frequency pattern of taste recall of the stimuli will be re-evaluated to see if it is stronger than the actual stimuli [4]. We will also observe the taste perception of acidic and alkaline pH stimuli's in comparison with each other, to see if there is a difference in frequency pattern and if it is distinguishable.

II. MATERIALS AND EXPERIMENTS

A. Participants

5 right handed subjects between age 20 and 30 participated in the experiment. All the participants who were chosen did not have any past neurological and psychiatric disease. They were instructed not to eat or drink anything except water 3 hours prior to the experiment. The Inter Stimulus Interval (ISI) was a brief (1 hour), and the subject rinsed his/her mouth with water after every stimulus presentation. The cognitive secession was conducted separately after a 24 hour time interval.

B. Device specifications

The EEG headset consists of 14 gold plated contact sensors, which make contact with the detachable electrode tips. These electrode tips have foam which is soaked in saline solution and sits on the scalp adapting to the scalp topology such as shape of the scalp and hairs because of its cushioning effect, the saline solution acts as a good conductive medium for small voltage fluctuations giving an efficient contact area. These electrodes are aligned in

TABLE I. GUSTORY STIMULI USED FOR AFFECTIVE EXPERIMENT

Serial no.	***Stimuli***	***State***	***pH***
1	Citric acid	liquid	2
2	Sodium bicarbonate	liquid	9

10-20 electrode placement technique pre attached to the headset with arms fixed to point the electrode at the locations: AF3, F3, F7, FC5, T7, P7, O1, O2, P8, T8, FC6, F8, F4, AF4, Common Mode Sense (CMS) at the left mastoid M1 and Driven Right Leg (DRL) at right mastoid M2. The electrodes when green which were in our case represent good signal quality as in Fig. 1. The hardware high pass filter and digital signal processing hardware performs various filtering and processing steps before the EEG signals are wirelessly transmitted to a USB dongle which can be connected to a computer to view and save the EEG data.

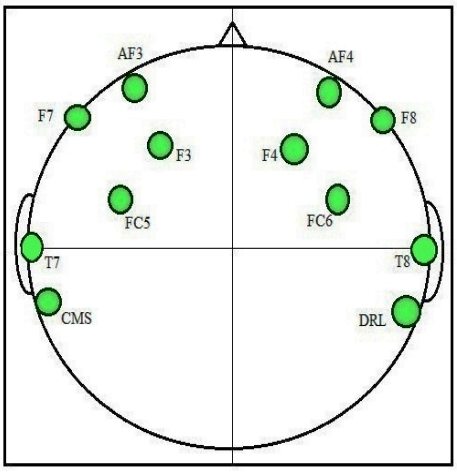


Figure 1. Scalp locations of the electrodes placed using Emotiv's gaming EEG headset

C. *Gustatory stimuli*

The taste stimuli used were 5% of citric acid extracted from lemon with pH 2; Sodium bicarbonate solution was made by dissolving baking soda in water, giving a solution of pH 9. Care was taken that harmless concentrations of standard consumables were used as the stimuli and quantity was maintained 10 ml.

D. *Experimental procedure*

The subjects were given detailed instruction about the experimental procedure and instructed to stay relaxed and with minimum physical movement throughout the experiment. The subject's hair was brushed properly to avoid external electrical interference. The wireless EEG headset from the Emotiv has been worn by the subject with electrodes placed according to 10-20 electrode placement system. The subject took 10 ml of the stimuli solution and the markers were given manually during the time when the stimuli were in subject's mouth.

After a brief interval of 24 hours the subject is instructed to observe the stimuli and remember the taste of the stimuli and markers were given manually.

E. *Data acquisition and analysis*

The EEG headset has M1 at the right mastoid which acts as ground reference point to which voltage of all other sensors is compared, and M2 at left mastoid reduces external electrical interference due to the DRL circuit design [5]. Signals from other electrodes were high pass filtered at 0.16 Hz cut-off, pre- amplified and low pass filtered at an 83 Hz cut-off. The analogue signals were digitized at 2048 Hz which is the fundamental sampling frequency of the device. The digitized signal is filtered and down sampled using a 5th order sinc notch filter (50-60 Hz) and low pass filtered. Finally the sampling frequency is close to 128 Hz, giving the effective bandwidth of 0.16-43 Hz. The above steps were pre-processing steps.

In order for the data to be suitable to our requirements we low pass filtered the data using a 30 Hz higher cut-off frequency and used a high pass filter of 5 Hz lower cut-off frequency. In the range of 3-4 Hz lie most of the eye blink artifacts which were eliminated during the high pass filtering. The signals were then epoched between -4 to 4 seconds at the marker and baseline corrected.

Data from all 10 channels have been processed with Complex Morlet wavelet transform where wavelet $\psi(t)$ is represented as in (1). It considers that high frequency data varies more rapidly in time domain. Hence we are able to differentiate between the high frequency bursts, which occur during the stimuli and other frequency variations [6].

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-i2\pi t} e^{-\frac{t^2}{2\sigma^2}} \qquad (1)$$

Where the width of the wavelet is $m = 2\pi\sigma$, σ being the parameter defining the number of oscillations.

Inter trail Phase Coherence (ITPC) measures the phase consistency over the epochs for the event [6]. Let $X(c, f, t, n)$ denote the time-frequency coefficient at channel c, frequency f, time t and epoch n of the EEG

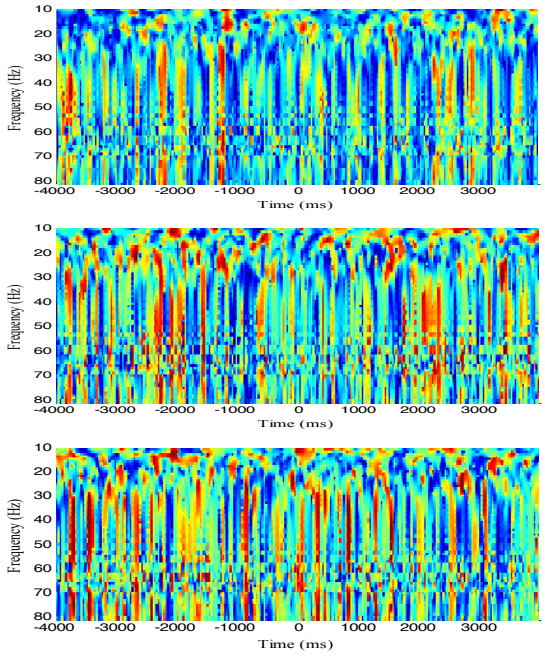


Figure 2. Time frequency representation for different stimuli (a) Normal state (b) Citric acid stimuli (c) Cognitive perception of the taste of Citric acid.

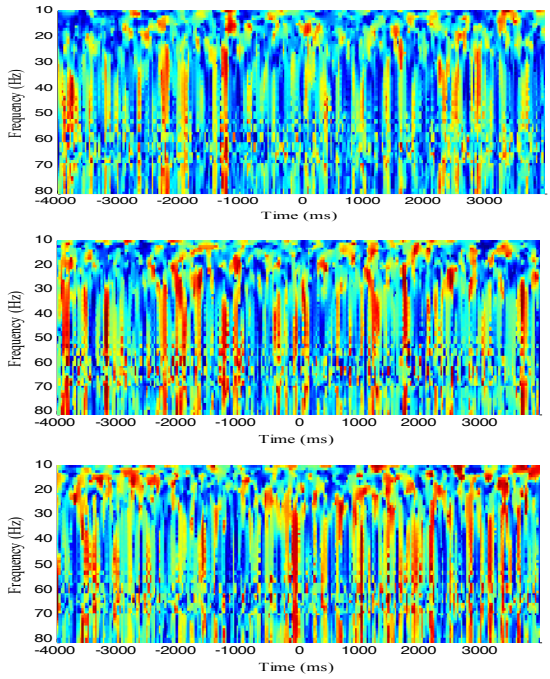


Figure 3. Time frequency representation for different stimuli (a) Normal state (b) Sodium bicarbonate stimuli (c) Cognitive perception of the taste of Sodium bicarbonate

signal represented in:

$$ITPC(c,f,t) = \frac{1}{N}\sum_{n}^{N}\frac{X(c,f,t,n)}{|X(c,f,t,n)|}. \qquad (2)$$

As all the epochs are given same weight for a particular stimulus, the measure of the phase consistency for different stimuli can be represented in time frequency spectrum individually, thereby allowing us to be able to compare them.

F. Classification results

The acidic and alkaline stimuli have shown a considerably greater amount of high frequency components in the time frequency spectrum than for the normal temporal activity. In Fig. 2 and Fig. 3 we show that acidic and alkaline solutions in time frequency domain changes throughout different states of the subject for one channel electrode. A very similar pattern of cognitive perception in comparison to actual stimuli or cognitive perception being more dominant than stimuli is visible. Also both have relatively higher frequency components than the normal temporal state. This is clearly visible at other channel locations considered for monitoring taste related EEG activity.

We can also observe that cognitive perception of acidic stimuli produces higher frequency components than cognitive perception of alkaline stimuli, which has not been discovered before as far as we know.

III. CONCLUSION

We have investigated EEG response to stimuli of acidic and alkaline pH levels, and were able to show with the time frequency spectrum that the taste perception of the stimuli produces more high frequency components than the actual stimuli. The cognitive perception of acidic stimuli produces relatively more high frequency components than cognitive perception of alkaline stimuli. By being able to conclude that Emotiv's Epoc headset can observe research quality EEG signals for taste related EEG patters, we show that taste stimulation and perception can be used as cognitive control tasks and be observed using a simple gaming headset. Our future work will aim towards implementing these cognitive perceptions into control tasks.

REFERENCES

[1] K. Ohla, N. A. Busch and J. N. Lundström, "Time for Taste: A review of the early cerebral processing of gustatory perception," Chemosencs Percept, Vol. 5, No. 1, pp. 87-99, 2012.

[2] M. Karami-Nogourani, R. Kowsari-Isfahan and M. Hosseini-Beheshti, "The effect of chewing gum's flavor on salivary flow rate and pH,"Dental Research Journal 2011;8(Suppl1):S71-S75.

[3] H. H. Jasper, "The ten twenty electrode system of the international federation," Electroencephalography and Clinical Neurophysiology, vol. 10, pp. 371-375, 1958.

[4] C.Park, D. Looney, and D.P. Mandic, "Estimating human response to taste using EEG," IEEE EMBS Annual international conference 2011.

[5] Winter, Bruce B.; Webster, J.G., "Driven-right-leg circuit design," Biomedical Engineering, IEEE Transactions on , vol.BME-30, no.1, pp.62,66, Jan. 1983

[6] Mørup, M., Hansen, L.K. and Arnfred, S.M. "ERPWAVELAB A toolbox for multi-channel analysis of time-frequency transformed event related potentials", Journal of neuroscience methods, Vol. 161, pp. 361-368, 2007.

People tracking using synthetically generated depth maps

L. Vretenar i K. Lenac
Faculty of Engineering – University of Rijeka, Rijeka, Croatia
lvretenar@riteh.hr, klenac@riteh.hr

Abstract - Modern 3D sensors using time-of-flight and structured light technology are increasingly being used in a number of interesting applications. One such application is people tracking where the available 3D information enables more robust human detection and tracking methods than those based on classic video cameras. A large data set is typically required both for testing and tuning of the solution.

We describe a methodology for efficient development of 3D applications using synthetically generated depth maps to complement the available data set obtained with real cameras. To demonstrate the effectiveness of the approach, a case study in the area of people counting with 3D sensors is presented.

I. INTRODUCTION

People detection and tracking is an important field of research. There are many applications in robotics, human computer interaction, video surveillance, home automation, pedestrian detection and counting in intelligent transport systems, just to name a few. Recent years have seen the advent and widespread diffusion of commercially available 3D cameras that greatly increased the interest of computer vision research in 3D vision.

In this paper we address the problem of robust people detection and tracking using depth map images from commercially available consumer grade 3D sensors. If depth data is available the task of detecting and tracking an object becomes much easier. In particular separation of background and foreground with objects that lack consistency in color and texture is not difficult any more. The same is true for the perception of the shape of an object. Another advantage over RGB images is that depth images are less sensitive to changes in color and illumination.

One of the key challenges of developing a people tracking algorithm is the availability of the dataset on which to base the development, testing and comparison with other approaches. This dataset must be most representative of the situations that may occur during real usage in target application. Ideally the developed algorithm is expected to properly handle nearly all possible situations, i.e. to detect and track people no matter the clothing, size, haircut, gender, walking style, followed paths and gestures performed. Moreover the environment may vary and the scene may include non-human things, like pets, suitcases and other objects. The problem is that with real recordings it is difficult to obtain such a dataset, and only a very small subset of possible situations is typically available instead. In this paper we address this and other major challenges with real data recordings for 3D people tracking applications by using synthetically generated depth maps.

The remainder of the paper is organized as follows: In Section 2, we briefly describe currently available depth sensors and the underlying technologies and then review recent works on people detection and tracking using depth maps. Section 3 proposes the usage of synthetically generated videos to address important challenges during development of people tracking solutions. In Section 4 the specific method for generating synthetic depth videos is described in detail. In Section 5 a people counting solution developed using the proposed methodology is described as case study. Conclusions are drawn in Section 6.

II. RELATED WORK

Traditionally, 2D computer-vision camera systems have been used to detect people, track them, count their passages through designated areas, estimate their pose and detect unusual or important events. Many different RGB image based methods have been developed typically using blob based tracking with or without additional classification schemes based on color, texture or other local image properties, and using 2D or full 3D models of human beings.

Early 3D range systems were costly and slow. They consisted either of systems which used a laser to scan the environment or systems based on stereo vision. A detailed review of stereo vision algorithms may be found in [1]. For real-time application in embedded systems consider reading [2] and for a recent overview of passive 3D imaging systems refer to [3]. However recently new types of 3D sensors became available using structured light and time-of-flight technologies (see Table I) and a number of new algorithms started exploiting the available depth information, either using only depth information or combining it with color information available from an RGB camera which is typically mounted alongside depth camera. In Table I we present a comparison and an overview of currently available 3D sensors and technologies and highlight important differences between them helping to select the best sensor for the problem.

In every people tracking solution it is important to be able to recognize and distinguish humans from non human objects in the scene. For this classification task many approaches use machine learning techniques applied on

The work presented in this paper is supported by the University of Rijeka research grant Grant 13.09.2.2.16.

processed depth maps or on derived point clouds. They rely on the ground plane assumption which assumes that people stand or walk on the ground. So first the ground plane is estimated and then classification is performed only inside of a small number of regions of interest which are likely to contain people. With a reduced search space a more robust classification can be performed without too much computation burden. The PCL library project [4] among numerous algorithms for processing point clouds contains a module providing a method for detecting upright people in RGB-D point clouds. Recently a project called OpenPTrack [5] was started extending the existing algorithms for real-time people detection and tracking from 3D data to a multi-camera scenario. Built on top of PCL and ROS [6] the software currently works with Microsoft Kinect v1 and Mesa Imaging Swissranger SR4500.

III. APPROACH

A. *Challenges with real data*

Developing a new solution to people detection and tracking using only live and/or recorded data with real people is challenging. The major difficulties that must be overcome are:

- lack of ground truth

Testing of the algorithms with real recordings is difficult because of the lack of an available ground truth. Recordings must be carefully inspected and labeled using another method (typically human visual inspection) to allow quantitative analysis and comparison with other algorithms. In particular machine learning techniques assume the training examples to be labeled. Preparing the data in itself is a very time consuming process. Moreover once the dataset is built the learner is typically run multiple times in an iterative process of training the algorithm, analyzing the results, modifying the data or the learner, and repeating the process.

- only a small sample of possible situations can be recorded

When preparing the recording setup one has to make decisions on camera type, position, field of view etc. After a recording setup is prepared for a specific problem, only a finite and relatively small number of recordings may be recorded. Recording is a time consuming process in itself and the process cannot last indefinitely. With respect to the above, the space of all possible situations that is of interest for the application is huge.

- non repeatability

With real people it is difficult to exactly repeat the recorded scene and behavior. For example it may be necessary or convenient to produce identical recordings but with different cameras, in different lighting conditions or in a different moment.

- hard to automate

Even with the availability of human actors that perform according to a predefined script it is difficult to generate the desired situation as often as required..

B. *Using synthetically generated data*

If we were able to synthetically generate the missing data in a programmatic way the above listed problems would be solved.

The usage of synthetic data is not a new idea. In machine vision, and people detection and tracking in particular, synthetic images have often been used to enlarge the training data on which classification algorithms are based. In this way it is possible to create arbitrarily large training datasets. For example pose recognition tasks alone require large datasets just to account for all possible human sizes, poses and shapes. A procedure which enables synthetic creation of camera images of humans in arbitrary poses may produce such a dataset. The most impressive usage of synthetically generated 3D depth images was in [7] where a large amount of synthetically generated images to train algorithms for super real-time 3D body pose estimation.

In our approach we further extend this idea by using synthetically generated data not just to train a classification algorithm, but also as a useful aid in several key stages of algorithm development. In fact synthetically generated data may also provide other important benefits to the researcher.

TABLE I. 3D SENSOR TECHNOLOGIES

Principle of operation	ACTIVE Projects electromagnetic (IR) radiation and analyzes the reflection of this radiation from the object.						PASSIVE Relies solely on ambient light; combines two 2D image sensors and uses the disparity in viewpoints to derive depth information.	
Important factors influencing the sensor choice	Distance range depends on illumination power. Very poor outdoor performance. Cannot simultaneously overlap the sensing area of two sensors of the same type.						Problems with smooth uniform surfaces. Poor performance in low light scenarios.	
Technology	Structured light		Time-of-Flight				Stereoscopic vision	
Typically used sensors								
	Microsoft Kinect v1	Primesense Carmine	Microsoft Kinect v2	Mesa SR4000	Creative Senz3D	Softkinetic DS311	Point Grey Bumblebee	custom designs

Gaining key insights during problem analysis: a researcher could identify and analyze problematic or otherwise interesting situations. Situations that would be difficult to produce with real people can easily be simulated.

Fine tuning of the algorithm: by systematically running simulations it becomes easy to fine tune various thresholds and algorithm parameters.

Testing the algorithm performance: the algorithm can be thoroughly tested across a multitude of situations. Tests can be automated. New datasets can be pro grammatically generated to test untested scenarios.

Comparing with other algorithms: With synthetically generated images, ground-truth camera positions and orientations at each frame in a video sequence are available. This allows explicit measurement of algorithm accuracy and comparison using the same data.

It is important for synthetically generated data to be realistic, i.e. the differences with real data should not interfere too much with the algorithm. This is indeed the case in people detection and tracking, which is an area of research where synthetic data is particularly beneficial. In fact many machine vision algorithms pre-process the data, filter the noise and apply some kind of foreground/background separation prior to detection stage. It is in these steps that differences between real and synthetic video are filtered away and the result is a video which is practically indistinguishable from video recorded with a 3D sensor with real people.

This is also true in our solution described in Section V. A comparison of pre-processing results for real and synthetic data is shown in Figure 1.

The downside of using a synthetically generated data is that it's generation may be a time consuming task.

From our experience building the framework and automating the tasks requires a substantial initial effort, however once all the tools are in place the generation of further datasets, when needed, is straightforward. This allows easy generation of new videos taken from different camera positions, or with changed people characteristics or paths they follow. New persons can be added, new types of sensors substituted in the exact position of current sensors (i.e. from structured light to time-of-flight sensor), additional cameras can be added to existing scene, and so on. The existing dataset can be extended in a later moment. The labeling of the ground truth is performed automatically in all cases.

Such synthetically generated scenes can simulate situations that would be difficult to recreate without simulations. Some examples of hard to recreate situations include environments that we don't have access to and scenes with very tall people that are not easy to find, people wearing strange clothes, hats, various bags, people walking in each other’s arms etc.

In the next section we describe in detail the framework and the tools that we used to generate synthetic datasets of depth recordings.

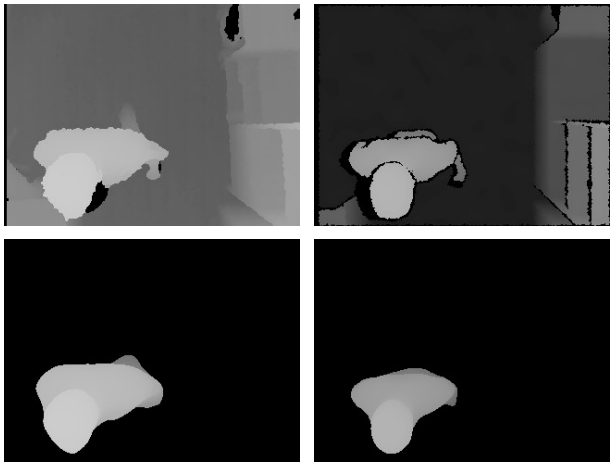


Figure 1. Comparison of real (left) and synthetic (right) depth recording of the same scene, with (bottom) and without pre-processing (top)

IV. Synthetic Depth Map Generation Framework

We evaluated a method for generating synthetic videos that would fit our problem domain easily and provide us with a work flow to simulate specific scenes and situations.

For this purpose we chose existing software solutions in the free/open source world for 3d modeling and sensor simulation.

Our method for generating synthetic depth images and video is as follows: 1. Model the scene we want to simulate; 2. Model the people in the scene; 3. Animate the people in the scene; 4. Position the sensors in the scene; 5. Generate depth images and video by simulating the sensors.

Central part of this work flow is the ability to model and simulate the depth sensors.

Among the listed tools for sensor simulation, one stands out for its focus on realistic sensor simulation and integration with professional modeling and animation tools. That is Blensor [9], an extension of the Blender project code and a plug in for the Blender modeling tool.

Our method centers on using Blensor as a tool to simulate the required 3d sensor images and videos, and as such extensively uses all Blender tools for modeling and animation.

Modeling of the scene, creating animations and positioning the sensors is all done inside Blender. In addition to using Blender, another tool was required to model people. For that we used a tool called MakeHuman [10] that can generate realistic models of humans in a format we can use inside Blender.

Modeling the scene: The scene is modeled inside Blender by using primitives such as planes, cubes and spheres. Those primitives are then deformed and subdivided to fit the shapes we desire. Such a model of a scene does not have to be very detailed. Since we are only interested in simulating the depth cameras we do not need the color information in a scene. That eliminates the need to add textures and color to the scene; in addition, depending on the depth resolution of a simulated sensor the geometry of the model does not need to be complex. Although the color is unnecessary for our simulation, we must still specify the reflectivity of the materials by setting the color intensity as Blensor uses the color intensity of an object to simulate the reflectivity of that object when simulating light rays from the sensor. In our simulations we used the default maximum reflectivity to simplify modeling and approximate the simulation.

Modeling the person: A model of a person is parametrically generated using the tool MakeHuman. The parameters that are of interest are the height and the weight of a person. The generated model is then exported in the native format of the tool and imported in the Blender scene using a Blender plug-in provided by MakeHuman. Such a imported model consists of a mesh and a skeletal rig. The skeletal rig is required for the animation and enables us to easily create our own poses and animations of the persons or map existing motion data to the rig.

Animating the scene: Animation of the scene consists of setting objects in specific poses by creating key-frames. Key-frames define the specific configuration of an object or the scene in a specific point in time. Animation is achieved by interpolation of poses and locations between two adjacent key-frames.

For complex animations, like human motion, existing databases of motion capture data exist. We chose to use CMU Graphics Lab Motion Capture Database obtained from mocap.cs.cmu.edu.

Positioning the sensors: Blensor uses existing camera objects in the scene to simulate the position and direction of the sensors. It is important to select the required sensor type inside Blensor and set the corresponding parameters for field of view, resolution and sensor specific functionality.

Generating depth images and video: For each sensor, Blensor can generate a depth map in Netpbm PGM format and a point cloud in PointCloud Library PCD format. Each scan consists of a ground truth scan, corresponding to the model from the sensors point of view, and a simulated scan containing simulated sensor noise dependent on the type and parameters of the sensor. If there is animation present in the model, such as the walking animation for a person, Blensor can generate a scan from a specified sensor for every frame of the animation.

V. Case Study: People Counting

A. System description

The purpose of the people counter application is detection of people in a scene and counting their direction of passage in the specified zones. The system consists of a depth camera positioned in a scene directly above a path we want to monitor and facing the floor, a networked embedded computer connected to the camera running the people counter algorithm and a networked desktop computer used to receive the counting data.

The depth camera we used was PrimeSense Carmine 1.08, a long range structured light sensor with a operating range of 0.8m – 3.5m, field of view of 57.5 x 45 degrees and resolution of 320 x 240 pixels. Embedded computer that is connected to the camera is an Raspberry PI, model B. The sensor provides depth images at the frame rate of 30 per second with each pixel representing a distance from the detected object in millimeters.

The people counter algorithm works by detecting people's heads and tracking their movement through the scene. This is a simplification possible by the mounting location of the sensor, which is mounted directly above peoples head. The advantage of this method is that an easy detection of a person is possible just by detection of a region that corresponds to the person's head. Visualization of the depth data, pre-processing and blob tracking are shown in Figure 3.

The following algorithm describes how the people counter application works:

```
for each depth frame:
  apply image filtering
  subtract background from the frame
  find blobs by finding all local minima
  match found blobs with previously found
    blobs
  check if a particular blob has crossed
    from one region to another
```

Image filtering: The raw depth frame is filtered with a median filter to eliminate the artifacts from the depth camera and down sampled to a 8 bit gray-scale image. Morphological operators like erode and dilate are applied to join regions with gaps do to occlusion of the sensor projector.

Background subtraction: A frame with an empty scene is selected as the base frame for the background subtraction. Since every pixel in the frame corresponds to the distance of the object to the camera, to subtract the background we only need to delete the pixels that are

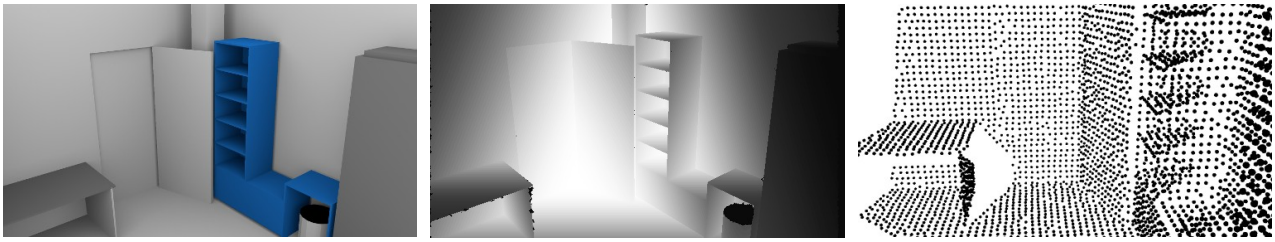


Figure 2. Scene model in Blender (left), depth map render (center), resulting point cloud (right)

close to the chosen background frame. All other pixels are closer to the camera then the background and are selected as detected objects.

Blob detection: Blobs are detected by finding all local minima points in the frame, which in ideal situations correspond with the top of a person's head. From those points a flood fill algorithm is performed to find all areas that correspond to a single persons head. Adjacent areas that did not contain local minima were discarded. All the found head areas are then marked as blobs to be tracked.

Blob tracking: To track the movement of people through time, new blobs corresponding to the heads of people are matched with the blobs found in the previous frame. The matching is done on the basis of two closest blobs within a minimum specified distance. If a new blob does not have a corresponding blob from the past frame, it eventually times out.

Passage counting: The scene is divided into three regions, corresponding to entry, exit and center. A tracked blob has to pass both region borders; the border between entry and middle, and the border between middle and exit; to be counted as an exit. The same logic applies in reverse to count an entry. The regions can be seen in Figure 3. Having the area be divided into three regions was necessary to eliminate possible flickering of a single blob around a single border, e.g. when a person stands in that place for more than a few frames.

B. Algorithm tuning

Parameters and tuning: Every step of the algorithm is parameterized in order to easily tune the algorithm to a specific environment and sensor.

Tuning is performed for following parameters: 1. Frame filter parameters; 2. Background subtraction threshold; 3. Local minima flood fill threshold; 4. Minimum and maximum blob size; 5. Minimum blob matching length; 6. Blob time-out duration; 7. Positions of region borders

Dataset: To get the best performance of the algorithm it was necessary to record a large dataset of depth videos and tune the parameters to the collected dataset.

A dataset of 38 depth video recordings was made using the long range and short range models of specified Carmine sensor, the recordings were taken of multiple scenes: various positions in the laboratory and of other rooms. The recordings were also taken on varying sensor heights from the floor and with different conditions: single person, multiple people, fast walking, waving hands, hugging together.

Additionally the laboratory was modeled as a scene and different scenarios were selected for simulation. Those scenarios consisted of a person walking through the scene in two directions, walking in circles and waiting, running through the scene and walking along the edge of the scene. Each scenario was modeled and animated using Blender, then rendered as a depth recording.

Since the real recorded dataset lacks ground truth measurements, the tuning was done by hand and was error prone. Creating synthetic video gives us an ability to generate large datasets with known ground truth, and with more control over the scene and the actions of people in the scene.

In the case of synthetically generated videos, the ground truth is known as the number of passages of a person through the scene. Additional information can be used as well, such as: the height of the person, the size of the head, the speed of the person, precise time the person enters the scene etc. Once the ground truth is known the tuning of the algorithm can be automated and only the situations in which the output of the algorithm is different from the ground truth can be analyzed.

C. Experimental results

Evaluation: Person detection is deemed successful if the algorithm's output, representing the counted number of passages through the scene for a selected synthetically generated depth recording, matches the known ground truth.

Additional scoring could be added by comparing the size of the persons head computed by the algorithm and known from the simulation. Other scoring metrics could include the tracking time of the person, proximity of the person to the border, and proximity to other people.

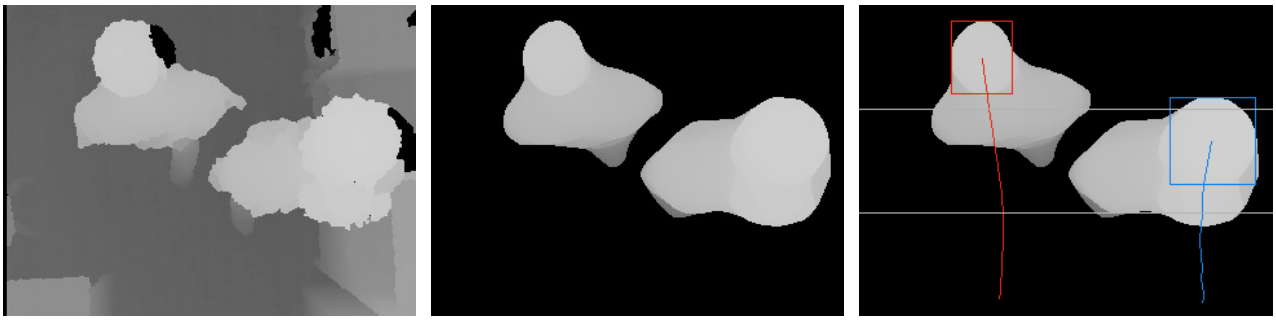


Figure 3. Raw depth frame from real sensor (left); pre-processed frame with background subtraction (center), head trackers and region separators (right)

Table II shows our results with the described scenarios when evaluated using the described passenger counter algorithm. Each scenario has a fixed number of passages of a single person through a scene. For each scenario we expect a certain number of passages to be detected, and we compare that value with the real number of detected passages by the algorithm. If the expected and real number of detected passages does not match, algorithm parameters are further tuned and the evaluation is re-run.

Qualitative results: The reader is invited to view the supplementary videos available at http://apaslab.riteh.hr/synthetic-depth-video/, or by mailing the authors.

Some successful detections and difficult scenarios are shown in Figure 4. The figure shows the detected person and its tracked path. In case of a successful detection and tracking, the rectangle represents a person's head, and the trailing line is the tracked path. Scenarios in which detection and tracking fail, such as a person running or walking along the edge help to further tune the algorithm until they are successfully tracked or detected.

TABLE II. Example run: simulated and detected passages through a scene

Scenario	Passages through scene in simulation (ground truth)	Passage counts by algorithm
Straight walk in a single direction	1	1
Straight walk, waving hands	1	1
Straight walk, loitering	1	1
Walk along the scene edges	1	0
Running in a single direction	1	0
Walk in a single direction, turn and walk back	2	2

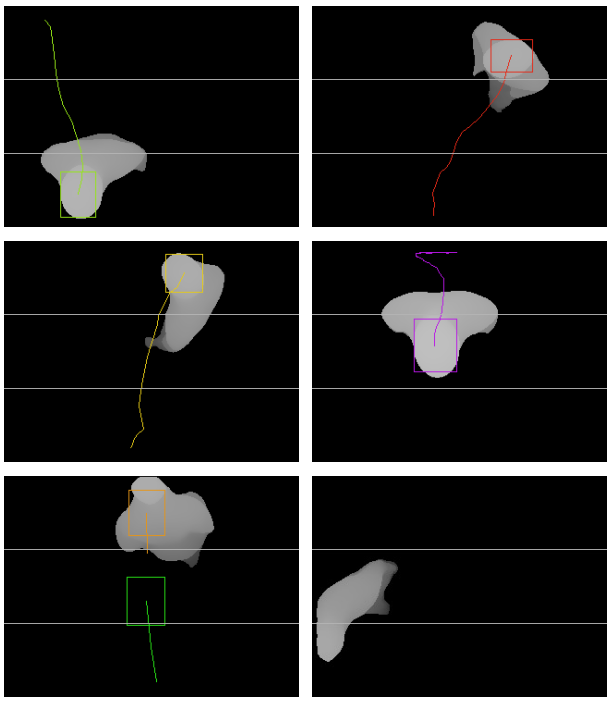


Figure 4. Results of applying people counter algorithm to synthetically generated depth data; successful tracking of a single person walking through the scene (top and center); person running (bottom left); person on the edge of sensor field of view (bottom right)

VI. Conclusions and Future Work

We have discussed the benefits of using synthetically generated depth maps in different key stages of the development of people tracking solutions that rely on 3D sensors. We have proposed and described a method and tools used for the generation of synthetic depth videos. The proposed method was used to develop a people counting solution. Its usage allowed for shorter development and testing times. We describe the algorithm and provide details of its implementation in a real embedded system. The solution was tested experimentally.

For future work we are planning to extend the described solution by implementing recent learning algorithms to improve the people detection from different camera positions. We are sure that the availability of large synthetically generated dataset and the capability to generate new data as needed will greatly benefit the development process.

Literature

[1] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. Journal Computer Vision, 47(1/2/3):7–42, 2002.

[2] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald. Review of stereo vision algorithms and their suitability for resource-limited systems. Journal of Real-Time Image Processing, 2013.

[3] S. Se and Nick Pears, Passive 3D Imaging, 3D Imaging Analysis and Applications, Springer ISBN: 978-1-4471-4062-7, 2012.

[4] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In International Conference on Robotics and Automation (ICRA) 2011, pages 1–4, Shanghai, China, May 9-13 2011.

[5] M. Munaro, A. Horn, R. Illum, J. Burke and R. B. Rusu. OpenPTrack: People Tracking for Heterogeneous Networks of Color-Depth Cameras. In IAS-13 Workshop Proceedings: 1st Intl. Workshop on 3D Robot Perception with Point Cloud Library, pp. 235-247, Padova, Italy, 2014.

[6] Morgan Quigley, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibs, Eric Berger, Rob Wheeler, and Andrew Ng. Ros: an open-source robot operating system. In International Conference on Robotics and Automation (ICRA), 2009.

[7] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient Human Pose Estimation from Single Depth Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 99, 2012.

[8] N. T. Siebel and S. Maybank, Fusion of Multiple Tracking Algorithm for Robust People Tracking, Proc. European Conf. on Computer Vision; LNCS2353, pp. 373-387, 2002.

[9] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "BlenSor: Blender Sensor Simulation Toolbox," in Advances in Visual Computing, G. Bebis, R. Boyle, B. Parvin, D. Koracin, S. Wang, K. Kyungnam, B. Benes, K. Moreland, C. Borst, S. DiVerdi, C. Yi-Jen, and J. Ming, Eds. Springer Berlin Heidelberg, 2011, pp. 199–208.

[10] M. Bastioni, S. Re, and S. Misra, "Ideas and Methods for Modeling 3D Human Figures: The Principal Algorithms Used by MakeHuman and Their Implementation in a New Approach to Parametric Modeling," in Proceedings of the 1st Bangalore Annual Compute Conference, New York, NY, USA, 2008, pp. 10:1–10:6.

Virtualizacija mrežnih funkcija u telekomunikacijama

Oliver Lovaković*

* Hrvatski Telekom/Sektor za razvoj sustava mreže i usluga, Zagreb, Hrvatska
oliver.lovakovic@t.ht.hr

Sažetak - Virtualizacija mrežnih funkcija (NFV, *Network Function Virtualization*) je novi koncept mrežne arhitekture u telekom industriji izgrađen na tehnologiji virtualizacije iz IT industrije s ciljem optimizacije troškova, bržeg razvoja te implementacije telekom usluga. Ovaj rad opisuje pogled na NFV od strane Hrvatskog Telekoma čija je osnovna arhitektura definirana DT projektom Terastream koji je prvotno implementiran upravo u HT-u. Također su prikazani status i planovi Hrvatskog Telekoma te preduvjeti koje je potrebno ostvariti za produkcijski rad u virtualiziranom mrežnom okružju. Posebni naglasak je na scenarij virtualizacije BRAS-a (*Broadband Remote Access Server*) kao posebno zanimljivog slučaja zbog specifičnih zahtjeva na kontrolnom i podatkovnom sloju za agregaciju xDSL korisnika. Virtualizacija vBRAS-a zbog toga predstavlja poseban izazov kako za proizvođače opreme tako i za telekom operatore.

I. UVOD

Situacija u telekom industriji zadnjih godina je takva da su telekomi suočeni sa kontinuiranim povećanjem IP podatkovnog prometa (i pratećih brzina) potaknuto širokopojasnim pristupom i novim Internet servisima te Internet poslovnim modelima. S poslovne strane postoje zahtjevi za kraćim vremenima uvođenja novih usluga te bitnim povećanjem broja usluga za lansiranje s mogućnošću ciljanja uskih korisničkih niša. Inzistira se na modelu koji bi omogućio mnogo pokušaja (i pogreški) pri lansiranju usluga, a koji s troškovne strane ne predstavlja bitan izdatak. S druge strane konkurencija u vidu OTT (*Over The Top*) igrača ili drugih telekoma negativno utječe na ostvarene prihode što rezultira da troškovi infrastrukture rastu brže od prihoda. Postojeće heterogene mreže i tehnologije teško se prilagođavaju gornjim zahtjevima i uz visoku cijenu. Povećanje troškova energije, navedena investicijska ograničenja, široki raspon ekspertiza potrebnih za upravljanje takvim kompleksnim mrežama upućuju na potrebu za promjenom postojećeg mrežnog koncepta u telekom industriji.

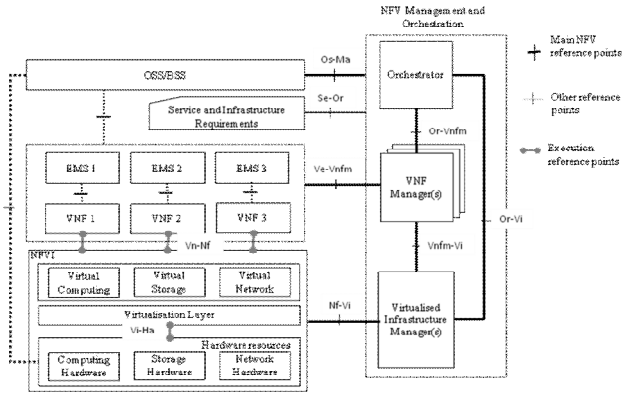
Sve to utjecalo je na razvoj novog koncepta mrežne arhitekture u telekom industriji - Virtualizacija mrežnih funkcija (NFV, *Network Function Virtualization*) izgrađenog pod okriljem ETSI-ja na tehnologiji virtualizacije iz IT industrije. Primarni cilj je povećanje efikasnosti kroz omogućavanje bržeg razvoja i implementacije telekom usluga uz optimizaciju troškova. Glavni pristup naslanja se na praćenje koncepta korištenja podatkovnih centara te računarstva iz oblaka (*Cloud Computing*) sa zadatkom da se isti proširi kako bi zadovoljio zahtjeve telekom industrije i to prvenstveno komunikacije u realnom vremenu (*Real-Time Communication*).

A. NFV organizacije

U tu svrhu je pri ETSI-ju početkom 2013. osnovana nova ISG (*Industry Specification Group*) grupa nazvana NFV (*Network Functions Virtualization*). Inicijativa je ubrzo privukla preko 200 kompanija (proizvođača i operatera) koji su doprinijeli u izradi osnovnih specifikacija NFV ekosustava u domeni arhitekture, zahtjeva, sigurnosti itd. Glavni moto grupe je korištenje standardne IT virtualizacijske tehnologije kako bi konsolidirali različite tipove opreme u standardne servere, preklopnike i medije za pohranu, koji mogu biti locirani u podatkovnim centrima, mrežnim čvorovima ili na lokaciji korisnika [1].

NFV arhitektura je prikazana na slici 1 .

NVF se u praksi često veže sa konceptom Programski upravljanih mreža (SDN, *Software Defined Networks*) kojem je cilj centralizirati i automatizirati upravljanje konfiguracijom mrežnih resursa. Dva su koncepta komplementarna prvenstveno iz razloga što SDN omogućuje dinamičko softversko mrežno povezivanje virtualiziranih mrežnih funkcija (*VNF, Virtual Network Function*) kreiranih u podatkovnim centrima sa korisnicima i ostalim funkcijskim blokovima potrebnih za realizaciju određene usluge. Dinamička priroda NFV-a se izvrsno nadopunjuje sa automatizmom i centraliziranim pristupom SDN orkestracije.



Slika 1. ETSI NFV arhitektura

NFV ekosustav se po ETSI NFV arhitekturi može podijeliti u slijedeće domene:

1. NFV Infrastruktura (NFVI)
2. Virtualizirane Mrežne Funkcije (VNF)
3. Upravljanje i orkestracija VNF aplikacija (NFV MANO, *Management and Orchestration*)

NFVI domena obuhvaća hardverske resurse i pripadajući virtualizacijski sloj koji definira na koji način se infrastruktura dijeli odnosno virtualizira.

VNF domena prikazuje softversku implementaciju mrežnih funkcija koja se izvršava na NFV infrastrukturi. NFV infrastruktura osigurava virtualne strojeve (VM, *Virtual Machines*) koji su zapravo virtualni kontejneri koje VNF aplikacije vide kao fizičke servere u kojima se izvršava njihov kod.

NFV MANO pokriva upravljanje životnim ciklusom i orkestraciju VNF aplikacija te hardverskih i softverskih resursa neophodnih za omogućavanje virtualizacije infrastrukture.

Upravljanje životnim ciklusom i orkestracija VNF aplikacija po ovom modelu je područje u kojem postoji znatan prostor za optimizaciju. Tako na primjer, ideja da svaka VNF ima svoje sustav za upravljanje (EMS, *Element Management System*) Postoji prostor za optimizaciju složene NFV arhitekture koja prepoznaje čak tri različita orkestracijska sloja – orkestraciju infrastrukture, aplikacija te sveobuhvatnu orkestraciju. Postoje inicijative u smjeru pronalaska zajedničkog načina za upravljanje VNF aplikacijama (npr. pomoću Netconf protokola uz prilagodbu Yang modela za svaku VNF aplikaciju) neovisno o dobavljaču softvera što bi omogućilo korištenje zajedničkog sustava za upravljanje konfiguracijom. Isto tako bi trebalo unificirati sučelje između VNF aplikacija i orkestracije istih kako bi pojednostavili i time unaprijedili robusnost arhitekture u orkestracijskom sloju (npr. također korištenjem Netconf protokola).

Glavni ciljevi i obećanja NFV inicijative su [1]:

- Smanjiti troškove opreme i energije (korištenjem komercijalno dostupnog hardvera, tzv. COTS, *Consumer of the Shelve*)
- Skratiti vrijeme razvoja i implementacije usluga (TTM, *Time To Market*)
- Omogućiti dijeljenje istih resursa za različite usluge i grupe korisnika te lako skaliranje resursa ovisno o trenutnom opterećenju
- Omogućiti dinamičniji ekosustav kroz razvoj i korištenje isključivo softverskih rješenja, tj. odvajanja razvoja softvera od hardvera (ulaze u igru mali inovativni igrači budući da nema potrebe za ulaganjem velikih sredstava u razvoj hardvera)

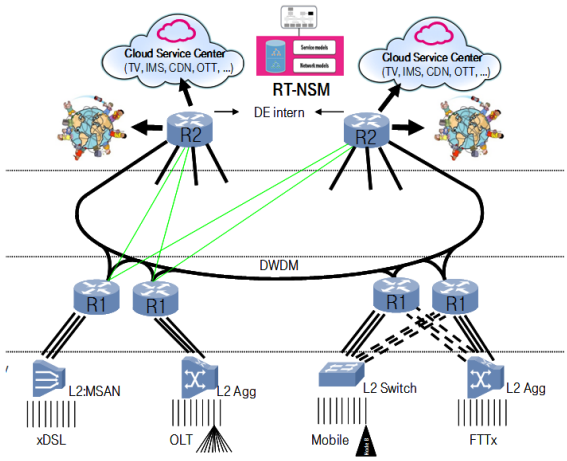
B. Status i planovi Hrvatskog Telekoma vezano uz NFV

Deutsche Telekom je jedan od osnivača NFV inicijative te kompanija koja snažno prati tehnološke inovacije. Hrvatski Telekom je kao dio DT grupe prepoznat kao telekom sa jakim tehničkim kompetencijama pa je kao takav izabran za pilot Terastream mrežnog koncepta. Terastream koncept je razvijen od strane DT-a s ciljem pojednostavnjenja mrežne arhitekture korištenjem slijedećih koncepata/tehnologija [2]:

- NFV (mrežne funkcije smještene su u posebnoj vrsti podatkovnih centara optimiziranoj za telekom potrebe - infrastrukturni podatkovni centri)
- SDN (popuno automatizirani sustav upravljanja mrežom i uslugama u realnom vremenu (RT NSM, *Real Time Network and Services Management*) upravlja životnim ciklusom kompletnog ekosustava)
- Pojednostavljena IP arhitektura (topologija s dva stupnja hijerarhije omogućava manji broj skokova između čvorova (*hop*) uz iznimno veliku širinu pojasa – 100 GE je osnova)

Glavni ciljevi Terastreama [2]:

- Smanjiti trošak transporta IP paketa zbog povećanja IP prometa pojednostavljenjem arhitekture
- Smanjiti broj tehnologija koje se koriste u prijenosu paketa (samo IP i optički prijenos)
- Integrirati IP i optičke mreže što je više moguće (integrirati OTN sučelja u rutere)
- Uvođenje IPv6 za sav interni promet (IPv4 je samo usluga koja se pruža korištenjem IPv6 infrastrukture L2TPv3 tunelima)
- Dimenzionirati mrežu da podržava sav IP promet bez gubitaka paketa



Slika 2. Terastream arhitektura

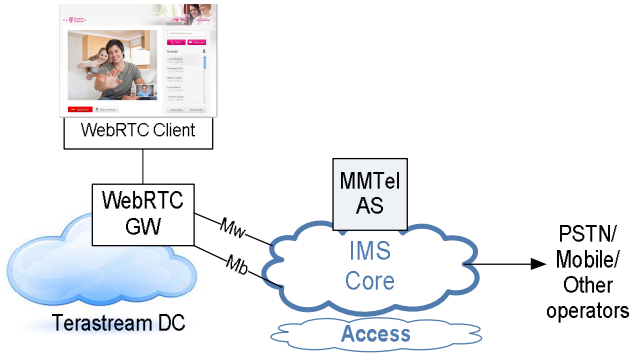
- Koristiti jednu mreže za sve usluge
- Potpuno automatizirani OSS sustavi (portal za upravljanje uslugom u realnom vremenu od strane korisnika…)
- Migracija mrežnih funkcionalnosti u DC (VNF aplikacije u infrastrukturnom podatkovnom centru)
- Iskoristiti *Open source* rješenja u najvećoj mogućoj mjeri (KVM, Openstack, OpenWRT…)

Testiranja u HT-u vezana uz virtualizaciju mrežnih funkcija vezana su uz Terastream pilote ili se odnose na testiranje VNF aplikacija instaliranih u Terastream podatkovnim centrima s ciljem dokazivanja ispravnosti koncepta (*Proof of Concept*). Posebno su izdvojeni slijedeći scenariji testirani u HT-u:

- Terastream pilot testiranja
- webRTC (*Real Time Communication*)
- virtualizacija IMS-a (*IP Mulitimedia Subsystem*)
- virtualizacija BRAS-a (*Broadband Remote Access Server*)

Terastream pilot je imao glavni cilj dokazati ispravnost koncepta za pružanje usluga rezidencijalnim korisnicima. Prva faza pilota (kraj 2012.) odnosila se na pružanje širokopojasne Internet usluge testnim korisnicima s ciljem testiranja VNF komponenti poput DHCP, DNS servera te AFTR (*Address Family Translation Resolution*). AFTR je bitna komponenta za uspostavu *Softwire* tunela (LW4o6, *Lightweigth IPv4 over IPv6*), neophodnog za prijenos IPv4 prometa preko IPv6 infrastrukture. Slijedeća faza pilota (kraj 2013.) ima za cilj testiranja pružanja 3play usluga u NFV Terastream ekosustavu.

WebRTC je tehnologija koja omogućava integraciju komunikacije u stvarnom vremenu sa *web* preglednicima. U Terastream podatkovni centar instalirali smo softver sa funkcijom webRTC *gateway*-a prema postojećem IMS sustavu s ciljem proširenja IMS usluga na web korištenjem scenarija „Učilice“. Implementacija i testiranje nisu prošli nezapaženo u telekom svijetu tako da je HT dobio nagradu Global Telekom Business za inovacije 2014. zajedno sa dobavljačem softvera – kompanijom Mavenir.



Slika 3. WebRTC shema

Virtualizacija IMS-a imala je za cilj pronaći rješenje za pružanje VoIP (*Voice Over IP*) usluge za fiksne korisnike spojene na Terastream mrežu. Cilj projekta je uz testiranje IMS funkcionalnosti i testiranje koncepta automatizacije i orkestracije same aplikacije.

Uz prethodno navedene scenarije HT provodi brojna testiranja NFV komponenti s glavnim ciljem istraživanja spremnosti pojedinih VNF aplikacija za produkcijski rad te istraživanja pouzdanosti i robusnosti Terastream arhitekture.

Terastream koncept se potvrdio u pilotima, ali ostaje još pokriti slijedeća područja kako bi se utvrdila stvarna spremnost za produkcijski rad:

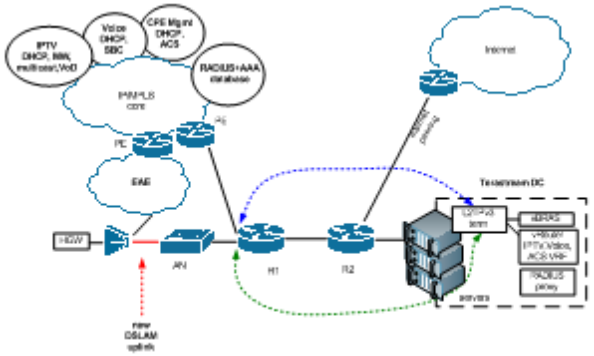
- Tehnološki (s ciljem istraživanja spremnosti dobavljača odnosno softvera za NFV)
- Operativni (s ciljem istraživanja kompleksnosti upravljanja životnim ciklusom NFV aplikacija te pratećih migracijskih planova)
- Ekonomski (troškovna usporedba scenarija migracije sa konvencionalnih mreža na NFV arhitekturu)
- Regulatorni (s naglaskom na veleprodajne modele)

Glavni ciljevi u slijedećem razdoblju vezani uz Terastream projekt te NFV implementaciju su dodavanje dodatnih VNF funkcija u podatkovni centar kako iste budu dostupne od strane dobavljača te podizanje stabilnosti i pouzdanosti svih komponenti Terastream infrastrukture (Infrastrukturni podatkovni centar, R1, R2…).

C. vBRAS kao poseban slučaj

Posebni naglasak stavljan je na scenarij virtualizacije BRAS-a potrebnom za agregaciju xDSL korisnika kao posebno zanimljivog slučaja zbog specifičnih zahtjeva na kontrolnom i podatkovnom sloju. Virtualizacija BRAS-a zbog toga predstavlja poseban izazov kako za proizvođače opreme tako i za telekom operatore. Glavni cilj je pronaći rješenje za migraciju xDSL pristupne opreme na Terastream mrežu. Ta oprema bazira se na PPPoE konceptu koji zahtjeva BRAS funkcionalnost. Scenarij je spojiti DSLAM uređaj na Terastream mrežu i osigurati BRAS funkcionalnosti u Terastream podatkovnom centru uz omogućavanje kontinuiteta usluge za korisnike sa starom CPE opremom (IAD, *Integrated Access Device*; STB, *Set top Box)*. Scenarij koji ne obuhvaća kompletnu zamjenu CPE opreme kod korisnika bitno pojednostavnjuje postupak migracije. Za IPTV, govornu uslugu te servis za upravljanje konfiguracijom CPE opremom koriste se servisne platforme smještene u konvencionalnoj mreži korištenjem vRouter VNF aplikacije koja se spaja na istu.

Sama Internet usluga se realizira preko vBRAS instance koja se izvršava u Terastream podatkovnom centru (osim Radius AAA komponente gdje koristimo



Slika 4. vBRAS scenarij

Radius Proxy server u podatkovnom centru koja šalje Radius poruke prema konvencionalnom Radius serveru).

Glavni arhitekturalni građevni blokovi rješenja su:

1. Terminator L2TPv3 tunela (terminira tunel od agregacijskog R1 rutera)

2. vBRAS (pruža iste funkcionalnosti kao i hardverski bazirani BRAS - PPPoE terminacija, AAA…)

3. vRouter (omogućava osiguravanje telekom VPN servisa – IPTV, VoIP, VPN za upravljanje CPE opremom)

4. Radius proxy (služi kao relej za Radius promet prema postojećoj centralnoj Radius bazi podataka)

Cilj testiranja bio je funkcionalno testirati kombinaciju rezidencijalnih usluga (3play) uz QoS sa kraja na kraj te testiranje performansi.

Zaključak je kako je samo rješenje zadovoljilo kriterije uz napomenu kako se ipak radi o kompleksnijem ekosustavu budući da je sama implementacija iziskivala prilagodbu (nadogradnju) operativnog sustava na samim serverima s ciljem poboljšanja performansi same vBRAS aplikacije (nadogradnja linux kernela). Isto upućuje na preveliku povezanost aplikacije sa infrastrukturom što se kosi sa osnovnim zahtjevom NFV-a. Procedure operativnog održavanja su također složenije budući da je samo snimanje prometa u svrhu detekcije potencijalnih problema bitno otežano činjenicom da vBRAS nema fizičkih sučelja na kojima je moguće snimati promet. U tu svrhu potrebno je redizajnirati koncept dizajna sustava za nadzor signalizacije i prilagoditi ga NFV okružju.

II. ZAKLJUČAK

NFV tehnologija doista pruža mnoge mogućnosti koje telekomi mogu pozitivno iskoristiti. Međutim, uz prethodno navedene prednosti postoje i brojni problemi koje treba riješiti ili barem umanjiti uzrokovani kompleksnošću ekosustava ili barem inkubacijskoj fazi NFV tehnologije [3]. HT-ova pozicija tehnološkog lidera u istraživanju NFV problematike otvorila je brojne nove probleme koje tek treba riješiti i kojih ostali nisu svjesni.

Tako implementacija podatkovnih centara odnosno pratećeg softvera (naglasak na Openstack instalacijama koje su prilagođene pojedinačnim potrebama operatera) postaje specifična pa je potrebna prilagodba raznim konceptima. Iako takva prilagodba optimizira ekosustav određenog operatora takav pristup otežava razvoj orkestracijskog i VNF softvera zbog bitnih razlika u infrastrukturnim postavkama od instalacije do instalacije.

Implementacija IPv6 protokola nije toliko raširena iako je standard već dugo vremena postoji. To je problem s kojim smo se susretali pri testiranju i istraživanju tržišta u potrazi za prikladnim VNF aplikacijama odnosno CPE opremom.

Nadzor prometa u oblaku je mnogo kompliciraniji nego u „starom“ svijetu fizičkih sondi. Promet između virtualnih strojeva (VM) na istom hardveru ne prelazi preko fizičkog sučelja tako da treba prilagoditi koncept nadzora prometa.

Open source softver demonstrira brojne prednosti u ovoj fazi implementacije. *Open source* zajednica se pokazala fleksibilnijom u razvoju NFV softvera od etabliranih dobavljača telekom rješenja. Oblak je za dobar dio tih proizvođača samo jedna nova platforma za instalaciju starog softvera uz manje prilagodbe. Nema posebne optimizacije softvera za oblak već se u pravilu softver kopira sa hardverske kartice sa određenom funkcijom (npr. signalizacijska) kopira u odgovarajući VM. Automatizacija i orkestracija te nadzor VNF aplikacija se nudi kroz proizvode/silose što se negativno odražava na cilj smanjenja ovisnosti o dobavljačima.

Operativni izazovi su bitno veći posebno zbog nejasne podjele odgovornosti u vritualiziranom okružju između organizacijskih cjelina i dobavljača (infrastruktura i aplikacija). Potreban je puno veći stupanj integracije IT i NT dijelova organizacije te poslova razvoja i operative. Isto tako u slijedećem razdoblju je potrebno staviti poseban naglasak na stabilnost i pouzdanost platforme koju tek treba verificirati na široj korisničkoj bazi i kroz dulje vremensko razdoblje.

LITERATURA

[1] ETSI-NFV, „Network Functions Virtualisation – Introductory White Paper“, October 2012, pp. 3

[2] P.Lothberg, „Terastream – A Simplified IP Network Service Delivery Mode”, October 2013

[3] O.Lovakovic, „Examining the ups and downs of SDN and NFV trials, November 2014.

C E

International Conference on

COMPUTERS IN EDUCATION

Steering Committee

Chairs: Marina Čičin-Šain, University of Rijeka, Croatia
Ivanka Sluganović, I. Tehnička škola Tesla, Zagreb, Croatia
Marko Čupić, University of Zagreb, Croatia
Jaak Henno, Tallinn University of Technology, Estonia
Hannu Jaakkola, Tampere University of Technology, Finland

Members: Michael E. Auer, Carinthia Tech Institute Villach, Austria
Snježana Babić, University of Rijeka, Croatia
Ivan Gerlič, University of Ljubljana, Slovenia
Marija Marinović, University of Rijeka, Croatia
Željka Požgaj, University of Zagreb, Croatia
Jadranka Šunde, Defence Science and Technology Organization, Australia
Slavomir Vukmirović, University of Rijeka, Croatia

PAPERS

The Relationship between Cultural Dimensions and the Quality of Communication in Education

B. Grobenski*,V. Vidacek-Hainš **
* Faculty of Organization and Informatics, Varaždin, Croatia
** Faculty of Organization and Informatics, Varaždin, Croatia
grobenski.bojana@gmail.com
vvidacek@foi.hr

Abstract - Culture is the collective programming of the mind distinguishing members of one group or category of people from others [1]. Each culture is different. Therefore, there are five dimensions of culture and this article explains and explores one of these: Power Distance (PD). This paper will outline the impact of these dimensions in educational institutions. In the study participated students from three countries (Croatia, Denmark, and Romania). The goal of this research is to determine how the students deal with a structured hierarchy and formal communication with a high level of Power Distance, rules and regulations. R**esearch shows what kind of communication and distance are in countries with low power distance (e.g. Denmark). A semantic differential instrument was used to process the survey results. We know that culture has an influence over education, i.e. Power Distance has an impact on the education of young people.** T**he paper also examines the influence of professors over students as well as the consequences of these impacts (level of formality of relations, the level of respect for student opinion, etc.). The possible models of improving the quality of communication and reducing the power of distance at the institutions of higher education will be explored. Keywords: Culture, Communication, Education, Hierarchy**

I. INTRODUCTION

Computers have had a big impact on interpersonal communication, particularly amongst young people and students, via computer-mediated communication (CMC). CMC can be also linked with the stress caused by intercultural differences [2].

Cultural differences do not disappear during technological modernisation. Students can develop international contacts and understanding of other cultures, nations and languages and therefore increase their intercultural competence.

In general, globalisation has presented the fact that modern society must learn to cooperate with each other [3]. The quality of communication in the international environment could be changed by using different information and communication tools. In spite of the fact that some implications of using ICT in education are positive, the negative implications of using ICT need to be considered [4].

The accessibility of ICT equipment is an important factor in enabling the quality of communication and is linked to the curricula of the higher education courses [5,6].

Educational attendance also relates to the digital development of countries. [7]. The combination of classical teaching methods supported by E-learning and computer mediated communication can be suitable for students and teachers in terms of their individual needs [8].

II. THEORETICAL FRAMEWORK

Culture represents a social pattern which includes behaviour in families, nations, organisations, etc. Also, culture includes tradition and behaviour in public and educational institutions and represents a way of creating ideas and views about other cultures and people from different cultures [9].

The number of projects about testing communication between speakers from different cultural backgrounds has grown in last few years.

Casrnir [10] says that communication should be approached with understanding and the realisation that people always change.

Because of that culture can not be seen as a static point. Traditionally, courses in intercultural communication that aimed to prepare university students to become effective intercultural communicators were designed to teach cognitive over affective learning, using lectures, readings and presentations [11].

Competence in intercultural communication has become a necessity for functioning effectively in our increasingly globalised and multicultural society.

A. Dimensions of Culture

The intercultural differences represent a challenge to all teams made up of individuals from different nations and cultures.

The literature review shows that several authors tried to operationalise culture through the use of multiple cultural dimensions.

The most famous and most widely used of these "models of cultural dimensions" was developed by Hofstede [12]. He described the five dimensions as follows: 1. Power-distance 2. Collectivism vs. individualism 3. Femininity vs. masculinity 4. Uncertainty avoidance 5. Long- vs. short-term orientation.

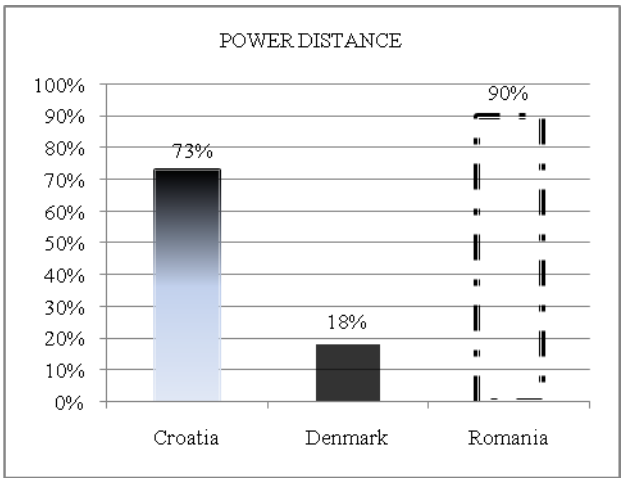
B. Power Distance

Here Power Distance (PD) will be explained. Power distance is the extent to which the less powerful members of organisations and institutions accept and expect that power is distributed unequally.

High PD cultures usually have centralised control. Low power distance implies greater equality and empowerment.

The cultures with high power distance retain a hierarchy, a decision-making structure, rules and regulations, and cultures with low power distance do just the opposite: they tend to balance [13].

According to Hofstede, the biggest power distance is in Romania, rank 2 is for Croatia and rank 3 for Denmark. Results are presented at Graph 1.



GRAPH 1: THE POWER DISTANCE FOR THREE EUROPEAN COUNTRIES

Source: http://geert-hofstede.com/croatia.html

II. RESEARCH

We know that culture has an influence over education, i.e. the dimension of power distance is in relation with the education of young people.

This research includes a representative number of students from Croatia, Romania and Denmark, in addition to analyses of the hierarchical division in the academic community, because when we observe PD all three countries are different.

A. Research goals

The research goal is to explore student opinion about structured hierarchy and power distance (PD) in communication in different countries (Denmark, Romania and Croatia). The case study of possible models of networking through informal ways of communication will be explored.

III. METHODOLOGY

A. Respondents

The first group of respondents were students of the Faculty of Organisation and Informatics (age 19-24, N= 109). In other groups were students from Romania (N = 43) and Denmark (N = 51). They represent the control group, i.e. a group which can be compared with the specified group tested by some empirical hypotheses [14].

B. Measuring Instruments

Author Grobenski B. developed a new questionnaire for this research. The items described power distance (PD) in a student's environment and the range of hierarchical divisions and rules.

The first four questions were demographic characteristics. The next three questions were designed as a scale for assessment of attitudes (from 1-5; e.g. 1- does not answer to e-mails, 5- answer the same day).

Questions from 8 to 11 are related to the student attitudes (answer Yes or No). The content of the items is related to the PD in communication.

The rest of the items are in the form of a semantic differential. The Semantic Differential (SD) is the psychometric tool in which bipolar scales are defined with contrasting adjectives at each end [15]. The respondent is asked to choose where his or her position is on the scale.

A questionnaire was implemented by a Google Docs tool in English and Croatian and sent to participants through closed groups on the Facebook social network. The survey was anonymous. The time required to fill out the surveys was 10 minutes.

IV. DATA ANALYSIS

The results for each item were analysed in terms of their basic statistic measurements: average /arithmetic mean (M) and standard deviation (SD).

Table 1 presents the answers to the Faculty's hierarchy: Does a hierarchical distribution of authority and responsibility exist in your academic community or college (Question 4), how long students do need to wait at the professors answers? (Question 5) how long do students need to wait at the assistants answers? (Question 6). These results indirectly measured the power distance in Croatia and other countries.

TABLE 1: POWER DISTANCE IN CROATIAN CULTURE

	N	Min	Max	Mean	Std. Deviation
4. Does a hierarchical distribution of authority and responsibility exist in your academic community or college?	109	1	5	2,14	1,28
5. Does it take a long time to get an answer from your professor?	109	1	5	3,66	0,81
6. Does it take a long time to get an answer from your assistant?	109	1	5	4,01	0,86
Valid N (list wise)	109				

The hierarchical division of powers and responsibilities (M = 2,14, SD = 1,28) shows that in the Croatian academic environment there exists a hierarchical division of powers and responsibilities: number 1 showing the existence of a hierarchy, and the number 5 showing a total absence of hierarchy.

A standard deviation was calculated for the scattering index. Deviation is less in the Question 5 (M = 3,66, SD = 0,81) and Question 6 (M = 4,01, SD = 0,86), which provides more relevant and accurate data.

This means that the time required for the teachers/assistants answers to students e-mail messages is "answers in a week" (3) and “answers in 2 to 3 days” (4). The results are acceptable for the Croatian academic community.

Table 2 shows the power distance in Romania. The results of the hierarchical division are shown in the Question 4 (M=1,93, SD=0,91). Considering that the arithmetical average is between 1 and 2, which means a very high level of hierarchy in the culture, i.e. a high level of power distance, it can be argued that the Romanian culture really stands out in its position in the academic environment.

Standard deviation shows a wide range of dispersion, i.e. a large deviation. Therefore the question remains about statistical significance.

When we observed Question 5 (M = 2,37, SD = 0,49) and Question 6 (M = 3,63, SD = 0,79), it can be concluded that the period of time to get an answer (e-mail) from teachers/assistants is in the mid-range (a month or a week).

In Danish culture (Table 3) the results are slightly different when compared with the results from the Croatian and Romanian cultures. Question 4 about the hierarchical division of powers and responsibilities (M = 3,53, SD = 0,64) shows a lower level of power distance. The arithmetical mean (M) is between 3 and 4, which means there are not less strong hierarchical divisions.

TABLE 2: POWER DISTANCE IN ROMANIAN CULTURE

	N	Min	Max	Mean	Std. Deviation
4. Does a hierarchical distribution of authority and responsibility exist in your academic community or college?	43	1	3	1,93	0,91
5. Does it take a long time to get an answer from your professor?	43	2	3	2,37	0,49
6. Does it take a long time to get an answer from your assistant?	43	2	4	3,63	0,79
Valid N (list wise)	43				

TABLE 3: POWER DISTANCE IN THE DANISH CULTURE

	N	Min	Max	Mean	Std. Deviation
4. Does a hierarchical distribution of authority and responsibility exist in your academic community or college?	51	2	4	3,53	0,64
5. Does it take a long time to get an answer from your professor?	51	3	5	4,24	0,65
6. Does it take a long time to get an answer from your assistant?	51	3	5	4,43	0,70
Valid N (list wise)	51				

The low level of power distance shows more equality and more balance in this culture. Answers to Question 5 (M=4,24, SD=0,65) and Question 6 (M=4,43, SD=0,70) showed that the period of time to get an answer (e-mail) from teachers/assistants is very short. The numbers 4 and 5 on the scale relate to a period of 2-3 days and an answer on the same day.

The results of the semantic differential in Croatia, Romania and Denmark are presented in Table 4. Croatian culture and the appropriate level of power distance is marked with red line, a blue line marks Romanian culture and a green line Danish culture.

TABLE 4: SEMANTIC DIFFERENTIAL FOR THREE EUROPEAN COUNTRIES

4	The existence of a hierarchy	Completely evident	1	2	3	4	5	Non-evident
5	The time of response from a professor	Did not reply	1	2	3	4	5	Replied the same day
6	The time of response from an assistant	Did not reply	1	2	3	4	5	Replied the same day

A. *Conclusions of the survey*

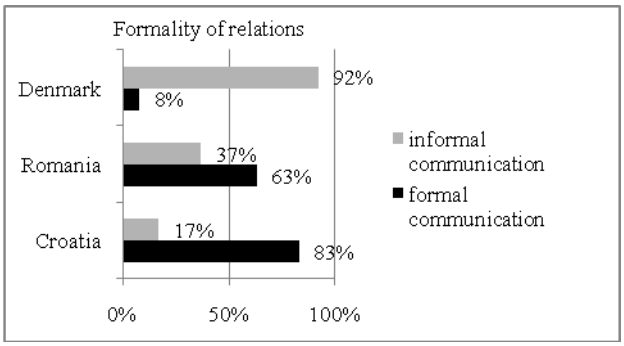
The research results show formality relations in the three analysed cultures and provide a clear display of formal or informal communication in the academic community of these particular cultures (Graph 2 and 3).

Croatian and Romanian students have slightly higher degree of formality comparing with Danish students (rank 3). Danish participants chose only a small proportion of formalities, and most of the answers speak of an informal and open communication.

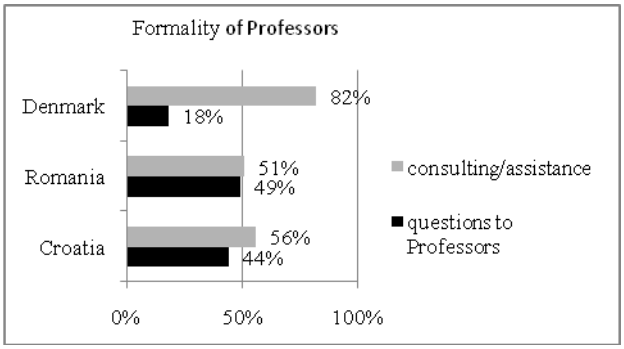
Therefore, the student can always ask for help and answers from professors or assistants. Croatian and Romanian students have more formal methods of communication with their professors and assistants. This can be connected with the final learning outcome of students working and learning.

Graph 4 shows how teachers and assistants in the relevant culture consider students opinions. Results show that level of openness in communication between students and professors is higher in Denmark (Rank 1).

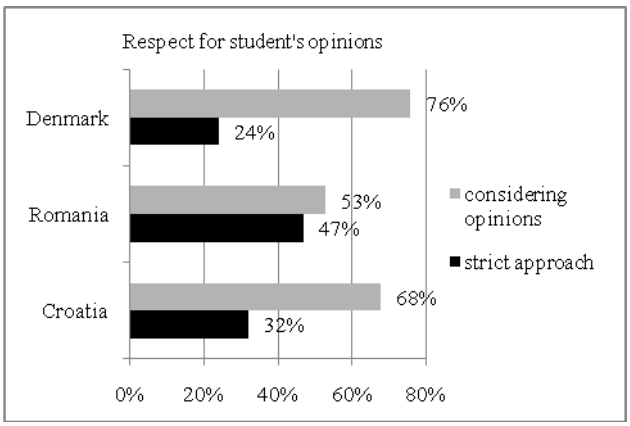
However, it is evident that in this part Croatian culture tends to be more similar with Danish culture than with Romanian culture.



GRAPH 2: THE LEVEL OF FORMALITY OF RELATIONS IN THREE EUROPEAN COUNTRIES



GRAPH 3: THE LEVEL OF FORMALITY OF PROFESSORS IN THREE EUROPEAN COUNTRIES



GRAPH 4: THE LEVEL OF RESPECT OF STUDENT OPINIONS IN THE ACADEMIC COMMUNITY IN THREE EUROPEAN COUNTRIES

V. IMPLEMENTATION OF RESULTS THROUGH CASE STUDIES

Possible ways of reducing the power distance in communication are through informal communication and networking. The possible networks between academic institutions and the general public can be volunteering projects, the popularisation of science, online student research symposiums, etc.

Practical examples showed that communication through volunteering project is less formal [16]. There is a lower power distance between teaching volunteers and student volunteers. The popularisation of science and communication with the public is an important way of informal communication between academic institutions and the general public. Scientific bars / SciBars are informal ways of presenting science in anon-academic environment, such as scientific cafes [17].

A student network in the international environment could be developed through international student online research symposiums. The main goal of this network is to encourage student creativity, ensure that students understand science better and develop their international communication competences [18]. Students work on a voluntary basis with their mentors. The power distance between students and mentors is lower and quality of communication is different.

VI. CONCLUSION

Foreign language, intercultural training and education programmes normally prescribe some manner of assessing participant performance / competence in a variety of academic and professional areas. It is clear that it is necessary to take into consideration cultural differences of participants and their different understanding. Results from this research show that Croatian and Romanian students have a much higher degree of formality compared to Danish students. This kind of relationship could be a barrier for developing openness in communication at academic institutions.

The quality of communication and the power distance at the academic institutions are different inside the academic curriculum, at projects linked to the local community and voluntary online communication. Case studies of volunteering projects organised by the students,

teachers and staff could reduce this distance in communication. Communication competences and social sensitivity could be developed thought different volunteering projects, online research symposiums and the popularisation of science. Further research will be focused on the compilation of case studies and examples of good practice at different countries

LITERATURE

[1] G. Hofstede, The Hofstede Centre, http://geert hofstede.com/ 2009

[2] L. Koskinen, K. Tossavainen, K., "Study abroad as a process of learning intercultural competence in nursing", International Journal of Nursing Practice, vol. 10, pp.111-120, 2004

[3] L. Samovar, R. Porter, and E. McDaniel, "Communication Between Cultures", Cengage Learning, Wadsworth, pp. 2-3, 2010

[4] H. M. Judi, N. Sahari, N. A. M. Zin, and Z. M. Yusof, "Framework of ICT Impact on Adolescent", Procedia Technology, vol. 11, pp. 1034-1040, 2013

[5] J. L. Fuertes, A. L. Gonzalez, and L. Martinez, "Including accessibility in higher education curricula for ICT", Procedia Computer Science, vol. 14, pp. 382-390, 2012

[6] K. Suzuki, H. Kondo, S. Yamaguchi, Y. Ohnishi, S. Tsumori, H. Wakana, and K. NishiNo, "Development of learning methods and system to practice problems using digital materials", Procedia Computer Science, vol. 3, pp. 1247-1254, 2014

[7] F. Cruz-Jesus, T. Oliveira, and F. Bacao, "Exploring the pattern between education attendance and digital development of countries", Procedia Technology, vol. 16, pp. 452-458, 2014

[8] S. B. Dias, J.A. Diniz, "Blended learning in higher education: different needs, different profiles", Procedia Computer Science, vol. 14, pp. 438-446, 2012

[9] R. Jeknić, "Cultures and organization cultures by Geert Hofstede", Zbornik radova Pravnog fakulteta u Splitu, vol. 1, pp. 103-123, 2010

[10] F. Casrnir, "Foundations for the study of intercultural communication based on a third-culture building model", International Journal of Intercultural Relations, vol. 23., pp. 91-116, 1999

[11] V. Milhouse, "Intercultural communication education and training goals, content, and methods", International Journal of Intercultural Relations, vol. 20, pp. 69-95, 1996

[12] G. Hofstede, "Culture and organizations", International Studies of Management and Organization, vol. 10, pp. 15-41, 1981

[13] A. Ergeneli, R. Gohar, and Z. Temirbekova, "Transformational leadership: Its relationship to culture value dimensions", International Journal of Intercultural Relations, vol. 31, pp. 703–724, 2007

[14] N. Larsen, D. Owen**,** "Denmark Business Culture Report" http://www.emita.org.uk/_downloads/Denmark%20Culture%20report.pdf , 2013

[15] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, "The measurement of meaning", University of Illinois, IL, USA, 1957

[16] V. Vidaček-Hainš, J. Bushati, A. Bencek, "The Usage of Information and Communication Technology in Promoting the Concept of Volunteering", International Journal of Ecosystems and Ecology Sciences, vol. 4, pp. 499-504, 2014

[17] B. Divjak, V. Vidaček-Hainš, and R. Horvatek, "Popularization of science and raising aspirations in the community", The 15th EAN Annual Conference "The Social Role of Universities: Reaching out to the Community", Thessaloniki, GR, 2006

[18] M. Aldarondo-Jeffries, M. Bowden, and V. Vidacek-Hains, "Enhancing Information Fluency in Undergraduate Students and Faculty through Study Abroad and Transcultural Communication", The Global Education Collaborative- on line Conference / computer tool Blackboard Collaborate, Orlando, FL. November 2011

Collaboration, Distribution and Culture – Challenges for Communication

Hannu Jaakkola*, Jaak Henno**, Bernhard Thalheim***, Jukka Mäkelä****
* Tampere University of Technology/Pori Department, Pori, Finland
** Tallinn University of Technology/Information Technology, Tallinn, Estonia
*** Christian-Albrechts-University Kiel, Computer Science Institute, Kiel, Germany
**** University of Lapland/Method Sciences, Rovaniemi, Finland
hannu.jaakkola@tut.fi; jaak@cc.ttu.ee; jukka.makela@ulapland.fi; thalheim@is.informatik.uni-kiel.de

Abstract - Work, to an increasing amount, is based on collaboration between different partners; collaboration emphasizes the importance of communication between the collaborating parties. Increasingly, work is also becoming distributed and carried out in different geographical locations; distribution underlines the importance of managing and organizing work. The third important aspect characterizing current work is globalization; this refers to the multicultural characteristics of work and the need to understand the behavioral patterns of different (national) cultures. The paper addresses these three challenges related to the current work context. The approach points out the importance of a deep understanding of the characteristics of work – collaboration, distribution, and cultural diversity. Adaptive learning provides one potential solution to the challenges.

I. INTRODUCTION

This paper is focused on the changing role of collaboration. Our experience lies in information systems (IS) development and software engineering (SE) work. However, the analysis is easy to generalize and can be applied in different situations, which expect collaboration between parties separated in some way. IS development is a typical sector that demands close collaboration between different interest groups and roles of developers. In spite of keeping our discussion in common level, we keep IS development as our application context.

In this paper we will not go into detail regarding complex process models and the role of communication in them. We will try to simplify the discussion to follow the evolution of IS development during the decades as follows. *The era of simple systems*: The first information systems were simple "one man" projects and communication needs were minimal, limited to the interaction between the developer and the client. *The era of large systems*: Growth of ISs forced the division of development work and sharing responsibility for the large system between several developers. Additional communication was needed to supervise the work of the development team. *The era of complex systems*: The growth of system complexity led to the situation where the work must be divided hierarchically first between teams and then inside teams between individuals. In practice this leads to a growing complexity of communication patterns. *The era of systems of systems*: Complexity has grown further and part of the development activity originates from such sources that are not under centralized (one's own) management. From the communication point of view, the complexity grows because of the unmanaged parts of the whole. The reasons for this are manifold: the tendency towards reuse, integration of individual ISs to the Internet using open interfaces, etc.

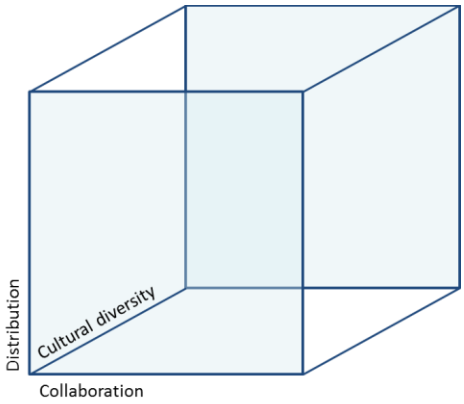


Figure 1. Communication complexity cube

The discussion above is slightly outside the scope of our paper, but included in this introductory part to clarify the progress that has led to the current state of the art. As mentioned earlier, in spite of using IS development (because it comes from our research field) as a case example, similar phenomena and progress are easy to find in many other sectors of society and business. Nor is the aim of our paper to focus on the communication patterns themselves – or only briefly, but to analyze three complexity factors related to communication. These are illustrated in Figure 1 in the form of a *communication complexity cube*.

The cube illustrates the space where (all) communication takes place (in any of the above discussed evolution phases and communication types). The three factors that interest us cover the degree of collaboration (x-axis), the degree of the distribution of work (in separate locations), and cultural diversity (multicultural teams). If we link this approach to the earlier evolution path, it is clear that when collaboration (amount of communication) increases, the consequence in most cases is first increasing distribution (transferring the work to such locations where (skilled) resources are available) and finally globalization (transferring work outside the geographical borders of the original location) is the response either to the still growing demand for resources or to benefit from the cost-effective use of globally available resources.

The communication complexity cube is used to structure our paper. The paper combines ideas derived from our cross-cultural collaboration related studies; results of this work are available in several earlier publications of the authors – some are also referred to in this paper. Section II provides a view to the typical aspects of IS development and software engineering work (our application context). Section III concentrates on analyzing the important aspects related to communication. An answer to the question "what actually is communication?" is sought in the collaboration context. Section IV gives a short overview of distribution ontology; this is a slightly more complex problem than just increasing the number of collaborating units, because part of it relates to how tightly they are bound together. Quite often distribution progress is also bound to the changes in the cultural base of the organization. Section V provides a short introduction to cultural differences and the means to manage the problems related to them. Section VI collects "tools" that are used to solve the problems encountered in the earlier sections. Section VII summarizes the paper.

II. ENGINEERING WORK – A PROCESS ORIENTED VIEW

The (very) simplified and generalized engineering process life cycle covers the following steps and communication needs:

- *requirements (elicitation, analysis and specification)*: communication between developers (system analysts), clients and other roles interested in the IS under development; the result is the answer to the question "what should the IS do?" and covers the functionality of the IS and related non-functional requirements (quality aspects);
- *design*: communication between different developers, in addition to the communication between the developing and client organizations; this step of engineering work transforms the requirements into the design that fills the architecture expectations of the system under development and provides the answer to the question "how should the IS be implemented?". The work covers two important tasks: (1) structuring the whole in architectural parts (top-down partitioning, covering both specification of components and interaction between them, with the focus on functionality) and (2) transformation of non-functional requirements into architecture, functionality or development process artefacts (goals);
- *implementation*: communication within the development team and in sufficient amount between the developer and client organizations; this step implements the IS requirements according to the defined architecture (principles) and results the implemented IS components; the focus is on the appropriate division of work and communication related to it; the role of client-developer communication is minimal;
- *integration (test)*: communication between developers having responsibility for separately developed IS components to guarantee their integrated operability; this step turns the development paradigm from top-down to bottom-up and puts the focus on actions guaranteeing the interaction between components; the role of client-developer communication is minimal;
- *system (test and) deployment*: the importance of client-developer communication is in focus again (as at the beginning of the engineering life cycle), where the client side is represented by different interest groups (and systems that interact with the system under development); after this the IS is deployed, the development part of the life cycle ends, and the *maintenance* phase starts;
- *(utilization and) maintenance*: this part of the IS life cycle, from the communication point of view, shares the features of the requirement step in an iterative manner and includes tasks related to change management in addition to the development tasks.

It is worth noticing that the engineering life cycle is just a part of a more comprehensive *process model* covering a wide variety of (grouped) processes. More details of these are available e.g. in [1;2] from the process structure point of view and in [3;4] from the process capability and organization maturity point of view. This material provides the means for a still more complex classification of communication needs as a part of IS development.

The implementation part of the engineering life cycle in practice may be done in several ways. Larry Boehm's conference paper [5] compares the traditional plan-driven development approach (discussed above) to the Agile approach. The Agile approach is based on the principles of the Agile manifesto [6]: "Individuals and interactions over processes and tools, Working software over comprehensive documentation, Customer collaboration over contract negotiation, Responding to change over following a plan". The traditional "plan-driven" approach points out the importance of following predefined plans, including the preplanned communication structure as part of it. The Agile approach, instead, points out the need for on-demand communication and flexibility, which means the growing importance of interaction between and within the interest groups and developer. The Agile approach on the organizational level is introduced by Andrew McAfee in his Organization 2.0 concept. In his book and www repository [7;8], McAfee introduces a novel communication and collaboration oriented organization culture, and compares it to the traditional approach called Organization 1.0 (as an analogy to the www context).

III. COLLABORATION AND COMMUNICATION

Jaakkola & Thalheim considered the characteristics of collaboration in connection with adaptive systems in [9]; Figure 2 summarizes the idea.

Collaboration - in connection with information systems development – is based on information processing by humans (H) and computers (C), the transfer of information between them and the transformations needed in the transfers. The nodes indicate individual sources of communication and the arrows between the nodes represent the communication (having direction).

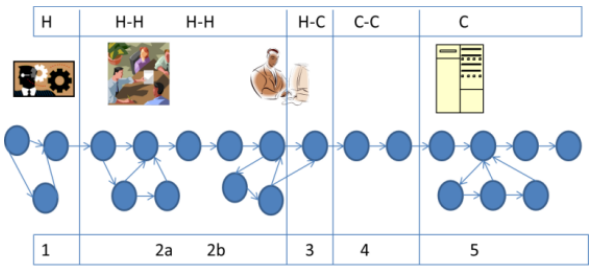


Figure 2. Characteristics of communication

According to Figure 2, information processing is done by humans (1) and computers (5). Individuals transfer their understanding either to other humans (2) or to computers (3). Teamwork is used in collaboration between humans (2a) and computers also collaborate in complex problem solving by dividing the processing responsibilities (4, 5). Typical processes include human and computer work, which requires the transfer of tasks from human to computer (systems) or vice versa.

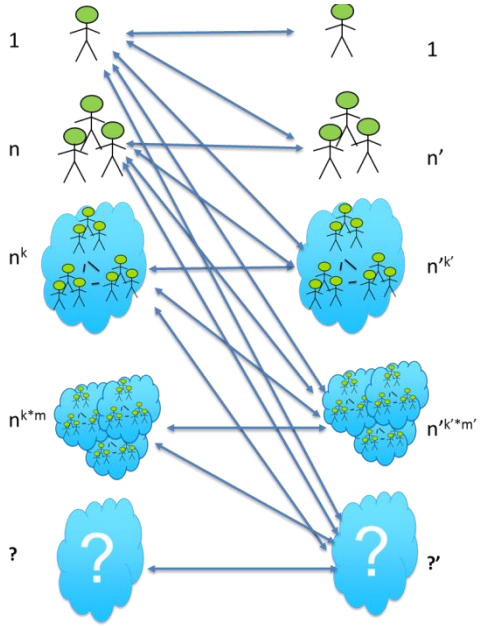


Figure 3. Degrees of collaboration complexity

Figure 3 illustrates different collaboration situations. The elements cover different collaboration parties (an individual, team, collaborative teams (in the cloud), collaboration between collaborative teams (cloud of clouds) and an unknown collaboration party (question mark cloud). The collaboration situations are marked with bidirectional arrows. The figure is self-explanatory: the complexity degree of the collaboration of one party increases in power when its complexity grows, and the complexity between parties is a multiplication of parties (1*1; 1*n'; $n^{k}*n'^{k'*m'}$). However, we agree that the idea described above does not fit reality: teamwork is organized (follows a pre-defined *teamwork process*), as is the collaboration between teams (follows a pre-defined *collaboration process*). These plan-driven (term adopted from [5]) work practices simplify the complexity and make it manageable (see [3; 4]).

The "question mark cloud" has not yet been explained. It covers the situation in which the collaboration partner is unknown; as a result this represents the most complicated situation – we have no opportunity to adapt ourselves to and learn the behavior of faceless partners. In software systems development, this situation is common when the system under development must be integrated to collaborate with existing common services; e.g., the utilization of open data as part of the information base provides in most cases only an API and interface specification for the developer without any opportunity to communicate with the data owner (see. e.g. [10; 11]).

Communication is an essential part of collaboration; the *information flow* between the collaborating parties binds the parties to work together. In [12], the essential communication elements are specified to cover: (1) the purpose of the communication; (2) the message transferred from the source (sender) to the target (receiver); (3) the interpretation of the message (meaning, understanding); (4) the structure of the message (structure of concepts and structural patterns); (5) the channel used for transformation (medium, connection) and (6) the context of the communication. Interpretation of the message is seen to be context-sensitive (i.e., in different contexts the interpretation may vary). The transfer of information (*message*) needs a *medium* to store the content and a *language* used for representation. The message itself is a *construction* of concepts. The conceptual *model* represents the structure of concepts from an individual collaborator's point of view. Figure 4 illustrates the role of the conceptual model in communication (for more details see [12]).

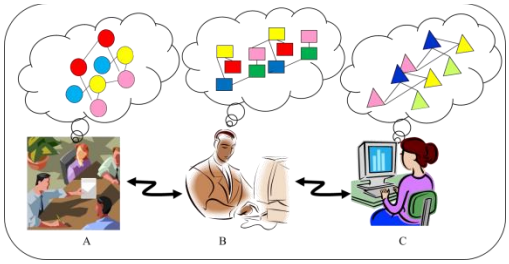


Figure 4. The role of conceptual models in communication

Every individual builds his/her own *private conceptual model* of the target. This is based on several personal properties; in this paper we especially want to emphasize the importance of (national) culture. In information flow the personal conceptual model is transformed (encoded) to the form accepted by the transfer medium (by the sender) and further transformed (decoded) to the conceptual model of the receiver. Figure 4 represents the differences of personal conceptual models using different shapes and structures. Our hypothesis is that one important source of misunderstanding and problems in collaboration has its roots in communication and especially in an inability to interact with conceptual models; transformations are not able to supply sufficient and exact content of the information (knowledge).

The discussion above contains somewhat philosophical aspects – which brings us slightly out of our

main competence areas. Therefore we will try to summarize shortly the main findings and link them to the next sections of this paper. The main point is the complexity of collaboration; this becomes an important issue in the organized distribution of work, which is handled in Section IV. Globalization (multicultural collaboration) increases the complexity further, as pointed out in Figure 1. Multicultural issues are discussed in Section V. Collaboration is represented as a set of information flows between collaborating partners (Figures 2 and 3). The transformations included in the flows need the adoption of a *common language* used to transfer the sender's private conceptual model to the intermediate model (encoding) for transfer and further interpretation of the intermediate model by the receiver (decoding). These languages as a solution to the co communication problem are discussed in Section VI.

IV. DISTRIBUTION OF WORK

Distribution of work is handled above (Figure 3) on a logical level. In practice, it is a question of organizational issues: an organization distributes the work based on either collaboration or on organizational needs. In both cases it is based on the strategic decisions of organizations and there are different motivation factors in the background: availability of resources, price of resources (work), location near important customers, extension of market, closer location to the sources of innovations.

Table 1. Organizational Distribution Taxonomy [13, p. 69].

	Intraorganizational		Interorganizational		Outside	
	One activity	Several activities	One activity	Several activities	One activity	Several activities
National	One site **(N1)**	Multiple sites **(N2)**	Out-sourcing **(N3)** Right-sourcing	Multi-sourcing **(N4)** Co-sourcing	Broker out-sourcing **(N5)** Right-sourcing	Broker multi-sourcing **(N6)** Broker co-sourcing
International	Off-shoring **(I1)** Near-shoring	Off-shoring **(I2)** Near-shoring	Offshore out-sourcing / near-sourcing **(I3)** Offshore right-sourcing	Offshore multi-sourcing **(I4)** Offshore co-sourcing	Broker offshore out-sourcing **(I5)** Broker offshore right-sourcing	Broker offshore multi-sourcing **(I6)** Broker offshore co-sourcing
	Insource		Outsource			

Distribution of work on the organizational level is always motivated by different business-related factors. As pointed out in Figure 1, distribution of work creates a need for increasing communication and increases the complexity of managing the work. In one-site distribution it is a question of teamwork; communication inside a team is easy and problems related to the division of work are solved by technical tools (e.g. in IS development by using version and configuration management tools, Integrated Development Environments – IDE; in a more general sense distributed work spaces, like Google Documents, Microsoft Office Online or by using different groupware platforms, like BSCW).

The taxonomy in Table 1 [13, p. 69] is divided into two rows (national and international (globalized) distribution), and six columns representing different types of distribution; the columns are grouped in *insourcing* and *outsourcing* (distribution inside and outside of an organization). Globalization (distribution in the international dimension) emphasizes a focus on the cultural differences between the global sites, which emphasizes the problems of cultural differences between the sites – multicultural teamwork. However, it is worth noticing that a multicultural work environment may also exist in one site and national organization-level collaboration in the case where employees represent different national cultures (which is becoming more and more common). The difference, however, is that on international level it is a question of *organizational* and *managerial* issues, on national level only *managerial* issues.

The organizational distribution categories in Table 1 indicate the following main types of distribution: *outsourcing* - transfer of work related to one activity outside one's own organization to an external contractor; *rightsourcing* - transfer of activities related to one activity to several external contractors; *multisourcing* - several activities outsourced to several contractors; *cosourcing* – several organizations have a permanent collaboration network to implement several related activities; the *broker* role indicates the situation in which the work is outsourced to independent external contractors that have loose connections to the "activity owner".

The aim of this section is to point out that distribution of work is no trivial matter but has several dimensions to be taken into account. The multicultural issues related to it are discussed in the next section of this paper. The author(s) have dealt with organizational issues in earlier papers ([14; 15; 16]).

V. CULTURAL ASPECTS

The culture-dependent features of human behavior are based on a layered structure: the lowest (innermost) layer represents the characteristics of human nature, the next culture, and the third (outermost) personality (Hofstede's pyramid model [17, p. 4-7]). National culture related practices are based on culture-dependent values, rituals, heroes, and symbols (onion model). The *cultural stereotype* is based on modeling cultural differences – typical features according to which people in different cultures differ from each other. There are several studies related to this topic; the most commonly used and widely known are the models of Hofstede [17] and Lewis [18]

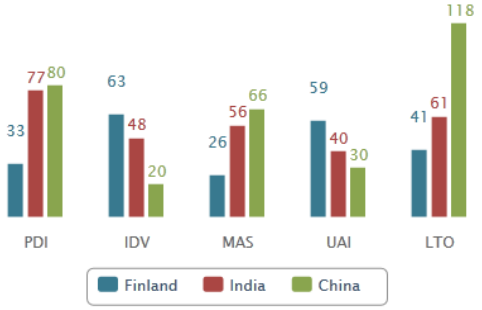


Figure 5. Stereotype analysis of Finland, India, and China according to the Hofstede model

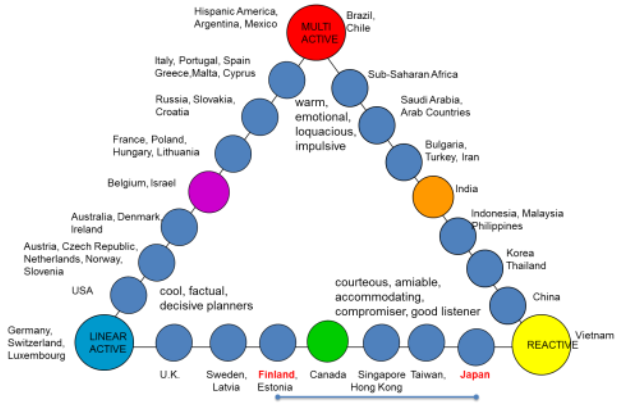


Figure 6. Cultural stereotypes according to Lewis' model.

The model of *Hofstede* [17] uses six dimensions (indices) providing a means to analyze behavioral differences between cultures:

- Power Distance (PDI): the extent to which power differences are accepted
- Individualism / Collectivism (IDV): the extent to which a society emphasizes the individual or the group
- Masculinity / Femininity (MAS): the general values in the society – hard / soft values
- Uncertainty avoidance (UAI): the extent that individuals in a culture are comfortable (or uncomfortable) with unstructured situations
- Long-term / Short-term orientation (LTO): the extent to which the delayed gratification of material, social, and emotional needs are accepted
- Indulgence / Restraint (IVR): acceptance of enjoying life and having fun vs. controlling life by strict social norms. This value is new and available only for few countries.

Figure 5 illustrates the cultural differences between three cultures: Finland, India, and China. It indicates the main sources of cultural collisions between Finland and other cultures in all dimensions and some similarity between China and India.

Table 2. Characteristics of cultural stereotypes according to Lewis' model

Linear-active	Multi-Active	Reactive
introvert	extrovert	introvert
patient	impatient	patient
plans ahead	plans grand outline	looks at general principles
does one thing at a time	does several things at once	reacts
punctual	not punctual	punctual
compartmentalizes activities	one activity influences another	sees whole picture
sticks to plans	changes plans	makes slight changes
sticks to facts	juggles facts	statements are promises
information from official sources	prefers oral information	information from official and oral sources
follows correct procedures	pulls strings	networks
completes action changes	completes human transactions	reacts to partners
likes fixed agendas	interrelates everything	thoughtful
uses memoranda	rarely writes memos	plans slowly
dislikes losing face	has ready excuses	must not lose face

The Lewis model ([18], Figure 6) defines three basic stereotype cultures (linear-active, multi-active, reactive) representing some of the cultures on the corners of a triangle and locates other cultures on the edges; location on the edge indicates the power of the influence of two of the basic stereotypes. Some behavioral characteristics related to the three basic stereotype cultures are shown in Table 2.

Reinecke [19; 20; 21] made an interesting study as an application of Hofstede's model. She studied user interface preferences in a multicultural context. The Hofstede index values are bound to the preferences of individual users in the context of information system usage. This aspect will be discussed in more detail in Section VI.

VI. MEANS FOR SUCCESS

A. *From problems to solutions*

The discussion above has mentioned the three-dimensional problem space in successful collaboration (Figure 1): communication, distribution, and cultural aspects. We will introduce five approaches that are beneficial as a partial solution related to these problems: adaptive processes, communication based on models, communication based on icons, adaptive user environments, and adaptation in a multicultural environment. We highlight the word "partial" in connection with the solutions, because the problems will still exist and we have to be able to find means for successful collaboration by understanding the sources of the problems and taking them into account in our behavior and reactions.

B. *Adaptive processes as a solution*

Statkaityte [22; 23] uses a process approach to support better quality in collaboration. The context of the studies is IS development; in the ICT business, mature organizations typically follow their predefined processes (e.g. following the principles defined by [1; 2; 3; 4] that are standardized in their quality management handbook).

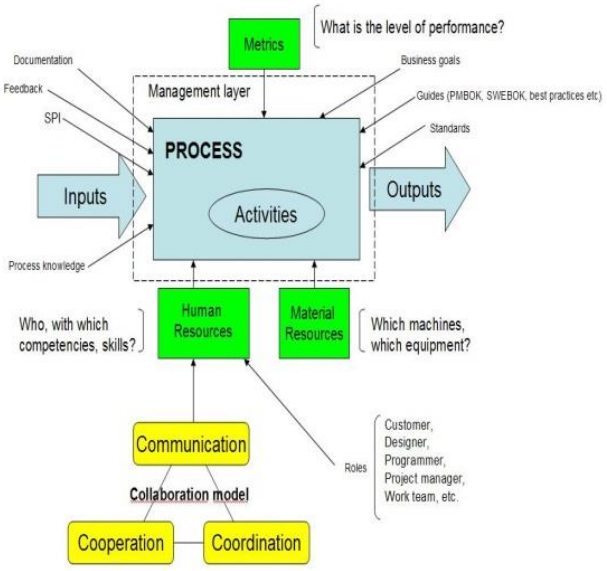


Figure 7. Culture-sensitive process definition

It is very typical that globally operating organizations apply the same processes all over the organization. In a way this helps to manage operations, but simultaneously

this approach does not take into account the differences of people from different cultures, nor does it take these differences as the driving force for some operations. There are differences in the culture sensitivity of processes: some engineering work processes are not culture-sensitive, but some human-related ones (organizing work, people management, communication-intensive processes, client interface related processes) are. In her thesis [22], Statkaityte has introduced an analysis tool called CSAM (Culture Sensitivity Analysis Model). The model extends a typical process model (components connected to the upper part of Figure 7) with human resources related factors derived from the collaboration model. The CSAM tool itself is based on Hofstede's dimensions, which are ranked for each process by the representatives of individual organizations.

C. *Models in communication*

Problems related to communication were discussed in detail in Section III. There are several sources of problems, among them the language used in communication and the conceptual models of the communicating partners. In the IS development concept, natural language is usually replaced by models – formal or semiformal. Using models in communication avoids the basic problems related to natural language: it is not exact, there are a lot of opportunities for misinterpretation, expression becomes long, and it does not provide means for abstractions.

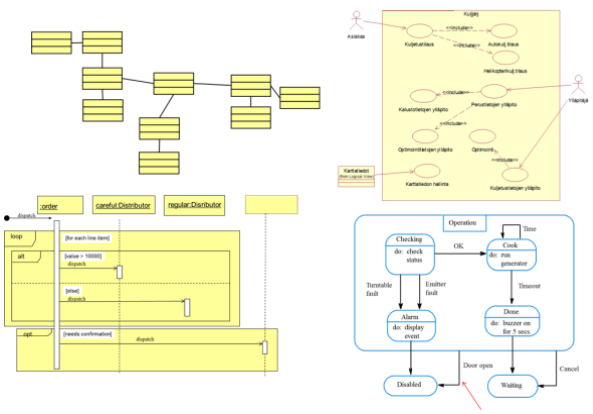


Figure 8. Semiformal model notations

The usefulness of languages in communication can be analyzed according to two factors: ease (naturalness) for users, and exactness. Unfortunately, these two factors are the opposite of each other. Natural language is easy, but not exact; formal language (Z, VDM, …) is exact, but needs special study to become useful. Semiformal languages are somewhere between these two extremes. The standardized (set of) communication language(s) in the IS development context is Unified Modeling Language (UML) [24; 25]. Figure 8 introduces some UML notations (class diagram, use case diagram, process diagram, and state chart diagram). The benefit of using this kind of visual notation, compared to the use of natural language, is its power for *abstractions*; only the elements in focus of the target system are illustrated in the model, non-important details are dispelled or used to complete the basic notation in the form of natural language explanations or using the OCL (Object Constraint Language) of UML.

D. *Icons in communication*

Khanom has studied communication in requirements engineering (RE) work.

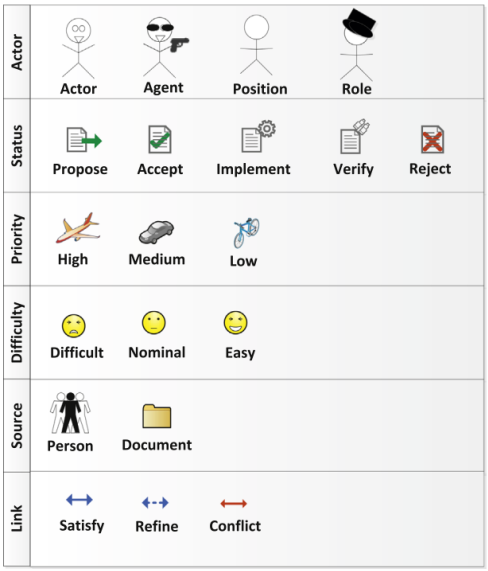


Figure 9. Elements of icon language used in RE [26]

RE is one of the most challenging phases of IS development [26; 27]. It is the starting point for further development and demands intensive communication between the client and the developer. Misunderstanding and the inaccurate specification of requirements is the main source of quality problems in information systems. In addition, wrongly interpreted requirements are also the main source of conflicts between customer and developer organizations. In her PhD thesis and related articles, Khanom has developed a basis for using icon languages as an essential part of communication in the RE context

Icons are used to visualize the requirements to make them more understandable for non-experts. The icon language proposed is culture-sensitive; the culture-related properties are based on Hofstede's dimensions according to an icon-based information modeling schema [27, p. 44] and User Preference Model [27, p. 45]. The metamodel binding the icons to the requirements specification are defined by a metamodel called Requirements Attribute Schema (RAS; [27, p. 41] developed a metamodel. The User Prefernce Model has some similarity with Reinecke's user model discussed below.

E. *Adaptive User Environments*

Reinecke's work combines user interface design with Hofstede's dimensions; in our case it is a question of human computer interaction related aspects (see Section III) in avoiding communication problems. Reinecke's work is based on two main components: the *User Model (UM)* and an *adaptive user interface* (called MOCCA; it implements 115 000 different culture-dependent user interface versions).

Reinecke's UM is based on the Cultural User Model Ontology defined in [21, p. 436]. It covers the user's

demographic data, like birth nationality, parents' culture, year of birth, political orientation, social structure, religion, educational level, and computer literacy. In addition, the model takes into account the time an individual has spent under the influence of a foreign culture. The "*user's extended national culture*" is a changed and enriched modification of the *culture by birth*; a mathematical model (simply length of stay / age in months; 21, p. 436]) is used to modify the parameter values as a consequence of the external cultural influence; we use the term "*adapted culture*" for this modified culture, because it is not only a question of extending the cultural base, but also "overwriting" (replacing), deleting, and modifying it.

Table 2. Relationships between Hofstede Dimensions and UI Design (modified from [21; 28])

Index	Low	High
PDI	• access variations, non-linear navigation • non-structured data allowed • most information at interface level, deep hierarchy not accepted • friendly messages • need for support is low	• linear navigation, few links, minimal amount of alternati • structured data expected • data in hierarchy, little data on top level • strict messages • support must be available
IDV	• traditional colors and images • image to text direction • high multimodality • colorful expression	• colors used to encode information • text to image direction • low multimodality • monotonous color map
MAS	• little saturation, pastel colors • exploration and different navigation paths allowed • personal content presentation and friendly communication	• high contrast, bright colors • restricted navigation possibilities • encouraging words in communication
UAI	• most information at interface level, complex interfaces • nonlinear navigation accepted • colors, typography and sound to maximize information	• organize information hierarchically • linear navigation paths, guided • redundancy and cues to avoid ambiguity
LTO	• reduced information density • content structured into small units	• most information at interface level • content can be arranged around a focal area

The studies of Myers and Tan [29] have a focus similar to that of Reinecke. They suggest that in the IS context the concept of "national culture" should recognize its emergent and dynamic nature: "It is something that is invented and re-invented and always in a state of flux". This points out an important aspect - people are forced/pressurized to change their traditions for several reasons and they are also constantly under the impact of foreign cultures. In the globalizing world, culture is becoming more and more global and does not follow national borders. This approach is handled in one new paper (Jaakkola & Thalheim) from the adaptive web information systems development point of view; not published, under work.

F. Adaptation of the Organization

The thesis of Harjamäki & Kostin-Harjamäki [30] analyzes the problems that may appear in an (IS development) organization between two groups from significantly different nationality backgrounds (India, Finland). In addition to the need for daily multicultural multi-site teamwork, there was the issue of acquisition, merger and organizational migration from Finnish ownership to Indian; there was pressure to adopt Indian culture at the Finnish sites of the company. The thesis argues that merging companies representing two different nationalities have everyday collisions due to culture differences. Problems relating to migration on a general level have been handled by (two of) the authors in [31]. Its findings also fit the company cases analyzed in the thesis.

The thesis proposes a *multicultural strategy development* to adapt the organizational and daily practices in this bi-cultural environment. It introduced (principles of) a tool for strategy development that guides adaptation strategy development in three steps: (1) Strategy goal analysis (ideas, vision, values); (2) Action points development (activities, processes, resources); (3) Implementation plan development. The thesis introduced a structured education/trainee package for the company to accelerate cultural adaptation. The package took into account the different roles and tasks in the organization and it covered needs of both trainers and trainees. The thesis introduces a five-level framework (maturity model) – Culture Intelligence Level (CIL), which defines the steps towards better management of multicultural aspects on the personal and organizational level.

VII. CONCLUSION

The paper focused on the analysis of problems that relate to collaboration. Three important aspects are identified for deeper analysis: communication, distribution of work, and cultural differences, which have increasing importance in the rapidly globalizing (business) world. All three aspects are handled in detail along with different problems related to globally distributed multicultural collaboration-based work. Our context comes from our own research area – information systems development – which is a typical example of work conducted by highly skilled experts (DCCI – Distributed Cross-cultural Collaborative Intellectual). The paper concluded the discussion by introducing five solutions for avoiding communication problems: Adaptive processes, Models in communication, Icons in communication, Adaptive User Environments, and Adaption of organizational practices. This paper does not go into all the solutions in detail. The solutions introduced are not possible to compare with each other – successful application is very context and situation dependent. Some of these have been handled in earlier publications by the authors, some belong to the further work planned on this interesting and important topic, which constantly raises new aspects for deeper study.

ACKNOWLEDGMENT

We would like to thank the Academy of Finland and the German Academic Exchange Service (DAAD) for supporting this research.

REFERENCES

[1] ISO/IEC, "ISO/IEC 12207/Amd2:2004 - Systems and Software Engineering - Software life cycle processes - Amendment 2 ", ed, 2004.

[2] ISO/IEC, "ISO/IEC 12207 - Systems and Software Engineering - Software life cycle processes ", ed. Geneva: ISO, 2008.

[3] SEI, "CMMI Version 1.3 ", Software Engineering Institute (SEI), 2010.

[4] ISO/IEC, "15504: Information technology -- Process Assessment – Parts 1-10," ed. Geneva: ISO 2006.

[5] B. Boehm, "A view of 20th and 21st Century Software Engineering," presented at the Proceedings of the 28th International Conference on Software Engineering (ICSE), Shanghai, China, 2006.

[6] The_Agile_Alliance. (2001, October 20th). *Manifesto for Agile Software Development*. Available: http://agilemanifesto.org/

[7] A. McAfee. *Enterprise 2.0*. Available: http://andrewmcafee.org/; ttp://en.wikipedia.org/wiki/Enterprise_2.0; http://andrewmcafee.org/2006/05/enterprise_20_version_20/.

[8] A. McAfee, *Enterprise 2.0: New Collaborative Tools for Your Organization's Toughest Challenges [Hardcover]*: Harvard Business Press, 2009.

[9] H. Jaakkola and B. Thalheim, "Multicultural Adaptive Systems," in *Information Modelling and Knowledge Bases XXVI*, B. Thalheim, H. Jaakkola, and Y. Kiyoki, Eds., ed: IOS Press, 2014, pp. 172-191.

[10] H. Jaakkola, T. Mäkinen, and A. Eteläaho, "Open Data – Opportunities and Challenges," presented at the Proceedings of the 15th International Conference on Computer Systems and Technologies - CompSysTech 2014, Ruse, Bulgaria, 2014.

[11] H. Jaakkola, T. Mäkinen, J. Henno, and J. Mäkelä, "Open^n," in *Proceedings of the MIPRO 2014 Conference*. vol. ISBN 978-953-233-078-6, P. Biljanović, Ed., ed Opatija: Mipro and IEEE [JUFO 1], 2014, pp. 726-733.

[12] B. Thalheim, H. Jaakkola, T. Nakanashi, S. Sasaki, and K.-D. Schewe, "Conceptual Modelling of Collaboration for Information Systems," in *Information Modelling and Knowledge Bases XXV*, T. Tokuda, Y. Kiyoki, H. Jaakkola, and N. Yoshida, Eds., ed Netherlands: IOS Press, 2014, pp. 272-305.

[13] E. Karttunen, "Ohjelmistopalveluiden tuottaminen monikulttuurisessa ympäristössä (Producing Software Services in a Multi-cultural Environment). ," MSc, Tietotekniikka, Pori, Tampere University of Technology, Tampere, 2010.

[14] H. Jaakkola, "Towards a Globalized Software Industry " *Acta Polytechnica Hungarica*, vol. 6, pp. 69-84, 2009.

[15] H. Jaakkola, J. Henno, and P. Linna, "From Local to Global - Path towards Multicultural Software Engineering," *International Journal of Knowledge and Learning (IJKL)*, vol. 7, pp. 5-24, 2011.

[16] H. Jaakkola, "Culture Sensitive Aspects in Software Engineering," in *Conceptual Modelling and Its Theoretical Foundations* A. Duesterhoeft, M. Klettke, and K.-D. Schewe, Eds., ed: Springer LNCS, 2012, pp. 291-315.

[17] G. Hofstede, G. J. Hofstede, and M. Minkow, *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. New York: McGraw-Hill, 2010.

[18] R. D. Lewis, *When Cultures Collide. Leading Across Cultures. Third Edition*. London: Nicholas Brealey International, 2011.

[19] K. Reinecke, "Automatic Adaptation of User Interfaces to Cultural Preferences," *IT-Information Technology*, vol. 54, pp. 96-101, 2012.

[20] K. Reinecke and A. Bernstein, "Tell me where you've lived, and I'll tell you what you like: Adapting interfaces to cultural preferences," in *User Modeling, Adaptation, and Personalization*, ed: Springer, 2009, pp. 185-196.

[21] K. Reinecke and A. Bernstein, "Knowing what A user likes: A Design Science approach to interfaces that automatically adapt to culture," *MIS Quarterly*, vol. 37, pp. 427-453, 2013.

[22] R. Statkaityte, "Culture Sensitivity in Software Processes," MSc Thesis, Information Technology (Pori), Tampere University of Technology, Tampere, 2011.

[23] R. Statkaityte and H. Jaakkola, "Modeling the Multicultural Issues in Software Engineering Processes," in *Information Modelling and Knowledge Bases XXV*, T. Tokuda, Y. Kiyoki, H. Jaakkola, and N. Yoshida, Eds., ed Netherlands: IOS Press, 2014, pp. 190-199.

[24] OMG, "Documents Associated with UML Version 2.0," ed. http://www.omg.org/spec/UML/2.0/: Object Management Group (OMG), 2005.

[25] ISO/IEC, "ISO/IEC 19505-1:2012: Information technology -- Object Management Group Unified Modeling Language (OMG UML) -- Part 1: Infrastructure," ed. Geneva: ISO/IEC, 2012.

[26] S. Khanom, A. Heimbürger, and T. Kärkkäinen, "Can icons enhance requirements engineering work?," *Journal of Visual Languages & Computing*, vol. 28, pp. 147-162, 6// 2015.

[27] S. Khanom, "On Enriching Requirements Engineering Work Using Icons," PhD Thesis, Faculty of Information Technology, University of Jyväskylä, 2014.

[28] K. Reinecke and A. Bernstein, "Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 18, p. 8, 2011.

[29] M. D. Myers and F. B. Tan, "Beyond models of national culture in information systems research," in *Advanced topics in global information management*, IGI Publishing, 2003, pp. 14-29.

[30] J. Harjamäki and E. Kostin-Harjamäki, "Monikulttuurinen yhteistyö ohjelmistoprojektissa." (Multicultural collaboration in software development projects). In Finnish. MSc Thesis, Tietotekniikka, Pori, Tampere University of Technology, Tampere, 2011.

[31] H. Jaakkola, O. Uusitalo, and J. Henno, "Adaptation in IS Integration and Enterprise Migration," presented at the MIPRO 2009, Opatija, Croatia, 2009***Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

Project- and Product-Oriented Course

J. Henno*, H. Jaakkola** and J. Mäkelä***
* Tallinn University of Technology, Estonia
** Tampere University of Technology, Pori
*** University of Lapland, Rovaniemi, Finland
jaak@cc.ttu.ee

Abstract – The current economic crisis and high unemployment rates especially among younger than 25 years old requires also re-thinking of teaching practices. Universities should produce not only professional skills, but also strong business-oriented mindset. Our graduates should be able to create new innovative products, start new enterprises and be prepared to apply all learned skills in business. This mindset can be developed using project- and product-oriented (PPO) approach in teaching, where whole course is built up as a series of projects; the course assessment is based on individual projects, which should produce a finished product with some (commercial) value. We have successfully used this approach for several years; in the presentation is described a PPO oriented game programming course.

I. INTRODUCTION - WHY ?

Many educators still see their task in preparing students with technical skills needed to accomplish some technical tasks. They suppose, that in the work life somebody (employer) gives tasks what employee (their former student) carries out. But - from where should appear such employers? Traditional work and workplace, which workers visited for fixed hours in order to produce tangible, physical values is more and more divided into discrete tasks; more complex tasks (which are not yet accessible to robots) are assigned to humans just when they are needed. More and more people are becoming self-employed workers, who should themselves seek for new tasks to implement. They should constantly look for new ideas, invent new products and be able to materialize their ideas as commercial products.

Advancements in technology and rapid changes in human society are constantly reshaping our teaching/learning/training practices, both their content and form. Computers and robots have made obsolete many traditional professions and the temp is all the time accelerating - there are already opening hotels without human personnel - the whole staff are robots [1], in the next decade cars will become driverless, and every day there appear new examples of professions which are not needed any more. The whole concept of 'work' is rapidly changing. Traditional work and workplace, where workers come for fixed hours in order to produce tangible, physical values is vanishing. Big enterprises are reducing staff and outsourcing, work is more and more divided into discrete tasks that can be assigned to self-employed people just when they're needed [2].

Humans work will more and more be intellectual and creative. Mechanical work - moving around things or tools - this can and will do robots. But robots are (not yet) good at creative tasks - creating poetry, stories, movies, games - all kind of 'content' - everything, what we seek in WWW. There are already robot rock bands [3], computer-generated rock-stars [4], even computer-generated video games [5], but robots/computers are (yet) not very good creators of intellectual content, they are (not yet) good in Science and Math. For our current students these fields - creating intellectual, artistic and/or entertainment content (games) will be the most promising engagements in their future work. Game Programming is probably the easiest way for IT students to start earning as independent (Indie) game developer.

The tendency to become self-employed, independent entrepreneur is especially visible in all Information Technology (IT) fields - you can work everywhere, if you have a laptop with (wireless) Internet connection. Currently 26.2 million people in the EU are unemployed [6], especially high is unemployment rate among youth – e.g. in Croatia 52.4% of younger than 25 are unemployed [7]. The situation is somewhat similar in all post-soviet countries – the big soviet-time enterprises with obsolete technologies vanished and it takes time for some replacements to appear. In Estonia, especially in the IT field the employment situation is already improving – we have many new IT-enterprises (Skype, Playtech) and new innovations - the digital ID-card and mobile ID, digital parking, E-Residency Digital Citizenship etc. And all these innovative ideas come from individuals, from people, not from some big faceless organizations. In IT terminology – innovative ideas appear/emerge bottom-up, from individuals, not top-down, from big organizations.

The best prospect for our students is also not to seek employment from some big international corporations (who are outsourcing to cheaper countries) or to move out in hope to find greener pastures (the 'brain draw'). If they care for their country and its future then they have to do something for this future. In practice this means self-employment, becoming independent entrepreneur, inventing the future.

The task for educators is creating necessary attitudes, prepare students to be not only professionals in their field, but also businessmen. This is even more important than providing technical skills, especially in the rapidly developing IT field. Many of the technologies which we teach today would be obsolete in 5-10 years (some even quicker). In order to be competitive, to make living in IT should everyone himself keep track of IT technology developments – the 'lifelong study' is necessity. You should be able to find everything new, but also create 'added value' yourself - create new products, services, something what you can sell. Bill Gates and Steve Jobs did not graduate, but they had very good entrepreneurial

and innovative attitudes, created lot of 'added value' and could sell their products.

Being your own boss requires that you also think like a boss - not like an employee, who does, what he is told to do.

You should constantly seek for new opportunities, invent new approaches, technologies, products, new services and learn everything new what have invented others. And you should always bring your projects to (successful) end - commercially valuable product; nobody needs a half-working demos.

You should be able to communicate, to market yourself and your ideas and always be on lookout for better opportunities to advertise your projects, products and services.

This is the mindset, what we have to inculcate to our students.

II. How ?

The task - combining teaching of IT skills with developing entrepreneur's and advertiser's mindset can be certainly achieved in several ways. By our experience the most useful is the project- product oriented approach.

The project approach: the course is presented as a series of consecutive projects which are introduced in lecture and implemented during one (some in two) 90-min exercise in computer class.

The product approach - every project should end with a finished product, implemented in class (or – with working students - in home). There should be a tangible output, not just a vague impression of hearing/learning something new. Hearing something new is not yet learning this 'new' – you have to go through this 'new' step-by-step in order to learn it.

From our experience, communication is for many IT students a very 'weak spot'. They believe, that their most important skill what they have to accomplish is programming. But nobody will know about their value, skills and programs if they cannot advertise those abilities and not to computers, but to people – people decide, what to buy, who to promote; we are (still) governed (mostly) by people. Therefore with every individual project should be presented also an explanation/help of the project (in natural language – humans still use it). Correcting language in these texts takes often more time than pondering the value/assessment of the whole project.

III. Learning by chunks

Every project should give students some new knowledge, some new (interrelated) skills. Presenting a course as sequence of self-contained projects which finish with some product having practical value inevitably causes division of course content into chunks.

Chunking is the main mechanism of human memory. Importance of chunking was recognized already in the 1940s and 1950s in perception, learning and cognition studies (see e.g. [8], [9], [10]) and the concept has often been used also as a principle of course design (see e.g. [11]).

Chunking is the natural presentation style for most teachers but unfortunately not always present. The first author has (still terrifying) experience of the first lecture as a young student (math), which began without any introduction with 'Cantor Cut' – a term difficult to explain to non-mathematicians and also to fresh high school graduate.

New information should always be introduced in consumable chunks and always considering the context - the previous knowledge of receivers. This requires careful planning – a topic for the new project (chunk) should not be too simple (mostly repeating some earlier experience with some new twists – this is suitable for additional exercise), but also not too complex (should allow implementation in one 90min exercise). The best is, if students themselves consider what is important to know/do next. Student's curiosity can be ignited with questions – "But if we want to do/achieve this…?" and they can produce some surprisingly good results and start investigate some totally new technologies, e.g. in this course one student become fascinated with SVG (Scalable Vector Graphics) and implemented his own game using SVG (hand-written, no SVG editors were used) graphics.

IV. Classroom Projects

In the following is brief description of projects which the first author uses in his course "Game and Virtual World programming" in Tallinn University of Technology. Every project is a game (or some improvement to a game produced earlier). All games introduce some new theme/topic (chunk) and have been selected for their simplicity, so that students could create finished product (game) during one 90-min exercise in computer class.

This course is for first year IT students, thus does not require good programming skills. The main technology used in this course is HTML5 – canvas, JavaScript, CSS3. This is currently the most popular technique for communication and distribution of digital content and also most convenient technology for students to become indie (independent) game developers. In their individual projects (games) themselves discovered and used also other technologies, e.g. Unity3D.

Course requires some digital image and sound-handling skills needed to manage game graphics and sounds and proficiency in handling computer file system - if for a game is downloaded an external library, which often has a complicated structure (e.g. three.js [12] - 195 MB, 6 levels of sub-directories), then students should be able quickly find from the library necessary files. Nobody had any problems and several used already some external storage techniques (GitHub, Microsoft Cloud) for hosting their projects.

First Game - Mushroom Picker: introduction to HTML5 and canvas, creating canvas background image with CSS, placing three sprites (tree, hero and mushroom) to canvas, moving hero with arrow keys, checking hero-mushroom collisions; on collision the mushroom appears in another random location. Player has in limited time collect as many mushrooms as he/she can, so with finished product students could already start competitions.

Figure 1. The first game - player should collect mushrooms.

Next - improvements to the previous: animating hero (making steps, but still only in one direction - adding just three frames to hero sprite), replacing tree sprite with graphically more pleasing randomly selected ones.

In the next project was created sprite sheet for hero with animated movement in 4 directions, sprite sheets for various mushrooms and obstacles (tree stumps, stones), JavaScript loader for external objects (sprite sheets, sounds), JavaScript objects for animated hero and animated (growing and deteriorating) mushrooms and introduced several canvases: the underlying canvas - for static trees, the upper canvas - for dynamically moving/changing game objects.

In the following projects were considered tile-based and isometric games: tile-based game data structures, various types of isometric grid, placing objects to grid tiles with script, tile-based collision checking, hero movement along diagonal grid, sprite mask; physics was introduced with the Box2D library and 3D - with the three.js.

Figure 2. Mario in isometric wood, the wood is procedurally generated

Step-by step were introduced more complicating elements (title screen, level selection, use of webfonts), and more demanding techniques. And the results become more and more playable as 'real' games. The map was in several projects procedurally generated labyrinth, thus students could in new projects re-use many parts from previous projects.

Figure 3. Cristmas Game - Santa has to collect his lost parcels .

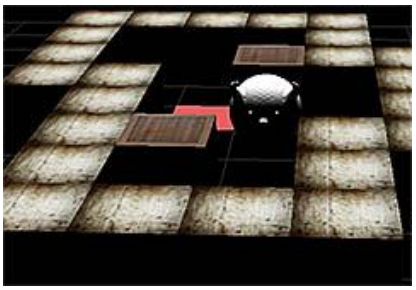


Figure 4. Sokoban - 3D game, created with tree.js library

V. HOMEWORK PROJECTS

Most of our students work, thus they prefer working at home. For all course (classroom) projects were created detailed tutorials, thus everyone could implement projects just inspecting the tutorial; even in computer class students asked very few questions, just studied the tutorials or googled to get answers for their problems.

In tutorials the main theme of a game was already set. Playing a variation of already existing game is not very exciting. Games are contemporary art, food for our intelligence; in order to become a successful game author one should introduce some new ideas, new twist create a totally new game. In order to pass the course all participants had to prepare two individual projects. The topic and implementation technique were free in both.

The first project was assessment of an existing game which students could freely select. For the assessment was prepared list of issues (game topic/flow, rating, implementation quality, graphics), which should be covered. This assessment (with pictures, video- and sound-clips from the game) should demonstrate communication skills and general understanding of games as a new media.

This task produced several technically very well implemented video trailers of commercial games.

The second (main) project was an individually designed and implemented game. This is not a simple task, but 96% of participants managed to present very interesting finished games. They had spent several weeks designing art, levels, and (in some projects) even complicated AI. The presented games covered very different themes: horizontal scrollers, racing, shooting (in deep space, defending villages in Ukraine), computer networks, kindergarten math education, exploring player's visual and mnemonic capabilities, fighting smoking etc. In the following are some images from presented games.

Figure 5. 'Lost Cat' (html5): player has to move a cat trough wilderness, full of angry dogs, who start actively pursue (AI) the cat.

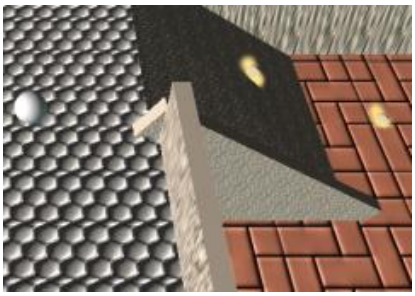


Figure 6. 'Vagabond Ball' (Unity) - player has to collect trophys and fight physics forces which try to move ball from playground.

Figure 7. 'Minimana' -(html5, all graphics - hand-written SVG), hero (red) has to collect mana (blue) but avoid falling mines (orange).

Figure 8. 'Smoking Game - Your way to the Grave' - player has to collect and destroy all sigarettes.

Several games produced in this course (in this and in previous years) have already been products with commercial value and have appeared in on-line shops (for Android or iPad) or in game sections of websites. This is the best achievement of this course.

REFERENCES

[1] http://www.nydailynews.com/news/world/japan-open-hotel-operated-robots-article-1.2103352

[2] Toward a Self Employed Nation? http://www.newgeography.com/content/003761-toward-a-self-employed-nation

[3] Comprssorhead. https://www.youtube.com/watch?v=Qlqe1DXnJKQ

[4] Concert with computer generated hologram lead singer. https://www.youtube.com/watch?v=LDMmcsH0fo8

[5] Angelina: An AI that makes computer games from scratch. http://www.extremetech.com/extreme/121633-angelina-an-ai-that-makes-computer-games-from-scratch

[6] IIEA. EU's Unemployment Puzzle Infographic. http://www.iiea.com/blogosphere/eus-unemployment-puzzle-pieced-together?gclid=CjwKEAjwu72pBRC9hsn2-e34vRUSJACnSYsT3Ny6hmIp92lk3elesrkfHtHLcquxx_Ojx_LI5Lo5thoC3yzw_wcB

[7] Hrvoje Butković. Zagreb – Croatia learns that EU membership is no panacea for unemployment. http://europesworld.org/2014/02/24/zagreb-croatia-learns-that-eu-membership-is-no-panacea-for-unemployment/#.VS9xRpOLylU

[8] Miller, G.A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97

[9] Simon, H.A. (1974) How big is a chunk? Science 183, 482–488

[10] Paul S. Rosenbloom. The chunking of skill and knowledge. Carnegie Mellon University 1987, http://repository.cmu.edu/cgi/viewcontent.cgi?article=1431&context=psychology

[11] Connie Malamed . Chunking Information for Instructional Design. http://theelearningcoach.com/elearning_design/chunking-information/

[12] https://github.com/mrdoob/three.js/

Personalization of e-tivities using Web 2.0 tools and ELARS (E-learning Activities Recommender System)

Jasminka Mezak*, Nataša Hoić-Božić ** i Martina Holenko Dlab**
* Faculty of Teacher Education, University of Rijeka, Rijeka, Croatia
** Department of Informatics, University of Rijeka, Rijeka, Croatia
jasminka@ufri.hr
natasah@inf.uniri.hr
mholenko@inf.uniri.hr

Abstract - This paper describes a collaborative learning process using Web 2.0 tools and ELARS (E-learning Activities Recommender System). The Web 2.0 approach stimulated the development of our own e-learning models that can be applied to different types of e-courses. Within the research project "E-learning Recommender System" a didactic model of using Web 2.0 tools in e-learning activities was developed. Combining LMS with Web 2.0 tools and using recommender system for personalization of activities make our courses at the Department of Informatics and Faculty of Teacher Education much more interested and challenged.

I. INTRODUCTION

New approaches in e-learning are based on teaching model using the constructivist theory of learning. This model is student-centred and it is using various Web 2.0 tools for collaboration, communication, sharing learning materials, tagging, etc. [8].

The learning process that includes social software in e-learning is called E-learning 2.0, due to Web 2.0 tools that are used to enhance the process of learning itself. Web 2.0 tools enable a shift from a distributive to a more collaborative mode in e-learning. In particular, the ease of use and simplicity of Web 2.0 technologies allow creating learning environments, which can realize activity-rich pedagogical models and facilitate students' development of competences [14].

Today students enter institution with some experience with Web 2.0 tools such as wikis, blogs, YouTube, Facebook, but they lack an awareness of how those tools can be used for learning [5]. Siemens and Matheos [15] suggested two trends in education: learners have freedom to access, create and recreate their learning content, and they have opportunities to interact outside of a learning system.

Student-centred didactical models presents the work in progress as part of the research project "E-learning Recommender System" with main goal to develop didactical models for realization of e-learning courses in Moodle based learning management system MudRi, supported by Web 2.0 tools and ELARS (E-learning Activities Recommender System). These courses will, besides the usual activities supported by LMS, include personal collaborative e-learning activities (e-tivities).

The second part of this paper describes the e-learning 2.0 process. The third part is explaining building parts of the model: e-tivities, Web 2.0 tools and ELARS recommender system. The fourth part gives overview of the "E-Learning Recommender System" project. The last part brings conclusions and future work plans.

II. E-LEARNING 2.0

While the first generation of e-learning has been oriented on distributing learning materials and modest communication between the participants of learning process, e-learning 2.0 includes changes that facilitate participation in learning as well as interaction and collaboration between students rather than continuous focus on subject matter knowledge.

E-learning 1.0 was based on the idea that teaching is transmissive task. Teacher present topics, distribute learning materials, and communicate with students about those learning materials. E-learing 2.0 emphasises the proactive role of the students. Learning is perceived as an interlinked social process, in which students use Web 2.0 tools to develop learning outcomes. They do not only use the already distributed learning materials, but also collaborate and extend their learning activities to open content on the internet [13, 14].

Many students did not traditionally participate in any decisions relating to their learning environment. E-learning 2.0 enables students to choose, create, share and publicize their learning materials [5]. Students are actually building their own personal learning environments not only with collecting (consuming) the learning materials but also with producing a significant amount of learning content [11].

While considering the pedagogical aspects of e-learning, it is necessary to include elements that comply

The research has been conducted under the project "E-learning Recommender System" (reference number 13.13.1.3.05) supported by University of Rijeka (Croatia).

with different theories of learning: behaviourism, cognitivism and constructivism. However, e-learning 2.0 promotes constructivism according to which students should be active participants who do not remember teaching materials literally, but create their own versions of the course content by exchanging views and opinions with their colleagues [6, 10]. The teacher is still an important participant of that process, although his/her main task is no longer transmitting the knowledge but guiding students in the process of acquiring knowledge [6].

The process of designing the e-learning 2.0 environments is grounded in socio-cultural learning, which is in a constant cognitive development with the power of dynamic social interaction. Cognitive learning focuses on learning processes and development trough creating, editing and remixing learning content socially and collaboratively [3].

From technological aspects, e-learning 2.0 can be supported with combination of functionalities of an LMS that are equal for all users, and appropriate third-party services available on the Web [6, 7]. Thus, the environment for e-learning 2.0 consists of a set of tools. Different students can use different tools and create their own personal learning environments (PLE) [6]. According to Attwell [1] and Costello and Shaw [4], PLE is not a specific tool, but rather a way of organizing resources. Although today Web 2.0 tools usually have a central place in PLEs, it can also be comprised of learning materials, peers, hardware, and other components that allow students to collect, process and share information and knowledge [8].

Today students need more than simply consuming the offered learning materials. They should be provided with challenging situations instead of ready-made instructions that are put in rigid and closed environments [16]. With PLE the learning outcomes change from knowledge acquisition to competence development, making students more motivated and satisfy.

III. USING WEB 2.0 TOOLS AND RECOMMENDER SYSTEM FOR E-TIVITIES

Didactical models developed during the research presented in this paper are in line with the e-learning 2.0 approach and are based on collaborative e-tivities. Web 2.0 tools enable realization of different e-tivities and therefore, represent one of the building parts of our didactical models. In addition, the personalization of e-tivities is achieved using the educational recommender system ELARS which extends the environment for e-learning.

A. E-tivities

An e-tivity is the interaction of a student with other students using specific tools [17]. It is achieved through completion of a task and oriented towards specific learning outcomes that should be achieved [8]. The components that constitute such activity are: the context within the activity occurs (the subject, the learning outcomes, and the environment within which the activity takes place), the pedagogy adopted (learning and teaching approaches) and the undertaken tasks. The task specifies the type of task, the resources (e.g. teaching and learning materials, tools), the interaction and roles of the participants involved (e.g. work in pairs or groups) [6].

Teacher, as an important participant of the e-learning process, designs a task. The task defines required learning materials, tools that can/should be used for e-tivity and mode of interaction between participants. For e-learning 2.0 more significant are e-tivities where the task is group-based and involves collaboration [6]. Besides defining a task, teacher should also support students during the e-tivity in order to encourage them for participation and collaboration with their peers.

B. Web 2.0 tools

Web 2.0 tools are shifting the e-learning process to personal learning environments (PLE) where students can organize and collect learning resources on their own way [1]. Depending on task, an e-tivity can be performed with one or more tools, which means that it is not necessary that all students use the same tools.

In the current stage of our research, developed didactical models include one of the nine Web 2.0 tools [8], described in the reminder of this section.

Blogger is a tool for publishing blogs. The characteristic of blog is commenting, tagging and linking. In the majority of blogs, visitors are allowed to write comments under each entry. This enables exchange of views between the author on the one hand and a group of people who all contribute by writing comments on the other. In order to facilitate subsequent retrieval of published records and searching similar content, records can be bookmarked by specifying keywords that categorizes the entry. Blogs can be individual or collaborative and connected in groups' topic. In e-learning environment, blogs can be used as a personal journal for learning. Students can independently express their opinions on learning materials as well as conclusions from learning.

Diigo is a tool for collecting and organizing bookmarks and other resources. In e-learning, it can be used for organizing and sharing favourite internet resources about learning topics. One of the goals is to create a common review on the teaching content.

Flickr is a tool for photo management and sharing. Some of the key elements of the Flickr are tags, common clustering and ranking of interesting facts. Students can store, organize and view photos within the group or individual e-tivities.

Google+ is a tool for social networking, sharing and communication. It can be used in e-learning with its *Circles* application that allows creating a group of users who can actively engage in the creation of content for the Web and participate in publishing and exchanging of learning materials working together. Furthermore, Google+ provides the tool called *Sparks*. This tool is looking for videos and articles that it considers interesting for the user. This service enables the teacher and students to monitor the current issues that can be applied in the

teaching process and update their knowledge of the potential new findings related to the topic.

Google Drive is a tool for creating and sharing documents (textual documents, spreadsheets, presentations, forms). Those applications allow students to create and edit documents online while collaborating with each other in real-time.

MindMeister is a tool for online mind mapping and brainstorming. A mind map is the most common teaching aid in all educational areas. During e-tivities, students can use this tool for presentations, for the purpose of allocating tasks, graphical explanation of concepts, for reviewing topics, to prepare for the assessment, the exchange of ideas and information and simultaneous work on the map.

SlideShare is a tool for sharing presentations. In e-learning, using this tool students can cooperate at the local level but also with a virtual community where they can make, share and collect teaching content.

Wikispaces is a tool for creating wikis. Wiki is very well suited to support collaborative online learning since it fosters collaborative writing. Therefore, it can be used in various forms to fit into e-tivities in which students need to summarize subject matter in a written form, present results of conducted WebQuest and similar.

YouTube is a tool for editing and sharing videos. As one of the most popular media tools YouTube has vast application in the teaching process. In e-learning environments, textual materials can be enhanced with animations and video content. In addition to the publication of teaching materials in videos, it can be used for displays of various examples, instructions, guest speakers recording, multimedia presentations, etc.

C. ELARS recommender system

Within this research a recommender system ELARS (E-Learning Activities Recommender System) was developed. It provides personalization in the context of collaborative e-tivities realized using Web 2.0 tools [7].

Recommender system is used in parallel to LMS and Web 2.0 tools [8]. Students use LMS to read previously prepared materials for learning, to get instructions, to communicate with each other and with the teachers and Web 2.0 tools to perform e-tivities. They use ELARS system, as a third component of the e-learning environment, to read recommendations before and during e-tivities.

Recommendations are generated based on several characteristics: student's learning styles, Web 2.0 tools preferences, knowledge level and activity level. Learning styles and tools preferences are collected explicitly using the feedback from users. Data regarding the other two characteristics are collected implicitly. Knowledge level is automatically calculated according to the results of (online) tests, while activity level is automatically estimated based on the data about students' interaction with the Web 2.0 tools. Activity level represent quantity and continuity of student's contributions relatively in respect to others students and group's contributions relatively in respect to other groups. The important part of our research are novel procedures for assessing student's (group's) activity level based on data collected from Web 2.0 tools.

Data regarding students' characteristics, including activity level, are used in recommendation rules in order to recommend possible students-collaborators, Web 2.0 tools or optional e-tivities, as well as to generate advices regarding participation in e-tivities [8].

IV. E-LEARNING RECOMMENDER SYSTEM PROJECT

Development of didactical models that can be applied to different types of e-courses was conducted within the research project "E-learning Recommender System".

A. Goals and Tasks

The purpose of the research was to improve the quality of higher education by introducing innovative computer technologies in the processes of teaching and learning and the promotion of new pedagogical approaches of e-learning 2.0 through the presentation of case studies.

The project included activities to meet specific objectives of the research:

- the development of algorithms and model of recommender system and its implementation
- the development of new models of learning based on collaborative learning and developed recommender system,
- the educational design with e-activities for different types of courses that will be realized in practice with the support of the recommender system,
- the designing the recommendations for the implementation of the developed didactical models and recommender system in the context of higher education.

B. Didactical models

Our didactical models are based on observing the process of e-learning as a sequence of (collaborative) e-tivities, on using Web 2.0 tools for their realization, and on using recommender system for personalization of e-learning process according to student’s characteristics [8].

Research provides scientific innovation in the field of e-learning with technological and pedagogical aspects. The recommender system, as one of the main results of the project, was used to determine personalized recommendations for the students in performing a sequence of collaborative e-activities with the help of learning management system and Web 2.0 tools.

In this way, with a new approach to the use of computer technology to support education, research promotes new pedagogical approaches that emphasize the need for collaborative learning and are based on the performance of e-activities.

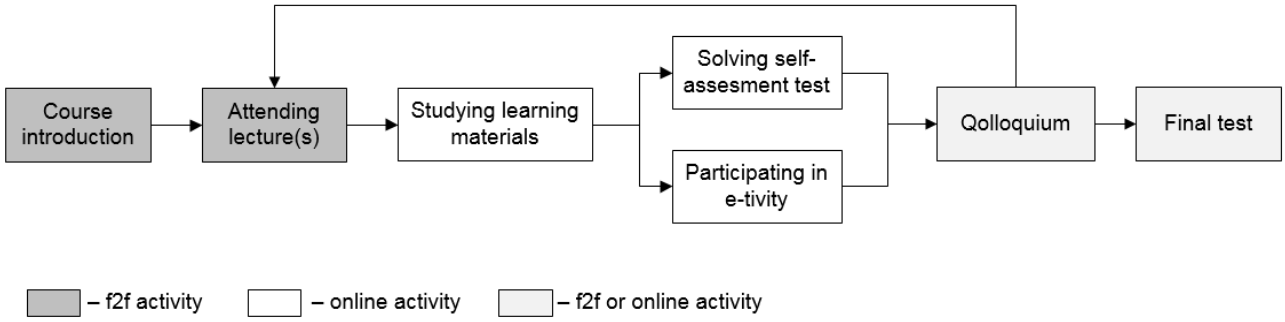


Figure 1. Course design by sequence of activities

It is expected that the application of the above mentioned learning model in practice result in increased performance and motivation of students which will be evaluated in the context of different types of e-courses that are conducted: face-to-face (f2f) courses with online available contents, mixed mode (blended learning) and fully online courses. For the purpose of the project some of e-learning courses in Moodle based LMS MudRi on University of Rijeka were adopted. The courses workflow includes individual and group-based e-tivities that are realized using Web 2.0 tools. The approach was used for two e-courses at the Department of Informatics ('Hypermedia in education' and 'Operational research') and currently is used for one e-course at the Faculty of Teacher Education ('Extracurricular informatics and technical activities'), University of Rijeka, Croatia.

During blended learning courses, a course introduction is followed by sequence of activities that are realized in f2f or online environment. These can, as shown in example presented on Fig. 1, include lectures in f2f environment in order to present subject matter to students. Students are usually required to study prepared learning materials that are available in LMS. Sometimes, students have opportunity to test their knowledge using online tests for self-assessment. In addition to self-assessment tests, e-tivities can also be used to ensure formative assessment before colloquiums. Depending on number of topics covered by the course, such sequence can be repeated. At the end, students' knowledge can be assessed using final test.

Personalization provided by the ELARS system is based on course learning design that describes the sequence of course activities. Between course activities, e-tivities (eLA) are the most important. Activities sequences can also contain f2f activities (FA), learning content activities (CA) and testing activities (TA). Besides that, it should be stressed out that planning e-tivities often requires planning additional e-tivities that are not directly oriented towards achievement of learning outcomes, but support the organization or personalization of e-tivities. Those are support activities (SA) and decision activities (DA).

ELARS system recommends collaborators if corresponding e-tivity is group-based. The system recommends the most appropriate tool for student or group to fulfil the aim of e-tivity if more than one tool is offered for its realization. Optional e-tivities are recommended when students are allowed to choose between several e-tivities and advice is offered during e-tivities, in order to encourage students to reach higher activity level. To enable this, the original recommendation model is developed.

Fig. 2 presents a didactic model with activities sequence that precedes the collaborative e-tivity personalized using ELARS recommender system. Before an e-tivity starts, students should get instructions with description of task they are supposed to do and other details relevant for the realization of the e-tivity (is it individual or group-based, duration, evaluation criteria, etc.). For that purpose, support activity is used.

In order to achieve personalization using recommendations, decision activities should be planned. If the e-tivity is group-based, collaborators recommendations are presented to students within corresponding decision activity in ELARS systems. Similarly, if there are several tools offered for the realization of the e-tivity, tools recommendations are presented. During e-tivity, additional support to students is provided in a form of advice regarding quantitative aspects of students' engagement. Therefore, students need to enter input data to the ELARS system in order to enable activity level calculation and advice generation. E-tivity can be followed by presentation of students' results in f2f environment.

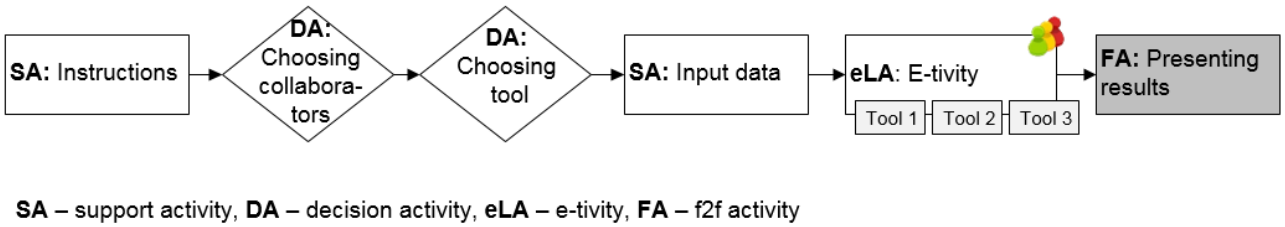


Figure 2. A didactic model of one e-tivity

C. Evaluation

The educational verification of e-learning model is in the progress, as well as the recommender system impact on student's performance in e-tivities.

We are now in stage of final evaluation that is conducted in order to determine the impact of the recommender system on student's performance during e-tivities and how collaboration affects the learning process itself. Before this final evaluation, several components were already tested. Selected Web 2.0 tools were previously used within e-learning courses. They found to be reliable, and that their performance matches the expectations and needs in practice [8]. Validation of the algorithms for assessing student's (group's) activity level was also conducted with aim to determine to what extent obtained values match the actual state [8].

Our goal is to determine whether proposed didactical models increase student's effectiveness and motivation for e-tivities included in e-courses. In addition, students' satisfaction with received recommendations will also be examined using questionnaires.

V. CONCLUSIONS

Development of didactical models described in this paper was enhanced with LMS, Web 2.0 tools and E-Learning Activities Recommender System - ELARS. Such e-learning environment is based on e-learning 2.0 approach that emphasize the need for including collaborative e-tivities in e-courses learning design.

Our future work will focus on creating didactical models for different types of higher education e-courses. In doing so, the extent of using f2f/online learning will be taken into account as well as number of students - a factor that could affect the realization of e-tivities. Since e-tivities included in the models are personalized with the ELARS recommender system further efforts will be made to improve the system. This include additional experiments to evaluate recommendation algorithms and making ELARS web application in line with the responsive web design, so it can be used on different devices, from desktop computer to smart phones. The recommender system will be open source coded and implemented to support large number of users via cloud service.

Project outcomes and results, including the examples of good practice, will be presented on the website. This will enable all interested parties to be acquainted with the developed models and facilitate their use in the broader academic community.

LITERATURE

[1] Attwell, G., "Personal Learning Environments - the future of eLearning ?", Lifelong Learn., vol. 2, no. January, pp. 1–8, 2007.

[2] Bubas, G., Orehovacki, T., Balaban, I. and Coric, A. "Evaluation of Web 2.0 tools in the e-learning context : Case studies related to pedagogy and usability," in University Information Systems - Selected Problems, pp. 259–277, 2010.

[3] Chih-Hsiung Tu, Laura Sujo-Montes, Cherng-Jyh Yen, Junn-Yih Chan AND and Michael Blocher, „The Integration of Personal Learning Environments & Open Network Learning Environments" IN TechTrends, Volume 56, Number 3, pp 13-19, May/June 2012.

[4] Costello, R. and Shaw, N. „Personal Learning Environments", in the proceedings of Learning Technologies workshop, pp. 4-12, 2014.

[5] Diaz, V., „Web 2.0 and Emerging Technologies in Online Learning" New Directions For Community Colleges, No. 150, Wiley Periodicals, Inc., pp 57-66, 2010.

[6] Hoić-Božić, N., Holenko Dlab M., and Mezak, J. "Using Web 2.0 tools and ELARS Recommender System for E-Learning." in International Conference on E-Learning Proceedings. 2014.

[7] Holenko Dlab, M. and N. Hoić-Božić, "An Approach to Adaptivity and Collaboration Support in a Web-Based Learning Environment," in International Journal of Emerging Technologies in Learning (iJET), vol. 4, no. 7. pp. 28–30, 2009.

[8] Holenko Dlab, M. and N. Hoić-Božić, "Recommender System for Web 2.0 Supported eLearning," in IEEE Global Engineering Education Conference (EDUCON 2014) Proceedings, 2014.

[9] Holenko Dlab, M., Hoic-Bozic, N. and Mezak, J., "Personalizing E-Learning 2.0 Using Recommendations," in The Proceedings of MIS4TEL conference, 2014.

[10] Pratt, D. D., "Good Teaching: One Size Fits All?," New Dir. Adult Contin. Educ., no. 93, pp. 5–16, 2002.

[11] Reed, T.V. „Digitized lives: culture,power and social change in the Internet era", Routledge, NY, 2014.

[12] Salmon, G., E-tivities: the key to active online learning. Psychology Press, 2002.

[13] Salmon, G., "Learning Innovation for the Twenty-First Century", in Changing Cultures in Higher Education, New York: Springer International, pp. 27–43, 2010.

[14] Schneckenberg, D., Ehlers, U. and Adelsberger, H. „Web 2.0 and competence-oriented design of learning – Potentials and implications for higher education" British Journal of Education Technology, Vol 42, No 5, pp 747-762, 2011.

[15] Siemens, G., & Matheos, K. „Systemic changes in higher education" Technology & Social Media in education, 16 (1), 2010.

[16] Valjataga, T., Laanpere, M., „Learner control and personal learning environment: a challenge for instructional design" in Interactive Learning Environments, Vol.18., No.3, pp 277-291, September 2010.

[17] Beetham, H „An approach to learning activity design", in H. Beetham & R. Sharpe (Eds.), Rethinking pedagogy for a digital age designing and delivering elearning, Routledge, pp. 26–40, 2007.

[18] ELARS - E-Learning Activities Recommender System (demo home page), http://161.53.18.114/elarsdemo. Accessed 2 Feb 2015.

Model of Information System for recognition of non-formal and informal learning in higher education

M. Balković *, G. Đambić * and Z. Jančić*
* *University college Algebra*
Ilica 242, Zagreb, Croatia
e-mail: mislav.balkovic@racunarstvo.hr
e-mail: goran.dambic@racunarstvo.hr
e-mail: zoran.jancic@racunarstvo.hr

Abstract – This article has been produced through active research within EU funded project tasked to implement recognition of non-formal and informal learning (RPL) to Croatian higher education, including all needed supporting mechanisms such as information system. Aforementioned project is part of wider national activities following recommendation of Council of European Union from December 2012, setting task of full implementation of RPL systems until 2018 in all EU member states. Research presented within this articles started from desk research of current state of play in RPL implementation across EU countries, following development of research questions and organization of survey in order to better understand requirements and attitudes of key national stakeholders. Main results of both research activities and specific stakeholder's requirements supported by statistical analysis are presented within this article, following proposal of potential information system model. The model has been built following theoretical understandings, practical implementation achievements from other countries and specific national stakeholders' requirements, and it consists of architectural recommendations, data model, and description of key "use cases", all useful as a foundation of future RPL information system.

I. INTRODUCTION

There are currently number of terms used in order to address process of recognition of prior non-formal and informal learning (RPL) such as: validation of non-formal and informal learning (VNFIL), prior learning assessment (PLA), prior learning assessment and recognition (PLAR), recognition, validation and accreditation of learning (RVA) and others. Although term VNFIL is mostly used in EU policy documents, term RPL is most common in scientific literature, and will therefore be used throughout this article as well. In order to get full understanding of the very process, one of the commonly used, self-explanatory definition of recognition is:

"*The confirmation by a competent body that learning outcomes (knowledge, skills and/or competences) acquired by an individual in a formal, non-formal or informal setting have been assessed against predefined criteria and are compliant with the requirements of a validation standard. Validation typically leads to certification.*" [1]

In number of countries recognition of prior learning has been implemented before implementation of national qualifications framework (NQF), while in others RPL has been introduced as part of NQF development. Very beginning of RPL practices could be found in USA and France after Second World War in order to provide access to labor market and qualifications to former soldiers. Stated motive of widening access to the labor market for workers which would be otherwise left jobless or doing less paid jobs is one of key economic drives to implement RPL. Still, there are number of other equally valid motives for the RPL's implementation such as: widening possibilities to access education for non-traditional learners, improving effectiveness of educational systems as well as catching fast pace of advancement in technology validating knowledge and skills obtained on the job. Regardless of the motives, RPL and its implementation become part of the policy development agenda in EU since 2000 and Feira European Council [2], following the communication from the Commission, "Making a European Area for Lifelong Learning a reality" in November 2001 [3] and being emphasized in policy documents since. Most important advancement in RPL implementation in policy sense occurred in December 2012 when Council of European Union issued Recommendation [4], indicating that EU member states should implement RPL by 2018. Stated recommendation formally introduced another strong policy obligation for member states due to the fact that Council insisted that implementation of RPL should be done using "equal value" principle:

"***Qualifications*** *or, where applicable, parts of qualifications* ***obtained by*** *means of the* ***validation*** *of non-formal and informal learning experiences* ***comply with*** *agreed* ***standards*** *that are either* ***the same as****, or equivalent to,* ***the standards for qualifications obtained through formal education*** *programmes".*

Aforementioned principle which originates from Recommendation puts significant development requirements even to the countries which already implemented RPL. In essence, stated recommendation calls for implementation of "standards" against which validation should be organized in order to present document (i.e. qualification) of equal value as one that

would be otherwise obtained through formal education. Stated two requirements in a nutshell call for implementation of national qualifications framework or similar instrument which holds national qualifications standards and for implementation of such validation process that is capable of producing validation decisions reliable enough so that same qualification certificate (document) could be presented to RPL candidates and graduates from formal programs [5]. In stated circumstances number of currently used validation practices (i.e. formative and declarative methods) are challenged and might be changed by more reliable methods based on summative assessments, similar to those used in formal education. This, in turn, put significant focus on Quality Assurance (QA) and transparency of the overall qualification award process in order to protect public interest and confidence in national qualifications. One of the potential tools to implement such RPL provision is development of RPL IT system linked to national NQF System.

II. STATE OF PLAY IN CROATIA

Republic of Croatia started development of its national qualifications framework in May 2006 and after 7 years of constant consultations with numerous stakeholders in February 2013 Croatian Parliament adopted Act on Croatian Qualifications Framework [6]. Stated process lasted longer than expected due to the fact that Croatia selected to develop its own solution instead of copying successful models from other countries, that most probably wouldn't fit to our culture and educational context [7]. By-low on the CROQF register has been introduced by Ministry of Science, Education and Sports (MSES) in May 2014, and it introduced required legal framework to start development of NQF ICT system (registers) as well as formation of Sector Skills Councils (SSC's) and National Council for Human Resources Development. Aforementioned by-low provided also blueprints for Qualification (QS) and Occupational (OS) Standards elements and set stage for implementation of Quality Assurance in respect to specific processes brought by NQF. Number of QS's and OS's are currently being developed and national NQF ICT system is in production. Still, RPL as an important element of the NQF is missing its legal establishment in the form of another missing by-low on RPL and its support system in form of RPL ICT system, QA recommendations for RPL as well as required implementation capacity in relevant public organizations (i.e. Agencies in area of education).

During development of NQF, Croatia purposely selected RPL to be implemented as last NQF's element in order to best protect its citizens against potential misuse of RPL such occurred in South Africa [8] and some other countries. In order to support policy development of RPL in higher education, Agency for Science and Higher Education (ASHE) initiated EU funded project: "Competitive Croatian Higher Education for better Employment" in order to propose RPL implementation model for higher education including RPL ICT system. Research presented in this article have been organized within aforementioned project as a potential blueprint for future by-low and overall national development in this area, while part of its results focused on ICT system are presented in this article.

III. RESEARCH METHODOLOGY

Research of requirements for ICT system for RPL focused on desk research of best practices in EU member states, as well as on design of research instrument (survey questionnaire) and pursuance of two surveys involving most significant stakeholders who should use proposed system in higher education: higher educational staff and accreditation agency experts.

A. RPL implementation and ICT system - desk research

In order to support implementation of RPL in EU member states, Council of European Union tasked CEDEFOP to observe achievements and to constantly support development by issuance of concrete guidelines [1] and publication of up to date Inventory on RPL practices which has been so far updated five times (2004, 2005, 2008, 2010 and 2014) [9]. Within the stated guidelines, there are no concrete recommendations on the need to implement RPL support ICT system, but Guidelines suggests that RPL process should be transparent, quality assured, cost effective and linked to NQF in order to use published qualification's standards. According to the Inventory [10] which reviewed achievements of 33 countries (28 EU member states + Iceland, Liechtenstein, Norway, Switzerland, and Turkey, currently only 3 countries (Finland, France and Spain) have comprehensive RPL strategy covering all educational levels (education subsystems), another 10 countries have national strategy for RPL with some elements missing, 9 countries currently develop their national strategy and only 7 countries (including Croatia) have no legal framework for RPL at all. One of the recognized challenges in implementing RPL is lack of data collection on RPL at national level and CEDEFOP indicates that much remains to be done to establish such systems [11]. Through research of achievements for each of stated 33 countries it can be seen that most advanced of them do have ICT system to support their NQF (i.e. qualifications register), but only one project have been found to develop IT solutions to support RPL. Namely, Vita project initiated by total of 10 partners from 9 countries in 2012 and financed by EU funds to develop e-portfolio approach that will support primarily RPL clients and through formative testing add "evidence" of their achievement to personal Europass portfolio [12]. With its focus on assessment methods and its primary goal to support RPL candidates, we found almost none of the achievements within the project useful in our case tasked to designing comprehensive RPL system capable of supporting, candidates, providers and QA authorities.

Although we could not get useful inputs from other successful implementations of ICT systems for RPL due to early stage of development in this whole area, we managed to build on recent implementation references linked to QA [13] and the use of ICT in building e-portfolios [14], [15] in order to develop research questions for potential RPL providers in higher education and structured interviews with policymakers in MSES, ASHE, Student Union, Employers Association and Trade Unions.

B. *Sample information and research instruments*

In order to get better understanding on requirements for RPL and its ICT solution, research questions for structured interviews with policymakers and survey of potential providers (key stakeholder's groups) were organized in four groups with total of 22 items using five point Likert type scale with numeric representation as follows: 5 - Strongly agree, 4 - Agree, 3 - Neutral, 2 - Disagree and 1 - Strongly disagree with additional 0 - Do not know, do not want to answer option. Additional two questions have been used to determine respondent's role within higher education and his/hers knowledge of the concepts of RPL raging through 4 levels from: "Expert in RPL", "experienced in NQF", "understand concepts of Learning outcomes (LO's)" to "know nothing". Survey has been organized using paper based templates while structured interviews with key stakeholders involved recorded and later transcribed focus groups meetings held in three separate occasions during 2014. Most significant items relevant for requirements for ICT system were as shown in Table I.

TABLE I. KEY ITEMS RELEVANT TO RPL ICT SYSTEM

Question ID	*Question text*
Q6	Following recommendation of Council of European Union (2012/C 398/01), Croatia should implement RPL using "equal value" principle.
Q10	In order for RPL process to be valid and transparent, it is necessary to validate (assess) and document result of each learning outcome within unit of learning outcome, regardless of its size.
Q16	Validation processes are much more endangered by potential misuse and "sales of diplomas" than classical studying process (one that includes both acquiring and validation).
Q17	Validation process should be significantly more "protected" with QA system than classical educational process and this justifies more documenting.
Q18	It is justified to archive complete RPL candidate's documentation for future audit.
Q19	It is justified to insist on use of national ICT system for RPL to enter all candidate's data including digitalized assessment materials.
Q20	It is justified to insist on use of national survey system for RPL candidates in order to track their satisfaction with the process.
Q21	It is justified for RPL authorities to have anonymized on-line access to candidate's documentation.

Research sample included total of 71 respondents. Out of stated number, 57 were relevant as representatives of potential providers in higher education, representing most of Croatian Universities and Universities of Applied sciences, as shown in Table II.

TABLE II. DESCRIPTIVE INFORMATION ON SAMPLE

Target group affiliation	*Frequency*	*Percent*
(1) Teacher in HE	47	66,20%
(2) Administrative person in HE	10	14,08%
(3) Trade union or Employer	3	4,23%
(4) Student	6	8,45%
(5) Educational authority	2	2,82%
(6) Non of listed groups	3	4,23%
Total	**71**	**100,00%**

IV. RESEARCH RESULTS

Results for Q6 show that most of respondents support the idea of RPL implementation using "equal value" concept, as per Table III. Further research in order to determine whether knowledge on NQF and RPL significantly affects this attitude is done using one-way ANOVA with Levene's test of variance homogeneity, showing no significant difference in such attitude ($F(2,52) = 0.478$, $p = 0.623$) between respondents' groups.

TABLE III. Q6 RESULTS

Question	N	Mean	Std. Deviation	Std. Error Mean
Q6	56	3,68	1,081	0,144

Results for Q10 show even more positive attitude towards thorough assessment and documentation of each LO within the Unit of LO's, as per Table IV. Further research in order to determine whether knowledge on NQF and RPL significantly affects this attitude is done using one-way ANOVA with Levene's test of variance homogeneity, showing no significant difference in such attitude $F(2,52) = 1.795$, $p = 0.177$) between respondents' groups.

TABLE IV. Q10 RESULTS

Question	N	Mean	Std. Deviation	Std. Error Mean
Q10	55	4,16	0,918	0,124

Results for Q16 show that most of the respondents find RPL processes to be endangered by misuse and should be protected more than formal education, as per Table V. One-way ANOVA with results ($F(2,50) = 0.005$, $p = 0.995$) show no statistically significant differences between groups or respondents' attitudes.

TABLE V. Q16 RESULTS

Question	N	Mean	Std. Deviation	Std. Error Mean
Q16	55	3,56	1,183	0,159

Next set of five items have been logically connected trying to determine level of quality assurance and documentation of RPL, with very high positive results per each item as shown in Table VI.

TABLE VI. Q17 – Q21 RESULTS

Question	N	Mean	Std. Deviation
Q17	52	4,12	0,983
Q18	52	4,44	0,698
Q19	52	4,42	0,871
Q20	52	4,31	0,853
Q21	52	4	1,01

With such results, we also tested reliability of 5 item scale instrument with Cronbach's Alpha with result of 0,76 and 0,768 (Cronbach's Alpha Based on Standardized Items). Although on the margin of reliability, stated results allowed us to calculate group result for the whole item set with mean of 4,2577, Std. Deviation of 0,63535 and Std. Error Mean of 0,08811. One-way ANOVA with results $F(2,47) = 1.227$, $p = 0.302$) show again no statistically

significant differences between groups or respondents according to their knowledge on the RPL.

Results of structured interviews with stakeholders mostly corresponded to the results provided by survey respondents, with additional relatively specific requirements which partly even contradict to main conclusions that would be drawn from presented statistical results. Main specific requirements are as follows:

- Ministry of Science, Education and Sports insisted that future ICT RPL system should not be mandatory for RPL providers but should instead allow, through use of web services or other interoperability options, data collection also through use of ICT solutions institutions already have. In such circumstances institutions will be able to choose what system they will use, but with obligation to provide required data set per each candidate / validation.
- Agency for Science and Higher Education raised question of data and privacy protection and would like RPL providers to hold and care for most of the data such as assessment materials, with data collection only for most basic RPL data.
- Both institutions as well as some of the potential RPL providers from higher education argued that assessment data could be collected but should not be uploaded to any system outside the provider's institution.

Taking into consideration all shown stakeholder's requirements and attitudes, RPL implementation guidelines, available national legislation in the field as well as already proposed data model for CROQF register [16] we suggest that national RPL system should be designed following key recommendations in Table VII.

TABLE VII. DESIGN RECOMMENDATIONS FOR ICT SYSTEM

ID	*Design recommendation*
R1	System should be designed for provider's internal usage, with assessment data never leaving the provider's system.
R2	System should allow export of only key information set to future national RPL register.
R3	System should be connected with CROQF database in order to use definition of Units of learning outcomes, Qualifications Standards and list of accredited RPL providers in order to map candidate's results to relevant Units.
R4	System should allow RPL candidates and potential RPL candidates to search RPL provision through full text search and through selection of types and sectors qualifications belongs to, by automatic connection to CROQF database.
R5	System should allow potential RPL candidates to apply for validation to accredited provider on-line.
R6	System should allow creation of e-portfolio for candidates (collection of data and documents required in the process).
R7	System should allow RPL providers to manage: their own staff involved in RPL, RPL candidates and guidance processes.
R8	System should allow RPL providers to publish key information on their RPL provision useful to candidates.
R9	System should support already available identity management systems in Croatian higher education in order to allow seamless use and single sign-on (SSO).
R10	System should be developed as on-line application.
R11	System should keep detailed information about all RPL processes in order to guarantee more transparency.

V. RPL SYSTEM MODEL

Following recommendations for future RPL system, it is apparent that there will be a separate instance of system for each provider. Each provider will manage its instance, keeping assessment materials private with only key information set published to future national RPL register. That key information set is yet to be defined, but it is reasonable to assume that it will contain at least the same set as current classical educational process: candidate's full name, personal identification number, date of birth, address and information about achieved qualification.

A. System Architecture

We suggest that the system architecture should be comprised of at least components shown in Table VIII.

TABLE VIII. ARCHITECTURE COMPONENTS

Component	*Role*
HTML5 web application	Used by candidates and provider staff, this application should offer features for every step of RPL process, from candidate's search for RPL programs, through submitting documentation, to assessor's upload of assessment materials and finalizing process.
Relational database	Should be used for storing all actions on every RPL process, both key information set and all other relevant data such as assessment materials, candidate uploaded documents, users, roles, etc.
Key information set publisher	Since key information set must be uploaded to future national RPL register by means of web service, this component will allow managing such upload. It can be implemented as separate and secure module of web application, or as a completely separate application.
CROQF search engine	Used by candidates for searching for RPL programs, this component is best implemented as separate module for web application. It should be built as a rich wrapper around CROQF registers, forwarding candidates search requests to CROQF web service and retrieving CROQF codes as responses.

In order to guarantee transparency and allow possible future audit processes, each operation on every RPL process should be recorded in form of action log that will allow auditors to retrace process from its beginning to its end. To further improve action log, users should be grouped into roles, shown in Table IX.

TABLE IX. MINIMUM SET OF ROLES

Role	*Available features*
Candidate	Uses search features to locate units of learning outcomes, qualification standards and RPL programs, applies for RPL and provides necessary information and documents. On successful completion, can print informal certificates.
Provider Coordinator	Manages provider's list of counsellors and assessors and assigns counsellors to RPL processes
Provider Counsellor	Validates candidate submitted information and documents, communicates with candidate, conducts interviews with candidates, assigns assessors to RPL processes and prints formal certificates.
Provider Assessor	Uploads assessment questions and corresponding results and reviews and decides on successful or unsuccessful completion of PL processes.

RPL system should use an external system for user authentication. Such system should provide authentication based on at least AAI@EduHr (Croatian higher education authentication infrastructure), with preferably other popular options available, such as Google, Microsoft, Facebook, Twitter, LinkedIn, etc.

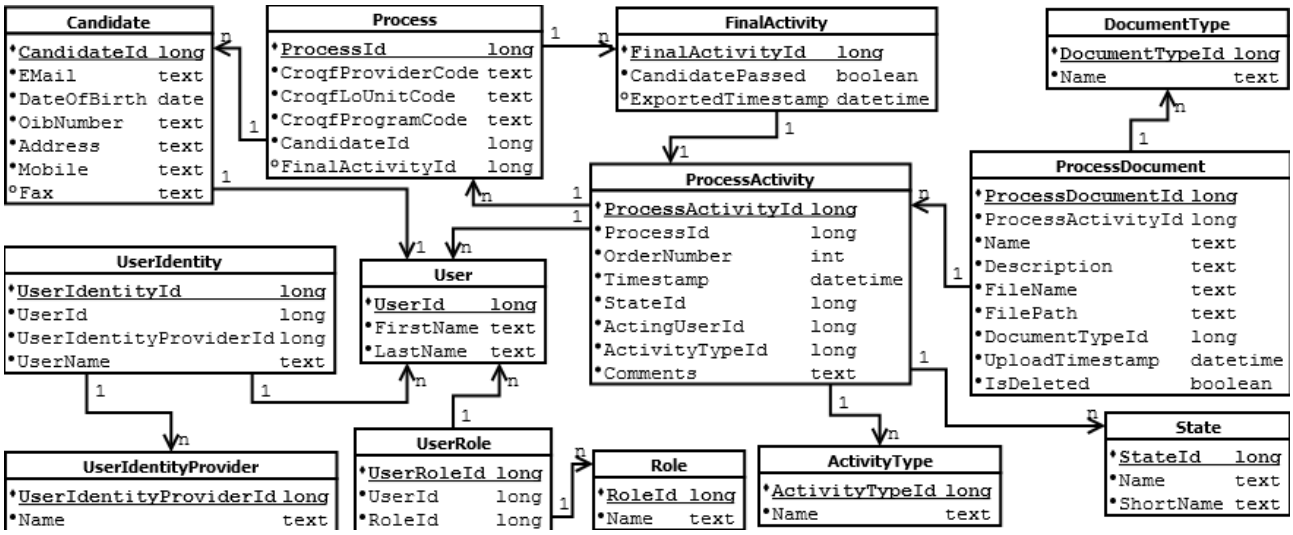


Figure 1. Relational model

B. *Relational model*

Relational model must be designed with action log as core requirement in mind. Figure 1 shows such relational model in its essential form. For brevity, only tables, mandatory columns with generic data types, primary and foreign keys and not null constraints are displayed. Actual platform-specific data types, additional constraints such as unique, default and check constraints, additional columns in existing tables and new tables should be added by implementer as deemed necessary, following good database design principles.

First core table is named Process and its purpose is to hold instances of RPL process, each linking CROQF codes for provider, unit of learning outcomes and RPL program to candidate. Each step of the process has its entry in table ProcessActivity. Each activity for the process has its type selected from a set of predefined types; it has an order number starting from 1 that should be used to sequence activities in timeline; timestamp when did the activity happen; state the process is in at the end of the activity, user who did the activity and user's comments. Table ProcessDocument is designed to hold uploaded documents, whether from candidate or from provider staff, with each document having a type set. Finally, fourth core table is FinalActivity which stands in 1:1 relationship with ProcessActivity. FinalActivity holds entries for only those activities that mark processes as completed and thus unchangeable. Stated is the reason why Process might point to such final activity, but only if the process is finished. Strictly, that foreign key is not necessary, but it is placed here in order to ensure easier querying. Lastly, FinalActivity has a nullable timestamp when this process' key information set was exported to national RPL register. Table X briefly explains other database tables.

C. *Important use cases specifications*

Although there are many more use cases applicable to a RPL ICT system, in this section only most essential ones will be outlined. Table XI describes them without going into much details.

TABLE X. EXPLANATION OF TABLES IN RELATIONAL MODEL

Table	*Purpose*
Role	A set of predefined roles from table IX
User	All users, regardless of their type
UserRole	Which user belongs to which role
UserIdentity Provider	A set of predefined providers, such as Aai@EduHr, Facebook, Google, LinkedIn, Twitter, etc.
UserIdentity	One or more identities for every user
Candidate	Only users that are candidates (1:1 relationship)
State	A set of predefined process states, such as CREATED, ACCEPTED, PENDING, VERIFIED, REJECTED, COMPLETED, FAILED, EXPORTED
Activity Type	A set of predefined activity types, such as: Initial Input of Candidate Data, Update of Candidate Data, Counsellor Assigned, Counsellor Messaged Candidate, Counsellor Verified Candidate, Assessor Assigned, Assessor Updated Assessment Documentation, Assessor Marked Process as Finished, Exported to RPL Authority
Document Type	A set of predefined document types, such as: Curriculum Vitae, School Diploma, Identity Card Scan, Assessment Question, Candidate Response, etc.

TABLE XI. MOST IMPORTANT USE CASES

ID	*Use case outline*
UC1	Candidate visits HTML5 web applications and uses browsing and/or searching application features to find appropriate Unit of learning outcomes, Qualification or RPL program from CROQF register.
UC2	Candidate signs in using any available authentication mechanism. Candidate initiates RPL processes for one or more Units of learning outcomes and provides necessary information and documents.
UC3	After signing in, candidate sees the state of her/his RPL processes, provides additionally information and documents requested by provider counsellors and prints informal certificates for successfully completed Qualifications.
UC4	Coordinator manages list of assessors and counsellors for provider, assigns counsellors to RPL processes and manages provider data available to potential candidates.
UC5	Counsellor validates candidate provided data and documents and communicates with candidate until they are validated. Once validated, assigns one or more assessors to process.
UC6	Assessor uploads assessment elements, such as questions, comments, results, and also marks RPL process as completed, either successfully or unsuccessfully.

VI. CONCLUSION

Research presented within this article, as well as key policy recommendations underpinning implementation of validation systems issued by Council of European Union, Cedefop, Unesco and others call for transparency, cost effectiveness and reliability of RPL processes in order to implement "equal value" based RPL provision. Starting from aforementioned wide and relatively nonconcrete implementation guidelines we recognized ICT support system for RPL as a means to an end, capable of supporting RPL candidates, providers and QA authorities as well as providing functional link to NQF and Qualifications standards required to support "equal value". Since we did not find any useful blueprints or ICT system's development recommendations through desk research of scientific articles and national RPL implementation documents in EU, we organized structured interweaves and survey research to better understand specific stakeholder's attitudes and requirements. As a result, we learned that predominant stakeholders' attitude goes in favor of strong documentation and transparency of RPL in Croatia in order to protect public interest and confidence in national qualifications (most of respondents strongly agree or agree mean=4,44 that all RPL data should be achieved for ex-post review, while similar majority mean=4,42 find RPL ICT system should be used also to store digitalized assessment results). Still, specific requirements of educational authorities (both ASHE and MSES) called for less centralized approach which would allow RPL providers to exchange only required minimum of data on each candidate and validation process with national ICT system, by use of its on-line interface or through interoperability with their existing applications. Furthermore, although most of the stakeholders through survey supported the idea that e.g. upload of digitalized assessment results for each RPL candidate is useful and justified in order for QA authority to have easy central access to relevant data for ex-post external audits, educational authorities as well as part of the providers do not share same idea mostly in order to protect candidate's privacy. In such circumstances use of RPL ICT solution is not rejected, but information and data that should be stored centrally and "controlled" by educational authorities is.

Proposed RPL ICT system design and its data model builds on results of presented research as well as other relevant recommendations, calling also for information provision for RPL candidates, importance of career guidance support and e-portfolios as an emerging tool. In order to produce future RPL ICT system for Croatian higher education, proposed model and blueprints could serve as a good starting point, but also one that lacks "real feel" test and inputs provided by very RPL candidates and RPL practitioners through interaction with the system. Due to the fact that Croatia still do not have legal framework for RPL and that there are currently no RPL providers nor candidates, further research on system usability from numerous roles is required in order to prepare documentation for full scale development of national solution. Still, with proposed approach focused on national database of all RPL relevant information, Croatia might be one of the countries fighting indicated EU level challenge of lack of RPL data for future development of national RPL provisions. For all aforementioned reasons, proposed ICT system design and data model, although developed through project in higher education and targeting mostly stakeholders relevant for this educational segment (level), requires no significant further development to cover all educational segments and become national RPL implementation infrastructure, if recognized by future by-low on RPL and future users.

ACKNOWLEDGMENT

This research would not be possible without team from Agency for Science and Higher Education who coordinated EU funded project within which work on implementation of RPL in Croatian higher education has been done. Thanks to Durdica, Emita and Aida. Thanks also to prof. Mile Dzelalija and University of Split for all the support.

REFERENCES

[1] European Centre for the Development of Vocational Training, *European Guidelines for Validating Non-Formal and Informal Learning*. 2009, p. 96.

[2] European Council, "SANTA MARIA DA FEIRA EUROPEAN COUNCIL 19 AND 20 JUNE 2000 CONCLUSIONS OF THE PRESIDENCY," 2000.

[3] European Commission, "Making a European Area of Lifelong Learning a Reality," in *COMMUNICATION FROM THE COMMISSION, COM(2001) 678 final*, 2001.

[4] Council of the European Union, *Council Recommendation*. Official Journal of the European Union, 398, Brussels, 2012.

[5] M. Dželalija, "Theoretical Basis for the Principle of Equal Value applied to the RPL and to Formal Learning," *Educ. UK*, 2011.

[6] Hrvatski Sabor, *Zakon o Hrvatskom kvalifikacijskom okviru*. Narodne novine 22/2013, 2013.

[7] B. Chakroun, "National Qualification Frameworks: From policy borrowing to policy learning," *Eur. J. Educ.*, vol. 45, pp. 199–216, 2010.

[8] S. Allais, *The rise and fall of the NQF: a critical analysis of the South African National Qualifications Framework.*, PhD thesis. Johannesburg: Graduate School of Public Management and Development, 2007.

[9] European Comission, Cedefop, and ICF International, "European inventory on validation of non-formal and informal learning 2014 - Executive Summary" 2014.

[10] European Comission, Cedefop, and ICF International, "European inventory on validation of non-formal and informal learning 2014. Thematic report: multi-level governance" 2014.

[11] Cedefop, *The validation challenge : how close is Europe to recognising all learning ?*, no. 2. 2014, pp. 2012–2015.

[12] "VITA - Software Development Concept and Report." http://www.vita-eu.org/uploads/media/03_1_Software_Development_Report, pp. 1–31, 2013.

[13] S. Sava, C. Borca, and L. Danciu, "Models of Quality Assurance in Evaluation and Validation of Competencies, for an Easier Access to Higher Education," *Procedia - Soc. Behav. Sci.*, vol. 142, pp. 176–182, Aug. 2014.

[14] R. Cameron and A. Miller, "Case studies in e-RPL and e-PR," *Aust. J. Adult Learn.*, vol. 54, no. 1, pp. 89–113, 2014.

[15] R. Cameron, "Recognizing workplace learning: the emerging practices of e-RPL and e-PR," *J. Work. Learn.*, vol. 24, no. 2, pp. 85–104, Feb. 2012.

[16] M. Balković, N. B. Lebarović, and H. Balen, "Prijedlog entitetsko-relacijskog modela podataka registara Hrvatskog kvalifikacijskog okvira," 2013.

Computer game Angry Birds in teaching physics

D. Robič* and R. Repnik*
* Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
robert.repnik@uni-mb.si, dominik.robic@student.um.si

Abstract -Information and communication technology (ICT) is very popular among children. A lot of children are playing computer games in their free time not knowing that some games are using game mechanics from actual physics. In this article we discuss how some computer games can be useful in teaching physics and which of them are appropriate for this purpose. Computer games are one of didactical tools that can make lessons more interesting for children and should not be excluded out of teaching. We check main characteristics of educational games and the main difference between none educational games. Our main focus is on popular Angry Birds in which we analyze the physics background of the game, where we try to determine different contents of physics and applicability in physics lessons on specific school level. We show an example proposal how to use Angry Birds in teaching physics in secondary school at topic projectile motion and which software is required to analyze data in the game. We also list additional examples with suitable physical content which are integrated in this computer game.

I. INTRODUCTION

Using computer in education and recreation is common nowadays [1]. Kids today are not only "digital natives" but are characterized with the term "always on generation". During the last decade many high quality e-learning materials were prepared in Slovenia for primary and secondary schools [2]. In those materials extensive use of multimedia and interactivity were used and even aesthetics of some learning materials was evaluated [3]. Should teachers produce such materials? From the analysis of the study programs they should [4]. Kids love such learning materials but as always it is evident that most of that technology is used for playing computer games [5]. For two decades people were trying to figure out why computer games are so motivating and why children spend so much time playing those [6]. The results were three features: challenge, fantasy and curiosity. Those three features are also very important in learning physics. Challenge helps at motivation to achieve our goals, fantasy helps at better imagination and explanation how things should work and curiosity drives people to figure out things that they didn't know. Some computer games use actual physics for their game mechanics and children are very eager to spend hours playing them. So why not use them as a teaching tools for physics [7]? A set of games that use physics as their game mechanics are very popular computer games Angry Birds. They are not specifically made for teaching though.

II. DIDACTIC ASPECT OF COMPUTER GAMES

As previously noted the use of ICT can be great motivational tool for teaching and learning, especially computer games can be great motivational tool. That leads to the games that were made specifically for teaching pupils. Computer games that were designed for learning are called educational games [8]. What make games »educational« are specific characteristics. Educational games should have educational content included in curricula that are based on level of users understanding. Educational games should also have appropriate methods for learning contents, which depends on nature of contents. For example we must distinguish between learning of knowledge, processes; procedures and casual principles. Each of those requires different learning methods that depend on content's complexity. Methods that are used for learning also affect game structure and game mechanics. Finally, what every educational game should have is a feedback system that provides players with information of learning success. In this description of key characteristics of educational games it is notable that game structure and game mechanics depends on game content and methods of learning [9]. This leads to another question. Why do children prefer playing classic computer games rather than educational computer games? Key element is that educational games are not primarily designed for fun only and major software companies do not develop educational games. Those companies have expert knowledge in computer graphics and games design but they have discovered that educational games are not commercial successful and revenue is too small for them. So to get good learning results with computer games we should either design and develop better educational games or find those computer games that are popular and could be used for educational purposes. In our case we focus on "Angry birds".

III. COMPUTER GAME ANGRY BIRDS

Angry bird's games are product from Finnish company Rovio entertainment. First game was released in year 2009 for Apple's iOS Android, Symbian and Windows Phone operating system. Since then Rovio entertainment upgraded the game that can also run on PC's and game consoles. It's addictive gameplay, comical style and low prize has increased its popularity irrelevant of the users age. Its popularity also encouraged the company to make

new sequels with different themes. From first game Angry Birds to latest sequel Angry Birds Transformers are listed eight more sequels, with different themes and game mechanics. Basic story in AB is about evil green pigs which stole eggs from birds which became angry. In game slingshot is used to shoot birds in a way to eliminate all green pigs (Fig. 1). Every level is limited with number of birds and their ability's which varies from bird to bird. Green pigs are sometimes protected with different building block which can be destroyed according to the different birds destructive properties. For example yellow birds can accelerate and are able to destroy wooden material. Each level increases difficulty and completion of some levels requires multiple trials and error. Despite increased difficulty on each successive level the game still remains attractive, addictive and motivates players to continue playing the game. Feedback of the success of the game is displayed with the stars. More stars you gain more successful you were in specific level. Final score is calculated based on your performance at specific level and is dependents on number of destroyed objects and number of birds used to eliminate all green pigs.

Figure 1: Different birds in the game have different properties, all we firstly shoot out from slingshot.

There are many sequels of Angry Birds and one is called Angry Birds Space where game environment is moved into space and gravity of objects affects bird's flight. At designing this sequel Rovio Entertainment worked with NASA which helped at programming gravity effects and also tested them in actual space [10]. Though games were not designed for educational use its game mechanics includes physics and makes them interesting for physics education.

IV. LEARNING WITH ANGRY BIRDS

We figured out that computer games Angry birds are appropriate for teaching physics so we will show an example how we can teach Physics of projectile motion with computer game Angry birds. But before we can check the physics of AB we have to make footage of AB gameplay and after that we have to analyze data in that footage. For that we will need some additional programs appropriate for classroom usage, for which we will give some examples.

A. LoiLo Game Recorder

Recommended software for making gameplay footage is a program called LoiLo Game Recorder. The program can be downloaded for free from their website (*http://loilo.tv/us/product/game_recorder)* and it can be easily installed on your computer. LoiLo Game Recorder is a program that enables us to record game sessions. It also supports Motion-JPEG file format that provides with the best balance between file size and image quality. For our purpose we recorded full HD videos and file size is still manageable. When program is installed simply start the program and the game in which recording will be made. Before playing a game press F6 on keyboard and program will start recording (Fig. 2). When playing a game is finished press F6 key again and the program will stop recording and save the footage in your PC's Video Directory. Now footage can be seen in program and it is ready for analysis.

Figure 2: User interface of program LoiLo.

At the benchmark test that was made using much known computer game Final Fantasy 14 the LoiLo Game Recorder showed the least performance loss compared to any competitor's non-free software [11]. To minimize the measurement errors at analysis of gameplay the footage must be smooth and without delays. To analyze the footage it is recommended to use program called Tracker.

B. Tracker

Tracker is a free video analysis and modelling tool built on the Open Source Physics (OSP) Java framework. It is designed to be used in physics education and can be easily run from USB drive. Requirements for using Tracker are small. For proper working it requires program Java 1.6 or higher which can be obtained on the internet. Tracker has a variety of tools to help user to analyze the data from recording where program shows what happened with physical quantity on the graphs. To analyze data from recording simply start program Tracker and open video that was made with program Loilo. Now different tools from program can be used to start the measurements. Tracker has a lot of tools to offer. Video analyzing tool is used to depart video frame by frame. Measurement analyzing tools is used for managing data in the graphs where we can set different graphs to see the final result of measurement. Calibration tools are used to set basic unit of measurement and set the starting point of measurements (Fig. 3) [12].

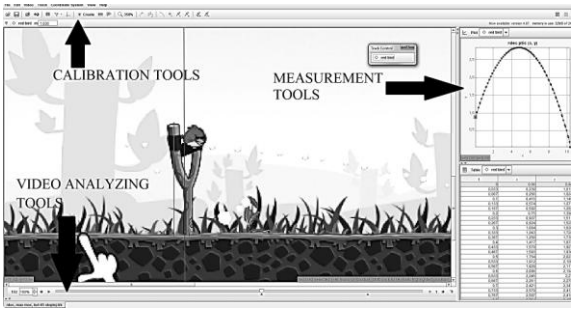


Figure 3: Program Tracker offers many tools for analyzing data.

Before acquiring the measurements for discussing physics problem, starting point of measurements needs to be appointed to define birds place in its space. This can be done with calibration tools where coordinate system is set in the foot of slingshot where also starting point is appointed. Setting a measuring unit is also mandatory and it can be also set in measuring tools. In this case it is recommend that measuring unit is set as slingshot size. When starting point and basic measuring unit is determined tracking tool can be used to track bird's movement in the footage. When tracking is finished the results of measurements are shown in the graphs.

V. PHYSICS OF ANGRY BIRDS

After analyzing the data with program Tracker physics of the game can be discussed. Projectile motion is a case of motion which can be described as motion in two dimensions, in vertical and horizontal way. In this particular case air resistance can be neglected because the game wasn't designed to include air resistance in projectile motion. The bird was shot with the slingshot where its initial speed ($\vec{v}_0$) is:

$$\vec{v}_0 = (v_{0x},\ v_{0y}), \tag{1}$$

where v_{0x} is size of the horizontal component of initial velocity and v_{0y} size of the vertical component of initial velocity [13].

Only gravity affects the birds trajectory that's why acceleration of angry bird is equal to the gravitational acceleration ($\vec{g}$) which is on Earth 9.8 m/s^2.
It's is known that the horizontal component of velocity is not changing in size because acceleration only got vertical component. That's why the movement of bird in time t can be defined as:

$$x_{finished} = (v_{0x}\cos\theta)t + x_{starting}\,, \tag{2}$$

where $x_{finished}$ is location where the bird has finished movement in time t in his horizontal path and $x_{starting}$ is initial location from where bird has started moving in horizontal path. θ is angle from which the bird was shot from the slingshot [13]. Result from the measurements are shown on the graph (Fig. 4) on which it is shown that body was moving in horizontal direction with constant velocity v_{0x} which is equal to 3.31 unit/s [14].

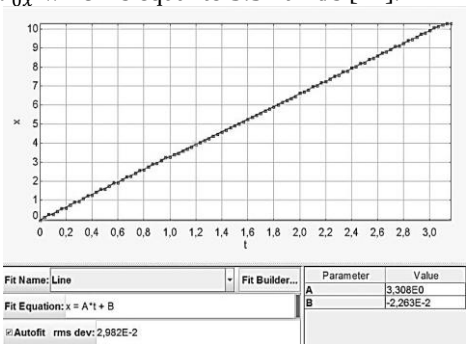


Figure 4: Measurements of horizontal movement.

For motion in vertical direction it is known that acceleration is equally accelerated, that's why motion in vertical direction can be defined as:

$$y_{highest} = (v_{0y}\sin\theta)t - \frac{1}{2}gt^2 + y_{starting}, \tag{3}$$

where $y_{starting}$ is starting hight from which birds was shot in vertical direction from angle θ and $y_{highest}$ is the maximum height that bird will reach in at time t. If the game happened on Earth acceleration would be equal to gravitational acceleration g. In the measurement can be noted that vertical motion fits to quadratic equation. From equation (3) we can see that acceleration (*A* in Fig. 5) in vertical direction is *a*/2. Which gives us as result acceleration which is equal to 1.9 unit/ s^2 . From measurment we can also read the vertical component of velocity (*B* in Fig. 5) which is 2.8 unit/s and height (*C* in Fig. 5) 0.8 unit [15].

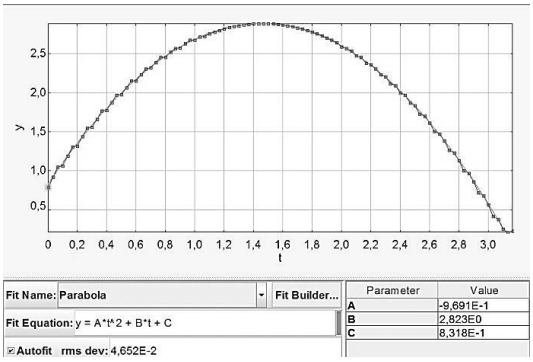


Figure 5: Measurements of vertical movement.

Size of games basic unit can be determined from measurements. In this case measured acceleration in vertical motion was 1.9 unit/s^2. Because we assumed that the experiment is happening on Earth, acceleration should be equal to gravitational acceleration which is 9.8 m/s^2 [14,15]. From this fact the size of basic unit can be calculated leading to result that slingshot was approximately 5.1 m high, because size of the slingshot was set as basic unit. Now that the basic unit of the game is known, velocity in vertical and horizontal direction can be also calculated by simply multiplying measured values with 5.1 m. This leads to results that v_{0y} is equal to 14.2 m/s and v_{0x} is equal to 16.7 m/s. Initial velocity can be also calculated using equation:

$$v_0 = \sqrt{{v_{0x}}^2 + {v_{0y}}^2}, \tag{4}$$

leading to result where v_0 was 21.9 m/s. Starting height $h_{starting}$ from which the bird was shot can be also read and calculated from measurements and is 4.2 m. When starting height is calculated also maximum height h can be calculated:

$$h = \frac{v_{0y}^2}{2g} + h_{starting} \tag{5}$$

resulting that maximum hight h was 14.5 m. It is also interesting from which angel the bird was shot:

$$\theta = \tan^{-1}\left(\frac{v_{0y}}{v_{0x}}\right). \tag{6}$$

Using previous equation for calculating angle the result is $\theta = 40.4°$. So we have all needed information to calculate also the range of bird's flight. For this purpose we use the following equation:

$$d = \frac{v_0}{g}\sqrt{{v_0}^2 + 2gh_{starting}}, \tag{7}$$

where range d equals 52.8 m. Similar results can be obtained reading measurements where range is 10.5 units which is 53.0 m. Range that we calculated is not exactly the same as the range that was measured. This can be can explained as an error of measurements.

VI. Applicability of Angry Birds

We have seen how we can analyze physics at red angry bird that has no special abilities. This type of analysis and understanding would be more appropriate for pupils in secondary school, but it can be also used in primary school. We could teach children in primary school how to use programs LoiLo and Tracker for simple analysis not only in Angry Birds but also in any other experiment footage. There is much more physics that Angry Birds can offer for physics lessons. For example it would be interesting to research, what is the initial acceleration and midair acceleration of yellow angry bird when we use its special ability. It would be also interesting to check physics background of blue angry bird where we would check what is happening with momentum when it splits in tree same sized angry birds if the mass of all birds is the same. Used materials in the game are also interesting for further analysis (Fig. 6). We could study how different angry birds affect the same material and on which material the influence of one specific bird is the most effective.

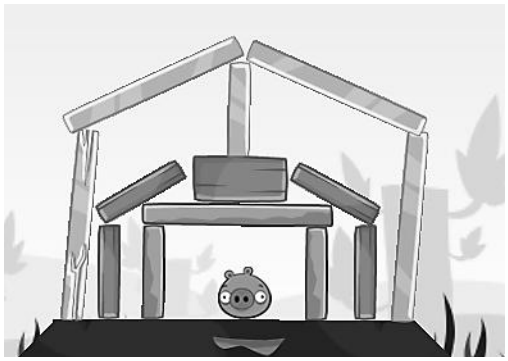


Figure 6: Different materials are used in the game Angry Birds, with different properties.

VII. Conclusion

In this article we determined that use of ICT in education is a necessary competence that every teacher should acquire during his education. In the present time and even more in the future, ICT *is*/*will be* one of the main teaching and learning tools. Collecting game experiences we also learned the difference between educational games and fun computer games. We found out that only some computer games (like Angry Birds), could be appropriate material for teaching because of its motivational fun factor. In educational games designers mostly forget about it because their focus is on teaching. Game mechanics is obstructed by learning processes, in which most children forget that learning new things can also be fun. We used two programs (LoiLo Game Recorder and Tracker) for analysis of selected sequences of the game. We showed an example how we could use computer game Angry Birds for teaching projectile motion where we explain physics of game mechanics. This can be done in secondary school but also in some cases in primary school physics course. At the end we checked applicability of Angry Birds and how it can be used for additional analyses in further physics lesions. Determine adequate games (high motivational fun factor, with real physics related game mechanics) and using them in teaching is of high importance, because the students will be more motivated for study in physics course and the knowledge will be better. Of course, nothing can replace the knowledge resulting out of real experimental work experiences in the teaching natural science subjects, especially physics. So the adequate educational games should be introduced in the teaching as supplement to experiment and not as substitution.

REFERENCES

[1] R. Repnik, B. Kaučič and K. Marjan, "E-Learning in the modern curriculum development," in *eLearning - Theories, Design, Software and Applications, Dr. Patrizia Ghislandi (Ed.)*, InTech, 2012, pp. 225-239.

[2] M. Krašna, "E-learning materials development projects," in *Proceedings of the 13th International Multiconference Information Society - IS 2010*, Ljubljana, 2010.

[3] M. Duh, B. Bedrač and M. Krašna, "Assessment of e-learning materials aesthetic design," in *MIPRO 2012*, Opatija, Croatia, 2012.

[4] M. Duh, T. Bratina and M. Krašna, "Elementary teachers competences for multimedia learning materials production," *Informatologija*, vol. 46, no. 4, pp. 333-342, 2013.

[5] M. Prensky, Digital Game-Based Learning, Minnesota: Paragon House, 2007.

[6] Ulicsak and M. &. Williamson, Computer games and learning: a handbook, London: Futurelab, 2011.

[7] R. Repnik, "Successfulness of traditional teaching methods in introduction of contemporary scientific discoveries within physics lessons in primary school- PhD Thesis," Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, 2012.

[8] I. Gerlič, "Fields of Computer Application in Education," n.d.. [Online]. Available: http://splet-stari.fnm.uni-mb.si/pedagoska/ivan/model98/eng/s1.htm.

[9] J. Enfield, *Designing an Educational Game with Ten Steps to Complex Learning - Doctoral dissertation*, Indianapolis: Indiana University, 2012.

[10] "Angry Birds,," Ltd, Rovio Entertianment, [Online]. Available: http://www.rovio.com/en/our-work/games/view/1/angry-birds. [Accessed 8 september 2014].

[11] LoiLo, "LoiLo Game Recorder," [Online]. Available: http://loilo.tv/us/. [Accessed 31 januar 2015].

[12] D. Brown, "Tracker: Video analysis and modeling tool," [Online]. Available: http://www.cabrillo.edu/~dbrown/tracker/. [Accessed 5 maj 2014].

[13] D. Haliday, R. Resnick and J. Walker, Fundamentals of Physics 8th edition, Cleveland: John Wiley& Sons, 2007.

[14] M. Rodrigues and P. Simeão Carvalho, "Physics Education," *Teaching physics with Angry Birds: exploring the kinematics and dynamics of the game*, 4 Ferbruar 2013.

[15] R. Allain, "The Physics of Angry Birds," WIRED.com, [Online]. Available: http://www.wired.com/2010/10/physics-of-angry-birds/. [Accessed 10 februar 2015].

Towards Blended Learning: A Case Study

Petri Linna, Harri Keto and Timo Mäkinen
Tampere University of Technology / Pori unit, Pori, Finland
e-mail: petri.linna@tut.fi

Abstract - University education is in the middle of a technological and pedagogical revolution. Traditional models are under considerable pressure to change. In this paper, we introduce a case where IT education was radically challenged by new possibilities. The case is a small university department, which runs a Master's degree program for an annual intake of about 60 students. The students study while working either full or part-time. Thus, the key challenge is to enable an efficient balance between work and studies especially by supporting distance learning. This means that the location and schedule of studies have to be flexible. One way has been to increase the multiformity of the pedagogical methods. For example, teaching is moving increasingly toward the flipped classroom method. In addition to the teachers' own video production, external academic Internet sources have been used and experiments have also been made of using external Massive Open Online Course materials for example in executing web development and database courses. This paper is a summary and analysis of the observations from the viewpoint of the students, teachers, and administration. Some plans and future intentions are also presented.

I. INTRODUCTION

During the last decade a lot has been invested in the utilization of different technologies in higher education teaching. However, now the turning point is being reached in which major changes can be seen in the pedagogy of teaching [1], in organizational structures, in the education needs of the labor force as well as in technologies.

In historical terms, society has just fairly recently moved from the industrial era to the information society. Now the direction is toward social, communicational, and technologically networked societies. Jan van Dijk [2] and Manuel Castells [3] use the term the network society, where communication and technological revolutions will affect societies, enterprises, and individuals in many ways (see also [4]). The world will be networked on many levels, which will also change the needs of education, such as more need for internationalizing with cross-cultural competency.

The education needs and structures have changed in different eras and attempts have been made to clarify and investigate future trends in many quarters, but mainly the trends are always somehow connected to the fast development of technology [5]. According to the NMC Horizon reports [6],[7], which focuses on the future of education, the trends of the near future in teaching are the rise of social media and social networks. This trend has also been noticed and tested in the case department [4]. The teachers' roles will change from traditional lecturer to personal education trainers and controllers. Teachers might take on new roles, for example lecturers who offer only mass teaching and producers of open and interactive educational material.

The trend of open education courses, materials, and their utilization will be huge in the near future. The multiformity of teaching, in other words blended learning will increase as different technologies and cloud services and applications especially in mobile devices develop continuously. Mobile devices will also change teaching in other ways, for example the ADP classes of schools will become a thing of the past when students bring their own devices (known as BYOD). Mobile devices also bring an opportunity for personalized studying and serious educational games in teaching. In personalized studying, the student's development is estimated, encouraged, and controlled with different techniques and only the necessary teaching materials are selected. The last future trend according to the report will be virtual and remote laboratories. Added reality and virtual environments bring new opportunities to teaching by removing the challenges of distance learning. [6],[7]

A strong increasing trend of digitalization is connected to companies. The viewpoint of universities is that it is important to know what the educational needs of enterprises are and also which are the most suitable techniques and pedagogies now and in the future. The Future Work Skills 2020 report [8], made by the Delphi technique, lists six main drivers which have an effect on the background of the need for future work skills: extreme longevity, the computational world, the rise of smart machines and systems, new media ecology, superstructured organizations, and globally connected worlds. These drivers lead to ten important skills, and examples of them are cross-cultural competency and virtual collaboration. One of the results is that workers will need to be adaptable lifelong learners.

Companies wish to have high-quality education for their employees, and that the education should correspond to their exact skills needs quickly. These educational needs can vary rapidly as companies react to global situations, as the Horizon reports [6], [7] had noted. Senge [9] introduces a five-step model of learning organizations, in which one of the steps is personal mastery. If organizations want to develop and grow, personal learning is always needed thus enabling an increase in human capital. The challenge is to educate rapidly and reasonably. Secondly, workers who are simultaneously studying would need to obtain as many different flexible methods as possible to combine company, home, and studying as well as possible.

Universities also have to pay attention to the challenges and strategic goals of the European Union [10] which include the global competition challenge where new skills are needed for the labor market. The goal is for 40% of the population to have a higher degree and for about 15% of adults to participate in lifelong learning.

The above-mentioned matters cause a global competition challenge between universities, when distances lose their significance and likewise the role of the implementer of courses (materials and the course itself).

The case university of this paper is quite a small department. The Tampere University of Technology, Pori unit, offers education leading to a Master's degree in Management and Information Technology. Postgraduate studies leading to the degrees of Doctor of Science in Technology and Doctor of Philosophy can also be completed. The major subjects are Software Engineering and Data Management, Network Management and Information Security, and Industrial Engineering and Management. The Pori unit is part of the University Consortium of Pori, which is a center of 2800 students, 170 experts and four universities operating networked in a multi-science environment. The focus is on multidisciplinary studying and adult education.

The Pori unit became a provider of multiform learning officially in fall 2013. This is relatively late in terms of the progress of the technique. Multiform learning was seen as a solution offering flexible education for students and a solution competition for new students. The solution seems easy, but multiform learning is not an actual solution yet, because it contains an innumerable number of different methods and ways of teaching. The right combinations must be found by trial and error.

In practice, the transition to multiform learning adds a large number of new questions to be solved. Furthermore, for the teachers the transition to multiform learning creates a great opportunity to develop new kinds of teaching pedagogies and use new technologies, but the resistance to change is a challenge. Here are listed some example questions, which should be solved: How are the courses scheduled? How are the physical distances solved? How are (remote, network, electronic) examinations arranged? How are the materials distributed? How are the training sessions and laboratories arranged? How is communication secured? How are student motivated to use eLearning? What is the commitment of students and teachers like? How are the needs for pedagogical change established and utilized?

The research question of this paper is how to enable multiform learning flexibly and effectively now and in the near future for students who are simultaneously working in enterprises.

In the second chapter the growing trend of open education is examined, especially MOOC (Massive Open Online Courses). Can it provide some solutions for multiform learning and teaching? Does MOOC enable the transfer to the flipped class pedagogical method? The Pori unit had already studied the possibilities of open education in the AVARAS project. Next, in the KASKU project, the utilization of open courses and of open materials in teaching will be studied in basic teaching as well as in continuing education.

In the third chapter of this paper, the case of the Pori unit is introduced. It describes students' and teachers' experiences of multiform learning and teaching, and analyzes the results.

The last chapter is conclusion and results of paper.

II. THIRD PARTY MOOCS AS PART OF CURRICULA AND MULTIFORM LEARNING

Massive open online courses (MOOC) are on online studying opportunities, which anyone can participate in despite their background, wealth, or place of residence. The courses usually last about ten weeks, consisting of short videos, exercises, and articles to study. In general, no mark is given when the course has been completed. Universities began to offer MOOCs on a large scale in 2012. [11]

The phenomenon began in the United States, and has led to several consortia for delivering their courses via the Internet. The best-known actors on the scene are Coursera [12] and Udacity [13], which are funded by a group of large investors. Their aim is to develop businesses around the courses. There are also non-commercial course operators of which the most famous is edX [14]. British universities have set up their own consortium, FutureLearn [15]. It was estimated that there were over six million MOOC students in the world by the end of 2013. In Finland, the pioneer is the University of Helsinki [16], which currently offers eight programming-related MOOCs. [11]

At the beginning of this year, the Pori Department of Tampere University of Technology established the three-year KASKU project to develop its educational processes. One of the work packages (WPs) of the project focuses on MOOCs. The aim of the WP is not to produce its own large-scale online courses but to utilize MOOCs of other universities as part of the department's curriculum.

The department in this case is relatively small. The annual intake is only about 60 students, who already have a bachelor's degree in some technical branch. An increasing number of students are working while carrying out their studies. Recently, the Pori Department has also offered international students the opportunity to conduct their studies on the campus. Despite the small number of students with a heterogeneous background, and with very limited chances to be present at lectures, the goal of the department is to offer a versatile selection of courses in its curriculum while keeping the costs at a reasonable level. Third-party MOOCs have been seen as a promising opportunity to achieve at least one part of the goal.

So far, four ways to apply MOOCs have been identified: 1) A third-party MOOC implements a standard course directly; 2) A MOOC replaces a standard course for international students; 3) A MOOC implements a course, of which the content is not completely fixed; 4) MOOCs or their modules are used as parts of a standard course. The flipped classroom is seen as a promising pedagogical approach when applying MOOCs. It

describes a reversal of traditional teaching: students first study new material outside of class, and class time is used to do the harder work of applying that knowledge through problem-solving exercises, for example [17].

In the Pori Department, an example of the first case is the Server Side Web Development course, which is based on the MOOC of University of Helsinki [18]. The MOOC is in Finnish, and it offers students a study book and a set of automatically checked exercises. Teachers who apply the course will also receive ready-made questions for examinations.

The example of the second case is the Databases MOOC offered by Stanford University [19], replacing Pori's Introduction to Databases (in Finnish) for international students. The Stanford course is modularized into several mini-courses that offer the flexibility to apply only certain parts.

The Pori curriculum includes a course, of which the content is not fixed, called Special Topics in Software Engineering (SE). Such a course can be implemented by any proper SE-MOOC, being an example of the third way to apply MOOCs. For example, the School of Information Sciences of the University of Tampere gives credits for the completion of any appropriate Coursera course [20].

There are standard courses that do not match any available MOOCs. However, there might be MOOCs or MOOC modules that cover parts of the course. In Pori, the basic course of Web Development represents such a case, discussing several topics of the theme briefly. For example, Udacity [13], Code School [21], and Codecamedy [22] offer introductory level web development courses online. The course material of Udacity can be accessed free of charge, but students have to pay for extra services such as guidance and a certificate. Code School charges for their courses with a few exceptions while Codecamedy's offering is free.

III. CASE, PORI UNIT

During the past two years, the Pori unit of the Tampere University of Technology has been making the transition towards the utilization of multiform teaching methods. From the viewpoint of an educational institution this can be seen as a focusing of the strategy with technological possibilities. To concentrate especially on students in working life means strong support for flexible studies that are independent of time and place. This means the deployment of modern communication technology but also a methodological revolution.

At the starting point, some fundamental decisions and presumptions were made. Firstly, a new official alternative path of studies was released. It included a promise to organize teaching so that studies alongside daily work would be well supported. It should be noticed that the decision and the announcement was made even before practical detailed plans had been made. Because this was a public marketing level promise and both new organizational and methodological practices should be implemented, a strong analogy to the radical process improvement approach can be drawn (see [23]). It was trusted that the practical arrangements would succeed if the vision was clear enough. This relied mostly on the view that the most critical part of the practical arrangements was to make a suitable schedule for the involvement of the students in the teaching. Secondly, an assumption was made that the pedagogical challenges could be worked out by individual courses and teachers. Later on this will be explained as some kind of experimental way to put theory into practice. This kind of strategy relies on the motivation and interest of individual teachers. Thirdly, the implementation had to start almost immediately. So it was not possible to acquire new technology or infrastructure during the preparatory work period. The teaching and the delivery of the material had to be based on the existing infrastructure and available tools.

The practical implementation started as follows. It was planned that an alternative path student would be able to graduate three or four years from the starting point. Because of the available time and resources it was clear that not all major subjects at the Pori unit could be offered on the alternative path. The same limitation concerned secondary subjects. Therefore the decision about major subject and secondary subject had to be made before launching the marketing activities. For those lectures and exercises where the attendance of the students was desirable or compulsory, the teaching was scheduled for Thursday evenings and Fridays. Saturdays were also originally expected to be required but during the planning this was left out. The idea was to leave enough time for personal homework and studies between contact sessions. The detailed schedule was then announced.

In comparison to the previous implementation of courses, fewer contact sessions were included in the new implementation. Careful analysis of content was needed. The decision on what to include in the contact sessions and what to leave to the responsibility of the students themselves was left to the teacher in charge.

The other main difference from the past was the increasing utilization of video material. Although first efforts to produce videos existed, there was no wider experience about using videos as a teaching material. The production of videos was started first by a small group of active teachers. The first videos were screen capture recordings with lecture slides and the teacher's voice. The free version of Screencast-O-Matic software was first used as a tool. The other technology was the Adobe Connect computer conference system, which was available via the Finnish University and Research Network (Funet). The latter technology was also used in real-time teaching. Very soon it was understood that a more powerful editing tool was needed. The commercial Camtasia Studio software was acquired for each teacher. The open source learning platform Moodle had been used in courses before, so it was natural that delivery of videos was implemented in Moodle.

After half a year, a summary of the teachers' experiences with the technology was made. The main findings were:

- Free and open source based tools are available. Their usability is good enough to be used as a starting point to capture experiences
- The limitations of free and open source based tools were the lack of editing properties and the limited length of video
- Professional video production software is recommended
- The encoding of video should be standardized within the organization

After the end of the first academic year, an overview of video material was made. Three kinds of videos were produced by five teachers for seven different courses:

- 7 short videos, length of 5 minutes or less
- 25 videos from 5 to 20 minutes in length
- 36 videos of over 20 minutes in length

The number of videos of over 20 minutes comes mainly from two mathematics courses with total coverage by video. In addition to the above, 15 external videos of different length were used and one course also utilized international MOOC material. Both of these categories are expected to grow in the future.

Also, after the first academic year a questionnaire was made about students' experiences. The main findings were:

- The educational videos were regarded as especially useful because of time- and place-independence and the possibility of repetition.
- The training done freely on the network were considered useful.
- More video material is hoped for.
- Giving of feedback should be developed.

Training in mathematics were also tried using Adobe Connect sessions. It was noticed that:

- The students are surprisingly passive, the relation of active and passive students was about 3/20
- The students expect ready-made solutions
- It is distinctly more difficult for the teacher to get a feel for the students' know-how compared to contact teaching
- The students do not utilize all the possibilities offered

During the discussion it was stated that it is recommended to keep contact training to ensure that the students will work. Feedback must be collected during the course and the student's know-how must be regularly ensured by measuring it somehow. Also, attention has to be paid to the students' activation.

IV. CONCLUSION

The universities are under a great deal of pressure to change. It is of the utmost importance to the case unit to offer the education which responds to the future skills need in companies and is flexible for students working at the same time. It is important to be able to utilize the supply of open teaching from major universities and find the means to cope with international competition by universities. It is also crucial to find the means with which to offer more tailored methods for studying.

The paper presented experiences of multiform learning and teaching in the Pori unit. The technology has not been found a problem. There are a lot of free or cheap tools available. Instead, the challenge lies on the pedagogical side. The teachers have had to find a suitable multiform teaching way for their own course. In some cases, the contact and situation awareness with the students has disappeared. Furthermore, the change in attitude has been a significant factor, for the students as well as the teachers. Changing from the old accustomed ways takes time.

The paper showed the development plans for the near future joining as a utilizer and provider of open course and materials. MOOCs can be seen as the transitioning to a networking society where multilevel networks help each other. Students and teachers around the world can help each other globally without ever seeing each other face to face. MOOC is not revolutionary technically but it is in terms of its massiveness and ideology. Its huge spread in the U.S.A and soon in Europe will be an amazing educational driver for change and it will probably change universities in very many ways. For example, it creates totally new possibilities for students to personalize their own study content and timetable, to become internationalized through the network, and at the same time to work flexibly in companies.

The results of transferring to multiform learning were successful. The numbers of applicants have increased. In the developing of the teaching it has been now invested considerably more. The KASKU-project is an example of the significant development projects which new teaching ways are developed as flipped class method and MOOCs. The lecture recording system, Echo360, and an electric examination are the next developing steps enabling more flexibility to the students. The teaching of the mathematics has received special praise from the students because the videoed lectures will make the viewing of lectures possible again many times. The mathematics has been a traditionally big stumbling block, but now this challenge can be facilitated. The transition to the multiform learning has helped to correspond to the case university unit to national but in the future also to an international competition for the students. After two year experiences the teaching is moving towards the flipped classroom method where the focus is on the ensuring of the learning and the studying is performed more communal using web-based collaboration tools.

REFERENCES

[1] Andy, H., Ann, L., Michael, F., & David, H. (Eds.). (2009). Second International Handbook of Educational Change. Second International Handbook of Educational Change. Springer.

[2] Dijk, J. (2006). The network society: social aspects of new media. SAGE.

[3] Castells, M., (1996). The rise of network society 1st, Blackwell publishers, Inc. Cambridge, MA, USA.

[4] Jaakkola H, Linna P, Henno J, and Mäkelä J, (2011). (Social) networking is coming Are we ready? MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, pp. 1133–1139.

[5] Jaakkola H, Brumen B, Henno J, and Mäkelä J, (2013). Are we trendy? MIPRO, 2013 Proceedings of the 36th International Convention. IEEE, pp. 529–537.

[6] Johnson, L., Adams Becker, S., Estrada, V., and Freeman, A. (2015). NMC Horizon Report: 2015 Higher Education Edition. Austin, Texas: The New Media Consortium.

[7] Horizon Report Europe: 2014 Schools Edition. Luxembourg: Publications Office of the European Union, & Austin, Texas: The New Media Consortium.

[8] Davies, A., Fidler, D., and Gorbis, M. (2011). Future work skills 2020. Institute for the Future for University of Phoenix Research Institute.

[9] Senge, P. M. (2014). The Fifth Discipline Fieldbook: Strategies and Tools for Building a Learning Organization. Doubleday Religious Publishing Group.

[10] Council conclusions of 12 May 2009 on a strategic framework for European cooperation in education and training ('ET 2020'). Document EUR-Lex 52009XG0528(01).

[11] Hiidenmaa P. (2013). Jos vastaus on MOOC, mikä on kysymys? (If the answer is MOOC, what is the question?). Report. University of Helsinki. In Finnish

[12] Coursera. www.coursera.org. (Accessed 22-Feb-15)

[13] Udacity. www.udacity.com. (Accessed 22-Feb-15)

[14] edX. www.edx.org. (Accessed 22-Feb-15)

[15] FutureLearn. www.futurelearn.com. (Accessed 22-Feb-15)

[16] MOOC.fi. (Accessed 22-Feb-15)

[17] Flipped Learning Network. flippedclassroom.org. (Accessed 22-Feb-15)

[18] Web-palvelinohjelmointi. (Web Server Side Development). University of Helsinki. MOOC in Finnish. mooc.fi/courses/2014/wepa/. (Accessed 22-Feb-15)

[19] Databases. MOOC. Stanford University. class.stanford.edu/courses/DB/2014/SelfPaced/about. (Accessed 22-Feb-15)

[20] Coursera courses. University of Tampere. www.uta.fi/sis/tie/coursera (Accessed 22-Feb-15)

[21] Code School. www.codeschool.com. (Accessed 22-Feb-15)

[22] Codecamedy. www.codecademy.com. (Accessed 22-Feb-15)

[23] Davenport, Thomas H., (1993). Process Innovation. Reengineering Work through Information Technology. Enst & Young, Center for Information Technology and Strategy. Harvard Business School Press. Boston, Massachusetts, USA.

The use of HILL online e-learning environment in language teaching

Mari Linna and Annika Valo
Information Technology, Satakunta University of Applied Sciences, Pori, Finland
mari.linna@samk.fi

Abstract - Due to an increasing number of adult students there is a need to create new forms of teaching. Blended or online learning is nowadays current practice also in language teaching, with new technical solutions constantly being taken into use. At Satakunta University of Applied Sciences (SAMK) in Finland, HILL, an online lecture system and e-learning environment is used in teaching. The aim of HILL is to provide adult students with an e-learning environment that facilitates combining of studies, work and family life. Students can follow lectures via HILL from anywhere and interact with the lecturer or participants. Sessions can also be recorded. The HILL concept was developed in cooperation between SAMK, IBM, Cisco and AKG. It has been used in teaching since 2010. For example, in November 2014 there were about 300 HILL sessions and 3000 users, and the number is increasing rapidly. Languages have been taught in SAMK via HILL since 2012. This is a case study concentrating on the possibilities and challenges that blended learning via HILL has brought into language learning.

I. INTRODUCTION

The use of electronic educational technology has brought many possibilities as well as challenges into teaching and learning. Studying online started to become popular in the end of 1990s, with the aim to make studying equal for all, despite of a student´s residential area [1].

The terms online learning, e-learning, multimedia learning, computer-based learning, Internet-based learning and virtual education can be used when talking about computer-assisted language learning. Blended learning can refer to any combination of different methods of learning, different learning environments and different learning styles [2], but in it, students learn at least in part via digital and online media with some element of student control over time, place, path or pace [3]. Students study at different times and in different places, but sometimes meet with other students and the teacher (e.g. lessons can be held via a video conference program and the teacher can be available online during certain hours) [1].

Blended learning offers new possibilities to education, especially to adult education [4]; it saves time and travelling costs, helps students to combine studies, work and family life and makes it possible for students to study according to their own timetable. [1] reminds that in order to succeed, online learning also requires a functional technical environment.

There is no one way to learn a language, just as there is no one way to teach it [2]. The rapid growth in the use of learning technologies, particularly the use of the Internet and Web-based communication, is providing language teachers and students with many more opportunities to explore the most suitable mix of teaching for a given task [2].

Blended language learning (i.e., integrating the use of technology into classroom-based learning and teaching) is still a relatively new concept, but when "appropriately" implemented, blended learning can significantly improve the learning experience [2].

In a study of students who have participated in Web-based distance learning in the United States (Noel-Levitz, Inc.,2006), students responded that the quality of instruction, the responsiveness of e-teachers to students' needs, and timely feedback from e-teachers are important factors in e-learning [2].

Indeed, in online and blended learning, communication is regarded to be an essential factor through which a teacher can support students, their motivation and commitment to studies [4]. The lack of key tools for human communication - including tone of voice, eye contact, body language, and so forth – create challenges for online learning [5]. Without communication, online learning can become simply a way of sharing information, but through communication it can become a real learning and working environment [1].

Blended learning can be said to provide a more individualized learning experience, more personalized learning support, support and encourage independent and collaborative learning, increase student engagement in learning, accommodate a variety of learning styles, provide a place to practice the target language beyond the classroom, provide a less stressful practice environment for the target language, provide flexible study, anytime or anywhere, to meet learners' needs and help students develop valuable and necessary twenty-first century learning skills [2].

The role of a teacher and a student has changed in online learning. [6] defines teacher as someone who supports and facilitates students' learning. Traditional teacher responsibilities such as generating resources and leading discussion shifts to learners in online learning. The focus has moved from passive transferring of knowledge from teachers to students taking a more active role in their learning [6]. E-learning also takes much

more self-discipline and self-motivation than a face-to-face class, and that blended learning requires a new way of mastering a teacher´s time management.

Students are seen as active participants who manage their own learning experience [1, 7]. They are not passive knowledge-*absorbers* but active knowledge-*generators* who assume responsibility for constructing and managing their own learning experience. However, it can be added that for many adult learners who are used to education in a lecture-based environment, interaction and collaboration is not intuitive and they may feel comfortable in a passive student role [7].

II. THE HILL CONCEPT

To support blended learning, HILL, an e-learning environment has been in development in the Huittinen unit of the Satakunta University of Applied Sciences (SAMK) since the 1990s.

SAMK's adult education in business administration was centralized in Huittinen in 2010. As 97 percent of the almost 400 adult students in the Huittinen unit study while working, SAMK was looking for a solution that would allow adult students to attend classroom teaching without being physically present at the school. This would make it easier for them to fit their studies, work and family life together [8].

The solution was developed by combining different technologies under the HILL online lecture system, which is the result of cooperation between SAMK's Faculty of Business and Culture Huittinen, IBM and Cisco. Based on a video conferencing system, the solution gives you a realistic feeling of being in the classroom without actually being there. The teacher can focus on teaching like in a normal classroom situation [8].

The first classroom utilizing the new technology was introduced at the Huittinen campus in August 2011 and the second at the beginning of 2012 [8].

The classroom where the teaching takes place is equipped with a Cisco video conferencing system including a display screen, HD cameras and microphones, as well as a SMARTBoard interactive whiteboard. The teaching situation is transmitted online by the Cisco Webex Meeting Center [8]. Sound technology is provided by AKG.

Only the teacher needs to be present in the classroom. Students need a multimedia computer and a working Internet connection. Through their browsers, students can see what is going on in the classroom, listen to the discussions of all parties, participate in the discussions and see the material that is being distributed or written on the whiteboard in the same way as a student present in the classroom [8].

In HILL concept, it is possible to teach online either in Meeting Center or Training Center. In Training Center, students can be divided into small groups, which is useful for practicing oral skills in a language (small group discussions, etc.).

Since January 2015, almost all language teaching except certain written exams is conducted via HILL. HILL enables work in pairs or in small groups as small group sessions for a desired number of people in each group can be formed. This is vital for language learning as students can practice their oral skills in small groups, just like in a class room setting.

III. STUDYING LANGUAGES ONLINE VIA HILL

In SAMK, the students study different languages at least three or five credits, depending on the degree programme. The obligatory languages are English and Swedish. Before entering the obligatory language course in English or Swedish, students' language skills are tested. If they do not reach the target level - B1 in Swedish and B2 in English in A Common European Framework of Reference for Languages - [9], a refresher course is recommended for them. In addition, there are courses in German, French, Russian and Spanish available. Most courses are studied in a traditional class room setting, but adult students in Huittinen unit study in the HILL e-learning environment. Also in other SAMK units, HILL is used in studying languages but not as widely as in Huittinen. The students in Huittinen are studying towards a degree in Business Administration, and they complete their whole education via HILL.

Until the beginning of the year 2015, HILL was used in language studies in the same way as in all other subjects. One course included three three-hour lessons. During those lessons, students were able to participate either by being present in the class room in person or online via HILL. It was also possible to watch a recording of the lesson later on. But, as students did not get enough practice in their oral skills this way, a change was needed. From January 2015, a new form of studying was introduced.

In the new mode, there are only two sessions that students attend in the classroom: an introductory lesson and a written exam. The student can participate in the introduction lesson also via HILL.

When studying online it is essential to have a good, functional learning platform. In SAMK, the virtual learning environment Moodle is used. After the introduction lesson, the students start studying in Moodle according to the instructions by the teacher. During the course the students have to participate in at least two or four, depending on the course, obligatory oral skills training lessons via HILL. In HILL there is a Training center for this purpose. During these oral skills training lessons, pronunciation, oral skills and communication skills are studied. The students are given several options for times of the lessons where they can choose the best one for themselves. The oral lessons are held mainly on Fridays or Saturdays or later on in the evening during the weekdays. This way the students can combine their studies and working life.

Besides these obligatory oral skills training lessons, students are given a possibility to enter voluntary lessons, so called "question lessons", where students can ask questions about for example exercises, or anything related to the course. Especially grammar topics can raise questions where students might need some support from the teacher. During these voluntary lessons, the teacher is available via HILL answering the students' questions and explaining unclear issues. HILL is also used for having oral presentations and oral exams. Written exams are still taken traditionally in a class room.

Studying foreign languages has many aspects: reading and vocabulary, oral skills and pronunciation, writing, listening comprehension and communication skills. Combining HILL and Moodle has created an efficient way of blended learning in foreign languages. Language teachers in SAMK have completed the Eurolta-Eurovolt education. Eurolta-Eurovolt concentrates especially on the problematic questions in teaching languages online [10]. All of these aspects mentioned above are taken into account when planning a blended learning language course. It is rather straightforward to create written exercises, e.g. reading comprehensions, vocabulary exercises and writing essays, in the Moodle, but for practicing oral skills, pronunciation and communication skills, HILL is ideal. Audio and video exercises in Moodle are also used. They are seen as a good way to practice pronunciation and also communication skills if worked in pairs. Different kinds of online language programs, grammar sites and other linguistic websites are used. When possible, such textbook is chosen which includes electronic material, e.g. listening comprehension exercises, read texts and the right answers that can be put online. These kinds of textbooks are very suitable for blended learning as they give students a freedom to choose when and where they want to study.

During spring 2015, four language courses have followed this new way of studying languages online via HILL: a refreshers' course both in English and in Swedish and an obligatory language course concentrating on specific vocational language, both in English and in Swedish. Besides students from Business Administration, students from Health and Welfare have participated in the refreshers' courses. The language courses concentrating on vocational language learning have been intended for Business Administration students in the Huittinen unit. They complete their whole education via HILL, but the students in Health and Welfare take only some courses or do parts of a study module via HILL. The extent of refreshers` courses is two credits and the obligatory vocational language courses are five credits. This spring there have been 154 students participating in these courses. The students were asked to fill in a questionnaire about blended language learning via HILL. 45 students returned their answers. In the questionnaire, the students were asked 1) to describe blended language learning via HILL, 2) to write about the similarities and differences compared to traditional studying in a class room, 3) to name pros and cons in blended language learning and also 4) to make suggestions for how to develop language teaching via HILL in the future.

Students described studying languages via HILL mainly in a positive way. Most students used phrases like efficient, easy, activating, interesting, pleasant, meaningful, excellent and good. They described studying being different and unusual but surprisingly easy, fun, relaxing, handy and inspiring. They felt that they had got teaching of good quality. The students explained that they liked independent distance studies and described HILL as a flexible and functional concept which makes it possible to combine studies and working life.

Language teachers see HILL as a very efficient answer to the challenges among blended learning. Using HILL requires that teachers are familiar with the technology to be able to take advantage of the different possibilities that the HILL concept offers. They also need to be aware of the pedagogic solutions that are suitable and meaningful for teaching languages online via HILL. In addition, they need courage to exploit new possibilities that they have not used before.

IV. STUDYING LANGUAGES ONLINE VERSUS TRADITIONAL STUDYING IN A CLASS ROOM

There are similarities as well as differences in blended language learning compared to traditional studying in a class room. Most students were of the opinion that studying online requires more from a student; more active involvement, motivation, self-discipline and skills of an independent way of studying. Some students felt they would have needed more support from the teacher. Some people wished for more feedback from the teacher; you can feel quite alone studying online. Especially if a long time has passed since a student has studied languages, he or she would need more support from the teacher and feel more comfortable with traditional face-to-face classes.

The students found studying online more efficient since you really concentrate on studying, rather than for example chatting with the person sitting next to you in the class room. You are able to concentrate better on the actual topic. Some also mentioned that they wanted to study more in order to be able to participate more actively during online HILL lessons. However, one student said that it is easier to skip lessons and stay in the background via HILL.

Several students mentioned that they are more courageous using their oral skills online via HILL than in a traditional class room setting. They found it easier to practice their communication skills online. They felt it was somehow more relaxing than in a class room. They could let go of needless self-criticism more easily; they felt it was easier to ask and make comments online and they did not have to be afraid or ashamed of failure in the same way. Studying online was also described as more democratic and flexible than studying in a class room. The students felt that online learning requires clearer instructions of the teacher, e.g. in Moodle. A student has

to be able to get the idea of an exercise independently without the help of a teacher.

Students saw also that one big difference was a possibility to get a fast answer to your question. When studying online you had to sometimes wait longer in order to get the answer while in a class room you could have got the answer right away from the teacher. One difference mentioned was also the fact that you did not get to know other students or the teacher as well as during traditional contact lessons. Forming personal contacts online is more challenging. Some also mentioned that it was sometimes difficult to concentrate on teacher´s pronunciation via HILL having headphones on than in person. Sometimes the voice is not clear enough and it makes listening more demanding.

Yet, some students saw that there are not so many differences when the technology or recording of a lesson is working. Also, according to some students in the study it is important that students possess quite good computer skills. Some students mentioned that technical problems have inhibited their learning sometimes; for example, sometimes the students' audio system has not worked properly. They find studying rather similar to traditional studying in a class room. Compared to a traditional class room a few described online learning being challenging and one found it lonely. Someone missed meeting other students and teacher in person. Also, someone felt it is a bit difficult to take part in common discussions naturally, so he or she rather stays silent. Also a student commented that it is just too easy to stay silent when online.

The students found several positive aspects in online learning. The most typical answers were related to time and place. It was very good that they could study when it is best for them. They can create own timetables for studying. They can also study where they want. They don´t have to use time or money travelling to school. It was also mentioned that studying at home, in a safe and familiar environment creates comfort. You did not have any extra voices during lessons via HILL like in a class room. Having headphones on made it easier to concentrate on the lesson without outside distractions. It was peaceful to study.

Studying online was also described as a more efficient way of learning. Since you have to take more responsibility of your own studies you also learn more. Studying online can be more rewarding because when you have to challenge yourself and then succeed, it feels better. Someone also mentioned the fact that you can participate in lessons if you are sick. This is true, you might be able to participate in e.g. one hour's lesson via HILL but you would not be able to drive to school. Students mentioned that you can also use online dictionaries and other online materials better all the time while studying online. It was also found as a positive thing that you can listen to recorded HILL lessons or other study videos several times, take breaks if you need, and then continue again. The students were very pleased with the HILL concept from the viewpoint of oral skills training lessons and "question hours" where they can ask linguistic questions.

For a teacher, the same kind of positive aspects are appreciated. Also, for a teacher HILL provides flexibility: you do not necessarily have to travel to school to give lessons but you can also give them from home, for example. Teachers can plan their own timetables more freely and work during the time of day that suits them best. Using HILL and blended language learning creates interesting challenges to a teacher: he or she constantly needs to search for better online learning solutions.

The cons from the point of view of teachers include that there is a lack of face-to-face communication. For example, teachers cannot notice it if someone does not comprehend something because they cannot see the students in the same way as in a class room as they are not in the same room. The students themselves need to be active and inform the teacher about any unclear issues and ask if they have questions. Online teaching also requires more time and effort from the teacher, and more time is needed also to answer email messages since students contact teachers more frequently than when the teacher and students regularly meet in a class room. Also, time has to be reserved for learning to master the online teaching system properly.

Getting to know students and forming teacher-student relationships is more challenging for teachers when teaching online. However, HILL helps in this as teachers can hear and see the students almost as if they were present in the class room.

V. CONCLUSION

There are many different online learning solutions for learning languages, and at SAMK, HILL has answered adult students' needs as students are able to combine their studies, work and family life. HILL has also enabled blended language learning, and particularly well the practice of oral and communication language skills. With HILL, adult students are able to stay in contact and in interaction with the teacher and other students, which increases the sense of belonging to the same group.

Due to HILL and the idea of blended language learning, better results are being achieved in learning languages. In blended learning, it is possible to take different learning styles and students' personalities into account. In that way, teacher can create different kinds of exercises and learning experiences and make it easier for different kinds of learners to learn. Also shy and sensitive students have a possibility to express themselves online and in that way achieve better results than before. Blended language learning requires surely more of a student but it can also give more.

The learning results have reached a better level also because of a change in the role of a student. They no longer get all information given, instead they have to work harder for it and take more responsibility for their

own learning. Through responsibility comes motivation and so, more learning takes place.

Blended learning requires more from a teacher than traditional teaching in a class room; active development and re-evaluation of one's own teaching methods and way of being a teacher; mastering the technology involved; initiative and willingness to develop methods for students to learn in the best way possible. At least in the beginning, teachers need additional resources for teaching online, and it is imperative that teachers moving into teaching online should get all the possible support and training needed.

The basis of blended learning is a stable technical solution, which also came up in many replies.

During spring 2015, the way languages are being studied online via HILL has been renewed. Students are still wishing for even more oral language lessons and "question hours". Also, according to students, more teachers (also teachers of other subjects than languages and also in other units in SAMK) should take small group discussions into use in HILL.

Blended learning offers various possibilities for language learning. The HILL concept and methods of teaching are constantly being developed to even better meet the needs of modern language learning.

REFERENCES

[1] E. Kalliala, "Verkko-opettamisen käsikirja", Jyväskylä, 2002, pp. 13, 18 – 19, 23 – 24, 30 – 31, 76, 89, 97, 128.

[2] D. Marsh, "Blended Learning: Creating Learning Opportunities for Language Learners", New York, 2012, pp. 3-6.

[3] Blended learning. In *Wikipedia*. Retrieved February 16, 2015, from http://en.wikipedia.org/wiki/Blended_learning.

[4] V. Meisalo, E. Sutinen and J. Tarhio, "Modernit oppimisympäristöt", Juva, 2000, pp. 154 – 56.

[5] R. Donnelly and F. McSweeney, "Applied E-Learning and E-Teaching in Higher Education", London, 2009, pp. xiv.

[6] H. Koli, "Verkko-ohjauksen käsikirja", Saarijärvi, 2008, pp. 15 – 16, 20.

[7] R.M. Conrad and J.A. Donaldson, Engaging the Online Learner: Activities and Resources for Creative Instruction, San Francisco, 2011, pp. 10.

[8] "Story: IBM Global Technology Services Case Study: The HILL online lecture system revolutionizes learning" , Helsinki, 2012.

[9] Council of Europe, "Eurooppalainen viitekehys. Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys." ("Common European framework of reference for languages: learning, teaching, assessment"), Porvoo, 2003, pp. 46 – 48.

[10] Eurovolt via VLE, 2015. Web. Retrieved February 6, 2015, from from http://www.icc-languages.eu/accreditation/eurovolt.

Information Literacy Assessment at the University of Zagreb: an Undergraduate Students' Perspective

M. Pavlovski* and I. Dunđer*
* University of Zagreb, Faculty of Humanities and Social Sciences/Department of Information and Communication Sciences, Zagreb, Croatia
mpavlovs@ffzg.hr, ivandunder@gmail.com

Abstract - Information literacy represents the set of skills that enables individuals to identify, locate, evaluate and utilise relevant information for a certain task. This paper presents results of a survey regarding information literacy, carried out on undergraduate students of the Department of Information and Communication Sciences at the Faculty of Humanities and Social Sciences, University of Zagreb. Previous surveys had asserted the need to include information literacy or information competencies, differentiated from already present tool literacy or technological competencies, in students' curricula on the University. The aim of this survey was to analyse the students' information competencies as well as examine the students' attitude towards introducing an undergraduate course, during which they would learn how to efficiently search for information in the electronic environment while preparing seminar papers and exams. Furthermore, to gain an insight into the students' information retrieval accuracy in course-related research, the authors of this paper shall evaluate their information behaviour, i.e. investigate whether the students develop an overall research plan or search in relevant information sources, e.g. peer-reviewed databases etc. The research results can be helpful to determine the importance of information awareness at this level of education.

I. INTRODUCTION

Information literacy or information competencies of students, their ability to discern relevant from irrelevant information during research used for personal and professional purposes, has been the subject of a numerous quantity of papers in the field of information and communication sciences [1]. Differences between tool literacy or technological competencies and information literacy, defined as a term that represents the set of skills that enables individuals to identify, locate, evaluate and utilise relevant information for a certain task, have been evaluated in different studies locally, in Croatia, and globally. Concerning the skills of undergraduate students at the Department of Information and Communication Sciences at the Faculty of Humanities and Social Sciences, University of Zagreb, in their information retrieval accuracy in course-related research, evaluations on the information behaviour have previously been done on previous generations. The main question is do undergraduate students of newer generations feel the need that information literacy or information competencies should be included in students' curricula?

II. RELATED WORK

According to [2], information behaviour represents the totality of human behaviour in relation to sources and channels of information, including both active and passive information-seeking, and information use, those activities a person may engage in when identifying his or her own needs for information, searching for such information in any way, and using it or transferring that information. In other words, it describes the user's information need, the way of seeking, disseminating and using information in different contexts [3].

The American Library Association states that information literate persons "know how to learn because they know how knowledge is organised, how to find information and how to use information in such a way that others can learn from them. They are people prepared for lifelong learning, because they can always find the information needed for any task or decision at hand" [4, 5].

According to a research, "information behaviour patterns are significantly influenced by the ever increasing sophistication of information resources and technological developments. A crucial element of information behaviour research is applied to the design of information literacy strategies" [6]. Furthermore, current educational systems thrive for student-centred active, flexible and problem-based learning, which emphasises the students' independent information seeking [6].

A survey shows that the majority of students are familiar with the concept of e-learning, consider it supplementary to traditional learning and that the contents they use most often are presentations used in lectures, submission of homework and seminar papers, course descriptions and reading lists. Furthermore, the authors conclude that they "believe that this type of research should be conducted regularly on an annual basis, thus enabling e-learning to be improved" [7].

Another research concluded, that "it would be wrong to claim that users are incapable of finding information by themselves as it would be wrong to claim users are

completely independent in forming their information queries. Accepting the fact that we have moved from system-centred to user-centred information environment would make a good start. Developing course curricula and library activities according to that would mean a great step forward, a step everyone involved in education process has to take in form of an interdisciplinary activity" [8]. The results that arose in this work concluded that information sources and the ways of retrieving information should be part of the curriculum, and that future information specialist (i.e. students of the Department of Information and Communication Sciences) should have more opportunities to learn about information, specifically in an electronic environment, as part of their formal education [8].

One article states that with the introduction and extensive use of electronic information in classrooms the need for information literacy competencies which are necessary for the learning process became more than apparent. "Hence, the ability to meaningfully interact with a wealth of information is deemed more important than ever. More specifically, this correlation has been proved by numerous investigations and studies confirming that academic success and the effectiveness of learning depends on competencies in accessing, evaluating, synthesising, communicating and ethically using information" [9, 10, 11].

The authors also affirm that "although the conceptual core of information literacy, with its emphasis on acquiring abilities that enable one to ethically seek, use and create information, has remained quite stable since the term was first introduced, theoretical foundations of information literacy as well as ideas related to its practical applications and processes have constantly been subject to revision and re-examination" [12].

Reference [13] states that information literacy can be approached as an object of teaching as well as an object of learning, and that it is interesting to evaluate the opinions of undergraduate students on the ways of implementing information literacy in the curricula and their high response percentage to introducing professional education of the subject in one or more courses.

III. RESEARCH METHODOLOGY

The research method was an online survey, which was conducted via the Omega system (https://omega.ffzg.hr/), a customised version of the Moodle Learning Management System (LMS) at the Faculty of Humanities and Social Sciences, introduced in the academic year 2004/2005 [6]. The research carried in February 2015. Two courses ("Information sources and service" and "Machine translation") were researched by means of the survey method.

Undergraduate students were asked to anonymously complete a voluntary questionnaire that consisted of ten questions: one closed question asked for a short answer, two questions provided multiple choice answers, one open-ended required the students to write down their own opinion, two questions used grading from 1 to 5, and the last four questions were used to gather data about the students' gender, age, study groups and year of study. The two questions (questions 3 and 5) that used grading from 1 to 5, 1 being the lowest and 5 the highest value, asked students to grade their competencies using the following information sources/tools: Google, Google Scholar, the library and databases (of scientific publications and citations), and to grade the corresponding availabilities.

The first question asked the students to choose which of the aforementioned sources/tools they use in preparation of seminar papers and exams, whereas the second question asked them to write which one of the four they use most often.

The fourth question asked the students who of the following they consult when looking for guidance in using information sources/tools: colleagues, teachers, family members, other. The option "other" was provided with a blank space in order to enable writing potential answers.

Question 6 was two-fold. It asked students if they did or did not think there is a need to include professional education of information literacy or information competencies as part of one or more courses at the undergraduate level of study. They had a choice between answering "Yes" or "No" and a blank space was added for the students to write an explanation of their answer, independently and regardless of their affirmative or negative opinion.

Questions 7 and 8 were about identifying gender, and age, respectively. Students' age was divided into three possible categories: from 18 to 20, from 21 to 25, and over 25. The last two (questions 9 and 10) had blank spaces in which students could write their study group and year of study.

IV. RESEARCH RESULTS

The questionnaire was finished by 40% of the students (i.e. 18, out of 45 students enrolled in both courses), of which all were undergraduates. The responses from graduate students were not taken into consideration for the results of this survey. The following table gives an overview of students' gender, age and year of study (Table I).

TABLE I. INFORMATION ON RESPONDENTS

Gender	*Male*	6 (33.33%)
	Female	12 (66.66%)
Age	*18-20*	1 (5.55%)
	21-25	17 (94.4 %)
	> 25	0
Year of study	*First*	1 (5.55%)
	Second	0
	Third	17 (94.4 %)
Total number of respondents	18	

The first question asked the students to choose which of the following information sources/tools they use in

preparation of seminar papers and exams: Google, Google Scholar, library and databases. Fig. 1 shows that Google is used by all of the respondents (18), Google Scholar by 9, the library by 16 and databases by 11 students. Google Scholar is therefore not being used by 50% of the respondents.

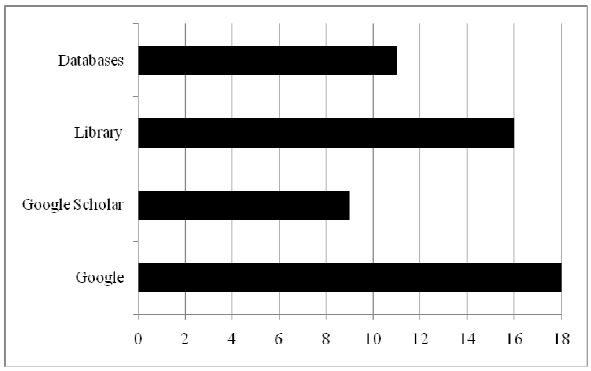


Figure 1. Usage of information sources/tools while preparing seminar papers and exams.

The second question asked which of the four aforementioned information sources/tools they use most often. Google is the most used tool for 10 students, followed by the library (6), whereas Google Scholar and databases are most often used only by 2 students, as shown in Fig. 2.

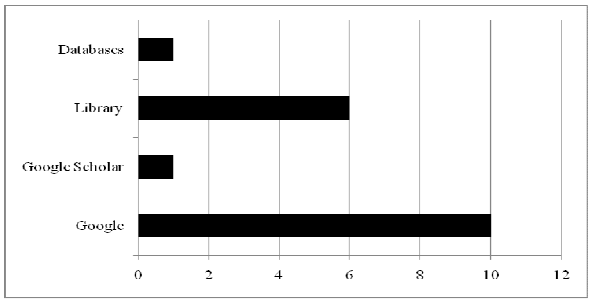


Figure 2. What information source/tool do you use most often?

The third question asked the students to estimate their competencies in using the aforementioned information sources/tools, on a scale ranging from 1 to 5. Fig. 3 shows average values of the respondents' answers.

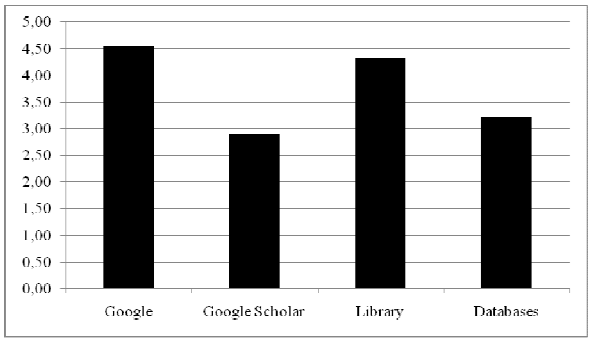


Figure 3. On a scale from 1 to 5, estimate your competencies in using the aforementioned information sources/tools.

For Google, 10 students graded their competencies with a 5, and 8 students with a 4. Grades lower than 4 were not given. For Google Scholar, 2 students graded it with a 5, 5 with a 4, 5 with a 3, 1 with a 2, and 5 with a 1. It is evident that Google Scholar is graded much lower, i.e. almost 30% of the respondents have no competencies whatsoever in using Google Scholar. Regarding the library, 9 students graded their competency in using it with a 5, 6 with a 4, 3 with a 3 and none with a 2 or 1. Concerning their proficiency in database research, 2 students graded it with a 5, 6 students with a 4, and also 6 students with a 3, 2 with a 2 and 2 with a 1. Highest average score is given to Google (4.56), followed by the library (4.33), databases (3.22) and then Google Scholar (2.89).

One multiple choice question (question 4) asked who is assisting them in using the information sources/tools. "Colleagues" had the highest response (17), followed by "teachers" (8) and "family members" (1). The option "other" had also one response: "persons I feel are experts in the particular field".

The fifth question asked students to grade the accessibility of the information sources/tools. For Google, 9 students graded its accessibility with a 5, 8 with a 4, 1 with a 3 and none with 2 or 1. For Google Scholar, 1 student graded its accessibility with a 5, 5 with a 4, 8 with a 3, 1 with a 2, and 3 with a 1. Three students graded the library's accessibility with a 5, 9 with a 4, 5 with a 3, none with a 2, and 1 with a 1. Regarding the databases, 2 students graded its accessibility with a 5, 10 students with a 4, 3 with a 3, 2 with a 2 and 1 with a 1.

Average values of the respondents' answers are presented in Fig. 4, showing an evident trend similar to the trend in question three, indicating correlation to a certain extent between competencies in using information sources/tools and their availability. Namely, again highest average score is given to Google (4.44), followed by the library (3.72), databases (3.56) and Google Scholar (3.00)

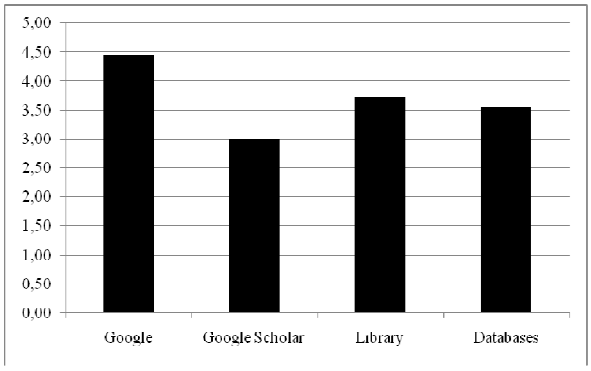


Figure 4. On a scale from 1 to 5, grade the availability of the following information sources/tools.

The sixth question was divided into two parts. For the first part of the question, 17 students answered "Yes" and 1 answered "No". This shows that almost all of the respondents think that inclusion of professional education of information literacy or information competencies as part of one or more courses at the undergraduate level of

study is necessary. Of the 18 students, 16 decided to explain their answer, whereas 2 students left the space for explanation blank.

In their explanations, they mostly agree that for their study, but also in terms of lifelong learning as a whole, information literacy skills are needed for easier finding of relevant information and the correct use of information. Some students believe that, regardless of the study group undergraduate students should obtain an "elementary knowledge" about information literacy and its use in the academic environment for a more efficient completion of scholarly tasks.

Furthermore, some of the respondents stated that education of information literacy should not be introduced only at the Department of Information and Communication Sciences, but at the level of the whole Faculty.

According to one student, information literacy should become an obligatory course for all students at the undergraduate level, starting from the first semester, as it is "obvious that information science students are more competent regarding information literacy, due to Department courses that impose the use of repositories, databases and the completion of different tasks that require a selection of relevant information."

Some students mentioned that colleagues from other study groups often refer to them and ask them for advice in searching for literature using library catalogues and different types of repositories.

As mentioned by one of the respondents, some of the students are not capable of forming correct queries to find relevant information on Google.

They also emphasise that in today's digital environment it is essential to correctly use, seek, find and evaluate information, and that those skills will be of use not only in further education, but in life and professional career as well.

One student pointed out that information literacy should be taught on a practical basis, during an internship in "relevant companies". This would possibly allow future employment at that company.

Students also mentioned that there are a couple of courses related to elementary skills in information literacy, but they assert the need for one course with the possibility of evaluating each students' information competencies, which offers individual, expert tasks to those more proficient users.

As already shown in Table 1, 66.66% of the respondents were female, 94.4% were in the age category "21-25" and in the third year of study. Only 1 student was in the age category "18-20".

All of the students study "Information Sciences" as single (10) or double major (8), e.g. Croatian language and literature (1), German language and literature (1), English language and literature (1), Philosophy (1), Japanology (2), Russian language and literature (1), Turkology (1). There was 1 student of the first study year, none from the second year and 17 of the respondents were students of the third year.

V. CONCLUSION

For the investigation and deeper analysis of information literacy, it is interesting to explore the students' perspectives on the need to integrate information literacy into curricula. Exactly this was examined in this research, which was conducted in form of a survey regarding information literacy on a sample of 18 undergraduate students of the Department of Information and Communication Sciences at the Faculty of Humanities and Social Sciences, University of Zagreb. Students at the undergraduate level of study show that key sources to course-related information are Google, the library and colleagues, rarely teachers, and almost never databases and family members.

They evaluate their information competencies mostly high for using Google and the library, and mostly average when using Google Scholar and databases.

The students graded the accessibility of information sources/tools mostly average, with Google being graded the most accessible, followed by libraries and databases.

Out of the 18 students who responded to the survey, 17 answered "Yes" to the question if they thought there is a need to include professional education of information literacy or information competencies as part of one or more courses at the undergraduate level of study.

The main default of this research is the relatively small sample. Still, representativeness is not always guaranteed by sample size. Namely, in this research the research focus was on identifying information behaviour trends and investigating whether the students develop an overall research plan or search in relevant information sources, databases etc.

This paper has shown that in order to optimise information behaviour, according to students, introducing one or more information literacy courses is of importance for students of all levels. Further research would include sample enlargement and shifting the focus to graduate students.

REFERENCES

[1] I. Irawati, "Information literacy competency of library and information science students at the Faculty of humanities University of Indonesia," Asia-Pacific Conference on Library & Information Education & Practice, 2009, pp. 483-493.

[2] T. D. Wilson, "Models in information behaviour research," Journal of Documentation, vol. 55, no. 3, 1999, pp. 249-270.

[3] K. E. Pettigrew and E. F. L. McKechnie, "The use of theory in information science research," Journal of the American Society Information Science and Technology, vol. 52, no. 1, 2001, pp. 62-73.

[4] M. Gross, and D. Latham, "What's skill got to do with it?: Information literacy skills and self-views of ability among first-year college students," Journal of the American Society for Information Science and Technology, vol. 63, no. 3, 2012, pp. 574-583.

[5] American Library Association, "Presidential committee on information literacy: Final report," 1989, available at: http://www.ala.org/ala/mgrps/divs/acrl/publications/whitepapers/presidential.cfm

[6] J. Lasić-Lazić, S. Špiranec, M. Banek Zorica and S. Žunić, "Discipline-related information behaviour research: educational policy implications," Proceedings of the 30th Jubilee International Convention : Computers in Education, MIPRO 2007, M. Čičin-

Šain, I. Turčić Prstačić and I. Sluganović, Eds. Rijeka: MIPRO, 2007, pp. 21-26.

[7] M. Tot and D. Živković, "The role of e-learning in LIS education : students' evaluations," Proceedings of the International Convention : Computers in Education, MIPRO 2011, M. Čičin-Šain, Ed. Rijeka : MIPRO, 2011, pp. 263-266.

[8] I. Hebrang Grgić and P. Miočić, "Searching for information sources: Students' behaviour in electronic environment," Proceedings of the International Convention, MIPRO 2012, P. Biljanović, Ed. Rijeka : MIPRO, 2012, pp. 1491-1495.

[9] L. Limberg, "Experiencing information seeking and learning: a study of the interaction between two phenomena," Information Research, vol. 5, no. 1, 1999, available at: http://informationr.net/ir/5-1/paper68.html

[10] C. C. Kuhlthau, Seeking meaning: A process approach to library and information services.Norwood, NJ:Ablex Publishing Corp, 2004.

[11] C. Bruce, "Information literacy research: dimensions of the emerging collective consciousness," Australian Academic and Research Libraries, vol. 3, no. 2, 2000, pp. 91-109.

[12] S. Špiranec and M. Banek Zorica, "Information literacy 2.0: hype or discourse refinement?," Journal of documentation, vol. 66, no. 1, 2010, pp. 140-153.

[13] L. Limberg, O. Sundin and S. Talja, "Three Theoretical Perspectives on Information Literacy," HumanIT: Journal for Information Technology Studies as a Human Science, vol. 11, no. 2, 2012, 91-128.

Tiny Operating System Kernel for Education Purposes

Samir Ribić*, Adnan Salihbegović*
* University of Sarajevo/Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina
samir.ribic@etf.unsa.ba, adnan.salihbegovic@ etf.unsa.ba

Abstract - The operating systems courses are often organized according to three basic approaches: as user-centric, theoretically oriented with the explanation of the principles of operation, or through analysis and modification of the kernel source code. While the third approach is the most illustrative, it requires the most of the time at the expense of broader analysis of implementation for different algorithms of the operating systems. Hence, the frequent approach is the development of small kernel, instead of analysing those that are in wider use. In this paper we will present a simple kernel, with a minimal size of source code, which can be fully described in 90 minutes of classroom teaching, but still large enough to be installed and started on a typical PC. This educational version of a kernel includes the basic functions of any OS core: video, keyboard, disk, file system, memory, processes and threads.

I. INTRODUCTION

Most of the faculties and collages that educate students in the field of computer science and engineering and information systems have in their curricula the course in operating systems. Within the variety of taught courses on this fundamentally important subject, three basic approaches have come up so far.

Within the study programmes oriented towards information systems and technologies, the most frequent are user-oriented courses, inclining towards the practical problems like installation, configuration and administration of the operating systems.

Theoretical analysis of algorithms that arise in the implementation of operating systems is an approach that is dominant in operating systems courses at the universities oriented towards computer science. The advantage of this approach is the bigger picture that students acquire on the structure of operating systems; the disadvantage being that the perception of the operating systems remains abstract.

The approach of introducing the principle of operation of operating systems through the explanation and source code development of an operating system kernel can be successfully applied if the students have sufficient preceding knowledge. This approach offers better understanding and acceptance by the student's operating systems fundamentals, but leaves less room for teaching alternative approaches and examples.

II. REVIEW OF RELATED COURSES

Size of the OS source code is a term that has a completely different meaning in the industrial applications and education. For example, one megabyte represents about 20,000 lines of source code. While stored on the disk is negligible quantity, when printed out on paper will contain about 500 pages, the amount of study content for whose explanation an entire semester will be required.

This of course, will be unacceptable as teaching content, because describing and explaining only the source code will produce rather monotonous way of teaching. Operating systems that are used in the real world applications (the Windows NT kernel and Linux), are much greater, for example Windows 7 has 2 million, and Linux 3.10 has 17 million lines of code, and it is only for kernel. Therefore, their source code cannot be suitable teaching tool for the study of the principles of operating systems. Some authors explain portions of complete kernels. Windows Research Kernel contains around 830,000 lines of code [1] and is available to educational institutions, with special agreements with Microsoft. Despite the mentioned size it cannot be executed independently. There are also textbooks centred over particular OS design, but use only short code snippets or pseudocode [2].

If it is desired that the source code of the operating system kernel is an addition to the theoretical consideration of the principles of operating systems, as an illustration of the kernel operation on a real machine, more practical and feasible are examples with smaller size of code. Frequently, older versions of successful operating systems are used or teachers develop their own version of small operating systems.

The most common approach is to develop the kernel, which resembles the UNIX kernel. This is not a surprise, considering the popularity of the UNIX having its roots among the university communities.

Minix [3] is the most famous educational operating system. Partly is UNIX compatible and explained in the cited book, together with the theory of operating systems accompanying the parts of source code itself. The OS length is around 16 000 lines, so the whole course can be organized based on this reference. Minix can be also used for high reliability systems, and its architecture is with micro-kernel.

Xinu [4] is also used as a kernel around which popular book on operating systems has been written. The first version was intended for the PDP-11 machine, followed by the versions for IBM PC [5], the Motorola 68000 systems, Mips, Spark and recently ARM based machines.

The approach to the subject in the book is easily acceptable. Xinu contains around 8000 lines of code, but does not enter into the scope of the file system. When installed on PCs it relies on MS-DOS calls to work with files.

Nachos [6] is written in object-oriented languages, there are C ++ and Java version, with the addition of C and MIPS assembly code. It contains around 5,000 lines of code, and runs in a virtual environment.

Old versions of Unix and Linux kernel are also used for educational purposes. At MIT university for teaching the design of operating systems, for long period of time, the old version of Unix V6, with 8300 lines of code has been used. After the students were wondering about the purpose of learning the outdated system on the old platform written in an ancient version of the C language, under name XV6 it was ported to the 80386 platform [7].

In the case of Linux, version 0.11 was the first version capable of independent execution; it had 14,000 lines of code. Annotated version of this kernel was mentioned in reference [8]. After code reduction down to the 5000 lines, new OS named Qutenix was created [9].

If the kernel is freed from design requirements to be similar to Unix, smaller kernels are also possible. Geekos [10], with the source code of around 3000 lines is an example of the kernel that is designed to be small and therefore easy to understand, and not burdened with a request for similarity to Unix or Windows. However, this operating system does not have its file system.

In reference [11] educational version of OS is described that similar to floppy based CP/M operating system, with considerable limitations (files with size up to 13KB and processes up to 64KB). The source code has only 1200 lines, which makes it the smallest educational version of kernel OS we have discovered during this research.

III. KERNEL STRUCTURE

A. *The principles of kernel design*

For students, to understand the principles of operating systems and yet not spend too much time in the analysis of the source code, the specific minimalist kernel of the operating system was developed. The design principles were as follows:

The kernel has to be able to run independently on real physical hardware, and not only on the virtual machine. It should not be dependent on any existing operating system. To make it runnable on wide spectrum of computers, kernel is designed for 386-compatible PCs in 32 bit mode, with peripherals that include an IDE drive, keyboard and text display.

Alternatively, it can run under qemu emulator, available at http://wiki.qemu.org/Main_Page. In order to clarify some technical specifics for this environment, we used information from the website http://wiki.osdev.org.

The kernel is written in C, except for the boot sector, part of the code to save registers and very short routines for input/output and interrupt stack structure that are coded in assembly language.

TABLICA I. KERNEL FILES, LAYERED

Files	Role	Line
bootsect.asm	Load kernel, switch to protected mode	85
string.c	String operation routine	41
ports.c	I/O mapping instructions	25
video.c	Display of text on the screen	58
hello.c	Greeting message	10
testvideo.c	Video subsystem test	
keyboard.c	Keyboard driver	147
testkbd.c	Keyboard test	
interrupt.c	Interrupt implementation	145
testinter.c	Interrupt test	
diskio.c	Direct access to IDE HD	47
testdisk.c	Disk test	
files.c	File system	278
testfiles.c	Files test	
memory.c	Linear page memory	104
testmemory.c	Memory test	
sched.c	Scheduler	138
syscall.c	System calls	90
testproc.c	Scheduler test	
Kernel total lines of code		1168

It was necessary to include in the kernel the range of functionalities like: inputs/outputs, the organization of memory, disk, file system and scheduling processes. Developed kernel can be used for simple algorithms in the realization of subsystems (first fit in memory, continuous allocation in a file system, cooperative round robin scheduling process), with the possibility that students through exercises implement more advanced algorithms, or get acquainted with them theoretically.

The kernel is organized in layers for didactical reasons. Each new source code file can use the previous ones, while with the addition of the main function, may represent a unified entity. The Table I lists the source code files with their length in number of lines. The table also represents the order in which these files are processed. The test files are called after explaining to the students the operation of individual subroutines, and also include kernel main program core. The acronym for kernel name is LEKOS (Layered Educational Kernel of Operating System).

With an average rate of kennel code explanation of 15 lines per minute, with all intermediate stages of compiling and displaying various demonstration examples, standard 90 minutes of laboratory exercises suffices for a complete presentation of kernel structure and operation. In the subsequent part of the paper, the implementation of kernel subsystems with code snippets for shorter routines will be demonstrated.

B. *Bootstrap subroutine*

This is the only subroutine written in assembly language. The boot sector has to be of the max. size of 512 bytes and finished with the signature. It consists of two main parts, 16 bit and 32 bit. Sixteen-bit segment of the code is executed at 000: 7c00 calling two times BIOS interrupt 13h. These interrupt calls pass the adequate parameters via registers. The first call resets disk controller, the second call transfers some sectors from the

disk into the RAM. With this transfer, the file with kernel is loaded with into the memory. In the second half of the sixteen bit segment GDTR register is loaded to point to a table that simply maps the memory segments for linear addressing. After changing the value of CR0 register, program switches to protected mode.

Thirty-two bit part of the Bootstrap subroutine sets the segment registers, stack pointer and jumps into the kernel. The boot sector contains descriptor table, required to transform Intel segment architecture into a linear one. The table consists of only three descriptors: one is empty, one for code and one for data.

C. String routines

The operating system is running on an empty machine. Since the programmer has no standard C library at disposal, several necessary functions for operation with strings have to be provided. These functions can be implemented purely in C language. The realized functions are: strncmp, strncpy and itoa.

D. Mapping of I/O ports

The standard C language does not have built-in access to I/O mapped peripherals. Therefore, it is necessary to use assembler routines, but they are easily encapsulated in the appropriate subroutines using inline assembler built into the GCC. For example, to read eight bit data, the following function is used:

```
unsigned char in(unsigned short _port){
   unsigned char result;
   __asm__("in %%dx,%%al":
        "=a" (result):"d" (_port));
    return result;
}
```

E. Video driver

In order to enable the greeting message to appear on the screen as the first step in work with the operating system kernel, the next step is to display message on the text screen. The first part of a simple video driver is positioning the cursor. The cursor position is held in the variable `cursorpos`, containing 80*row+column memory location. Its value can be read and changed with the following simple functions `SetCursor` and `GetCursor`. In subroutine `hardwarecursor`, the registers on ports 0x3D4 and 0x3D5 are used, with which the text cursor is switched on.

Textmode video buffer on PCs is located on the absolute address 0xB8000. Access to the textmode memory is in the form of a matrix with 80 columns, in format of ASCII characters for every character on the screen. The screen deletion is accomplished by inserting a character with the code 0 or 32 at desired location in video memory. Printing of the text on the screen is done with subroutine `Print` containing a loop that inserts ASCII codes into the message, skipping the attributes in video memory, while subroutine `PrintNum` prints integer value calling atoi function before `Print`.

```
void PrintNum(int num) {
    char text[20];
    itoa(num,text);
    Print(text);
}
```

```
void Print(const char * msg) {
    unsigned long i;
    unsigned char *vmem;
    vmem= (unsigned char *)0xB8000;
    vmem += cursorpos<<1; i = 0;
    while (msg[i] != 0) {
     *vmem = msg[i++]; vmem += 2;
    }
    cursorpos += i;
    hardwarecursor();
}
```

What is left for laboratory exercises or homeworks are: scrolling functions, printing special characters such as CR / LF etc.

F. Keyboard driver

The relatively easy part of this driver works directly with scan codes. When the key is pressed (0x64h port returns on zero bit the value of 1), the value of the pressed key can be read from port 0x60h.

```
unsigned char GetScanCode(){
  static char code;
  if(in(0x64) & (0x01)) code = in(0x60);
  else code = 0;
  return code;
}
```

Larger part of this driver is a subroutine that converts the scanned code in ASCII, in cases with SHIFT, CTRL combination and direct keypress. Three tables are used: `kbdshift`, `kbdctl` and `kbdnorm`, which enable the conversion of scanned code into ASCII codes.

```
char * kbdnorm=
"\0\0331234567890-=\b\tqwertyuiop[]\n,\0asdfghjk
l;\''`\0\\zxcvbnm,./'\0\0\0 ";
```

The subroutine `ScanToAscii` is looking whether the keypress is one of the: Shift, Caps Lock, Ctrl and Alt keys, etc., and on that basis, include or exclude certain flags. According to a combination of flags, the selection of the adequate conversion table is carried out. Reading of the key at this stage is performed by subroutine `WaitKey`, and demo program displays pressed keys on the screen:

```
char WaitKey(){
  int keycode = 0;
  while(!keycode) keycode = GetScanCode();
  return ScanToAscii(keycode);
}
```

For the lab. exercises and homework assignments the implementation of other keyboard layout, support for function keys and numeric keypad can be added.

G. Interrupt driver

The next task in kernel development was to avoid busy wait. For this purpose the interrupts are used. In protected mode IDTR register points to the interrupt vector table. Each element of this table consists of eight bytes. Segment part of the descriptor is always set to the value of 8, since the memory was previously linearized, and flags are set to the same value for all interrupt descriptors. Therefore, only the lower and higher 16-bit offset values are set. Setting of the beginning of this table is done with machine instruction command `LIDT` that is incorporated in the C subroutine. In a similar way, the routines for switching on and off of interrupts are included, for example

```
void disable_interrupts () {
  __asm__ ("cli");
}
```

Early interrupt controllers on the PC computers were setting vectors for processing hardware interrupts starting from IRQ0, on interrupt vectors from 8 onwards. This is still implied configuration. Since the 80386 processor, these vectors are used for other purposes, on these addresses processor exceptions are generated. Therefore, it is necessary to remap the interrupt vectors so that the hardware requirements are processed from vector 32 onwards, what is implemented in routine `PIC_remap` consisting mainly of `out` instructions.

Format of the C subroutine is not the same as format of interrupt routines. Therefore, the routines that are executed in the interrupts have certain assembly instructions at the beginning and the end, in order to have the correct format of the stack.

The first of these is related to keyboard keypress reading. Received scan code of the pressed key is placed in a circular buffer. The routine for reading the keyboard takes over the character from this buffer. These functions are implemented as in the following section of code.

```
char kbdbuffer[256];
  unsigned int readchar,storedchar;
  void kbdirq() {
  static unsigned char c;
  __asm__ __volatile__("pushal; ");
  c=GetScanCode();
  if (c != 0) {
   storedchar=(storedchar+1) % 256;
   kbdbuffer[storedchar]=c;
  }
  out(0x20,0x20);
  __asm__ __volatile__
     (" popal; leave ; iret");
}

unsigned char readintscancode() {
  if (readchar!=storedchar) {
    readchar=(readchar+1) % 256;
    return kbdbuffer[readchar];
  }
  else
   return 0;
}
```

Second implemented interrupt is the timer interrupt. Its task is simple, increases the counter by one every 10 ms.

```
void Clock() {
  __asm__ __volatile__("pushal");
  elapsedtime++;
  out(0x20,0x20);
  __asm__ __volatile__
   (" popal ; leave ; iret");
}
```

The initialization of interrupt system sets the interrupt table and activates interrupt vectors for these two interrupts. Demo program shows the keyboard presses while time passes. For the sake of exercises, critical section to control the access to keyboard buffer can be added, processing of the interrupts generated by the serial port, or implementation of the processor exceptions.

H. Disk driver

Although their use in new PCs increasingly rare, as an external memory device an IDE hard drive is used, because it is the simplest kind of disk access at a low level. With these types of disk drives the ports 0x1f1 to 0x1f7 are used to send LBA or CHS address of the first sector in a group that user wants to transfer to RAM memory, with the information how much sectors wants to transfer. By sending the values to port 0x1f7, the command is activated which starts the transfer in one direction or another. After that, the values can be read or written. The subroutines for writing or reading blocks of data from disk are very similar, they send the LBA addresses to appropriate ports and read or write data.

```
int WriteHDBlock(void* buf,int first,int num ){
  int i;
  unsigned short* dat = (unsigned short*) buf;
  while(num > 0 ){
    while( in(0x1f7) & 0x80 ){} //Wait
    out(0x3f6,2);
    out(0x1f2,1);  //sector count
    out( 0x1f3,first & 0xff  );
    out( 0x1f4,(first >>8) & 0xff  );
    out( 0x1f5,(first >>16) & 0xff );
    out( 0x1f6,((first >>24) & 0x0f)|0xe0
    out(0x1f7,0x30);       //start write
    while( in(0x1f7) & 0x80 ){}
    for(i=0;i<256;++i, dat +=1)
      outw(0x1f0,*dat);
    --num;
    ++first_sector;
   }
   return 1;
}
```

Demonstration example of this subsystem is displaying on the screen the bytes on disk where the kernel of the operating system is stored.

I. File system

The file system was designed to be as simple as possible. Therefore, the directory was implemented with only one level, with a continuous allocation of files on disk. Nevertheless, data structures that describe the elements of the directory, beside the data on the name and the position of the file on disk, include the field directory, which allows extension on directories with two or more levels. Directory structure contains information on all the files on the disk, and about open files. These structures are resident in tables `dirtable` and `opentable`.

Opening of the file directory copies element with forwarded location number in the directory table, to the free location in the table of open files, while filling the elements of open files table that do not exist in the table directory. Separate routine `OpenFile` searches for file with the specified name and call `OpenFileByNumber`.

Modifications in the file content result in writing of the element directory on the disk, which is done by procedure `SaveDirEntry`. This procedure calculates the position of the specific element of the file directory structure on disk, and writes this in the block.

```
int SaveDirEntry(unsigned int num) {
  if (num<0 || num>=maxfiles)
    return -EINVALID;
  int dirsperblock=512/sizeof(struct direntry);
  int relposondisk=num/dirsperblock;
  int direntrytowrite=relposondisk*dirsperblock;
  WriteHDBlock( &(dirtable[direntrytowrite]),
  dirtableondisk+ relposondisk,1 );
  return OK;
}
```

Directory structure can recognise as the file type: free region, allocated regular file, and file that is type of directory. Deleting of the file is simply achieved by conversion of its content into free zone, to convert its directory content into the free zone, as seen in the following part of the code:

```
int DeleteFile(char * name) {
  int i;
  struct direntry * cde;
  cde=dirtable;
  for (i=0;i<maxfiles;i++,cde++)
    if(cde->entrytype==REGULARFILE)
      if(!strncmp( cde->name,name, NAMELEN) {
        cde->entrytype=FREEZONE;
        return SaveDirEntry(i);
      }
  return -EFILENOTFOUND;
}
```

Subroutine for file creation, due to the continuous allocation of space on disk, needs to contain its maximum size as parameter. After that, the best fit algorithm finds free block on the disk among previously allocated files. Closing of the file simply changes the status field in the table of open files. Positioning is also easy to implement:

```
int SeekFile(int fd,char* buffer,int pos) {
  if (fd<0 || fd>=maxopenfiles)
    return -EINVALID;
  if(opentable[fd].status!=OPEN)
    return -EINVALID;
  if (pos>=opentable[fd].actualsizeinbytes)
     pos=opentable[fd].actualsizeinbytes-1;
  opentable[fd].currentseekpos=pos;
 return pos;
}
```

Writing to a file begins with finding of its position blocks on the disk, and then the first of them is loaded into the buffer. After that, the data is transferred from the user buffer in the buffer disk block. When the cursor reaches the end of the buffer, the block is written to the disc. Here the care should be taken, that recorded blocks do not exceed the limit set for that file. If more blocks are written than the current length, the corresponding element in the directory stricter is updated to the value of new length. Reading of the file from disk also uses the value taken from the position in the open file, to calculate its position on the disk. Thereafter, the function is called to read sectors from the disk into memory.

Initialization sets the initial values of the relevant variables and loads into memory the directory structure. Demo program includes writing, deleting, additions and displays of several created files. Further lab. exercises can include addition of subroutines for directory search and modification of starting location of partitions.

J. Memory management

In the operating system, memory between 0 and 640 KB is reserved for kernel and its tables, while the custom applications are allocated memory space above 1 MB. The memory management module is implemented with two strategies, one for the physical memory and one for virtual memory. The physical memory block is represented by memTable element.

This element comes with a series of simple structures containing start address and bit valued fields indicating the length and population percentage. Initialization of the memory subsystem sets the table of the physical memory elements as blocks of 128 megabytes, beginning with address 0x100000.

`AllocMemory` will assign the first available block. The physical memory is allocated through continuous allocation, first fit algorithm, and unused portion of the memory block hole becomes the new block for allocation.

```
void * AllocMemory(int size) {
  int i;
  if (maxused>=TABLELEN)
    return (void *) EOUTMEMORY;
  for (i=0; i<=maxused; i++) {
    if ((memtable[i].length>=size) &&
    (memtable[i].busy==0)) {
       memtable[i].busy=1;
       if (memtable[i].length>size) {
         memtable[++maxused].busy=0;
         memtable[maxused].startaddress=
           memtable[i].startaddress+size;
         memtable[maxused].length=
           memtable[i].length-size;
         memtable[i].length=size;
       }
       return (void *) memtable[i].startaddress;
    }
  }
  return (void *) EOUTMEMORY;
}
```

The routine for freeing occupied block of physical memory takes care of recombination of neighbouring memory holes.

Although we implemented simplest algorithms, the power of Intel processors allows us to implement some more advanced topics, virtual addresing and relocation. Processes can see the memory differently, as it is the case when mapping of the virtual memory. Routines `ActivatePaging` and `LinearMode` change registers CR0 and CR3 and hereby change paging mode to be enabled or disabled. Paging mode is on two levels.

The address space of each process up to 4 megabytes is seen as being between the addresses 0 and 4MB, and is pre-mapped to the corresponding physical address above 1MB. As Intel processors are using paging on two levels, requiring two tables, one representing the directory page (with only one element used) and the other representing the page table (with 1024 elements). The first megabyte in that table is mapped equally for all processes, while the remaining three megabytes of virtual memory space is mapped into memory separately for each process.

K. Processes

For the sake of simplicity, the scheduler is of cooperative type. That means that processes or threads in the kernel must periodically call the system calls, which carry out scheduling. To save pre-empted context of the process or thread, the `jmp_buf` structure is used, to preserve the statuses of the registers and prepare for jump. The command `setjmp` prepares the return point and with command `longjmp` jumping back to one of predefined return points is performed.

The structure `task_t` contains process control block. It consists of the process status, the pointer to its memory and the pages of `setjmp` content registers. Process control blocks are arranged in variable sized array. The scheduler works in the cooperative version of Round Robin

principle. When the scheduler function is called, `setjmp` saves the state of currently executing process, the table of memory pages is changed and then with `longjmp` command advances to the next process.

```
void schedule(void) {
  unsigned prev;
  static int cr0;
  prev = current;
  do {
   current++; // round-robin switch to next task
   if(current >= NUM_TASKS)
      current = 0;
  } while (g_tasks[current].state!=READY);
  idlecpu=(prev==current==0);
  if(setjmp(g_tasks[prev].regcontext) != 0)
     return;
  ActivatePaging ( g_tasks[current].pagetable);
  longjmp(g_tasks[current].regcontext, 1);
}
```

Starting a new process allocates memory with the size of 3 MB, sets registers, pre-maps the memory, and if it is not a system thread, loads the file with the process code, and adds the values in the process table. The completion of the process frees allocated memory, and calls the scheduler.

```
void Terminate(int pid) {
  if (pid >=0 && pid < NUM_TASKS ) {
    setprocstate(pid,DELETED);
    LinearMode();
    FreeMemory(g_tasks[pid].memoryblock);
    schedule();
  }
}
```

L. System calls

Since the scheduler is of the cooperative type, the system calls are, beside the basic function, extended with the call to scheduler, as in the following example:

```
int _OpenFile(char * name) {
  int i=OpenFile(name);
  schedule();
  return i;
}
```

IV. CHANGES IN EDUCATIONAL PROCESS

In the academic year 2014/2015, at the faculty of Electrical engineering, University of Sarajevo, described minimalistic kernel was introduced as part of lab. exercises. After explaining each kernel layer group and writing different versions of main program, gcc compiler and nasm assembler compile previously processed layers and execute them under QEMU. Then it is ported to any PC with IDE drive, to be executed on real hardware.

In previous years, on the operating systems course, as an example of operating system source code, the older and newer versions of Linux or Minix were used. However, very small number of students was able to follow the source code and modules of these OS systems, while the lecturing time for their complete description was not sufficient.

Experimenting with developed kernel, showed that kernel structure and code can be completely described within laboratory exercises lasting two teaching hours (90 minutes). This demonstration has followed the completion of the theoretical part of the course on operating systems, so the students were prepared to follow the topic. However, if it is desired to include more independent work of students on working with kernel and making their own modifications, more teaching and laboratory hours are required.

V. CONCLUSION

Minimalist operating systems do not strive to be used in real world applications, but can be of great importance as an educational tool to display the principles of operating systems structure and organization. Making use of the simplest algorithms, the developed operating system kernel has fulfilled its objective and purpose, to illustrate the practical side of the operating systems role and inner working, while not spending too much time on its explanation, at the expense of theoretical considerations and approaches.

REFERENCES

[1] S. Diomidis. "A tale of four kernels." Proceedings of the 30th international conference on Software engineering. ACM, 2008.

[2] M. J. Bach, The design of the UNIX operating system (Vol. 5). Englewood Cliffs, NJ: Prentice-Hall (1986).

[3] A. S.Tanenbaum, and A. S. Woodhull. "The Minix Book–Operating Systems." (2006).

[4] D. Comer, "Operating System Design: The XINU Approach, 1984." Bell Telephone Laboratories, Incorporated

[5] T. Fossum, "PC-XINU features and installation." ACM SIGOPS Operating Systems Review 21.3 (1987)

[6] C. Wayne., S. J. Procter, and T. E. Anderson. "The Nachos instructional operating system." Proceedings of the USENIX Winter 1993 Conference Proceedings.

[7] R. Cox, M. F. Kaashoek, and R. Morris. "Xv6, a simple Unix-like teaching operating system." 2013-09-05]. http://pdos. csail. mit. edu/6.828/2012/xv6. html (2011).

[8] Z. Jiong, and T. Jarvi. "Adopting and Commenting the Old Kernel Source Code for Education." Linux Symposium. 2005.

[9] Qu, Bo, and Z. Wu. "Design and Implementation of Tiny Educational OS." Recent Advances in Computer Science and Information Engineering. Springer Berlin Heidelberg, 2012.

[10] H. David, J.K. Hollingsworth, and B. Bhattacharjee. "Running on the bare metal with GeekOS." ACM SIGCSE Bulletin. Vol. 36. No. 1. ACM, 2004.

[11] M. Black, “Build an Operating System from Scratch: A Project for an Introductory Operating Systems Course,”, Proceedings of the 40th Technical Symposium on Computer Science Education, SIGCSE 2009

Usage of Compulsory Content in Secondary School Education

Renato Barisic
University College Algebra, Zagreb, Croatia
renato.barisic@racunarstvo.hr

Abstract - Modern information and communication technologies suited to be used in the educational process have to bridge the gap between the desires of today's students for modern educational content and compulsory literature in the form of printed textbooks. This paper presents the results of research on use of printed textbooks in the Humanities, Social and Natural Sciences and Engineering in the third and fourth grade of secondary school. The survey lasted for three years and was conducted among 3.217 students from 79 secondary schools from 16 Croatian counties. An analysis of research on the use of printed textbooks in the final years of secondary school shows that a significant number of textbooks are used on a small scale during the day at school and even to a lesser extent during independent work at home. The aim of this paper is to highlight the need to find alternative and complementary solutions whose application should modernize the educational process and present courses to students in an attractive and motivating way on a variety of devices that today's students have at home, at school and on the move.

I. INTRODUCTION

To meet legal forms, all schools at the beginning of the school year prescribe the necessary literature for each class, but its usage in the class is constantly getting lesser. Teachers, who want to improve the teaching process, in order to adjust the class to the thoughts and needs of the contemporary students, are forced to use various additional formally approved or not approved teaching resources or to create new content that will best represent the specific teaching unit to the students. It is hard to say if these kind of innovative teachers are currently a majority or a minority, but we must urgently think about the changes in the approach to teaching, especially in the introduction of new interactive content into everyday teaching process.

The printed textbook in the new school remains almost the only means of creative communication mediation between teachers and students. Since educational material is investigated in the teaching process, as the teaching process means drawing up an inventory of ideas, and teachers mediate this investigation, both the teaching process and teachers are direct and necessary participants in textbook creation. The teaching process can in this context be seen as the symbiosis of the overall education. Qualitative restructuring and redesign are done by those who mediate textbook content – teachers and students [1].

It is not easy to understand the increasingly complex global events, to see facts within the framework of objective context and to think critically about means of information. This requires huge efforts in creating a strategy to train students for life-long learning. Computer-based instruction (including the use of e-textbook) surely provides flexible individualized learning. It increases motivation and students' self-esteem because it helps them earn their knowledge. Nevertheless, the computer itself does not necessarily mean that the children will develop divergent thinking, creativity, collaboration, responsibility in decision making, democratic behavior, etc. The priority task of educational sciences is to investigate ways to bring educational media under certain standards and to suggest criteria for their design, so as to make them carriers of society progress [2].

Daily creation and sharing of unlimited amounts of information is no longer possible to be followed with the traditional printed resources due to the slowness of their occurrence and the inability to change content in the time of new knowledge creation and technological achievements. Even if we do not raise doubts about the accuracy, completeness or timeliness of the classic printed textbooks, we need to rethink whether a teaching aid in that form is suitable for the education of modern technology minded young generations and consider whether the modern tech-savvy teachers would use classic printed teaching tool as the first choice for high-quality transmission of knowledge in educational institutions of the 21st century [3].

This paper will show results of the research that lasted for three school years and included third grade and fourth grade secondary school students from technical schools and gymnasiums from almost all Croatian counties. The aim of research was to collect information directly from the secondary school students and to find out do they use printed textbooks at school and at home. The aim of this paper is to show research results and to highlight the obvious need for finding alternative and complementary solutions whose application should modernize the educational process and present courses to students in an attractive and motivating way.

II. METHODOLOGY

A. Method

Survey using a printed questionnaire was a selected data collection method. Prior to conducting the survey the

permission was sought from the responsible persons: principal, school counselor, psychologist or teacher. The survey was carried out in groups, on a voluntary basis, using paper format questionnaire, for one or two classes that were present in the classroom at the same time. At the beginning of the survey the interviewer demonstrated the questionnaire to students and they were given filling instructions. Interviewer emphasized that the survey is completely anonymous and does not collect any personal or school data. Time to fill in the questionnaire was limited to 10 minutes.

B. Instrument

Data collection instrument was a questionnaire. A questionnaire completed by the students consists of general questions and questions about the use of printed textbooks and digital content.

General questions about age, gender, type of school, and the direction of the class are primarily used for the analysis and presentation of the structure of the respondents according to various criteria. Given that the target sample were students of third and fourth grade of secondary school, the question about respondents' age offered a choice between 13 and 19. Question about gender could be responded by circling M for male or F for female. When asked about the type of school, the students could respond with T for technical school or with G for gymnasium. The question about the direction of education is set up as an open question which could be responded by free writing. The last question is a general question about the current grade with possible answers 1 to 4. It is axiomatic that the target population was fourth grade and third grade at times. Still, there was an option of circling first grade or second grade if someone from the lower grade appears.

Questions about the use of printed textbooks are set as a closed set of multiple choice questions. Question about the use of printed textbooks during the school was constructed in the form of a table where the columns highlight three main divisions: subject, at school, at home. In the subject column is a list of subjects in the Humanities (Croatian, Foreign language, History, Religion, Ethics), Social Sciences (Politics and Economy), Natural Sciences (Mathematics, Physics, Geography) and Engineering (Computer Science, Information Science). In the group of columns at school and at home, there are subdivisions offered intensities: never, rarely, often or very often. With this approach the students were able to immediately designate the usage intensity of printed textbooks at school and at home for one course.

C. Sample

Data was collected during three school years: 2012/2013, 2013/2014 and 2014/2015. First school was included at 22nd November 2012 and the last one at 6th March 2015.

Numbers of included students, secondary schools and counties by each school year could be seen in Fig. 1.

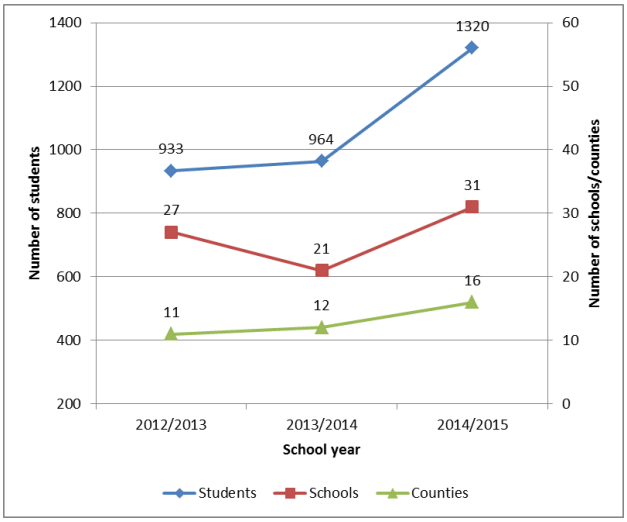


Figure 1. Number students, secondary schools and counties included in the research

III. RESULTS

Fig. 2. shows number and share of students by their age.

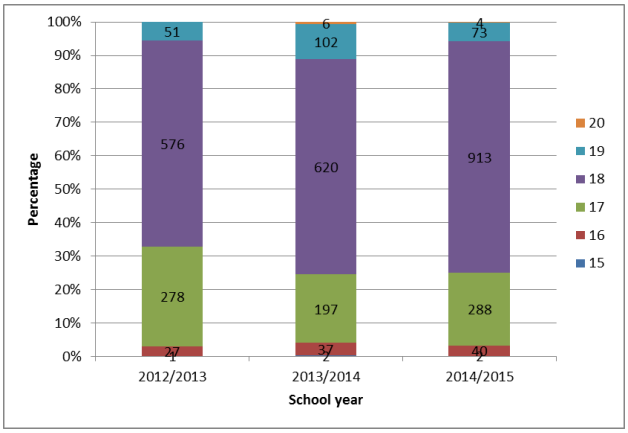


Figure 2. Number and share of students by age

Fig. 3. shows number and share of students by their gender.

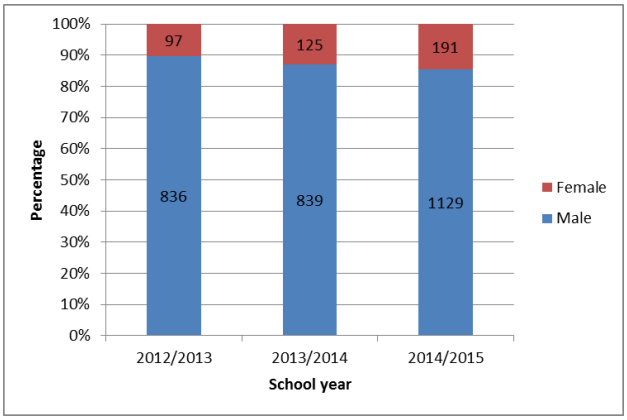


Figure 3. Number and share of students by gender

Fig. 4. shows number and share of students by their grade.

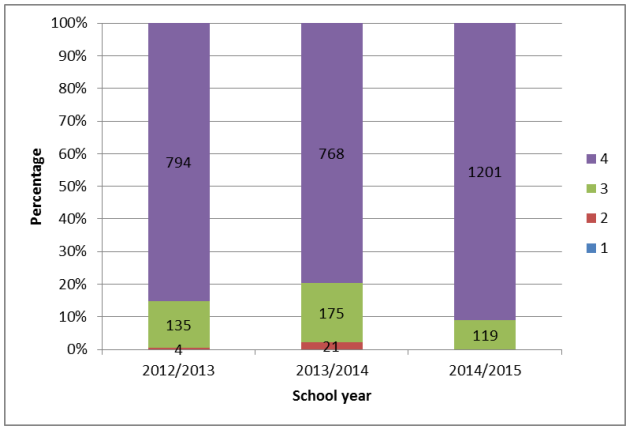


Figure 4. Number and share of students by grade

Fig. 5. shows number and share of students by the type of secondary school.

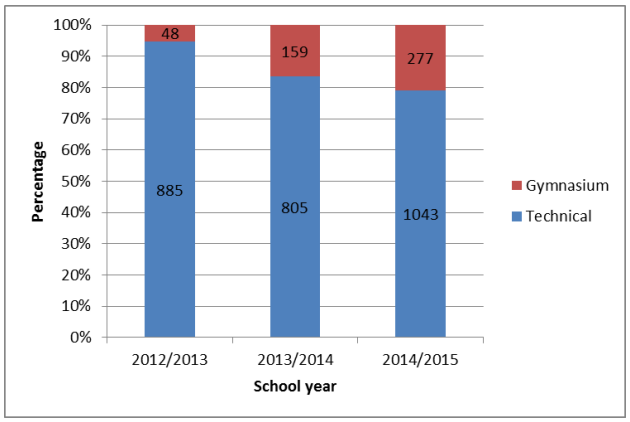


Figure 5. Number and share of students by school type

Next figures show summary of student answers about using printed textbooks during study at school and at home. Data is sorted descending by the number of answers for intensity Very often so the first column shows subject with the best percentage of Very often use, second subject have smallest percentage and so on.

Fig. 6. shows data about printed textbooks usage during work in the classroom at school collected during school year 2012/2013, fig. 7. shows data collected during school year 2013/2014 and fig. 8. shows the most recent data from school year 2014/2015.

Fig. 9. shows data about printed textbooks usage during students' independent work at home collected during school year 2012/2013, fig. 10. shows data collected during school year 2013/2014 and fig. 11. shows the most recent data from school year 2014/2015.

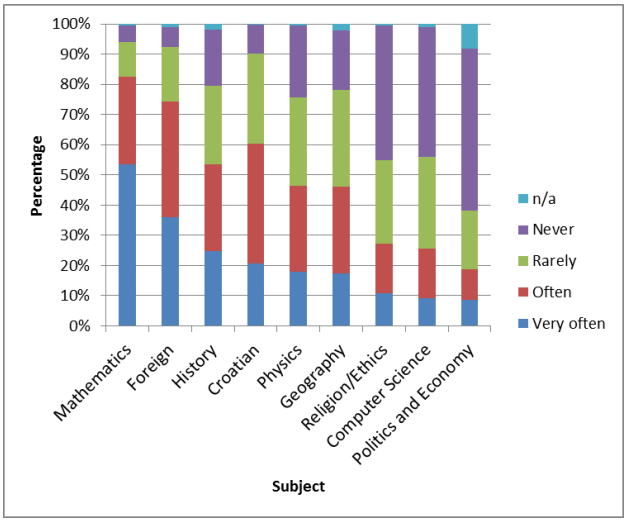


Figure 6. Printed textbook usage at school – school year 2012/2013

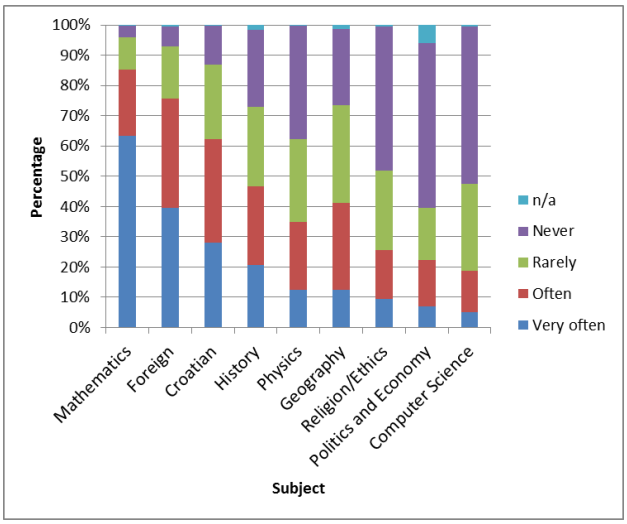


Figure 7. Printed textbook usage at school – school year 2013/2014

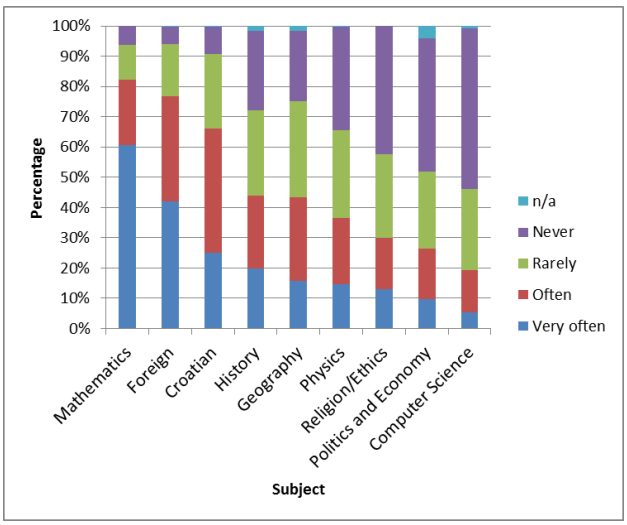


Figure 8. Printed textbook usage at school – school year 2014/2015

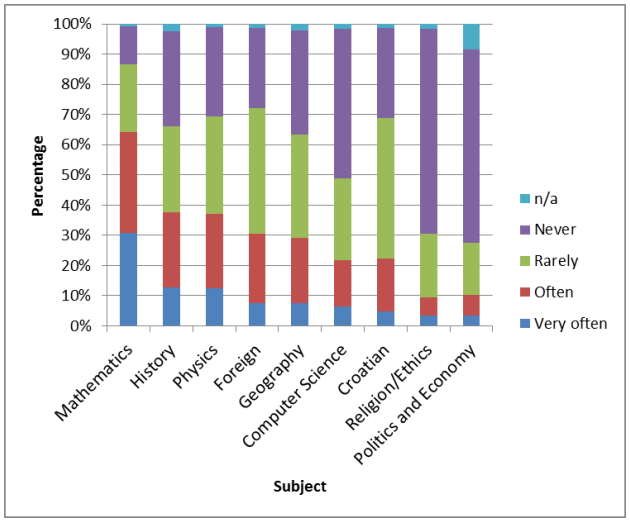


Figure 9. Printed textbook usage at home – school year 2012/2013

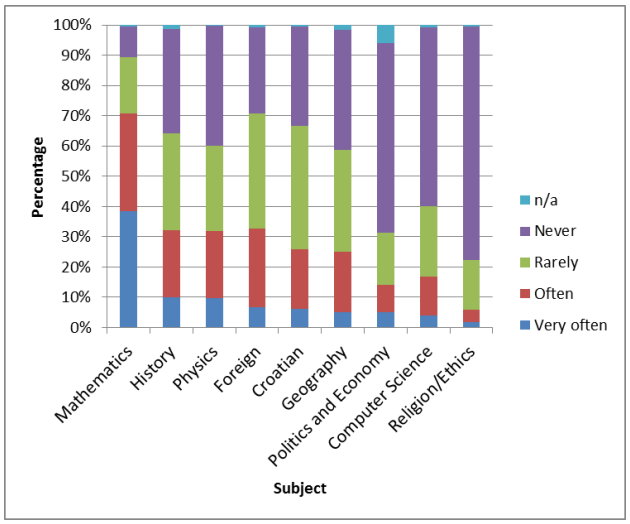


Figure 10. Printed textbook usage at home – school year 2013/2014

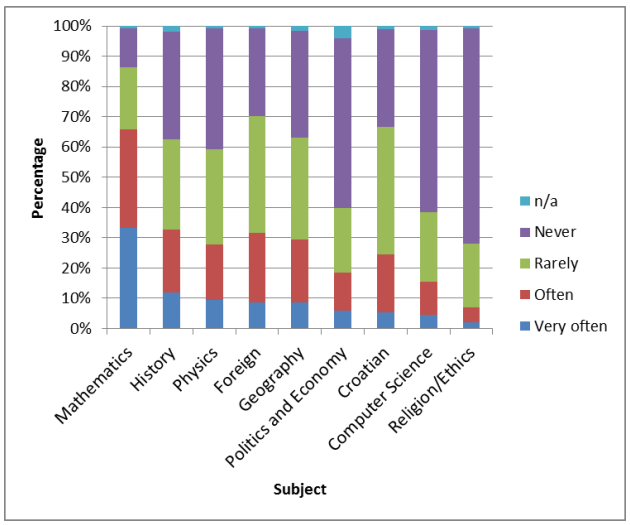


Figure 11. Printed textbook usage at home – school year 2014/2015

IV. CONCLUSION

Modern students need new approaches, motivation and learning processes that will align the teaching practices with new technologies, modern educational approaches and reform tendencies. In today's world of intense use of top quality audio-visual, communication and computer technology achievements, the acceptability and usability of printed textbooks is questionable because they are no longer a medium that is acceptable to students and that motivates them to acquire new knowledge and new skills, and are no longer a medium of choice in the search for new, complete and updated information. Information and communication technologies have enabled that with the use of the Internet there are no longer limits to publishing, retrieving and exchanging information. Search and access to information is instantaneous, and the sources of information can be multiple and thereby contribute to authenticate the information found. Even more important than the search are the speed of publishing new information and the speed of updating existing information. The methods of publication or updating information on the Internet are measured in seconds and minutes while the process of publication or updating information in printed textbooks and the release of new editions can wait for months and often years.

Results of shown research, conducted among students of third and fourth grade of secondary schools on use of printed textbooks from nine subjects in Humanities, Social and Natural Sciences and Engineering, show that textbooks in only three (Mathematics, Foreign language, Croatian) of nine courses are regularly (sum of Very often and Often is more than 50 percent) used in the classroom at school, and in only one (Mathematics) of nine are used regularly during independent work at home.

Research has shown that there are a significant number of subjects for which the classical printed textbooks are very little used and some hardly used (Religion/Ethics, Politics and Economy, Computer Science). This especially applies to use of those textbooks at home, but it is evident that the usage at school for the majority of analyzed subjects are not much better.

Shown results indicate that students are not motivated and do not want to significantly use printed textbooks and that maybe teachers do not insist on printed textbooks usage at school and at home. As could be seen, the data for all three years of research are very similar and in majority almost identical, so these trends warn us and call us for urgent action.

REFERENCES

[1] A. Zuzul and D. Vican, "Textbook in new school", in Life and school, no. 13, vol. 1, 2005, pp. 50–55.

[2] S. Spanovic, "Pedagogical Aspects of E-textbooks", in Educational Sciences, no. 12, vol. 2, 2010, pp. 459-470.

[3] R. Barisic, "Use of Printed Textbooks and Digital Content in Secondary School Education", in Proceedings of the 5th International Conference on Information Technologies and Information Society (ITIS 2013), 2013, pp. 169-176, ISBN: 978-961-92509-9-0.

Online Learning Efficiency in the Humanities

Smiljana Gartner, Marjan Krašna
Faculty of Arts, University of Maribor, Maribor, Slovenia
smiljana.gartner@um.si, marjan.krasna@um.si

Abstract – Online learning or web-assisted learning has become common over the past decade. Various studies have established that it is appropriate for learning in the areas of technology, natural sciences and skills. The arts and humanities are more reluctant to use the new technology, since it seems less naturally suited to their disciplines. Is it possible that the arts and humanities can not find such an approach effective, or does it require a mental change on the part of the of lecturers? We tested three different approaches to acquiring student feedback: (1) the traditional classroom lecture, (2) studying with documents from Moodle, and (3) studying with interactive multimedia learning materials.

In our research (in the course The Basics of Philosophical Thought) we have discovered that students who are digital natives still perceive traditional classroom lectures as more useful in the sense of understanding the reading materials than the other two approaches. This was evident in the students' homework, grades and survey analysis but not in time management. Given assembled data, we suggest that the optimal approach would be to provide e-learning materials prior to the lectures because this would allow students to participate actively in the lectures. If we use online interactive multimedia learning materials, most students still require live communication, which is not always feasible when online learning is the whole method. The online multimedia study approach could therefore be effective in the analytical philosophy course, but it would require a huge mental change on the part of all stakeholders.

Keywords: online learning, humanities, educational method comparison, e-learning materials.

I. INTRODUCTION

In the last decade Slovenian universities accepted the Bologna guidelines and reformed the higher education system into a three-level system. The background ideas are inspiring: an individual approach to students, an active student, interactivity, student-centered study etc. However, the reality of the present financial crisis has meant less money for education. Economic reality, also known as rationalization, is manifested in larger student groups in seminars and for laboratory practice (from 30 to 70), minimized contact hours, transfer of hours to individual student work, and a great many assessments. This has had the following consequences:

- a teacher-centered approach and passive students (in the classroom);
- student and teacher off site preparation, production and verification of coursework and homework;
- many small tasks without comprehension of the topic or of the course as a whole and
- a leak of time for deep understanding of specific topics or themes.

If learning is a constructive process (the constructivist view), where a student builds or deepens an internal illustration of knowledge and is therefore an active process involving individual experience, then we are failing in this objective. The following claims about study objectives sounds like a science fiction at public universities in Slovenia [1]:

- that objectives should be negotiated with students based on their own needs,
- that learning activities should emerge from within the contexts of the student's life,
- that students should work together with peers in the social construction of personally significant meaning, and
- that evaluation should be personalized.

A solution to this discrepancy and a potential benefit to all stakeholders lie in computer-based learning, online learning or web-assisted learning. Although it is evident that such techniques are appropriate in natural sciences and for learning skills, we can also find positive results from the social sciences (especially psychology and language learning) [2] [3] [4] [5] [6]. Nevertheless, the humanities are different, thus posing a dilemma. Can philosophy (the most fundamentally abstract science) benefit from ICT? Benefits are obviously greatest in the sciences, where the level of abstraction is lower. Therefore, we want to know how to incorporate communication, interaction and cooperation (the key elements of philosophy courses) and to analyze how students react to different didactic approaches in Philosophy.

II. METHODOLOGY

We tested three different approaches to acquiring student feedback.

The first approach was *traditional classroom study*. Although this approach is usually understood as lectures with textbooks, notes, exercises and an examination at the end, courses in analytical philosophy, including ours, also comprise the response-based approach, argumentation, class debate, group discussion, active reading, solving case studies and using computer supported learning (presentations and video material). Because of

rationalization, however, these methods have been curtailed.

The second approach was *studying with documents from the electronic classroom* (hereafter called ***online elementary study***). Students were given reading material, online notes and presentations.

The third approach was *studying with interactive multimedia learning materials*, where students had webpages and links, online notes, reading material, online case studies, video, online interaction with a lecturer and quizzes (hereafter called ***online multimedia study***). We use an approach that was previously developed for other students in the design of multimedia learning materials for our course [7] [8] [9].

The data presented here were gathered during the winter semester of 2013/2014 among 1st year students (N=29) in the course "*The Basics of Philosophical Thought*". Students were divided into three groups, each of which used all three of these approaches.

In the study quantitative methods were used, including a questionnaire (N=22, twenty one female, one male) about student self-evaluation of the three approaches in terms of study hours and grades received for their homework. The questionnaire for students consisted of nine questions/claims (two independent variables and seven dependent). Three questions/claims were simple closed questions with a single answer:

- How often do you raise questions?
- I prefer classes where I can participate;
- In lectures and seminars I prefer … (activity in lectures)

We also provided several closed questions:

- Working on-line enables me …;
- The role of a lecturer should be …; and
- a question with eleven sub-questions about the preferred approach to studying).

One question on the questionnaire was open-ended: What do you think are the advantages and disadvantages of the three approaches to studying?.

Students had to submit their homework after every theme, and in their homework they needed to specify the time taken to grasp the theme.

Where applicable, students used a three-level grading scale, responses to which have the following meanings: 1 (mostly true, max); 0 (sometimes true, average), and −1 (mostly not true, min). We decided to use this scale for psychological reasons. Anyone can immediately see the positive and negative aspect of the grade, and the grading process involves more thorough reasoning. For the statistical analysis, we later transcoded the values -1, 0 and 1 to the values 1, 2 and 3.

Our research was based on the hypothesis that multimedia is one of the most useful methods for a response-based approach and that it enables independent learning. Students are considered "digital natives"; they are therefore familiar with ICT and do not need instructions for using such multimedia learning materials. Such an approach facilitates independent work habits, where a student can build an internal illustration of knowledge, while being actively involved. The most important feature, especially in the case of Philosophy, is that the student has the ability to "remove the four walls of the classroom, which seems to be form of prison to reluctant learners and open up a borderless world" [2].

III. RESULTS AND DISCUSSION

A. Most attractive aspects of online learning in a Philosophy course

The most positive aspect of this approach is time management. More than 95% of students think that online learning approach enables them time self-management (Table I). The open-ended question yielded the same results about the positive and negative aspects of these three different approaches. (Table II)

Table I: Working on-line enables me...

Row Labels	Self-managing time	%
0 (No)	1	4.54
1 (Yes)	21	95.45
Grand Total	22	100.00

Table II: Self managing time (open question)

	classroom	Online - elementary	Online - multimedia
time	1.24	2.33	2.76

There are differences (Table II and Table III) between the online multimedia study and online elementary study approaches. One reason for this could be time constraint with the online multimedia approach. Both quizzes and online lecturer access were available for a limited period of time. In the classroom approach, students had to be present one hour per week, with additional two optional hours per week. The rest of their study hours were up to them, and any additional time - even the smallest amount - added to their assignments was perceived as a restriction and consequently as an inability to self-manage their time. This is interesting because, when one is reading demanding philosophical texts, having a lecturer by your side is always an advantage. One explanation could be that students are unfamiliar with online multimedia study. a hypothesis which is compatible with the finding that it is important to give students time to familiarize themselves with the tools, the work and the whole online multimedia study environment [10].

Another positive aspect involves access to course material. Students can access learning materials at all times and also have more time for studying (reading articles and watching video material). Moreover, on-line multimedia study is more attractive and offers more interaction than the online elementary study approach (*see* Table III). Therefore, if we want to talk about the positive aspects of the on-line approach, this must mean the on-line multimedia study approach.

Table III: Positive aspects of on-line approaches

	Count of online multimedia	%	Count of online elementary	%
Time-managment	8	40%	12	60%
Reading material	3	15%	3	15%
Communication, interaction	4	20%	1	5%
Interest	4	20%	2	10%
Remembering	1	5%	2	10%
Grand Total	20	100%	20	100%

B. Least attractive aspects of on-line learning in a Philosophy course

Activity and motivation

The results of students' preferences about being active are interesting. Namely 81.82% of students answered that they liked lectures where they could participate, 18.18% that they did not (Table IV). At the same time, more than half the students (59.09%) said that they would rather just listen during classes, 18.18% of students chose listening and raising questions (2); 22.73% of students chose listening and debating (3) as preferable activities (Table V). Even more, 72.73 % of students claimed that during the class they rarely ask anything, 18.18 ask occasionally, 4.45% very often, whiled 4.55 never (Table VI).

Table IV: Participating in class

	responses	%
1 (yes)	18	81.82
2 (no)	4	18.18
Grand Total	22	100.00

Table V: Activity during the lecture

	responses	%
1	13	59.09
2	4	18.18
3	5	22.73
Grand Total	22	100.00

Table VI: Asking questions during the lecture

choices	responses	%
1	1	4.54
2	16	72.73
3	4	18.18
5	1	4.55
Grand Total	22	100.00

We can infer that they would like to be active but not in such a way as to be singled out in front of the group by asking questions or debating. This is especially because, in class Philosophy, the most common replies to student assertions are "Why?" and "How come?", with which they are learning to justify every claim and to make arguments supported by evidence. For some students these queries could be, at least initially, quite stressful. This is good information for the multimedia approach in Philosophy class because we can prepare interactive exercises and learning methods that avoid forcing a student to endure needless individual scruting.

At the same time, students think that being in the classroom is more motivating; there are fewer chances to skip the study material, and they have higher levels of concentration (Table VII). By comparing the online multimedia study approach and the online elementary study approach, we can infer that students have lower levels of concentration in the multimedia approach but are still more motivated. Taking into account the answers to the on open question, they are more motivated because the video material and quizzes are more interesting; they have the feeling that they are more active, but their thoughts can easily be diverted to something else on the internet. With these factors in mind, it is understandable that with online multimedia there are fewer opportunities to skip the study material.

Table VII: Concentration, activity and motivation

	classroom	Online - elementary	Online - multimedia
Concentration	2.57	2.00	1.81
being inactive	1.71	2.33	1.95
motivation	2.62	1.76	2.10
skiping the studying material	1.71	2.57	2.29

When we considered whether concentration is influenced by the online-elementary and online-multimedia types of study as compared to the classroom type, we discovered that there are no significant differences between the classroom vs. online-elementary groups ($\chi^2 = 6.33; P = 0.176$), and the same is true for classroom vs. online-multimedia ($\chi^2 = 7.55; P = 0.110$) . There are however significant differences between the groups classroom vs. online-elementary ($\chi^2 = 10.72; P = 0.030$) and also classroom vs. online-multimedia ($\chi^2 = 11.43; P = 0.022$) in the level of student activity. Interestingly, we discovered no significant difference between classroom and online-elementary in motivation ($\chi^2 = 4{,}84; P = 0.304$) , but there are significant differences between classroom and online multimedia ($\chi^2 = 13.11; P = 0.011$) in motivation. There are no significant differences in the area of skipping learning materials between types of study ($\chi^2 = 2.76; P = 0.598$), ($\chi^2 = 6.32; P = 0.177$).

Understanding, knowledge retention and depth of knowledge (course material, instructions)

Students still think that the most effective way to study philosophy, i.e. to understand (100%), to memorize (77.27%) and to acquire depth of knowledge (77.27%) is in the classroom (Table VIII). This is especially true in cases where the classroom experience means having to read demanding articles. The main reason is the presence of the lecturer and interaction with her and with other students (Table IX). From Table IX, we can again find differences between the online multimedia and online elementary study approaches. Thus, with the multimedia approach, they also had the chance to have an interactive conversation with the lecturer at a predetermined time, but no one used it. The reason could be, again, unfamiliarity with this approach because asking questions, peer teaching, seeing how and what others are doing and collaboration among peers are some of the main elements of the online multimedia study approach. [11]

Table VIII: Working on-line enables me to ...

	Understandi material	%	Acquire depth of knowledge	%	Remember	%
0 (No)	22	100.00	17	77.27	17	77.27
1 (Yes)	0	0.00	5	22.73	5	22.73
Grand Total	22	100.00	22	100.00	22	100.00

Table IX: Understanding and remembering

	classroom	Online - elementary	Online – multimedia
Remembering	2.62	1.86	1.95
Interaction with lecturer	2.76	1.48	1.67
Understanding the reading material	2.76	1.76	1.90
Understanding instructions	2.81	1.57	1.76

In the case of remembering, we discovered significant differences between the groups classroom vs. online-elementary ($\chi^2 = 16.63; P = 0.002$) and classroom vs online-multimedia ($\chi^2 = 18.27; P = 0.001$) . In interaction with the teacher, significant differences are expected and confirmed between classroom vs. online-elementary ($\chi^2 = 22.30; P = 0.000$) and classroom vs online-multimedia ($\chi^2 = 11.47; P = 0.022$) . In understanding the reading materials, we discovered no significant differences between classroom and online-elementary ($\chi^2 = 7.89; P = 0.096$) , but there are significant differences between the classroom and online-multimedia groups ($\chi^2 = 10.84; P = 0.028$). Using non-parametric Friedman test we discovered that there are significant differences between groups in understanding of the reading materials (Table X).

Table X: Non-parametric Friedman test

	$\bar{R}$	$\chi 2$	P
classroom	2.67	17.82	0.000
online elementary	1.60		
online multimedia	1.74		

Interestingly, there are also significant differences between groups in understanding instructions classroom vs. online-elementary ($\chi^2 = 7.38; P = 0.025$) and classroom vs. online-multimedia ($\chi^2 = 15.33; P = 0.000$)

Elapsed time and degree of effort put into the work

According to most students, working online did not enable them to work less (86.36%), and they had to invest the same or more time and effort in online learning, while at the same time they think their work has less depth (Table XI). However, there is a huge difference between the online multimedia study approach and the online elementary study approach. What we can see in Table XII is that the time elapsed while finishing a task depends on the theme and approach. "Exercise 2" was about contemporary ethical dilemmas like euthanasia, abortion and genetic engineering, including considerable video material; therefore, students needed less time to complete the tasks. "Exercise 4", on the other hand, was about abstract theories and comprised only reading material. Students in the classroom used half the time in comparison to those that working with online multimedia, and a third of the time compared with students using the online elementary method. Therefore, when a lesson includes considerable reading material, a lecturer is irreplaceable for directing the reading of articles, extracting the most important parts, explaining complicated issues and making analogies with contemporary (ethical, political) dilemmas (Table XII, column four, i. e. "Exercise 4"). She can do this by asking questions and so engaging students even more. If this works, then the same thing could be used in the online multimedia study approach and consequently form a positive aspect of its use. This is compatible with a research study which stated that "*active involvement of the facilitator is needed in the process of interaction, communication, and discussion, especially in the forum because this activity is a priority for learning via website*" [10], as well as with the assertion that online teaching will generally result in a significant increase in teacher effort per student [11].

Table XI: Working on-line enables me...

Row Labels	Least amount of work	%
0 (No)	19	86.36
1 (Yes)	3	13.64
Grand Total	22	100.00

Table XII: Average time taken for the work in hours

	Exercise 2	Exercise 3	Exercise 4	Exercise 5
classroom	3.28	3.74	1.68	4.40
online elementary	3.06	4.16	6.23	6.47
online multimedia	2.82	2.56	4.04	4.80
Total Average	3.05	3.59	4.09	5.14

IV. CONCLUSION AND RECOMMENDATIONS

The main question we tried to answer was: "*Is it possible that the humanities (specifically Philosophy) cannot find online multimedia study approach effective, or will it require a mental change on the part of the lecturers and students?*" This is especially because the constructivist approach is crucial in courses of analytic philosophy; i.e., a student is constructing or deepening an internal illustration of knowledge and it is therefore an active process through her own experience. Students work together with peers in the social construction of personally significant meaning, and the program is very individualized. In our research (in the course The Basics of Philosophical Thought), we discovered that students who are digital natives still perceive traditional classroom lectures as more useful in the sense of understanding, extending and comprehending the course materials. This was evident in the students' homework and survey analysis but not for their time management or motivation. Although there is a discrepancy between what students think would work for them or what they would like to have and what they are prepared to invest, the online elementary approach is the least effective. But because the positive aspects of working in the classroom were understanding, asking questions and having guidance from the lecturer, all of these could be accomplished through the online multimedia study approach. Therefore,

the following conclusions can be drawn for a class in analytic philosophy:

(i) the online course must be the multimedia type because, according to our research, although the online approach allows students time self-management, freedom and autonomy, only with multimedia study are students active; they participate, they try to understand and comprehend material and they can experience almost all instructional elements as in the classroom;

(ii) even though these students are digital natives, they do not know how to use online multimedia materials and services. Therefore, we need to be aware that, for success in studying, they must learn how to use multimedia or at least must become familiar with it (for example, using online consultation, online interactivity with other students etc.);

(iii) if we use the online multimedia approach, it is better to apply to the whole course (or in combination with classroom instruction i.e., blended learning), otherwise, it could be difficult for them to switch between different approaches.

From the assembled data, we suggest that the optimal approach at this moment would be to provide e-learning materials prior to the lectures because this would facilitate active student participation in lectures. If we use online interactive multimedia learning materials, most students still require live communication, which is not always feasible when the online learning is the only option.

Thinking that the university would save a lot of money by implementing an online multimedia study approach is the wrong way to go; if we want students to develop, to achieve the same or better results than with using the traditional, classroom, approach, we need to invest money. As with any other approach, to achieve any goal, it needs to be properly done, and this means allowing for additional education for students as well for lecturers. Preparation of an online multimedia course and its effective implementation means considerable knowledge and a great deal of work outside the classroom. Therefore, we would need new criteria and new standards for judging the work of lecturers, ones where lecturers would not be judged on hours in the classroom but on hours invested in the successful accomplishment of the goals and aims of the course [12]. Therefore, implementing an online multimedia study approach, even if this could be a very effective approach for a course in philosophy, we should be very careful. First, people might think that teachers are available 24/7, which would jeopardize their private time and raise stress levels to the maximum and, second, because teachers would not be in the traditional classroom, people could think that they were doing no work, so – who needs lecturers anymore?

Of course, we can overcome these obstacles by implementing the idea of an online classroom in the sense of a virtual online classroom, with everyone present online at a certain time, but then (i) everyone would need to have a complete set of multimedia equipment. However, (ii) at the same time we would be reducing what the students stated as an absolute preference: i. e., time self-management while (iii) debating, the sharing of opinions and dealing with opponents through verbal and non-verbal communication would still be missing.

Therefore, our answer is, yes, it is possible for a philosophy course to find the online multimedia study approach effective, but it will requires a huge mental adjustment on the part of all stakeholders.

V. REFERENCES

[1] S. Hackbarth, The Educational Technology Handbook, New Jersey: Englewood Cliffs, 1996.

[2] E. I. Allen in J. Seaman, „Class Differences: Online Education in the United States, 2010," Babson Survey Research Group, 2010.

[3] H. Mushin, „The Using of E-Learning Techniques to Improve the Medical Education," v *Information and Communication Technologies: From Theory to Applications, 2008*, Damascus, Syria, 2008.

[4] S. Richards, „Computer-based learning in psychology using interactive laboratories," *The Journal of the Association for Learning Technology (ALT)*, Izv. 2, št. 2, 1994.

[5] N. Radzi, S. Sivapalan, W. Ahmad in A. Abdullah, „Incorporation of Multimedia Approaches to Literature Online via E-learning for foundation level literature courses," v *Information Technology (ITSim), 2010*, Tronoh, Malaysia, 2010.

[6] C. A. Twigg, „Improving Learning and Reducing Costs: New Models for Online Learning," *EDECAUSE review*, pp. 28-38, 2003.

[7] M. Krašna and T. Bratina, "E-learning materials for social science students," *Philosophy of mind and cognitive modelling in education - 2014*, vol. 61, pp. 77-87, 2014.

[8] M. Krašna, M. Duh and T. Bratina, "E-learning next step - learning materials for students," in *MIPRO 2014*, Opatija, Croatia, 2014.

[9] T. Bratina, M. Duh and M. Krašna, "E-learning controversy in practical application," in *MIPRO 2013*, Opatija, Croatia, 2013.

[10] A. M. Mahamad, "Students' opinions on the usage of online learning," in *ICT4M - Information and Communication Technology for the Muslim World*, Sarawak, Malaysia, 2014.

[11] G. K. T. Stahl and S. D., "Computer-supported collaborative learning: An historical perspective," in *Cambridge handbook of the learning sciences*, Cambridge, UK, Cambridge University Press, 2006, pp. 409-426.

[12] M. Krašna, "E-learning materials development projects," in *13th International Multiconference Information Society - IS 2010*, Ljubljana, 2010.

Designing m-learning materials for language learning

Dejan Pukšič, Marjan Krašna
University of Maribor, Faculty of Arts, Department for Slavic languages & Department of Pedagogy Maribor, Slovenia
dejan.puksic@student.um.si, marjan.krasna@um.si

Abstract: There are hot debates if e-learning and m-learning are the same thing. This is a dilemma only for those who were now participate in the m-learning learning material production. From technical perspective we might see this just as viewport constraint. But those constraints have high implication into the background thinking processes. Classical windows navigation cannot be used and if we want m-learning materials to be successful we need to pay special attention to the just-in-time learning. Much effort needs to apply to the context when and where information is searched and displayed to statistically predict following user requests. Only such materials that are successful in prediction of user's actions are successful as learning materials and do not becomes "digital rubbish" on mobile device. The article is focused on the design of m-learning for Slovene language on level A1 and A2 which is of paramount importance for further motivation for language learning.

Keywords: language learning, m-learning, haptic user interface, design and development,

I. INTRODUCTION

It is common belief that Google will solve language barriers in the near future. Promising video can be seen on YouTube but reality is somehow different. It is possible to translate between languages if text in one language is written in specific way [1]. Such translations are not perfect but sufficient. For the purpose of understanding the text all linguistic subtleties are not required. But there are also another topics in the language. Each living language has phrases and understanding the phrase and proper translation/interpretation is still domain of the human being. According to these facts it is now the right time for the question: "Should I learn another language?" Answer to this question is not the easiest in today's fast moving World. But if the answer is "Yes" than it becomes our business.

There are many levels of competences of foreign language. In this article we are going to apply the "Common European Framework of Reference for Languages". As an educational institution for teachers and even language teachers we have many foreigners who attend language courses for different purposes. Foreign students from the students exchange come for one semester, some student want to study in our country and there are even some who want to live in our country. Their motivation is different but they all want to become familiar with the language on different scale. The first contact with the foreign language is the most important for the perseverance and further learning. If frustration level reaches certain level people generally give up and stops with the attempt to learn further. We wanted to reduce frustration level for Slovenian language learning on the level A1 and A2. Slovenian language is one of the oldest language since it still has dual and cases and as such is not easiest for learning. Especially it is hard for translation from foreign language into Slovene since some information for translation is missing and brute force approach misses the context. From Google translate we can observe much better translation from Slovene to other language than in the opposite direction. Other online translation has the same problems. How is it possible to help learners to decrease their frustration level? There are many possibilities. From the past we know dictionaries with the translation of most common sentence. In contemporary mobile world we can use mobile devices to ease the old principle and such software are available on the software markets. From the analysis of such software we discovered that we could do better. Applying the principle missed in the other software – "*just in time*" language assistant we could provide learners with the right information at the right time. This could be achieved in two distinct principles: applying artificial intelligence module or statistical analysis for the best scenario. It is possible that artificial intelligence would give us the best results but it would require much more computational power than contemporary portable devices can offer. On the other hand statistical approach is much more energy efficient and suitable for portable devices but it requires intelligent multilevel planning. It is necessary to recognize the situation where the user is when he/she needs the language help and what are the best possible next language helps he/she will need. In general this is not feasible therefore we need to compromise. The compromise is the intelligent user interface that minimizes the search of the right linguistic help. Designing such interface from the scratch is difficult therefore we set our initial aim to provide students with the tool to ease the accredited language course from the (government?) which are the mandatory course for all foreigners who wants to learn Slovene for different purposes.

II. M-LEARNING VS. E-LEARNING

It is not without the reason to say that today new countries are first visited on the web. Later it continues in the real environment among the people communicating in that countries language. Today people in the western world are multilingual and visiting foreign countries is not such linguistic and mental shock of being shut out from the conversation. English is speaking all over the World is a phrase that becomes reality and basic communication is feasible. But we all know that speaking English and "*spekin Englis*" can be a problem in anything but fundamental communication. Therefore anyone who travels frequently knows the phrase "*spikin inglish good*" used in any response and in all situations. At that time technologically advanced humans resort to the Google all mighty on its portable "*smart*" device. What should be simple is not always the case in foreign country with inadequate ICT infrastructure or "*kidney priced*" Internet services. Therefore despite the 65% of world population have smarphones (and growing) [2] free Wi-Fi is generally not available and roaming costs are simply too high to be useful. In such cases we start to understand the differences between e-learning and m-learning. Language e-learning projects are mainly reserved for controlled learning environment in school or at home but m-learning modules are included into the other real life situations where talking and language understanding is natural part of the learning. General aim of mobile learning known from all its definitions is "*anywhere and anytime*". M-learning materials developer should foresee and recognize potential learning environments. He/she should pretty well know the users of language application (MALL – Mobile Assisted Language Learning), and what are their users' habits. On the other hand it is paramount to know where the mobile application will be used, in domestic or foreign environment, is the web available or not, and which functionality of the phone should be included into the application. All those constraints were taken into account when we design the pilot application for mobile learning of Slovene language as foreign language – application BlaBla™.

III. MOBILE ENVIRONMENT

In the mobile application development many questions arose. M-learning materials design and development is no exception in this process. From the perspective of the e-learning materials development it is just another device and no big deal. But deeper understanding the mobile environment shows us that we cannot dismiss this topic so lightly [3]. In the recent years we can find the changes in the m-learning perspective where m-learning materials should be distinguished from e-learning materials in general. Mobile environment is full of selective disturbances, from physical contacts to sound and light oscillations that are not present in controlled e-learning environment (home or in school) [4]. Therefore the transition of e-learning materials to the mobile device is not the valid option since they are meant to be used in controlled environments. M-learning is still considered the extension of the e-learning but the developer soon discover that this is not the case. User interface needs to be designed differently for mobile devices and in totally new mobile context.

A. Mobile context

Mobile context has changed since smart mobile devices are not only portable versions of other stationary devices. They presents specific technological segment of devices. The main perpetrator for this is the touch sensitive display that defined the mobile phone devices as devices that have phone app on them. Therefore the only correlation between previous mobile phones and smart mobile phones are three apps (phone app, messaging app and ringing app) with addition of speaker and microphone. Smartphone is therefore phone only by its name and does many other things that are far beyond the phone's primary purpose. It was proven that haptic user interfaces are more suitable for learning than interfaces with the proxy (mouse, touchpad, or keyboard) because the system react to the direct sensomotoric coordination i.e. moving the body part (touch). Interaction with the e-learning content is direct and does not need mental translation [5] [6, p. 2]. This can be best observed with the young kids that master tablets and smartphones without any training. The keyboard, mouse and other input devices are strange to them completely and they are not capable to establish cognitive link and transform the movement of proxy devices to the virtual world (mouse forward/backward is screen up/down) [5]. Any deviation from this natural interface proves to be unsuccessful on the long run and application that were not designed with that in mind were soon forgotten. Applications for smart devices need to be designed for smart devices or face consequences – loss of effectiveness and user frustration.

B. Touch screen

Basic guidelines for handheld user interfaces design derive from Shneiderman's fundamentals rules for effective interaction between human and PC assurance [7, pp. 74, 75]. Some are general guidelines like: 1. Enable shortcuts to the regular users, 2. Assure information feedback, 3. Cascade design of user interface etc. As we see these guidelines are of no use in the m-learning application for language learning design. Even after its redesign in 2004 [8, p. 3751–3754] they are still obsolete since in 2004 were no contemporary smart mobile devices. Mobile ideology has changed significantly since 2004 and users behavior perspective is now fully focused on haptic displays. New design proposals were prepared for m-learning materials development for haptic user interfaces that are good foundation for our language learning m-learning materials [9, p. 227]:

1. Navigation needs to be simple, clear on all levels and consistent throughout the application (#consistency)
2. Application needs to be user friendly and enable user to master its user interface in minutes (#ergonomics)
3. Learning topics needs to be designed and organized to prevent frequent scrolling (#scalability)
4. Only relevant information for learning processes needs to be presented (#relevance)

5. Text should be limited to the least possible volume (but not more) and information should be presented with images, animations, and videos (#minimalistic)
6. Control over learning processes is the users domain (#control)

IV. BLABLA™

Application BlaBla™ was designed and developed according to the previous guidelines to support the intensive language course of Slovene language for foreign exchange students – EILC (Erasmus Intensive Language Course) [10]. The intensive Slovene course is designed for students to motivate them for language learning and support them in the predetermined situations. Our pilot research was to prove the concept only therefore not all topics are implemented. But those that were implemented were carefully selected to get the best possible response form the users using m-learning materials for language learning the first time. Application has four (4) thematic parts from the beginning of language learning for foreigners on the level A1/A2 (Alphabet & Numbers, Introduce yourself, Describe your family and What are you wearing). These topics were selected because of the two years' analysis of the most probable scenarios where students encounter themselves in the communication with the natives who do not speak foreign language. Our goal was to help foreign student to retrieve their knowledge in the routine situations and speech environments with interactive multimedia elements and learning games.

To achieve #consistency and #ergonomics we situate the navigation elements on the permanent locations where Android operating system would naturally put them. Application was developed for Android system since the analysis shows that most smartphone users that came to the Slovenia use Android (79%) (for our courses 54,2%). Application has menu, buttons, icons and interactive contents. All interactive e-learning content was located in the ellipsoid thumb location reach – primary control area. Others less frequent elements (system commands, exit, menu, background theory explanation) were located outside this primary area.

Figure 1: Phone control area.

Putting less frequent commands and exit button in the secondary control area we minimize the possibility to accidentally trigger them. Small screens are still challenge to effective display of educational content the application of the #minimalistic rule and #scalability was used for chunking the content into fundamental parts – learning pads (least possible size m-learning module).

Figure 2: Chunk of lesson.

Such small chunks present one situation (question and answer) where both speech positions are presented to better understand the situation. Chunks of knowledge are in the coherent relation with other chunks and they are

designed not to overload the learners. Learning chunks include only #relevant content presented in such way that they seems natural and generic in all mobile operation systems (i.e. calendar view for number presentation, communication windows for SMS and social interaction). Wherever possible we use pictograms and/or single words. Transition between chunks were designed to encourage learners to continue learning even they have full #control over the application and can always exit. At the end of lessons learners have the ability to enter the learning game to verify their acquired knowledge or they can continue with the next lessons. All lessons are in the thumbs reach and navigation enables quick search between stored lessons to retrieve relevant content in the specific situation.

Figure 3: Educational game and quiz.

Figure 4: Navigation in the language learning application.

V. RESEARCH

In our pilot research 24 students (12 male and 12 female) from Czech Republic, Estonia, Finland, Lithuania, Hungary, Norway, Portugal, Spain, and Turkey were included. Students' age was between 21 and 25 years. We prepare two groups of 12 students based on their devices. Other 12 students were the control group. Students from the experimental group were given the application to use it during the intensive Slovene course. Small groups render more complex analysis questionable but for the purpose of our pilot research descriptive part is sufficient.

Table 1: Usefulness of application

	No.	%
Not at all	0	0
Ok, but too few topics	3	25%
Useful	7	58,3%
Very helpful to me	2	16,7%

Research has shown that more than half of attendants assess the application as useful (58.3%). Even those who decide that application have too few topic actually did not assess that application was not useful but just need to be filled with the additional lectures/topics. Since no one have assessed application as not useful at all we can confirm that our approach was the right one.

Table 2: Which visual mode you have found most effective in the individual options?

	No.	%

Tap-on-word	4	33,3%
Tap-on-pictogram	2	16,7%
Tap-on-picture	3	25%
Allareequally effective	3	25%

Mobile devices require extensive attention in the visualization of the content. Small screen prevents the display of anything but essential for given learning topic. Learning "*just in time*" requires prediction for the situations where learners would require assistance and fast mental flow from device to the user. Despite pictograms and images presents least cognitive load users decide to use text (reference language of the application) as favorite navigation. The answer to this question could be the topic of another research. Processing of images is always in the primary language (mother's thong) of the learner but in the multilingual environment (system notification in another language) possible require high mental load in the language switching which is not always simple even for the fluent speaker. This phenomenon would really require another research.

Table 3: Do you find the language theory section useful?

	No.	%
Not at all	2	16,7%
Ok but too complex	3	25%
Useful	7	58,3%
Very helpful to me	0	0

Theoretical grammar background can be a problem for heterogeneous learning groups. Learners originate from different language groups and socio-environmental background and topics in language they learn may be totally alienated from their perception of language. Learners have also different attitude to the grammar. Language specialist want to understand the background of mental structure of the sentences while others want to use language in the day to day activities only (for the beginning). To understand why the grammar is not useful for (some) students we need to ask different questions and expand research in that area. Probably this research would be more useful in the higher level of language learning (B1/B2). But even at this lower level of language learning we receive positive feedback from 83.3% of students. It is evident that grammar should be explained accordingly to the capabilities of the learners and that is always problematical since grammar is the basic building blocks that has only one explanation.

Table 4: Do you find the native sound support useful?

	No.	%
Not at all	0	0
Ok but I don't really need it	4	33,3%
Useful	3	25%
Very helpful to me	5	41,7%

Sound of the language (pronunciation) gives the learner full immersion into the language learning. Only correct pronunciation enables foreigners to join the conversation in one language space. Analysis show that speaking language is "ok" but depends on learning group. Slavic learning group rarely used language pronunciation but students from another language groups did. This give us the feedback that we could include students' primary language as the differentiation factor into the design of the m-learning language application.

Table 5: Do you find learning games useful?

	No.	%
Not at all	1	8,3%
Ok but too easy	4	33,3%
Useful	5	41,7%
Very helpful to me	2	16,7%

Learning games were the easiest and most effective motivation for drill and practice in knowledge acquisition. This known fact was confirmed by 91.7% of our students. In our test application we include only one type of the game/quiz ("pick the right answer"). We would probably get much better results if more games' types. From the 33.3% of students we also receive the feedback that quizzes were too easy. Therefore we could prepare more complex quizzes for them but there is always a fine line between too easy and too difficult. We never wanted to get into too difficult area since this would frustrate learners and probably move them away from the language learning.

Table 6: What would you add to the application?

	No.	%
More topics	10	83,3%
More sounds	1	8,3%
More theory	1	8,3%
More games	0	0

As expected most users would like more topics in the application. Therefore m-learning has the bright future and should be considered the valuable paradigm in the future TLS (Teaching, Studying, and Learning) processes. Students' feedback from this application was of a high importance for language learning. We were able to verify our expectations and acquire new knowledge about language learning students. Probably most important consideration in the application design and development is the "*content for the users*" and not "*users, here you have content*".

VI. CONCLUSION

Mobile learning (m-learning) is the way where "anywhere and anytime" changes the "map" of learning. It introduces the new paradigm to the TLS (teaching, learning, studying) processes what were not possible 10 years ago. In the past and even today we recreate proxy real situations to teach some topics but with the m-learning we can challenge the learning in all life situations. This is particularly useful in the language learning – where "just in time" gets effective meaning. Learning of language with the book and sound recordings in the classroom is still not dead but it is a dead way to learn the language. Smartphones as the learning aid have the ability to change that to the live language learning. It is not the problem of lack of technology because at least half of the people own at least one smartphone. Key to

the success is therefore not the classical infrastructure for language learning but mobile language learning. Mobile learning application for language learning needs to be designed accordingly to the new quality standards where haptic user interface have to be fully respected. Haptic user interfaces enables direct contact with the knowledge and multitude of new opportunities how to structure learning materials. Despite multiple tries to define rules for m-learning design this area remains virtually "untouched" in the expectation of the next big thing to ease the language learning.

LITERATURE

[1] M. Krašna, B. Bedrač in V. Jesenšek, „Computer support for phrases translation," v *IEEE Global Engineering Education Conference (EDUCON)*, Berlin, 2013.

[2] eMarketer, „Smartphone Users Worldwide Will Total 1.75 Billion in 2014," 16 1. 2014. [Elektronski]. Available: http://www.emarketer.com/Article/Smartphone-Users-Worldwide-Will-Total-175-Billion-2014/1010536. [Poskus dostopa 4 2. 2015].

[3] M. Krašna, M. Duh in T. Bratina, „E-learning next step - learning materials for students," v *MIPRO 2014*, Opatija, 2014.

[4] P. Tarasewich, „Designing mobile commerce applications," *Communications of the ACM – Mobile computing opportunities and challenges*, p. 57–60, 2003.

[5] N. Charness, P. Holley, J. Feddon in T. Jastrzembski, „Light pen use and practice minimize age and hand performance differences in pointing tasks," *Human Factors*, p. 373–384, 2004.

[6] A. Collins McLaughlin, W. A. Rogers in A. D. Fisk, „Using Direct and Indirect Input Devices: Attention Demands and," *ACM Trans Comput Hum Interact*, p. 1–15, april 2009.

[7] B. Shneiderman, Designing the User Interface: Strategies for Effective Human-Computer Interaction, Menlo Park: Addison Wesley Longman, 1998.

[8] J. Gong in P. Tarasewich, „Guidelines for handheld mobile device interface design," v *Proceedings of the 2004 DSI Annual Meeting*, Boston, 2004.

[9] A. S. Hashim, „A Study of Design Principles and Requirements for the M-Learning Application Development," v *User Science and Engineering (i-USEr), 2010 International Conference*, Shah Alam, 2010.

[10] CMEPIUS, „CMEPIUS - Center RS za mobilnost in evropske programe izobraževanja in usposabljanja," 2013. [Elektronski]. Available: http://www2.cmepius.si/en/files/cmepius/userfiles/razpisi/era13/EILC/A_EILC_2013_UL.docx.

Comparison of ICT in education between Slovenia and Serbia

Marina Popović*, Dragan Cvetković**, Marjan Krašna*
* Faculty of Arts, University of Maribor, Slovenia
** Faculty of Education, University of Novi Sad, Serbia
marina.popovic@student.um.si, dcveles@gmail.com, marjan.krasna@um.si

Abstract: Cooperation today is on a global scale and anyone is welcomed to participate. Global standard in education means that ICT becomes backbone of the education infrastructure and incorporated to the educational paradigm. Europe dictates that student and teacher mobility is almost mandatory. Our intention was to see the difference between two countries that should start using their ICT in education concurrently. With the historical overview, contemporary research on comparison of study programs, and students' opinion we provide the comparison that shows potential for cooperation and mobility. Today's students are digital native in both country and in the article we are present the differences through their views and opinions. We constraint our research on students of educational study programs in three universities in Slovenia and one in Serbia.

Keywords: ICT, education, university study programs, special didactics, student survey

I. INTRODUCTION

ICT in education is the driving force to change curricula at all levels of education. New generations grow with it and use them on a daily basis. These generations, one defined as "*digital natives*", are now more "*always on*" generation. Whoever lectures the students know that students who supposed to listen the lecturer are actually always present on their social networks. Their ability to check the validity of the knowledge they get is instant.

New generation of student require better learning materials. Books are fine but too time consuming. The changes are constant and all present. Though we have introduced ICT in education of teachers in the previous century [1] it is still evident that we have not done enough. From the past project reports "Status and trends of application of ICT in Slovenian schools" it is evident that growth of ICT is not self-evident and require funding [2] [3] [4]. But in the last decade we have huge development in e-learning materials production where university level of education was excluded. [5]

It is evident that we need to prepare teachers to this new reality and give them instruction how to use and prepare their didactically suited and aesthetically pleased learning materials [6] [7] [8] [9] and according to their study discipline [10].

II. HISTORICAL OVERVIEW

The majority of developing countries, among them also former Yugoslavia started the implementation of computers in the educational system relatively late and uncontrolled. The process of implementation only started in the era of microcomputers. Usually it was a centralized system of equipping mostly with imported computers due to the fact that there was no computer industry of their own for these countries in the early eighties. In Yugoslavia, including Slovenia, programmed courses came into effect in the seventies. [11] Until 1970 computers were too expensive and too, while teachers had to make their own teaching materials, sadly without the wanted effect and result. As a consequence there was a minor usage of computers in schools. [12] In every educational institution following the SFRY there was only a few qualified teachers able to teach ICT and computer technology and a few of those using computers for their work. [13] [11]

In the time line below, we will expose events that were crucial for the development of computer science and technology in individual areas:

Table I: Historical presentation of events

1968	First meeting of the Initiative group for the usage of cybernetics in pedagogy in Zagreb.
1970	Joint meeting of representatives of SFRY`s Institutes of Education in Čatež.
1970	SLO: Implementation of computer technology as a course in 3rd and 4th year of secondary schools.
1971	Federal agreement (document) titled The Suggestion of long-term Modernization of Educational Activities in SFRJ, international conference of innovations in educational technology in Subotica, the annual school of andragogy with a seminar themed by PUK and learning with the help of a computer in Poreč.
1971	First international research center for programmed courses and implementation of computers – Multimedia center of the Referral center of University of Zagreb.
1972	Start of computerization of schools in Slovenia and other republics.
1972	The establishment and operation of Computer center for programmed learning (RCPU) led by Aleksandra Kornhauser.
1974	I. Bratka and V. Rajković write the first textbook titled Introduction into Computer Technology and Collection of Tasks by a group of authors.
1976 – 1979	RCPU is leading and coordinating the project Computer Technology Courses in Guided Education.
1979	SLO: Computer technology becomes a regular subject in 4 year programs of technical schools.

In a period of ten years a lot of events took place that contributed to the development of computer technology, although we have to emphasize that considering the time of the early start of implementation of computer technology as a study course in secondary schools, there are still schools today with no ICT or computer technology courses.

Table II: Historical presentation of events

1983 - 1984	SI Introduction of optional courses, extracurricular activities and supplementary education for teachers in the field of computer technology.
1985	SI: The establishment of a work-group i.e. a computer technology commission led by V. Rajković. Ivan Gerlič was in charge of educating teachers. The objective was to develop a course of computer technology and encourage its usage in Slovenia.
1985	SI: The establishment of the Center for computer technology in education at the Faculty for Pedagogy in Maribor, where the objective was to increase the usage of computers with didactical courses.
1986	An organized and mass implementation of computers in education starts in other republics and countries
1986	SI: COMPUTER IN SCHOOL – designing a politics of equipping Slovenian schools. The first appendix, Computers in School is published within the magazine Upbringing and Education that deals with problems that occur with the usage of a computer, and a book collection Computer Horizons.
1987	SI: The subject of Computer technology in secondary schools is added a chapter of informatics and therefore a new subject arises, Informatics and computer technology – ICT.
1987	Yugoslavian conference titled Computer in Education which conclusions and observations were a basis for further politics and strategy in implementing computers and contemporary informatics technologies into education. There was also a debate of the politics of modernization of educational technology.
1987	SI: Project COMPUTER DAY – intended for all teachers of primary schools and guiding counsellors of Institution of RS of Education.
1987	SI: Implementation of the following projects: CIRCLING CIRCLES, LOGO in primary schools, and MEASUREMENT COMPUTER in secondary schools, ROBOTICS in primary and secondary schools.
1989	SI: Within the project of DUCK-89 (RAČEK-89) an action of an organized purchase of computers and publishing of a book collection The Duck`s Little Book (containing mostly manuals for educating teachers) is carried out.
1989	SI: A new publication NIT is published – New Informatics Technologies, which substitutes the dropped appendix Computers in School of the magazine Upbringing and Education.
1990	SI: Project DUCK-90 – computers and licensed software is bought by a large number of primary and secondary schools.

In the early nineties the first implementations of projects of unified equipping of secondary schools and HR secondary schools with personal computers have started. It was already known at that time that the action of equipping schools with computers is not a one-time action but a long-term process, which is supposed to develop continuously.

The development of computer industry has picked up speed and it still shows no decline while the hardware is getting even more functional, more powerful and what is most important more financially available to masses. [13]

Following the disintegration of Yugoslavia, Slovenia led a new path of developing computer science and computerization of schools as an independent country. Towards the end of the nineties, Slovenia has moved away from others previously more or less equal Yugoslavian republics and continued to develop for herself, which has gained her advantage over others.

Table III: Slovenian project for ICT in education

1992	With the project PETRA a process of computerization of the fifth grade of primary school is implemented. Courses such as Slovene, Art and Technology are taught with the help of a computer. Beside that the project brought novelties such as team teaching and learning, team learning for teachers and the development of didactics of ICT.
1994 – 2000	Realization of a six year program Computer literacy - RO – creates opportunities for computerization of schools on the educational as well as on the organizational level. It included all stages of Slovenian education from pre-school to university.
1994	Project SCHOOL TOLAR (ŠOLSKI TOLAR) – enables the creation and implementation of the RO program.
1999-2000	The Institution of RS for Education produces a project of Computerization of subjects. It consists of three projects. PIKA - computerization of educational activities in kindergartens and lower grades of primary school. TIMKO – team learning and team teaching with informatics technology from 5th to 9th grade of primary school, in gymnasium and in four year secondary technical education. VESNA – searching for new ideas and fresh incentives of Computerization in kindergartens, primary and secondary schools along with their education within the project`s frame.
2001-2003	IKT strategy in Slovenian educational system is based upon the results of the RO program – Computer literacy (Računalniško opismenjevanje). (Čampelj, 2001)

In the time of Yugoslavia the development of computerization was on a different path, for it was slower in Serbia than it was in Slovenia, which has few advantages in this area. The disintegration of Yugoslavia came as a great blow for Serbia because there was a status quo at that time in the field of computerization. After the disintegration there were other factors that impeded the development of computer science.

The reason for Serbia starting to develop the IKT technology later, at the beginning of 21st century was in its long-lasting period of economic and political isolation at the end of 20th century.

Table IV: Serbia's project for ICT in education [14]

1992-1999	The biggest decrease of the development of IKT.
2000 -	Intensive and fast development of the IKT.
2004	Ministry of Education contributes the IKT equipment for more than 2.000 teachers of primary and secondary schools.

Table V: Internet speed.

	SI		SRB	
	F	f(%)	f	f(%)
from 1 Mbit/s	12	7.74	16	16.67
from 1 Mbit/s to 5 Mbit/s	40	25.81	10	10.42
from 5 Mbit/s to 15 Mbit/s	47	30.32	28	29.17
from 15 Mbit/s to 50 Mbit/s	30	19.35	38	39.58
from 50 Mbit/s to 100 Mbit/s	22	14.19	3	3.13
more than 100 Mbit/s	4	2.58	1	1.04

It is evident from the table that the average data transfer rate from the internet among Slovenian students is at its peak from 5 Mbit/s to 15 Mbit/s (30%), while the average data transfer rate among Serbian students is from 15 Mbit/s to 50Mbit/s (39%), which shows that on average Serbian students have a faster data transfer rate. More Slovenian students found themselves in the ratio from 50 Mbit/s to 100 Mbit/s (14%), while there are only three Serbian students in that level. The results were not expected and they show that infrastructure is not the obstacle to the use of ICT in education. According to the results it can be said that Serbia, especially the area of Vojvodina, has been developing fast in the field of computer science in the last few years.

Table VI: The application of ICT in education within didactical courses.

			Country		Total
			SI	SRB	
Application of ICT	YES	Count	88	42	130
		% within country	45.8%	31.3%	39.9%
	NO	Count	104	92	196
		% within country	54.2%	68.7%	60.1%
Total		Count	192	134	326
		% within country	100.0%	100.0%	100.0%

Regarding the Pearson Chi-square test it is evident that there are significant differences between countries ($\chi2$=6.911, p= 0.009). Data also shows that 54% of students of Slovenian universities think ICT is not present enough within the didactical courses Therefore we could not show that Slovenian students have unified opinion about the presence of IKT. In Serbia the difference is more obvious, 68% of students think there is not enough of ICT within their courses. It can be concluded that the ICT is more present in Slovenian study programs but obviously not enough. We believe the result is expected because there is often a lack of interest with the lecturers for improving their own teaching strategies that would require ICT.

Table VII: ICT knowledge between study programs

			Country		Total
			SI	SRB	
ICT knowledge	YES	Count	76	116	192
		% within country	39.6%	85.3%	58.5%
	NO	Count	116	20	136
		% within country	60.4%	14.7%	41.5%
Total		Count	192	136	328
		% within country	100.0%	100.0%	100.0%

The table shows results we have not expected. In Serbia 85% of students claim they get enough information on ICT in their study programs, while Slovenian students are not even near this numbers (39.6%). Statistical differences is significant ($\chi2$=73.743, p=0.000). The result was certainly not expected but we are pleased to see that despite the late development of computer science and introduction of ICT it has extended among study programs. Slovenia has to improve certain areas renovations of study programs with a more ICT inclusion into the curricula.

Table VIII: Use of ICT equipment

	SI		SRB	
	F	f(%)	f	f(%)
Computer	188	98.43	129	94.85
Tablet PC	59	30.89	20	14.71
Smartphone	168	87.96	77	56.62
Internet	184	96.34	86	63.24
Video	62	32.46	17	12.50
Sound	35	18.32	8	5.88
Other	2	1.05	0	0.00

Students could chose more options in this research question. Slovenian and Serbian students use computers most of the time (more than 90%). Internet is firmly in the second place and smart phone in the third though Slovenian students show a higher use of those (smart phones). The table shows that Slovenian students are more distributed among the variety of ICT equipment (or services) than the Serbian students. These elements are influenced by a factor of availability to a variety of equipment.

Table IX: Use of ICT equipment among lecturers in classes/practice for special didactics

			Country		Total
			SI	SRB	
Use_of_ICT_lectures	mostly no	Count	14	1	15
		% within country	7.4%	0.7%	4.6%
	rarely	Count	79	122	201
		% within country	42.0%	89.1%	61.8%
	often, not always	Count	83	13	96
		% within country	44.1%	9.5%	29.5%
	always	Count	12	1	13
		% within country	6.4%	0.7%	4.0%
Total		Count	188	137	325
		% within country	100.0%	100.0%	100.0%

Students in Serbia were very critical of their lecturers because 89% of them think their lecturers do not use enough of ICT equipment, while in Slovenia opinions are divided, which causes a significant statistical difference ($\chi2$= 82.597, p= 0.000). In Slovenia 42% of students think the ICT is not used often enough and another 43% think opposite. Regarding that it can be concluded that Serbian students demands more use of IKT equipment among their lecturers, although Slovenian lecturers have a lot to do to reach a satisfying level of everyday use of IKT equipment. The lecturers are the ones that can teach students how to use certain learning strategies in classes because students will not learn them by themselves for it is not systematic and it would take too much time.

Table X: Use of ICT for study deliverables production.

	SI		SRB	
	F	f (%)	F	f (%)
Web classrooms	88	46.81	22	16.06
e-materials	147	78.19	80	58.39
Web encyclopedias	108	57.45	50	36.50
Specialized data bases	100	53.19	38	27.74
Texts from other educational institutions	98	52.13	89	64.96
Multimedia elements	138	73.40	26	18.98

Student use ICT to make materials needed for their studies. Slovenian students mostly use multimedia elements (73%) and e-materials (78%), interestingly on the

last place is the use of web classrooms (46%) despite Moodle is used on every educational institution in Slovenia. In Serbia the most used elements represent texts from other educational institutions and e-materials, while at the last place is also the web classrooms. According to these results we can conclude that students should probably be more informed on the use of those elements and introduce them as a very practical and interesting tool with the production of various works.

Table XI: Respect of copyright materials

			Country		Total
			SI	SRB	
Respect of copyright materials	DA	Count	147	21	168
		% within country	78.6%	15.3%	51.9%
	NE	Count	40	116	156
		% within country	21.4%	84.7%	48.1%
Total		Count	187	137	324
		% within country	100.0%	100.0%	100.0%

Due to the fact that students produce many seminar works during the study years and should respect the legal authors' and copyright agreements. The biggest differences among Slovenian and Serbian students have been noticed here as 78% of students think that faculties in Slovenia respect those rights, while 84% of students in Serbia think that their faculties do not have an adequate system for that issue. That caused a significant statistical difference ($\chi 2$=137.206, p=0.000). Every faculty needs to respect reserved rights with no exception and therefor a lot of efforts need to be put into solving this issue.

Table XII: uniformity of citation system

			Country		Total
			SI	SRB	
Uniformity of citation	1	Count	169	124	293
		% within country	90.9%	90.5%	90.7%
	2	Count	17	13	30
		% within country	9.1%	9.5%	9.3%
Total		Count	186	137	323
		% within country	100.0%	100.0%	100.0%

Different study programs use different citation systems of citing the other work in their deliverables (seminars work, diplomas, master and doctoral thesis). A high percentage of students from Serbia as well as Slovenia, even 90% think that there should be a uniformed citation system, with no statistically significant difference ($\chi 2$=0.011, p= 0.915) between countries. This shows that a uniformed citation system would solve many issues and avoid inconveniences when students use different citation system for each discipline or even each lecturer. The result was expected because the variety of citation' systems causes many frustration between students. To speculate in that matter a little we do not want only one citation system. But it would be nice to have such standards (e.g. APA, IEEE) which are supported in contemporary Microsoft Word and solve those problems and inconveniences forever.

III. CONCLUSION

Based on the historical review of events and survey, which was conducted in Slovenia and Serbia, we got surprising and unexpected results. According to the data from the 90th where the difference between the two countries was the largest, is now steadily decline. It becomes evident that the use of ICT in education in Slovenia declines as country is griped in the constant economic crisis. On the other hand Serbia, whose economic situation is no better, slowly and steadily grows. Professors and students must be aware of the importance of ICT whose purpose and use is versatile. On the other hand technology itself is not enough and countries must educate professors to use it. It is also wrong perception that young generation will know how to use ICT efficiently and effectively without any help and instructions. No matter what we do "digital natives" will use the ICT technology in ever greater extent and we ought to be prepared for them.

From the results of our research it is evident that the advancement in ICT is not self-sustained. If there are no incentives the advancement stall and eventually even decline. There should be multilevel incentives from funding the hardware and infrastructure to the increased payment to those who use ICT in education more often than the others. Young generation of teachers are ICT literate and capable to use ICT in the whole educational spectrum. But environment where they are not stimulated rapidly decrease their competences. Ever present changes of the educational system where each new government almost completely changes policy is no incentive for the teachers in general. ICT is inherently fast changing and policy changes create almost unimaginable and in most cases unwanted consequences.

We are always bound to the past in all aspects. It is now becomes evident that some society can advance faster than the other if they build on the solid foundations. Elaborate systems are not easy to change and they become increasingly rigid by established rules for almost everything. At least one aspect in the rigid rules is good – established procedures assures general compatibility of the study results. Any student who graduates in all levels of education in Slovenia will comply with the standard of respecting the copyright materials and their work should never be categorized as copyright infringement. Slovenian students are well aware of this problem but Serbian students will have to learn this in the very near future. For the future research it would be very useful to include more countries and study disciplines to the research. We suspect that results published in the yearly self-evaluation reports in the matter of quality assurance of faculties and universities in Europe and around the World are not equally understand among their students and in some cases even their employees.

IV. LITERATURE

[1] M. B. T. Krašna, „Trends of e-learning in the education of teachers," v The fifth International Conference on Informatics, Education Technology, and New Media in Education, Sombor, 2008.

[2] I. Gerlič, T. Bratina, N. Veček in T. Pungartnik, „Stanje in trendi uporabe informacijsko komunikacijske tehnologije (IKT) v slovenskih osnovnih šolah," Faculty of Natural Science and Mathematics, Maribor, 2011.

[3] I. Gerlič, T. Bratina, M. Zver in R. Pučko, „Stanje in trendi uporabe informacijsko komunikacijske tehnologije

(IKT) v slovenskih osnovnih šolah," Faculty of Natural Sciences and Mathematics, Maribor, 2010.
[4] I. Gerlič, N. Jaušovec, T. Bratina in T. Jarc, „Stanje in trendi uporabe informacijsko komunikacijske tehnologije (IKT) v slovenskih osnovnih šolah," Faculty of Education, Maribor, 2005.
[5] M. Krašna, Multimedia in education, Nova Gorica: Educa, 2010.
[6] M. Krašna, M. Duh in T. Bratina, „E-learning next step - learning materials for students," v MIPRO 2014, Opatija, Croatia, 2014.
[7] M. Krašna in B. Bedrač, „ICT didactics: the new study discipline is needed," v MIPRO 2013, Opatija, Croatia, 2013.
[8] T. Bratina, M. Duh in M. Krašna, „E-learning controversy in practical application," v MIPRO 2013, Opatija, Croatia, 2013.
[9] M. Duh, B. Bedrač in M. Krašna, „Assessment of e-learning materials aesthetic design," v MIPRO 2012, Opatija, Croatia, 2012.
[10] M. Krašna in T. Bratina, „E-learning materials for social science students," Philosophy of mind and cognitive modelling in education - 2014, Izv. 61, pp. 77-87, 2014.
[11] I. Gerlič, Sodobna informacijska tehnologija v izobraževanju, Ljubljana: DZS - Državna založba Slovenije, 2000.
[12] V. Rebolj, E-izobraževanje skozi očala pedagogike in didaktike, Radovljica: Didakta, 2008.
[13] I. Gerlič, Računalništvo v interesnih dejavnostih osnovne šole, Ljubljana: Zveza organizacij za tehnično kulturo Slovenije, 1987.
[14] SCORE, „Jačanje strateške saradnje između Evropske unije i regiona Zapadnog Balkana u oblasti istraživanja IKT," n.d.. [Elektronski]. Available: http://consultations.score-project.eu/attach/ictcr_rs_sr.pdf. [Poskus dostopa 9 2 2015].

The Application of Formative E-Assessment Data in Final Exam Results Modeling Using Neural Networks

Jasna Gamulin[a], Ozren Gamulin[b], Dragutin Kermek[c]
[a] University of Zagreb School of Medicine
Šalata 3, Zagreb, Croatia, Phone:+385 1 4590 292, E-mail: jasna.gamulin@gmail.com
[b]University of Zagreb School of Medicine, Department of Physics and Biophysics
Šalata 3, Zagreb, Croatia, Phone:+385 1 4566 797, E-mail: ozren@mef.hr
[c]University of Zagreb Faculty of Organization and Informatics, Department of Computing and Technology, Department of Theoretical and Applied Foundations of Information Sciences, Pavlinska 1, Varaždin, Croatia, Phone: +385 42 390864, E-mail: dkermek@foi.hr

Abstract – After introducing e-assessment into laboratory and seminar teaching of physics in a biomedical university study program the improved final practical exam passing rate was observed. The final practical exam passing rate increased from approximately 60% to more than 80%. Encouraged by these results, in this paper we will try to show a correlation between the results of e-assessments collected during continuous monitoring of student performance and their success in final tests using neural networks as modeling method. The models will be built using various sets of data collected during three academic years. Aside from the impact on the final practical exam passing rate, the models should help us in dealing with the missing data problems. Some data are missing due to technical problems and some due to errors in student handling of e-application. In this paper we have observed data collected by two types of e-assessment. The first type of e-assessment is carried out during laboratory exercises and it influences the formation of the final grade. The second type of e-assessment is performed during seminars and has no impact on the final grade.

I. INTRODUCTION

The e-assessment results collected during continuous monitoring of student performance could be used for modeling and predicting student success in final exams. Similar models are described in [1] and [2]. During the process of data preprocessing we will always observe that there are some data missing. The question is which method would be optimal for dealing with the missing data in these specific models.

The general rule in dealing with the missing data is that the missing data could be simply ignored if they occur randomly and the conclusions could be drawn based on the observed data only. If the data are not missing randomly, ignoring the missing data leads to a systematic bias [3]. In that case some imputation techniques should be used.

Intuitively, it could be presumed that students obtaining better results in final exams would have less missing data, meaning that they are more conscious regarding their obligations but then it could be the pure technical clumsiness or the data could be missing due to some technical reasons. The thing is that we don't really know why the data are missing and are they missing randomly or not in this specific situation.

In order to deal with this issue we will build models in two ways: in the way that the missing data will be simply ignored and in the way that null value will be imputed instead of the missing data.

This paper will begin by the digest of the up-to-date papers elaborating the handling of missing data in prediction models. Afterwards, the dataset and the methodology will be described. The paper will be concluded by discussing the results and presenting conclusions, limitations and guidelines for the future work.

II. LITERATURE REVIEW

According to [4] and [5] there are three possible modes for missing data. The data could be missing completely at random meaning that the missing data are not related to the study variables. In the second mode, data are missing at random and the probability of an instance having an attribute with a missing value may depend on the known values, but not on the value of the missing data itself. In the third mode, data are not missing at random, meaning that the probability of an instance having an attribute with a missing value could depend on the value of that attribute.

Imputation is a term describing a procedure of replacing the missing values in a dataset with some presumable values. The missing data treatment is independent of the learning algorithm used. This allows the user to select the most suitable imputation method for each situation [6].

As stressed by [7] many works in the literature have demonstrated that the benefits of imputation methods depend mainly on the problem domain, the size of the available dataset and missingness patterns.

Clinical trials databases are the most frequently used and described when researching the missing data problems. Substantial instances of missing data are a serious problem that undermines the scientific credibility of causal conclusions from clinical trials. The assumption that analysis methods can compensate for such missing data are not justified, so aspects of trial design that limit the likelihood of missing data should be an important objective [5].

In the work of [7] several imputation methods based on statistical techniques were applied, for instance mean, hot-deck and multiple imputation and some methods based on machine learning techniques like multi-layer perceptron (MLP), self organization maps (SOM) and k-nearest neighbor (KNN). The data on demographic, therapeutic and recurrence-survival information from women with operable invasive breast cancer were used. The accuracies of predictions on early cancer relapse were measured using artificial neural networks (ANNs), in which different ANNs were estimated using the datasets with imputed missing values [7].

The industrial data are also submitted to imputations. Reference [8] analyzed one database with instrumentation maintenance and test records for an industrial process considering the selection of appropriate imputation methods.

In the paper of [9] ten clinical trials databases were used to explore and compare nine different approaches to missing attribute values.

Reference [10] reviewed the routines to incorporate observations with incomplete variables in regression models in the context of a motivating example from a large health services research dataset.

The approaches to characterize patterns of bias in missing data were considered by [11]. They sought to determine the specific attributes that predict the missingness of data values [11].

In the work of [12] the authors strived to understand better the problem of dropouts in higher education. They processed the data using data mining methods: logistic regression, decision trees and neural networks. The models are compared to select the one which best predicts the student dropout. They applied the method of missing values replacements called "Most frequent value" and "Tree imputation".

The authors generally agree that there is no the universal rule how to handle the missing data problem and that the approach should be attuned to the specific problem. This paper will consider routines for handling missing data problem in educational prediction models.

III. METHODOLOGY

In this paper we have observed data collected by two types of continuous e-assessment. The first type of e-assessment is carried out during laboratory exercises and it has the minor influence the formation of the final grade. The second type of e-assessment is performed during seminars and has no impact on the final grade. The continuous e-assessments are accomplished by a web-application named *Pitalica*, which has been developed in order to allow the teacher to quickly assess students at the beginning of each exercise. It is assumed that a student will spend approximately 5 minutes on a computer answering to 4-6 questions related to the exercise that he/she is about to work off. This way, a student has the feed-back on his performance in terms of number of points achieved for a specific matter covered by an exercise [13].

The second type of e-assessment is performed during seminars and the idea is to motivate students to prepare for the seminars and to participate actively during seminars which would hopefully lead to better exam performance [14].

The topics elaborated through lectures and seminars are examined either by means of three written midterm exams (12 multiple-choice questions each) and by a final exam.

The practical exam is performed after the course and it is graded as pass/fail. During the practical exam a student has to carry out the practical assignment which only involves elements of the practical exercises performed during the course. In order to pass the practical exam the calculated result has to be +/- 10% correct and a student is expected to synthesize the knowledge acquired during laboratory practicals. Passing the practical exam is the prerequisite to take the oral part of the final exam.

After introducing e-assessment into laboratory and seminar teaching of physics course in a biomedical university study program the improved final practical exam passing rate was observed. The final practical exam passing rate increased from approximately 60% to more than 80%.

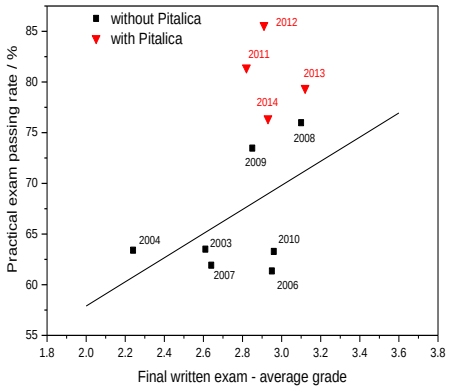


Figure 1. Linear regression fitted on Practical Exam passing rate/ Final Written Exam average grade (no Web-quiz) data, 2003-2014

Encouraged by these results we built models for predicting student performance using continuous e-assessment data as attributes and exam success data as class labels.

A. *Dataset preprocessing*

TABLE I. Number of students enrolled in Physics course, study program Medicine

Year	All instances	Reduced dataset (instances with missing data are removed from the dataset)
2014	284	217
2013	291	184
2012	262	191
Total	**837**	**592**

The datasets used in this paper are described in Table I.

The student performance prediction model is proposed using the results achieved at the first midterm exam (FME), the second midterm exam (SME), the third midterm exam (TME), all three midterm exams (AME) and the final practical exam (FPE) as the class labels.

The attributes for specific models are described in Table II.

All models are calculated using 2 datasets: reduced dataset (m_R), where instances with missing data are removed from the dataset and complete dataset (m_A) where nulls are imputed instead of the missing values. Nulls have been chosen for imputation since it is presumed that it would be the result if the student would have responded. Also, students who failed a certain question have null registered also for the time attribute.

B. *Modelling*

Tables with results accumulated by means of *Pitalica* during e- assessment on seminars and practicals and mid-term and final exam results were stored in a database. The data were exported to the format which can be processed by Rapid Miner 5.3 tool [15].

Neural Networks (NN) have been used for prediction models by [16, 17, 18, 12 and 19]. NN are computational models inspired by neuropsychology where a nerve cell functioning principle has been used [20]. A multi-layer neural network consists of large number of units (neurons) joined together in a pattern of connections. Input units receive information to be processed, output units present the results of the processing and units in between are called hidden units.

There is an issue is how to properly determine the size of the hidden layer since an underestimate of the number of neurons can lead to poor approximation and generalization capabilities. Too large number of nodes can result in over fitting and eventually make the search for the global optimum more difficult [21].

NN are used in this paper for modeling since the non-linearity of the neural network is able to overcome poor embeddings resulting from zero and mean [3] and there are reported good results in educational prediction models using NN [12].

IV. RESULTS

TABLE II. DESCRIPTION OF MODELLS, ATTRIBUTES AND CLASS LABELS

Model	**Atributes Description**		**Values**
m_1	Results - *Pitalica* – seminars 1 and 2 – for each student and for each question	binominal	=[0, 1]
m_2	Results - *Pitalica* – seminars 1,2 and 3 – for each student and for each question	binominal	=[0, 1]
m_3	Results - *Pitalica* – seminars 1,2,3 and 4 – for each student and for each question	binominal	=[0, 1]
m_4	Results and time - *Pitalica* – seminars 1,2,3 and 4 – for each student and for each question	binominal and numeric	=[0, 1] =[0 – 30]
m_5	Results - *Pitalica* – seminars 1,2,3 and 4 – sum of scores for a seminar for each student	numeric	=[0, 15]
m_6	Results and time - *Pitalica* – seminars 1,2,3 and 4 – sum of scores and total time for a seminar for each student	numeric	=[0 – 450]
m_7	Time - *Pitalica* – seminars 1 and 2 – for each student and for each question	numeric	=[0 – 30]
m_8	Time - *Pitalica* – seminars 1,2 and 3 – for each student and for each question	numeric	=[0 – 30]
m_9	Time - *Pitalica* – seminars 1,2,3 and 4 – for each student and for each question	numeric	=[0 – 30]
m_{10}	Total time - *Pitalica* – seminars 1,2,3 and 4 – sum of time for each student	numeric	=[0, 450]
m_{11}	Results - *Pitalica* – practicals- for each student for each question	binominal	=[0, 1]
m_{12}	Time - *Pitalica* – practicals-for each student for each question	numeric	=[0 – 120]
m_{13}	Results and time - *Pitalica* – practicals	binominal and numeric	=[0, 1] =[0 – 120]
Class label	**Class Labels Description**		**Values**
y_1	FME = result of first midterm exam	binominal	=[0, 1]
y_2	SME = result of second midterm	binominal	=[0, 1]
y_3	TME = result of third midterm exam	binominal	=[0, 1]
y_4	AME = result of all midterm exams	binominal	=[0, 1]
y_5	FPE = result of final practical exam	binominal	=[0, 1]

When comparing accuracy results of models using the reduced dataset (m_R) and models using complete dataset (m_A) with imputed nulls instead of the missing values it could be observed that models using incomplete data have consistently better accuracy for a few of percents.
Regardless of the dataset used the most successful models are the ones having as attributes the results of *Pitalica* at seminars 1,2,3 and 4 – sum of scores for a seminar for each student.

The best model (m_{R5}, accuracy 92,13%) has as attributes the results of *Pitalica* at seminars 1,2,3 and 4 – sum of scores for a seminar for each student using reduced set of data and as the class label has the results of the first midterm exam (FME) as shown in Table III.

TABLE III. ACCURACY FOR SELECTED CLASS LABELS AND MODELS, TWO DATASETS USED

Dataset: reduced					
accur %	FME_R	SME_R	TME_R	AME_R	FPE_R
m_{R1}	86,52	66,29	75,84	67,98	80,90
m_{R2}	84,83	67,42	72,47	66,29	79,78
m_{R3}	84,27	64,61	73,30	64,61	82,00
m_{R4}	85,96	65,17	74,72	62,92	79,78
m_{R5}	92,13	70,22	83,71	66,29	87,08
m_{R6}	91,01	68,54	83,71	67,42	87,64
m_{R7}	83,15	64,04	73,60	64,04	83,15
m_{R8}	82,58	61,80	71,91	61,24	78,09
m_{R9}	85,39	62,36	70,79	61,80	78,65
m_{R10}	91,57	69,66	83,15	67,98	87,08
m_{R11}	87,64	67,54	75,28	67,98	80,34
m_{R12}	84,83	64,61	71,91	66,86	80,34
m_{R13}	85,96	70,22	76,40	65,17	80,34
Dataset: all instances					
accur %	FME_A	SME_A	TME_A	AME_A	FPE_A
m_{A1}	85,52	66,29	74,90	61,53	80,88
m_{A2}	84,83	63,75	73,51	62,55	75,30
m_{A3}	84,86	64,61	73,03	62,55	76,89
m_{A4}	84,86	65,74	73,31	65,54	75,70
m_{A5}	89,24	71,71	83,71	66,53	87,08
m_{A6}	89,24	69,32	80,08	67,42	83,72
m_{A7}	81,67	61,75	73,60	64,04	79,68
m_{A8}	80,08	63,35	71,92	59,76	74,50
m_{A9}	83,27	63,36	69,32	60,16	75,30
m_{A10}	83,24	68,92	78,88	67,98	83,27
m_{A11}	82,47	68,92	72,91	63,35	80,34
m_{A12}	80,88	64,61	68,53	58,57	84,27
m_{A13}	82,87	67,33	76,40	64,54	80,34

It could be also observed in Table III that adding time attributes (total time used at *Pitalica* seminars, m_6 compared to m_5) the accuracy of models decreases slightly. It could point to the greater impact of the factors not connected to knowledge and ability to the time spent for solving e-quizzes. The time consumed for e-quizzes by a student could be affected also by technical problems due to communication to the server. It could be also determined by a student strategy of solving problems since some students answer to questions as soon as they perceive the true answer and some students mark the chosen answer only when the time is up.

When considering results of Pitalica at seminars for each student and for each question the best accuracy results are achieved when modeling using the results of the first two seminars irrespective of the dataset used (reduced or imputed). The first two seminars have greater number of questions (28 out of totally 48) and the matter included in those two seminars is mostly questioned at the first midterm exam (FME).

The influence of the *Pitalica* results of the third and the fourth seminars to the accuracy results when modeling to SME (result of the second midterm exam) and TME (result of the third midterm exam) is respectively smaller and it is connected to the smaller models accuracy.

The accuracies of models m_1 and m_2 with SME and TME as class labels respectively, using both datasets, confirm the tendency.

In order to rank the models according to accuracy indicator the Wilcoxon Matched Pairs Test was calculated using Statistica 12.5.

Table IV contains the Wilcoxon Matched Pairs Test scores calculated for models using the reduced dataset (FME_R, SME_R, TME_R, AME_R, FPE_R) and the dataset with imputed nulls (FME_A, SME_A, TME_A, AME_A, FPE_A). It could be concluded that there is a statistically significant difference in accuracy for models using the reduced dataset and the dataset with imputed nulls.

Exceptionally, when modeling to SME (result of the second midterm exam) the accuracies of models using the reduced dataset (SME_R) compared to accuracies of models using dataset with imputed nulls (SME_A) do not show significant difference (Table IV). This exception could be explained by the fact that the second midterm exam questions the field of electricity (which is supported by *Pitalica* seminar) and the field of thermodynamics (not supported by *Pitalica* seminar). The thermodynamics questions always outnumber the electricity questions in the second midterm exam and the thermodynamics also proves to be more difficult for students.

TABLE IV. WILCOXON MATCHED PAIRS TEST

	Wilcoxon Matched Pairs Test Marked tests are significant at p <.05000	
Pair of Variables	statistical significance	p-value
FME_R & FME_A	significant	0,002874
SME_R & SME_A	not significant	0,878482
TME_R & TME_A	significant	0,021825
AME_R & AME_A	significant	0,036659
FPE_R & FPE_A	significant	0,036659

V. CONCLUSION

Since reduced models gave significantly better prediction accuracy than the models with imputed nulls instead of the missing values we could conclude that in this specific situation, the data are missing at random or that null value is not a good substitute. In order to resolve

this ambivalent inference, in a future research one could use some other technique of imputation of the missing values as for example hot-deck.

Hot-deck imputation is a method in which each missing value is replaced with an observed response from a "similar" unit. This theory is extensively used in practice, but it is still not developed as well as some other imputation methods [22]. Methods based on machine learning techniques like k-nearest neighbor (KNN) could be the further possibility.

REFERENCES

[1] J. Gamulin, O. Gamulin and D. Kermek, "Data mining in hybrid learning: possibility to predict the final exam result", in Proceedings of the 36th International Convention, Conference Computers in Education (MIPRO 2013/CE), Opatija, Croatia, IEEE Conference Publications, pp. 591 – 596, 2013.

[2]…J. Gamulin, O. Gamulin and D. Kermek, "Comparing classification models in the final exam performance prediction", in Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2014/CE), Opatija, Croatia, IEEE Conference Publications, pp. 663 – 668, 2014.

[3] B. M. Marlin, Missing Data Problems in Machine Learning, PhD Thesis, Graduate Department of Computer Science University of Toronto, 2008.

[4] D. B. Rubin, "Inference and missing data", Biometrika 63, pp. 581–592, 1976.

[5] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, and H. Stern, "The Prevention and Treatment of Missing Data in Clinical Trials", New England Journal of Medicine, 367(14): pp. 1355–1360, 2012.

[6] G. Batista, M. C. Monard, "An analysis of four missing data treatment methods for supervised learning", Appl Artif Intell 17, pp. 519–533, 2003

[7] J.M. Jerez, I. Molina, P. J. Garcıa-Laencina, E. Alba, N. Ribelles, M. Martın, L. Franco, "Artificial Intelligence in Medicine", 50, pp. 105–115, 2010.

[8] K. Lakshminarayan, S. A. Harp and T. Samad, "Imputation of Missing Data in Industrial Databases", Applied Intelligence 11, pp. 259–275, 1999

[9]] J. W. Grzymala-Busse and Ming Hu, "A Comparison of Several Approaches to Missing Attribute Values in Data Mining", W. Ziarko and Y. Yao (Eds.): RSCTC 2000, LNAI 2005, pp. 378−385, 2001

[10] N. J. Horton, and K. P. Kleinman, "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models", Amer. Statist. 61, pp. 79–90, 2007

[11 M. C. Tremblay, K. Dutta and K. Vandermeer, "Using data mining techniques to discover bias patterns in missing data", Journal of Data and Information Quality, 2(1), pp. 23–55, 2010

[12] M. Jadric, Z. Garaca and M. Cukusic, "Student Dropout Analysis With Application Of Data Mining Methods", Management, Vol.15, pp. 31-46, 2010

[13] Gamulin and O. Gamulin, "The application of web-based formative quizzes in laboratory teaching in higher education environment", in Proceedings of the 35th International Convention (MIPRO 2012/CE), Opatija, Croatia, IEEE Conference Publications, pp. 1383 – 1388, 2012

[14] J. Gamulin, J. Gugić and O. Gamulin, "Improving classroom teaching in higher education environment using web-based formative assessment", in Proceedings of the 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2010/CE), Opatija, Croatia, IEEE Conference Publications, pp. 1001-1006, 2010.

[15] Rapid Miner, 2014, http://rapidminer.com/

[16] M. Jovanović, M. Vukičević, M. Milovanović and M. Minović, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study", International Journal of Computational Intelligence Systems, 5:3, pp. 597-610, 2012

[17] S. Kotsiantis, C. Pierrakeas, P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques", Applied Artificial Intelligence, 18: pp. 411-426, 2004.

[18] M. Delgado, E. Gibaja, M.C. Pegalajar and O. Pérez, "Predicting Students' Marks from Moodle Logs using Neural Network Models", in Proc. International Conference on Current Developments in Technology Assisted Education, (Sevilla, Spain), pp. 586-590. 2006

[19] M. Paliwal, U.A. Kumar, "A study of academic performance of business school graduates using neural network and statistical techniques", Expert Systems with Applications, Volume 36, Issue 4, pp. 7865–7872, 2009

[20] G. Klepac, L. Mršić, Poslovna inteligencija kroz poslovne slučajeve, Lider press d.d., Tim press d.o.o., Zagreb, 2006.

[21] S. B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas, "Machine learning: a review of classification and combining techniques", Artif Intell Rev 26, pp. 159–190, 2006.

[22] R. R. Andridge and R. J. A. Little, "A Review of Hot Deck Imputation for Survey Non-response", Int Stat Rev., 78(1), pp. 40–64, 2010.

Knowledge Generating House – a tool for education based on Open Innovation principle

I. Korobaničová*, D. Paľová**, N. Urbančíková*
* Department of Regional Science and Management
** Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
iveta.korobanicova@tuke.sk, dana.palova@tuke.sk, natasa.urbancikova@tuke.sk

Abstract - Creativity and innovation play a key role in the present. Program Europa2020 declared a need of strengthening the knowledge triangle - education, research and innovation. Based on this, the changes are happening in education from the perspective of the knowledge society, focusing on one of the basic elements of innovation - creativity and its role in education. In the paper the OpenInn project and its outcome – OpenInn portal, an on-line source of information concerned creativity enhancement and tool for creative cooperation among its users – Knowledge Generation House, will be presented.

I. INTRODUCTION

Globalization, competition, rapid changes in technology and other factors have prompted companies, institutions such as schools, to ensure that knowledge is being created, transferred and shared in their organizations. Based on [1], [3], [10], [22], [38], [31] organisations are organizational spaces in which the learning processes on the individual, group and organizational levels are enabled and taking place. This suggests a new way of thinking about how people work together and the need for the greater emphasis on reviewing past practices and experiences [16].

In 2010, the "Creative Learning and Innovative Teaching" [4] has been published, where the strong emphases on broad, flexible and motivating education that recognises the different talents of all children and delivers excellence for everyone is highlighted rather than just improving literacy and numeracy skills. We need to support the creative learning process and try to connect it with practice.

Torrance and Myers [34] provide an excellent framework for describing the creative learning process: „Becoming sensitive to or to be aware of problems, deficiencies, gaps in knowledge, missing elements, disharmonies, and so on; bringing together available information; defining the difficulty or identifying the missing element; searching for solutions, making hypotheses, and modifying and retesting them; perfecting them: and finally communicating the results.". Treffinger [35] proposed a practical model for describing three different levels of creative learning, with consideration of both cognitive and affective dimensions at each level. The three levels are divergent functions (level I), complex thinking and feeling processes (level II), and involvement in real challenges (level III) as illustrated on the Figure 1.

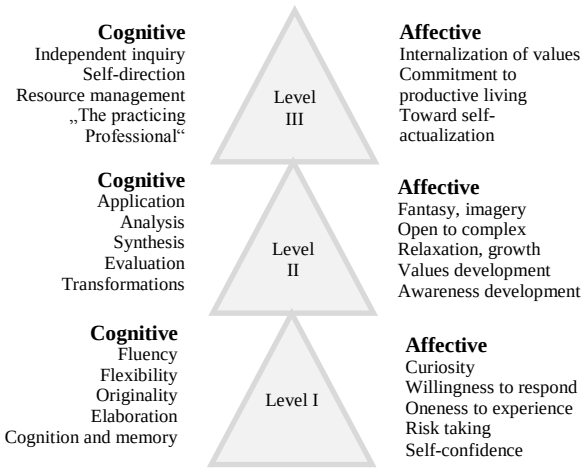


Figure 1. Creative learning model [35]

Teaching for creativity must consider the new opportunities presented by information technology. New technologies offer unprecedented opportunities for young people to broaden their horizons; to find new modes of creativity and to deepen their understanding of the world around them. The number of Internet users are rapidly increasing as mentioned in the Figure 2. Internet user penetration has reached 40% globally, 78% in developed countries and 32% in developing countries [30].

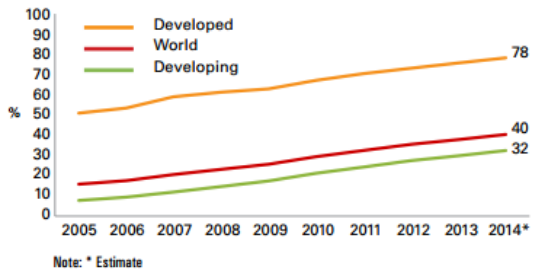


Figure 2. Percentage of Individuals using the Internet [30]

The similar situation is also in Slovakia. In 2013, 74% of Slovaks reported using the internet at least once weekly (regular users, internet users going on-line weekly), above the EU average of 72%. Slovakia exhibited slightly lower rates of daily use of the internet (frequent users, internet users on a daily basis), with 61% of the population reporting going online every day, compared to an EU average of 62%. In 2013, 15% of the population had still never used the internet; lower than in 2012 and lower than the EU average of 20% (individuals who have never used the internet).[8].

The growing number of Internet users has strong relationship to the younger generation [5] which grew up surrounded by digital media – roughly referring to individuals born in the early 1980s or later – displays significantly different learning styles from previous cohorts. Several terms have been used to describe this generation of learners, for example "digital natives" [20]. Bayne and Ross [2] have argued that the persistence of this image of the "Netgeneration" or Digital Native student has a strong relationship to the logic of the market and a culture of enterprise and it is evident in advertising that these ideas are clearly reinforced by marketing aimed at the education sector. New Millenium Learners [5] are using a mix of ICT elements, digital media for their own personal learning needs, mixing and matching different tools and capitalising on social computing applications to build networked, extended communities of interconnected learners who exchange ideas, query issues, provide support and check progress.

In Slovakia, in 2012, 43% of Slovaks had low or no digital skills, compared to 47% for the EU average. Disadvantaged people (i.e. individuals who are aged 55-74, low educated and/or unemployed, retired or inactive) in Slovakia are less digitally skilled than is average in the EU. Within the workforce, skills are higher; 34% of the workforce has low or no digital skills. On average in the EU, the figure is 39%.

Furthermore, the Commission Communication "New Skills for New Jobs" [28] calls for the education, training and employment policies of the Member States to focus on increasing and adapting skills and providing better learning opportunities at all levels, in order to develop the workforce that is highly skilled and responsive to the needs of the economy. The use of ICT must be supported by the strategy of the learning organisation. In higher education classrooms a variety of software tools and free web applications should be used to enhance learning, communication, and engagement, the use of social media, and other content such as Web 2.0. Based on [7], [15] Web 2.0 is a platform on which innovative technologies have been built and also a space where users are as important as the content they upload and share with others. It includes social networks, such as MySpace, Facebook, social bookmarking, creative works, such as podcasts, videocasts, blogs, content aggregation and organization, and remixing or mash-ups of content from different content providers into new forms [15].

Schools also have resources available through these technologies to transform methods of teaching and learning. OpenInn project has created new frontiers linking together education and the corporate world of innovation and supported the creative learning process in the schools by using Web 2.0. The SP4CE project (Strategic Partnership for Creativity and Entrepreneurship - grant agreement 2014-1-PL01-KA200-003341) is widening this approach for further utilisation especially in business environment.

II. OPENINN PROJECT

OpenInn project is aimed at sketching a new Pedagogical and Organisational Model for communities and individuals as an example to explore their innovative potentials proceeding the strengthening of their self-esteem. The project approached the great challenges of bridging the implementation gap in ICT-use for enhancing creativity in lifelong learning [27].

Nine institutions, Technical University of Košice from Slovakia, Univeridade do Porto (Portugal), Università delle LiberEtà del Fvg (Italy), Petko Rachov Slaveikov Secondary School (Bulgary), Aarhus Social and Health Care College (Denmark), Bremen Institut für Produktion und Logistik GmbH (Germany), TREBAG Vagyon –és projektmenedzser Kft. (Hungary), Fondo Formación Euskadi (FFE - Spain), LiNK MV (Germany) decided to promote and prepare the project OpenInn with following aims [27]:

- Bridge the implementation gap in the use of ICT in a creative way to support learning and social coherence linking together different sectors.
- Support the evolvement of a stable positive self-esteem both for learners and for facilitators in the learning environment. By reassuring the self-esteem through autonomous learning, it also generates inner motivation for learning.
- Increase opportunities for utilisation of intellectual capital of young talents and groups at-risk.
- Produce a comprehensive e-Assessment Model containing the theoretical concept and didactics for creativity, with topics including: ways to support the learner to introduce new ideas, evaluating and rewarding creative results, methods for giving instructions to reach innovative solutions, providing autonomy for creative expression both for not-yet ICT users and advanced level users in their specific sectors.
- Create the 'Knowledge Generating House' as an innovative social network for exploring creative potentials. Users could use the platform when looking for ideas expressed by others and by utilization of their own inner potentials as well. This generates a new learning experience, where not only young talents, but groups at-risk from falling behind the educational system can be actively engaged in creating and inventing. By collaborating with others, they can attain new perspectives for understanding a case and developing it, and by the new perspectives they are able to understand themselves better. This reassures their self-esteem and makes the base for exploring their inner potentials.

This project is aimed at implementing the Open Innovation concept in the educational sector to develop a new learning paradigm which utilises the potentials of ICT for learners aged 15 and above. There is an importance of leaving the model open for more sectors of the educational field because the field of use can be still in a process of evolution depending on the end users. This concept was targeted to the following groups [23], [27]: talented youth in higher and secondary education in the partnership institutions and their international network; groups at risk of falling behind the educational system in secondary, higher education, vocational training and adults wanting or needing to further explore their capabilities; teachers and facilitators of talented youth and groups at-risk in secondary, higher education, vocational training or adult education.

The results of this project could be used in the long-term for the following target groups: students with a specific research or scientific interest, experts, SMEs, managers of large enterprises, non-profit organisations and groups that need more support in order to be able to use the portal for enhancing creativity.

III. OPENINN PORTAL

As mentioned above, main outcomes of OPENINN project are gathered in OpenInn portal, available via openinn.eu .

To main project results [24] could be integrated:

- *The Guide to Creativity* (an e-Assessment Model) - a complex electronic educational material, which can serve as a "guide" for teachers using KGH, respectively for teachers interested in innovation trends in education and in innovative didactics for enhancing creativity.
- *The Knowledge Generating House* - on-line ideation tool,
- *Demo video and video guide* to OpenInn portal and Knowledge Generating House,
- *OpenInn Guidebook* - user manual for the Knowledge Generating House with explanations of all features developed,
- *OpenInn project website* with all information about the project, partnership, activities done, etc.

The portal is multilingual – all information is available in nine languages: Bulgarian, Danish, English, German, Hungarian, Italian, Portuguese, Slovak and Spanish.

A. Guide to Creativity

The Guide to Creativity (GtC) presents on-line material for teachers and also for managers interested in innovation trends in education and in innovative didactics for enhancing creativity. The material is divided into modules containing:

- ways of how to support the learner when introducing new ideas,
- ways of how to evaluate and reward creative results,
- methods for giving instructions to reach innovative solutions, providing autonomy for creative expression, both for teachers who are not-yet computer users to those who are advanced internet users.

Every module includes:

- the self-learning material called "Pedagogical background" that provides the reader with comprehensive theoretical background about the Open Innovation principle and its possible usage in the field of addressing and assessing creativity in an educational context,
- the ICT Tools section - a list of ICT tools, their descriptions and the ways how they can be used in an educational context. Every item contains three sections based on the level of a potential user's digital literacy: without the necessity to use ICT tools; using basic level ICT tools; using innovative and advanced level of ICT tools.
- The Practical examples section – presents a collection of particular scenarios describing how to apply ICT for teaching and learning to increase creativity. The 25 scenarios were developed for different educational contexts as well as for the different digital literacy of the users. Rather than consecutive order, the scenarios are presented as single independent items, assigned to two different categories: particular educational sector (secondary school student, vocational student, university student, and adult learner), level of ICT skills (non-ICT user or beginner, basic ICT skills, advances ICT skills).
- The Resource library contains more interesting teaching and learning material about enhancing creativity which is not used in the previous sections.

Because GtC is an on-line material, each part can be used independently from others, but it is cross linked to draw connections from one material to another.

B. Knowledge Generating House

The Knowledge Generating House (KGH) [25] represents an on-line ideation tool based on the principle of crowd sourcing (the practice of engaging a 'crowd' of external individuals for a common goal, for example solving problems and producing things [26]). This principle is usually used by companies in process of designing the new products or services. The use of external resources for increasing innovative potential is generally called open innovation. In connection to this principle, KGH supports the development of new ideas, implements a kind of online brainstorming, offering commenting, and voting for ideas.

These activities are divided into three main parts (Figure 5): place for collaborative learning ('Innovation e-Learning Rooms' for brainstorming and mind-mapping), space for uploading documents (like academic studies, thesis, and portfolios) and individual 'creativity assessment' (based on the level of activity, provided ideas etc.).

Using KGH after the registration to the system, users can actively contribute by adding public or private challenges, ideas to challenges, votes for ideas, and comments on ideas.

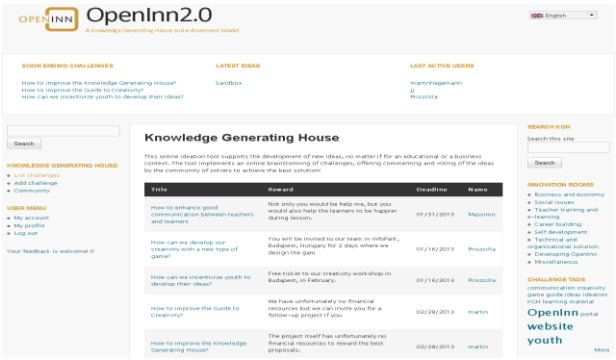


Figure 3. Knowledge Generating House home page

C. *Pilot testing of the portal*

One of the phases of development any software product is pilot testing of the pre-final/beta version. Pilot testing was aimed at testing the functionality and quality of every part of the OpenInn portal. In connection to target groups and main objectives of the project, the partnership decided to provide tests with three pilot-group tests with the portal - 2 pilot-groups with 10-30 students/learners including one teacher/facilitator and one pilot-group only for teachers with 5-20 persons [26].

Besides 25 national pilots done in Danish, German, Italian, Portugal, Slovak, Spanish language, one international pilot involving 4 partner-countries (Denmark, Hungary, Germany and Bulgaria) was realized in English, so that students were able to use the portal with peers online from other countries. Together 357 persons participated on the pilots involving different educational institutions, based on project's target groups: secondary schools, vocational schools, universities and adult education institutions and with different ICT skills level - no ICT, basic ICT and advanced ICT.

TABLE I. SUMMARY OF PILOT TESTING ATTENDANTS DIVIDED ON THE BASIS OF THEIR ICT SKILLS AND EDUCATIONAL SECTOR [26]

Education sector	**ICT skills**		
	No ICT users	***Basic***	***Advanced***
Secondary schools		49	42
Vocational trainings	23	33	28
Universities		21	119
Adult education		23	23

During the pilot testing users created 113 different challenges included in various Innovation rooms (Business and economy, Social issues, Teacher training and e-learning, Career building, Self-development, Technical and organizational solution, Developing OpenInn, Miscellaneous), added 165 new ideas to these challenges and 155 comments.

During pilot testing the testers were asked to use different parts of OpenInn portal and at the end of testing to fill out the on-line feedback questionnaire. In context of the project objectives and our interest in users' opinion about the portal, the partnership decided to divide the feedback into following parts:

- Basic information about participants – general information about recipient such as his/her sex, ICT skills of participant and her/his opinion on creativity and previous experience with some tool for enhancing creativity.
- OpenInn portal evaluation – contained questions about OpenInn portal itself (e.g. navigation on the site, graphic design, browsing tools of the portal, …).
- Evaluation of KGH – questions concerned especially about evaluation of KGH and its navigation, functionality like adding new challenges, forums, ideas, comments, etc. and quality of developed scenarios.
- Evaluation of the Guide to Creativity, its concrete parts – the pedagogical background, practical examples, ICT tools.

291 users took part in the final feedback, what represents 82% of all pilot test participants. The achieved results could be briefly generalized as follows:

- About 80% users agreed or strongly agreed, that developed OpenInn portal is easy to use and navigate, has clear layout and format, has easy-readable content and has sufficient response time. Concerning OpenInn portal structure and functionality, the pilot tests participants appreciated sharing ideas with others and possibility to get opinion from other users as also the overall layout, with the different areas visible in the top and menu to the left, placement of different articles.

  Almost 75% of users agreed or strongly agreed, that KGH part is also easy to navigate, easy to add new challenges, ideas, votes or comments, easy to find existing challenges, ideas or comments. The majority of respondents thought the possibility to read ideas from others and sharing the contributions and the simple way to add challenges, ideas, comments and votes and possibility of get response from several different countries to be the best feature of the KGH for the supporting creativity itself.
- The last specialized part of OpenInn portal - Guide to Creativity – and its subparts were evaluated only by teachers, what means 28% of all respondents' feedback. The majority of them agreed or strongly agreed that all developed parts are clear and easy to understand, well and logically structured and suitable to support creativity in teaching and learning (Figures 5-7).

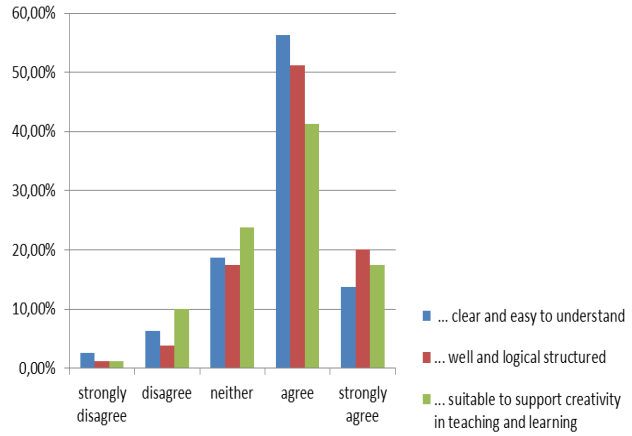


Figure 4. Evaluation of the GtC: PEDAGOGICAL BACKGROUND

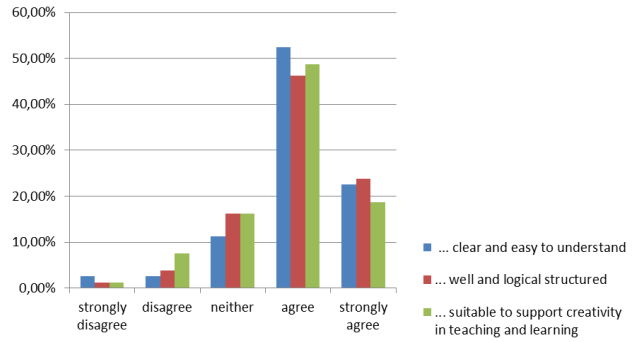


Figure 5. Evaluation of the GtC: PRACTICAL EXAMPLES

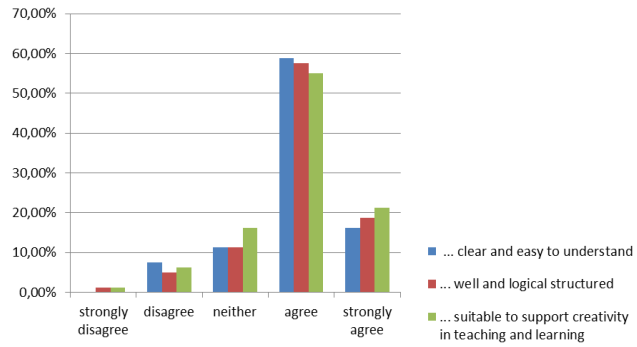


Figure 6. Evaluation of the GtC: ICT TOOLS

IV. CONCLUSION

Organizations and companies at the forefront of development are already experimenting with the „open innovation" concept, inviting experts outside the company to take part in developing new products or services. New connections are established using social networking sites. Innovation, research and development (R&D) is driven by new ideas and new perspectives which are vital characteristics of the fresh mind of youth.

Their activity could be marked as a key factor for the development of innovations in practice. That is why it is important to introduce new, creative approaches in education that would allow people to educate a generation of critical, but also innovative thinking. The developed OpenInn portal and its part Knowledge Generating House could be a tool which helps to build not only creativity of teachers and students, but also the bridge between educational institutions and practices. By saying that, the Knowledge Generating House supports the development of new ideas, whether in an educational or a business context. It offers possibility to create challenges for public or private ideation, in school or business environment, with material or non-material reward.

The OpenInn approach contains a new assessment model and innovative didactics for enhancing creativity with topics including ways to support the learner to introduce new ideas, evaluating and rewarding creative results, methods for giving instructions to reach innovative solutions, providing autonomy for creative expression both for "not-yet" ICT users and advanced level users in their specific sectors. OpenInn enhances creativity and self-realization as well as learning in formal and informal education. It builds on the method of open innovation to empower young talents to be visionaries and contribute to ~~a~~ new R&D knowledge in the future [27]. The viability of the portal could be confirmed by increasing numbers of the registered users (495), challenges (133) and ideas (205), created after the end of the project.

The OpenInn project outputs are used and further utilised within SP4CE project which addresses directly aims and needs identified in Bruges Communiqué on enhanced European cooperation in vocational education and training, especially defined there "improving the quality and efficiency of VET and enhancing its attractiveness and relevance" and "Enhancing creativity, innovation and entrepreneurship". All project results and project actions will be connected with promoting take-up of innovative practices in education, training by supporting personalised learning approaches, collaborative learning and critical thinking, strategic use of Information and

Communication Technologies (ICT), Open Educational Resources (OER), open and flexible learning, virtual mobility and other innovative learning methods. The mentoring & coaching scheme will be introduced. Teams will have support provided by mentors (from teachers/trainers staff) and coaches (from the enterprises staff.

ACKNOWLEDGMENT

This work has been kindly funded by the European Commission through project SP4CE - Strategic Partnership for Creativity and Entrepreneurship *(Grant agreement 2014-1-PL01-KA200-003341).*

REFERENCES

[1] M. Armstrong, Human Resource Management Practice. London and Sterling, VA: Kogan Page. 2005.

[2] S. Bayne, J. Ross, The 'digital native' and 'digital immigrant': a dangerous opposition. paper is presented at the Annual Conference of the Society for Research into Higher Education (SRHE) 2007. [online], Available at: <http://www.malts.ed.ac.uk/staff/sian/natives_final.pdf>. [Accessed 19. December 2014].

[3] J. Burgoyne, As reported in Personnel Management Plus.

[4] N. Butcher, ICT and Higher Education: A Summary of Key Issues for Consideration. Unpublished discussion prepared paper for the first meeting of the conference committee for the UNESCO 2009 World Conference: 'The New Dynamics of Higher Education. 2008.

[5] Romina Cachia, Anusca Ferrari, Kirsti Ala-Mutka and Yves Punie: Creative Learning and Innovative Teaching; Final Report on the Study on Creativity and Innovation in Education in the EU Member States, 2010

[6] G. Conole, Maarten de Laat, T. Dillon, J. Darby, Disruptive technologies', 'pedagogical innovation': What's new? Findings from an in-depth study of students' use and perception of technology", *Computers & Education, 50 (2) (2008), 511-524.*

[7] G. Cormode, B. Krishamurthy, Key differences between Web 1.0 and Web 2.0. First Monday, 13 (6). Retrieved October 5, 2008, from http://www.uic.edu/btbin/cgiwrap/bin/ojs/index.php/fm/article/view/2125/1972

[8] Digital Agenda Scoreboard 2014 – Slovakia, "Digital Performance of Slovakia." [online], Available at: <http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=5695>. [Accessed 24. January 2015].

[9] J.P.J. De Jong, W. Vanhaverbeke, T. Kalvet, H. Chesbrough, Policies for Open Innovation: Theory, Framework and Cases. Research project funded by VISION Era-Net, Helsinki: Finland, 2008, on-line: http://www.eurosfaire.prd.fr/7pc/doc/1246020063_oipaf_final_report_2008.pdf, [10..2015].

[10] M. Easterby-Smith, L. Araujo, Organizational Learning: Current Debates and Opportunities. In M. Easterby-Smith. L. Araujo. (Eds.) Organizational Learning and the Learning Organization. London: Sage. In: P. Juceviciene, I. Ceseviciute, Organizational Creativity as a Factor of the Emergence of Learning Organization. Social Sciences /Socialiniai Mokslai. Nr. 3 (65). 2009. ISSN 1392-0758.

[11] European Commission, Information Society, „i2010 in context: ICT and Lisbon Strategy. [online], Available at: < http://ec.europa.eu/information_society/eeurope/i2010/ict_and_lisbon/index_en.htm >. [Accessed 14. January 2015].

[12] European Commission, „Europa 2020", [online], Available at: < http://ec.europa.eu/europe2020/index_en.htm>. [Accessed 14. January 2015].

[13] European Commission, „Kľúčové údaje vzdelávaní a inováciách prostredníctvom IKT v európskych školách 2011". Education, Audiovisual and Culture Executive Agency, 2011, pp. 122 , ISBN 978-92-9201-203-8.

[14] D.A. Garvin, Building a learning organization, Harvard Business Review, pp. 78-84, 1993.

[15] Ch. Greenhow, B. Robelia, J. E. Hughes, Learning, Teaching, and Scholarship in a Digital Age. Web 2.0 and Classroom Research: What Path Should We Take Now? Educational Research, Vol. 38, No. 4, pp. 246-259. 2009 AERA: http://er.aera.net

[16] M. I. M. Hamzah, F. Maidin, S. Rahman, Supporting and Inhibitting Factors of Creativity, Innovation and Wisdom among Teachers in a Learning Organization. World Applied Sciences Journal 15 (Innovation and Pedagogy for Lifelong Learning), pp. 56-62, 2011. ISSN 1818-4952.

[17] H. Chesbrough, Open Innovation: Researching a New Paradigm. Oxford University Press, USA, 2006, pp. 392, ISBN-13: 978-0199290727.

[18] H. Chesbrough, Open Innovation: The New Imperative for Creating And Profiting from Technology. Harvard Business Press, USA, 2003, pp. 227, ISBN 1-57851-837-7.

[19] ICT Cluster, "Learning, Innovation and ICT (Lessons learned by the ICT cluster Education & Training 2010 programme)", 2010, [online], Available at: <http://www.kslll.net/Documents/Key%20Lessons%20ICT%20cluster%20final%20version.pdf>. [Accessed 10. October 2011].

[20] P. Juceviciene, I. Ceseviciute, Organizational Creativity as a Factor of the Emergence of Learning Organization. Social Sciences /Socialiniai Mokslai. Nr. 3 (65). 2009. ISSN 1392-0758

[21] S. McLester, Technology Literacy and the MySpace Generation: They're Not Asking Permission. *Technology & Learning, 27 (2007), 16-22.*

[22] R. Miller, J. Stewart. Opened University. People Management, 5, 12. In: P. Juceviciene, I. Ceseviciute, Organizational Creativity as a Factor of the Emergence of Learning Organization. Social Sciences /Socialiniai Mokslai. Nr. 3 (65). 2009. ISSN 1392-0758.

[23] OpenInn consortium, "Project official web site" [online], Available at: <http://project.openinn.eu/>. [Accessed 19. December 2014].

[24] OpenInn consortium, "The OpenInn Portal", unpublished internal project document, [Accessed 24. January 2015].

[25] OpenInn consortium, "The Knowledge Generating House", unpublished internal project document, [Accessed 24. January 2015].

[26] OpenInn consortium, "Report on the results of pilot test", internal project document, [Accessed 24. January 2015].

[27] OpenInn consortium, "Detail action plan", internal project document, [Accessed 24. January 2015].

[28] Ch. Redecker, K. Ala-Mutka, M. Baugalupo, A. Ferrari, Y. Punie, Learning 2.0: The Impact of Web 2.0 Innovations on Education and Training in Europe. JRC European Comission. 2009.

[29] F. Rennie, T. Morrison, E-Learning and Social Networking Handbook. Resources for Higher Education. Abingdon, Oxon: Routledge, 2008, pp. 208, ISBN 978-0415426077.

[30] B. Sanou, The World in 2014. ICT Facts and Figures. International Telecommunication Union. Switzerland 2014.

[31] P. Senge, The fifth discipline: the art and practice of the learning organization. NY: Doubleday Currency. In: P. Juceviciene, I. Ceseviciute, Organizational Creativity as a Factor of the Emergence of Learning Organization. Social Sciences /Socialiniai Mokslai. Nr. 3 (65). 2009. ISSN 1392-0758.

[32] SI presidency, "Lifelong learning for creativity and innovation (A Background Paper)", 2008, [online], Available at: < http://www.sac.smm.lt/images/12%20Vertimas%20SAC%20Creativity%20and%20innovation%20-%20SI%20Presidency%20paper%20anglu%20k.pdf>. [Accessed 25 January 2015].

[33] J. M. P. Tornero, "Promoting digital literacy - Understanding digital literacy". [online], Available at: < http://ec.europa.eu/education/archive/elearning/doc/studies/dig_lit_en.pdf >. [Accessed 14 January 2015].

[34] E. P. Torrance, R. E. Myers, Cerative learning and teaching. NYC: Dodd. Mead. 1970

[35] D. J. Treffinger, Encouraging creative learning for the gifted and talented. Ventura County Suprintendent of Schools, LTI Publications, 1980

[36] D. J. Treffinger, S. G. Isaksen, R. L. Firestien, Theoretical Perspectives on Creative Learning and Its Facilitation: An Overview. Handbook of Creative Learning: Volume I. 1982.

[37] UNCTAD: Creative Economy Report 2008. ISBN 978-0-9816619-0-2 [online], dostupné na internete: http://www.unctad.org/en/docs/ditc20082cer_en.pdf

[38] J. Walton, Strategic Human Resource Development. Great Britain: Pearson Educational Limited. In: P. Juceviciene, I. Ceseviciute, Organizational Creativity as a Factor of the Emergence of Learning Organization. Social Sciences /Socialiniai Mokslai. Nr. 3 (65). 2009. ISSN 1392-0758.

FASTER platform – an on-line tool for EU accountants education

D. Paľová*, M. Vejačka*
* Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
dana.palova@tuke.sk, martin.vejacka@tuke.sk

Abstract – The European Commission developed the Europe 2020 strategy for sustainable and inclusive growth and it was adopted by the European Council in June 2010. The EC agreed on five main objectives; one of them is to raise the employment rate of citizens 20 to 64 years old to 75 % by 2020. EU member states implemented these objectives into their national policies. In that way, the employment is at the heart of many EU policies. Based on data for 2013, the employment of persons aged 15 to 64 was at 64.1 % in EU-28. By gradual integration of new members into the EU the process of opening common marketplace becomes continuous and brings new businesses. From the perspective of accounting, in case of operating of company at new markets is appropriate that accountants are able to keep track of accountancy of other countries. Our experience shows that complications are common, so extensive consortium of educational institutions acting in the field of the education decided to create a tool facilitating and integrating accounting information. The aim of the paper is to present FASTER platform an on-line platform for accountants' education and job seeking – the result of FASTER project.

I. INTRODUCTION

Reporting and accounting in general has important role in market economy, since it enables tracking and enterprise financial planning. In the field of accounting, the importance of practical experience is indubitable. Exactly the experience allows new graduates to increase their value on the job market. Therefore, allowing of practical experience acquiring must be one of priorities in accounting education across EU states. Furthermore, accounting skills and knowledge are often important for the people with an education in other fields. Information on differences in area of accounting procedures and standards is often not available and unclear. In many cases education is provided mainly at theoretical level with lack of practical accounting skills training. Deficiency of praxis is one of reasons of young people high rates of unemployment [1].

Within EU-28 states accounting standards and procedures still differ, complicating the collaboration of various international companies and projects the accounting practice. Education in the field of accounting often lacks proper, extensive and suitable use of ICT [2]. Educational institutions offering cross-border accounting courses focus mostly only international standards, but rarely cover these differences between particular accounting systems in EU countries. Training and education in accounting is also provided by lifelong learning institutions. EU supports further practical training even after graduation (at secondary or tertiary schools), basically on lifelong time basis. In addition to that, using the flexible working conditions, for example part-time work, distant working or work from home, improvements in the availability of childcare facilities, providing more opportunities for lifelong learning, or facilitating job mobility, the labor market should be stimulated. Therefore lifelong learning, mobility and improvement of the education quality and efficiency are the basic objectives of the European Strategic Framework for Education and Training [1].

FASTER was an international cooperation EU project with aim to provide address these issues in accounting education area by creating online platform allowing education in accounting, standards and procedures comparison and job seeking and mobility in area of accounting across EU states. Consortium of five universities from five EU states proposed and developed online platform to address problem of praxis deficiency and job seeking in the field of accounting. This platform serves for users in all countries of consortium with the prospect of expansion to serve also other EU-28 states.

II. FASTER PROJECT

Project FASTER (Financial and Accounting Seminars Targeting European Regions) is a project on distance education and lifelong learning, funded by the European Union. The project started in January 2012 and ended in December 2013. The project product, FASTER platform, may be considered as an answer on EU Priorities on Knowledge Society and Innovation with objectives such as improving the quality and intensity of investments in human capital for the upgrade of the education system, investing in the future, promoting reforms in the education system and improving access and basic skill levels for all [3].

As was mentioned above, the project aims were based on Europe 2020 strategy objective, especially support long-term career advancement, international mobility of employees in accountancy branch, as well as graduates with degrees in disciplines other than accounting, or people employed in business, with the foundation studies to become a professional accountant. Special attention was placed on mature and motivated student seeking a

career in accounting for which no prior professional experience will be required. This group of people represents the group at risk due to their high unemployment [4]. With respect to these priorities, the project objectives could be summarized into following:

- to produce a modern technological e-learning & knowledge-sharing environment on finance and accounting which allows both students and professionals in pursuit of lifelong learning to gain practical knowledge on their discipline, but not restricted to national borders. The platform proposes besides e-learning environment also a job seeking mechanism.
- to link studies & working environment and widen transnational pursuit of jobs in finance & accounting.
- to offer new incentives and produce a prototype ICT environment expandable to all languages and to all disciplines.

Connected to the project focus – to help people to understand accounting and practice it at the same time in a friendly, step-by-step learning virtual accounting office environment - the final platform represents an interesting tool for wide groups of users [5]:

- educational institutions – academics and students in the field of accountancy,
- people interested in vocational training,
- public institutions and private companies, potentially looking for employees in area of accounting,
- jobseekers
- and consultants operating in the field of accountancy.

On the project cooperated five educational institutions from different EU member states: Athens University of Economics and Business (Greece), Sofia University "St. Kliment Ohridski" (Bulgaria), City University, London (United Kingdom), University of Warsaw (Poland), Technical University of Kosice (Slovakia). The consortium developed a European e-learning tool supporting European collaboration in the field of Finance and Accounting – FASTER platform described below in more details.

FASTER project activities could be shortly described via following timeline table:

TABLE I. BRIEF TIMESCHEDULE OF FASTER PLATFORM DESIGN

Date	Activity	Outcome
January – March 2012	Collection of users' requirements via questionnaires and interviews	Requirement analysis & specifications
June – July 2012	Collection of sources for Part A – literature (e-books, books, journals, legislations, etc.) on accounting, glossary, interesting and important links from particular states and EU as whole	FASTER Platform - Part A
August – October 2012	Development of on-line study material content for 8 different modules Analysis of platform processes and functionality	FASTER platform – Part B Platform first (pre-alpha version) architecture design
February – July 2013	Development of jobseeking part – searching for the employers, job positions, functionality for uploading CVs, employer-employee communication, arrangement of appointment Translation of all parts of platform Pilot using of platform by partnership	FASTER Platform - Part C Multilingual platform / Alpha version Bug report on Alpha version
September - November 2013	FASTER platform pilot testing by target groups members	FASTER platform pilot users' feedbacks Bug report on final version of platform
November – December 2013	Improvement of platform based on pilot test outcomes	Final version of FASTER platform

Besides platform development activities we did also many different exploitation and dissemination activities in order to provide the appropriate ground of exploitation of the product after the project's lifetime. We used mostly:

- transfer of results to target groups - in order to successfully transfer information about FASTER project we have set several communication channels and conducted several dissemination activities:
  - directly via pilot tests,
  - printing and distributing the newsletters (in 2012, 2013), poster (Figure 1) and brochures (Figure 2) via mail,

Figure 1. FASTER project official poster

Figure 2. FASTER project brochure

- different publications in the press - advertisements (electronic media, newspapers, TV interviews); press releases about opportunities, ideas and exploitation of FASTER platform in teaching accounting,
- publishing multimedia advertising multilingual spot,
- organization of two information days and various on-line events,
- distribution of information about project events and activities connected with project via e-mail,
- providing LinkedIn social communication channel with aim to establish contacts to accounting professionals,
- developing official FASTER web-site and its national versions (Figure 3) available at http://www.fasterproject.eu.

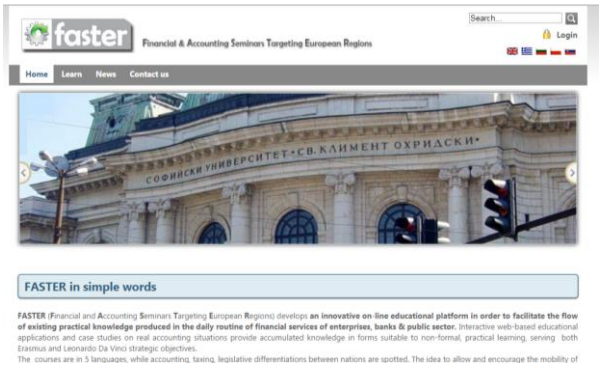


Figure 3. Official FASTER project web-site

- integration of FASTER platform in partners' universities' curricula - during the pilot testing period and afterward the partners integrated the FASTER platform to their curricula, for example: Accounting Information Systems, Financial Accounting, Financial reporting standards, Basic Accounting, Accounting, Accounting, Auditing and Fiscal Consulting, Accounting of Public Administration, Economic Information Systems,
- organization of FASTER track on international conference - 6th Conference EuroMed 2013. Track title was "e-learning web based models for business and accounting: Requirements, procedures, implementation experiences and implications on lifelong learning".
- setting up ICT – Accounting Network and National Support Groups:
  - *ICT - Accounting network group* of people from members the FASTER target groups created with aim to serve the project as the initial international pool of exploitation and dissemination.
  - *National Support Groups (NSG)* formed in each country from scientists and professionals from universities, accounting offices, financial organizations, from both private and public sector. The intention of the NSG is to provide support to the project and to represent the interests of each key player and country. The NSGs had a role to help steer and advise the project team and the implementation of the FASTER platform.

To achieve quality and fulfill requirements of further exploitation in real life, partnership used different evaluation of all sub-processes done during project lifetime:

- *during design phase* - analysis of future user requirements via on-line questionnaires and interviews,
- *during pilot testing phase* – analysis of feedbacks of pilot users.

At the same time the partnership developed different supporting groups of people, helping to partners fulfill project objectives in accurate way for defined target groups:

- *Accounting Network and National Support Groups*, described above. These support groups had to follow the project team activities and outcomes and to give an advice how to improve it.
- *External Evaluation Committee* - an external mechanism of results and outcomes evaluation throughout the project implementation.

III. FASTER PLATFORM

The FASTER platform is the main outcome of the FASTER project. User requirements on FASTER platform were investigated by a survey conducted in Bulgaria, Greece, Poland and Slovakia at the beginning of the project (see Table I). Four potential target groups (graduates and undergraduate students, academics, employers, people interested in vocational training) were addressed in this survey. Respondents evaluated usefulness of planned features of FASTER platform and suggested any additional functionalities of a platform [6].

Following numbers of respondents participated on user requirement analysis:

- 354 graduates and undergraduate students,
- 40 people interested in vocational training,

- 33 academics and
- 24 employers.

All groups of potential users in average indicated that on-line platform would be very useful for teaching accounting in form of web-based educational applications, case studies on real accounting situations and also for job seeking. Detailed responses are in following table:

TABLE II. USEFULNESS OF PLATFORM PARTS RATED BY POTENTIAL USERS

All countries	Students	Academics	Employers	People interested in vocational training
Educational part	3.87	4.39	4.33	4.03
Jobseeking part	4.27	4.27	4.33	4.20

1-5 Likert scale (1: not useful at all – 5: very useful), Source: [7]

Based on data presented in Table II, the FASTER platform has potential to fill out a large gap existing in education and practice in the field of accounting. The respondents of this survey demanded the opportunity of practical training by exercises and applications as the most important and useful feature of future platform. The FASTER project consortium took results of this analysis into account during creation of FASTER platform.

The platform now includes laboratory with bibliographical resources and glossary with 734 terms in all five consortium national languages (English, Greek, Bulgarian, Slovak and Polish). The bibliographical sources contain 182 journals' titles, 371 book's titles, 255 legislation related internet links, 164 e-books, 69 Web sites.

The FASTER platform educational core consists of eight e-learning modules in following accounting sections:

1. Financial Accounting,
2. Accounting Plans,
3. Accounting for Public Limited Companies (S.A.) – Limited Liability Companies (Ltd) – Groups of Companies,
4. Tax Accounting,
5. Balance Sheet Issues,
6. Analysis of Financial Statements - Business Valuation,
7. Accounting ERP,
8. International Accounting Standards.

Content of each of these 8 e-modules consists of succinct summary of relevant theory in given accounting area, practical applications and examples of use, exercises including self-assessment testing, questions (yes/no or multiple choice answers) and additional suggested reading material for given topics.

Third main pillar of FASTER platform is a jobseeker feature that allows advertising jobs positions or posting CVs of people looking for job in the field of accounting. This feature of the platform facilitates connection between education and labor market, where requirements of employers can be met by skills and competences of job candidates. So the jobseeker part of FASTER platform supports job mobility in the field of accounting across EU states [8].

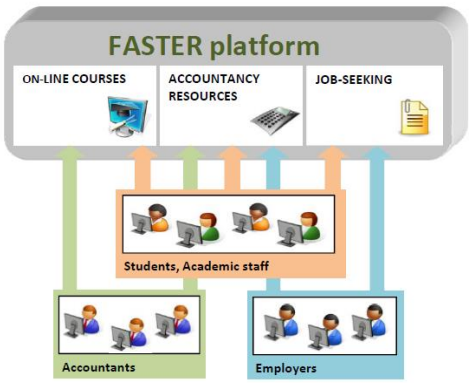


Figure 4. Schematic representation of FASTER platform

After establishing initial version of FASTER platform in 2013, its pilot testing was conducted. The platform's compliance with user requirements and its technical and educational qualities were tested. Testers were from project target groups and relevant institutions such as the academic and student community, SMEs, employers and accounting offices, financial services of governmental agencies and local authorities [9].

The pilot testing was realized in three stages:

1. Initial stage – selection of testing subjects through application forms in each consortium country
2. Pilot testing of the FASTER platform by selected testers.
3. Evaluation by filling out evaluation forms

TABLE III. NUMBER OF PILOT TEST PARTICIPANT IN PROJECT COUNTRIES

State/Target group	***Greece***	***Poland***	***Bulgaria***	***United Kingdom***	***Slovakia***	***Total***
Academics	5	4	5	18	9	41
Students	46	97	24	19	53	239
People interested in vocational training	10	16	5	17	37	85
Employers	11	4	5	13	6	39

Pilot testing participants evaluated functionality of the platform, clarity of its layout and format, easiness to use, content of the platform, accuracy of information and usefulness for teaching, learning and job seeking.

Pilot testers highly evaluated the most of functionalities of the FASTER platform including job seeker feature. They highlighted practical exercises and case studies. Mainly technical problems were denoted as

negatives, for example slow reaction time of servers, exercises solving issues, very long process of creating exercises or problems of displaying in some web browsers. All problems reported in external pilot testing were addressed by the project consortium before releasing final version of platform to the general public. The FASTER platform fulfilled users' requirements according overall evaluation of pilot testing.

The final version of FASTER platform (Figure 5) is public and open accessible via following link: http://platform.fasterproject.eu/Homepage.

Figure 5. Initial page of FASTER platform

IV. INTEGRATION OF FASTER PLATFORM TO EDUCATION

Integration of FASTER to education was and still stays one of most important FASTER platform exploitation tool. During the pilot testing almost every partner integrated FASTER to one or more subject curricula educated at their institution. In more detail the integration took place in the following universities and subjects (Table IV) [9]:

TABLE IV. LIST OF SUBJECTS REALISED BY PARTNERSHIP USED FOR FASTER INTEGRATION

Subject description	Modules
Athens University of Economics and Business	
• Accounting Information Systems	
Presentation of basic principles of Accounting Information Systems. It provides the necessary theoretical and technical background required in order for students to understand the interconnection between business and accounting processes and the use of information systems.	All 8 modules
University of Warsaw	
• Financial Accounting	
The course aims at acquainting students with regulatory framework and accountancy rules, financial accounting system, which is a logical, complete and compact system, describing assets, sources of their financing, business activities.	M1, M2, M3
• Financial reporting standards	
The course aims at familiarizing students with the principles of financial reporting standards. At the end of the course students will be able to prepare balance sheet, profit and loss account, notes to the financial statements, cash flow statement and statement of changes in equity, based on information from the books and records of the company	
• Basic Accounting	
The course aims at acquainting students with regulatory framework and accountancy rules, financial accounting system, which is a logical, complete and compact system, describing assets, sources of their financing, business activities.	M1, M5
Sofia University "St. Kliment Ohridski"	
• Accounting	
The course aims to teach the students about the main accounting principles and rules, the accounting objects – assets, equity, liabilities, revenues, costs and financial result, the preparation of the financial statements.	Partially all 8 modules
Technical University of Kosice	
• Accounting	
The content of the subject covers chapters: The nature and importance of accounting, The structure of the accounting information system, Business asset, Sources of property, Balance Sheet, Economic operation, Active marketing operations, Account, An accounting documentation, Synthetic and analytical evidence, Checking the formal correctness of the accounting records, The financial statements, Financial statements.	Partially all 8 modules
• Auditing and Fiscal Consulting	
The content of the subject covers: Auditing profession, auditor independence and auditor liability and the client; Fundamentals of auditing, other assurance services, activities prior to contract and audit planning; Internal control system. Sampling in auditing; The application of audit procedures in the audit cycles; The basic objectives of the audit cycle, the main business cycles, control objectives and control procedures; Influence the results of the audit to the auditor's report; Completion of the audit and the audit opinion formulation, the auditor's report; Administration of taxes and fees, auditor's code of ethics, code of conduct tax adviser.	M4, M5, M6, M7, M8
• Accounting of Public Administration	
The content of the subject cover accounting procedures and framework chart of accounts to budget organizations, municipalities and higher territorial units and accounting for the accounting group 0, group 1, group 2, group 3, group 4, group 5, group 6 and group 7 and closing the books and preparing of financial statements.	All 8 modules
• Economic Information Systems	
The goal of subject is using information and communication technologies to learn to operate a small virtual company from its inception to annual financial statements. The course provides practical experience in accounting, marketing and building economic information system of a modern company.	M1, M2, M3

* The abbreviations of modules: M1 - Financial Accounting; M2 - Accounting Plan; M3 - Accounting of Public Limited Companies (S.A), Limited Liability Companies (Ltd.), Groups of Companies; M4 - Tax Accounting; M5 - Balance Sheet Issues; M6 - Analysis of Financial Statements - Business Valuation, M7 - Accounting ERP; M8 - International Accounting Standards

As is presented above, in Table IV, partnership integrated 8 modules in different combinations developed under FASTER project into 9 subject curricula concerned on various parts of accounting, auditing and economic information systems.

Integration process should be generally described by following phases [10]:

1. Initially, a lecturer explained to students the aims, usefulness and functionalities of the FASTER platform. During this lesson, instructions for the registration demonstration of the usage of the platform were provided.
2. Students used FASTER platform during laboratorial part of the courses and asked to study particular part of subject using platform courses and sources at their homes, to exercise themselves what they have learned during the lectures solving the tasks and activities provided by platform. In case of problems they have possibility to consult the problems via e-mail or on other face-to-face lessons.
3. All students were asked to evaluate platform and its parts using evaluation forms.

Presented integration validates the fact, that the FASTER platform should be useful in education as a whole but only a part. Students may use only those modules and their content that they currently need within the particular subject. During pilot testing particular partners used the platform within education process of full-time students as well of external students, who usually miss a teacher's explanation communicated during the face-to-face lessons.

V. CONCLUSION

Based on previous information can be concluded, that FASTER consortium developed an innovative on-line educational platform for wide group of people: graduate and undergraduate students of economics, people interested in lifelong learning. FASTER platform presents multilingual tool for facilitating the flow of existing practical knowledge which is not limited by national boundaries. The platform also provides connectivity between study and work experience and expands the interstate job-seeking in the field of accounting.

In the context of employment and social cohesion the gradual increase the practical training in education FASTER supports the development of those skills and abilities which are needed to cope with the ever-changing labor market. In that context users of FASTER platform will be able to appropriate integrate into a professional and business environment.

As final conclusion it could be stated that versatility of educational part of the platform ensures more sustainability of FASTER platform in educating of future accountants, economists but also of people only vocationally interested on the field of accounting.

ACKNOWLEDGMENT

This research has been kindly funded by the European Commission through the LLP Project FASTER: Financial and Accounting Seminars Targeting European Regions (No. 518894-LLP-1-2011-1-GR-KA3-KA3MP). I also wish to acknowledge my gratitude and appreciation to all the FASTER project partners for their contribution during the development of research presented in this paper.

REFERENCES

[1] European Commission, "Education and Training Strategic framework for education and training," [online], Available at: <http://ec.europa.eu/education/lifelong-learningpolicy/doc28_en.htm>. [Accessed 25. January 2015].

[2] L., Révészová, "How to integrate ICT into the education?", In: Information and Communication Technology in Education proceedings: Rožnov pod Radhoštěm, Czech Republic, September 2012, pp. 219-228.

[3] European Commission, "Europe 2020 in a nutshell," [online], Available at: <http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/priorities/index_en.htm>, [Accessed 24. January 2015].

[4] FASTER consortium, "Project official web site" [online], Available at: <http://www.fasterproject.eu>. [Accessed 19. December 2014].

[5] FASTER consortium, "FASTER Progress Report – public part" [online], Available at: <http://eacea.ec.europa.eu/llp/projects/public_parts/documents/ict/2011/mp_518894_ict_PR_FASTER.pdf> , [Accessed 24. January 2015].

[6] FASTER consortium, "User requirements on FASTER platform: Survey outcomes," [online], Available at: <http://www.fasterproject.eu>. [Accessed 27. January 2015].

[7] FASTER consortium, "Newsletter 2," [online], Available at: <http://fasterproject.eu/attachments/article/103/Newsletter_Issue_2_en.pdf>. [Accessed 28. January 2015].

[8] FASTER consortium, "Newsletter," [online], Available at: < http://fasterproject.eu/attachments/article/91/newsletter_EN.doc.> [Accessed 26. January 2015].

[9] FASTER consortium, "Newsletter 3," [online], Available at: <http://fasterproject.eu/attachments/article/115/FASTER_Newsletter_3_EN.pdf>. [Accessed 28. January 2015].

[10] FASTER consortium, "Integration of FASTER platform in partners' universities curricula", Internal project document, [Accessed 30.October 2013]

User Knowledge Development in the Field of Electronic Services and Their Security in Slovakia

M. Vejačka*, J. Bucko*
*Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
jozef.bucko@tuke.sk, martin.vejacka@tuke.sk

Abstract - Computerization of society and electronic government are current problems of society and its development is the part of all governments' program declarations in the European Union. Despite of the well-built internet infrastructure, Slovakia is still tail ranking, when considering the level of development of electronic services and their use by the general public. The precondition for improving this situation is to increase the level of knowledge about electronic communication, its security and trust in these services. Theoretical knowledge and practical skills in the area of electronic services are developed during undergraduate studies at the Faculty of Economics. The aim of this paper is to present the development of knowledge and skills in this area among Slovak students of third grade at our faculty, during the years from 2009 to 2014. Students' knowledge is evaluated and compared by newly constructed index. Data were collected by a questionnaire survey during each year in the period. Results of evaluation and comparison are briefly presented and assessed.

I. INTRODUCTION

Globalization lays high demands on the EU member states in the effective use of modern information and communication technologies (ICT) in almost every area of human activity. Active use of ICT and the global computer literacy of the EU citizens in general is one of these demands. Digitalization of offline processes in various life aspects is a necessity. Effective use of electronic communication is one of requirements of digitalization.

In the center of activities is a secure electronic communication between population of state and institutions of public administration, so called e-government. The concept of e-government is the subject of many studies and different definitions of this term. The phrase "Electronic government and governance" as a synonym for e-government covering all administrative functions of public administration and civil performed electronically by using information and communication technologies [1].

The primary objective of electronic delivery of public services is to replace or minimize necessity of personal visits of public administration offices, shorten reaction time for processing and issuing the required documents and to streamline public services provided [2]. According to Means and Schneider, e-government is defined as the relationship between the government, the "customers" (businesses, citizens, but also other governments), and its "suppliers" (well, businesses, citizens and other governments) realized using electronic means [3]. In this area Slovakia still lags behind the developed EU countries.

The last decades have brought a rapid increase in the form of information sharing systems based on cloud technologies. The basis of these technologies is electronic services directly enabling their use. The degree of safety and trust in these services determine the rate of active usage of cloud solutions. Trust and security aspects in the development of these cloud platforms were and are at the aim of the research activities of the European FP6 and FP7 projects (e.g. FLUID-WIN, SECURE, SERENITY, SWAMI, HYDRA, SEAMLESS etc.).

One of the reasons for the electronic services usage in Slovakia is lower literacy populations regarding security of the services and the related lack of confidence to use these services. Building trust in any new technology is a very complex process. Our goal is to raise awareness in the area and increase the number of given technology users. The best way to achieve this goal is the active inclusion of these issues in the teaching process at universities. An important part of the course is theoretical knowledge of the issue as well as practical experience with technology enabling increasing the security of electronic communication, such as the active use of electronic signatures. In line with this fact, we have integrated into the education process the course Electronic Services in Banking. The course is included in the schedule of teaching third year of study in Finance, Banking and Investment field of study.

II. EDUCATION IN THE FIELD OF ELECTRONIC SERVICES

The Faculty of Economics at Technical university of Kosice provides education in specializations of Finance, Banking and Investment and Public Administration and Regional Development. Electronic services in banking and e-government are in the scope of studies at this faculty in the course of Electronic Services in Banking (ESIB). It

provides combination of skills acquisition in e-learning environment with theoretical seminar on topics related with electronic services in banking and also in public administration.

Additional topics are discussed in essays which are integral parts of this course. Essays are aimed at students' real-life experience with chosen topic related to electronic services and informatization. Students get to know various forms of electronic banking in both theoretically and practically. Great importance is attributed to security of given electronic banking form and building trust in its use. Our previous research showed that security of electronic banking perceived by its users has significant influence on its acceptance by its users [4]. Significant part of course represents the electronic signature topic. Course participants acquaint with principles of electronic signature (in form of digital signature) and with its practical use in secure electronic communication [5].

For teaching support of this course, Learning Management System (LMS) Moodle is used, which is used for e-learning in general at our faculty. Second important tool in this course is Electronic Online Bank. It is virtual bank administered to our faculty by company UNICOM for free for educational purposes. Its software is antiquated, but it might be used for technological demonstration and testing of multiple forms of electronic banking and their authorization [6]. Virtual bank supports digital signature authorization and the most of e-banking forms up to modern, but excluding mobile banking for smartphones. Third tool supporting education process within ESIB course is the faculty's certification authority based on EJCBA open source authority. It provides the possibility of digital signature certificates issuance and their management.

Thus, the main themes of ESIB course are as follows:

A. Electronic signature and secure communication

B. Internet Banking,

C. Home Banking,

D. Mobile Banking and Smart Banking.

A. Electronic signature and secure communication

Practical skills as long as theory in the field of electronic signature are main aim of this part of our course. Main topics of electronic signature introduced are symmetric and asymmetric encryption (RSA), digital and electronic signature, legislation, software applications and Public Key Infrastructure (PKI). Signing, encryption and following decryption and signature verification is practically demonstrated and tested by students. File and email encryption is supported by Mozilla Thunderbird and freeware GPG4win. Possibilities of electronic signature usage are theoretically outlined and discussed.

Faculty's certification authority is used for Public Key Infrastructure (PKI) theoretical as well as practical introduction. Students apply for certificate and acquire it from our certification authority and they further use it in for all eligible purposes practically. Manuals for certificates usage are available at certification authority webpage.

B. Internet banking

Second part of our course is dedicated to the Internet banking as the most used technology in area of electronic banking. Emphasis in this topic is on security and safety of internet banking using. Therefore multiple ways of authentication and authorization are discussed. In praxis basic login and password are often complemented with additional security element. Grid card, One-Time password and digital signature are supported in our virtual bank. Students try various forms of payments, collections, electronic statements in internet banking of our virtual bank.

C. Homebanking

Homebanking in our course is understood as electronic banking form, primarily intended for institutional clients and enterprises where communication between client and bank is supported dedicated homebanking software from the bank, while the actual data is transferred via network. Homebanking is often integrated with accounting software or firm's ERP system. Students test about active and passive account homebanking operations, importing and exporting banking operations [5].

D. Mobile banking and Smart banking

Mobile banking represents group of electronic banking forms. In this part of Electronic Services in Banking course students get to know multiple forms of mobile banking from WAP banking and GSM banking up to Smart banking. In recent years, smart banking (e-banking in smartphones) recorded very strong growth in usage thanks to rapid spread of smartphones. Students may practically test WAP banking and GSM banking in our electronic online bank. Smart banking is practically tested on demo application.

III. METHODOLOGY OF VECTORS COMPOSITION

Improving and increasing the efficiency of public administration and the transition to computerization is closely related to the promotion and development of personal skills of citizens, businesses and public officials at all levels. Universities should ensure the development of skills across the informatics society to the issues of electronic signature and secure electronic communications, as well as handle the safe use of electronic services. Not only by teaching appropriate courses and training of future professionals and savvy users in this area, as well as the implementation of scientific and development projects focused on his issue.

The weakest link of each system in terms of security is still human factor and unsafe behavior when working with ICT. Knowledge of the issues of secure communication and level of trust in these technologies has been the subject of research by an electronic questionnaire. Research, the knowledge and use of electronic services is realized since the year 2009. The questionnaire is distributed electronically to students of ESIB course using Google documents. It consists of 3 areas covered 18 questions and the demographic section containing 4 questions. Over 97% respondents were from Slovakia in age range from 20 to 22 years, because our survey was

distributed only to students of our course. In each year approximate number of respondents was 120 students. The first part of the questionnaire examines the issue of knowledge on electronic services in general. The second part of the questionnaire is dedicated to electronic signature and the third part is devoted to the issue of electronic communications with public administration (e-government).

For expression of the degree of the issue knowledge we have defined the vector of knowledge of electronic security services and electronic communications.

We have labeled this vector with symbol $\vec{V}_k$. This vector pf knowledge consists of three individual vectors (components):

$\vec{I}_{es}$ - individual vector of electronic services knowledge

$\vec{I}_{sig}$ - individual vector of electronic signature knowledge

$\vec{I}_{eg}$ - individual vector of e-government knowledge

Individual vectors are an expression of partial sections of the questionnaire. In terms of questionnaire results processing methodology, individual vectors have maximum of 1. Consequently, knowledge vector electronic services and electronic communications we express in a natural way, as a linear combination of the individual vectors:

$$\vec{V}_k = a \times \vec{I}_{es} + b \times \vec{I}_{sig} + c \times \vec{I}_{eg} , \qquad (1)$$

where *a, b, c* are calculated values of the knowledge of the examined area by the processing of the questionnaire.

$a,b,c \in \langle 0,1 \rangle$, (or 0 - 100%)

The length of the resulting vector $|\vec{V}_k|$ can be expressed by expression:

$$\left|\vec{V}_k\right| = \sqrt{a^2 + b^2 + c^2} , \qquad (2)$$

it will express the level of knowledge of electronic security services and electronic communications, $\left|\vec{V}_k\right| \in \langle 0, \sqrt{3} \rangle$, (or 0,100%)

Then we defined the index knowledge of electronic security services and electronic communications and identified by the symbol I_k. Index may have values from ⟨0, 1⟩ and it can be expressed as the following ratio:

$$I_k = \frac{|\vec{V}_k|}{\sqrt{3}} , \qquad (3)$$

IV. RESEARCH SURVEY AND RESULTS

The vector of knowledge $|\vec{V}_k|$ was from answers on 18 questions of the questionnaire, from which 16 were multi-choice and self-assessment questions and two questions were direct responses formation. The respondents were students of ESIB course. For processing the questionnaire we used scoring method of estimating weights of criteria, which is a known method of multicriterial evaluation for questionnaire processing [7]. Based on the quantitative methods we judged different types of response to these issues questionnaire (b_i) and then we calculated the corresponding weights (v_i) the individual responses according the following formula:

$$v_i = \frac{b_i}{\sum_{i=1}^{k} b_i} , \qquad (4)$$

Based on the method of scoring, we rated initial ratings answers of individual issues and expressed the corresponding weights of the responses as follows:

TABLE I. QUESTIONAIRE ANSWERS AND THEIR WEIGHTS

Question/ Answer	b_i	v_i
1. Have you ever used any of the electronic services?		
Email	1	0.07
Electronic banking services	2	0.13
Purchase in e-shop	3	0.20
Electronic payment	4	0.27
Card payment on the internet	5	0.33
2. Which service of these is the safest according to you?		
Email	1	0.08
Electronic banking services	4	0.31
Purchase in e-shop	3	0.23
Electronic payment	3	0.23
Card payment on the internet	2	0.15
All of above are safe	1	0.08
3. Which of following services do you trust at least?		
Email	5	0.23
Electronic banking services	2	0.09
Purchase in e-shop	3	0.14
Electronic payment	3	0.14
Card payment on the internet	4	0.18
I don't trust all of above	5	0.23
4. What would increase your trust in given services?		
Reading short publication on this topic	2	0.11
Viewing website on this issue	2	0.11
Watching popularisation TV show about it	3	0.16
Visiting expert lecture on this topic	3	0.16

Listening to my friends' experience	4	0.21
Trying out on demoversion myself	5	0.26
5. Have you ever used electronic (digital) signature?		
No	0	0.00
1 – 2 times	2	0.20
Less than 5 times	3	0.30
I use it more frequently	5	0.50
6. Do you know, what is electronic signature used for?		
Yes	5	0.71
No	0	0.00
I am not sure	2	0.29
7. Do you know, what are prerequisites of electronic signature using in Slovakia?		
Yes	4	0.40
Yes, probably	3	0.30
I am not sure	2	0.20
No, probably	1	0.10
I do not know at all	0	0.10
8. Do you know the difference between certificate and electronic signature?		
Yes	3	0.50
Yes, probably	2	0.33
I am not sure, but I heard about it	1	0.17
No, I never heard about it	0	0.00
9. Where you can acquire personal certificate?		
Local administration office	0	0.00
Registration authority	2	0.20
Certification authority	5	0.50
Court or notary	0	0.00
Bank	3	0.30
10. Which of electronic signatures has nore general validity?		
Commercial	3	0.33
Warranted	5	0.56
Does not matter, both are equal	1	0.11
11.How much does commercial certificate cost in Slovakia?		
Less than 10 EUR	0	0.00
10 – 20 EUR	2	0.25
20 – 30 EUR	3	0.38
30 – 40 EUR	2	0.25
40 – 50 EUR	1	0.13
More than 50 EUR	0	0.00
12. Would you like to use electronic (digital) signature?		
Yes	5	0.71
No	0	0.00
I am not sure	2	0.29
13. What form of communication with public administration do you prefer?		
Personal contact	0	0.00
Written form	1	0.11
By phone	3	0.33
Electronic (email, web form)	5	0.56
14. If you would need electronic signature for electronic communication with public administration, would you use it?		
Yes	5	0.71
No	0	0.00
I do not know	2	0.29

Investigation of skills and knowledge using electronic is underway since 2009 to the present by a questionnaire, which is distributed via Google docs [8]. In this paper we investigated the evolution of the calculated index during period of 2009 - 2014. The results are summarized in the following tables and graphs.

TABLE II. RESEARCH RESULTS FOR PERIOD 2009 - 2014

Year	**a**	**b**	**c**	**V_k**	**I_k**
2009	0.6198	0.5344	0.8750	1.1980	0.6917
2010	0.7262	0.5599	0.8198	1.2300	0.7101
2011	0.6997	0.5145	0.8065	1.1852	0.6843
2012	0.7366	0.5586	0.8125	1.2307	0.7106
2013	0.7471	0.5510	0.8224	1.2402	0.7160
2014	0.7982	0.5869	0.7540	1.2450	0.7188
Total	0.7364	0.5532	0.8034	1.2223	0.7057

Table II shows the values of a, b, c components of vectors individual vectors and calculated magnitude of the vector of knowledge (V_k) and index of knowledge (I_k). The detected values show a small increase in knowledge and use of electronic services (17.8%), as well as small increase in knowledge of the issues of electronic signatures (5.25%). Surprising result however is the decrease in trust and willingness to communicate with the public administration electronically almost during whole period. This decrease is probably an indicator of a lack of awareness in this field among young people, as well as the lagging implementation of the objectives of e-government building.

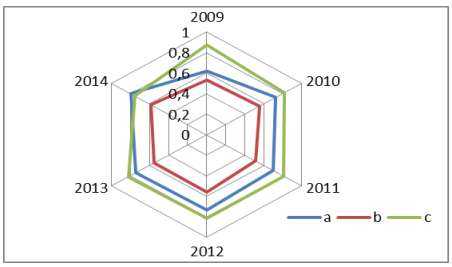


Figure 1. Index of knowledge components during period 2009 - 2014

Figure 1 presents these data graphically. For these values, we expressed the change of index of knowledge in security electronic services and electronic communications I_k (Figure. 2) from 2009 to 2014. The index reached the 2014 level of almost 0.72, what is a slight increase of almost 0.03 since 2009. Surprisingly, the level of knowledge fell in 2011. It suggests lower level of knowledge or less effort of our course students in 2011.

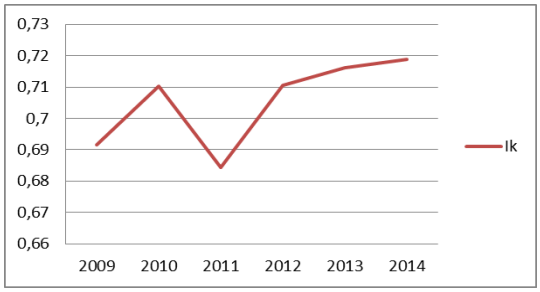


Figure 2. Index of knowledge during period 2009 - 2014

Another reason of decrease of knowledge index in 2011 might be also decrease of trust in e-government in Slovakia, because of the case of dysfunctional electronic tax system introduction named KONS at the time. This system had to be replaced by the old one virtually immediately after its implementation. This case is still not resolved by the Slovak court [9].

Furthermore respondents of our survey might be skeptical about functionality of e-government processes, because these processes are electronic mirror of paper processes without consideration of their optimization for electronic environment.

Since 2011 respondents had to respond to questionnaire twice, before and after completion of our course. The total average values (for all years in total) of knowledge index are presented in the following table and figure.

TABLE III. TOTAL AVERAGE INDEX OF KNOWLEDGE VALUES BEFORE AND AFTER COURSE

Period	a	b	C	V_k	I_k
Before course	0.736	0.533	0.803	1.222	0.706
After course	0.804	0.738	0.800	1.353	0.781

As we expected, after the course completion values of knowledge vector V_k and knowledge index I_k are higher than before course completion. It is obvious that the global awareness of the issue among students was increased by almost 0.08 and reached 0.781. However, the individual vector of e-government knowledge (I_{eg} - parameter c) slightly decreased. It shows the same trend as the decrease of this parameter in the overall development of the knowledge index (see above).

Vector of electronic services knowledge (I_{es} - parameter a) grew by almost 7%, which roughly corresponds to an increase whole index of knowledge. The highest increase was recorded by knowledge vector of electronic signature (I_{sig} - parameter b) by 0.20, which refers to the fact that the issue of electronic signature, the safety of this technology and its use is still not well known.

Results of our research suggest that our course helps to promote knowledge in areas of electronic services in banking and electronic signature in the conditions of Slovak republic. However in the area of e-government we should improve course contents to better support knowledge and skills necessary for e-government usage.

The following chart shows a comparison of the different vectors and the index of knowledge. Left column always represents status before the course completion and right column status after the completing the course.

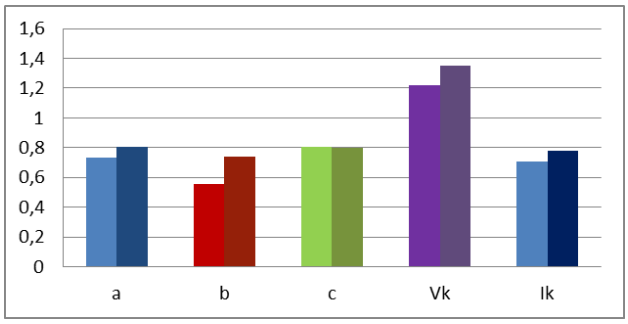


Figure 3. Total average vector components and final index before and after ESIB course completion during perioad of 2009 - 2014

V. CONLUSION

The aim of our paper was to highlight the importance of education in the field of electronic services security and electronic signature in Slovakia, as well as their use for communication with the public administration as one of the important indicators of the successful development of e-government. The necessity of raising awareness and active inclusion in the education process was tested questionnaire research conducted among students of third year of bachelor study at Faculty of Economics, Technical university of Košice in Slovakia during the years from 2009 until 2014.

The results of the questionnaire were processed by the method of points assigning and then calculating the weights of individual answers based on knowledge of the area in each year of research. By defining appropriate vectors of knowledge and the overall knowledge vector (V_k), overall index of knowledge was calculated. The results correspond with our expectations, although paradoxically a decline in student interest in the development of e-government was recorded. It indicates declining trust in the success of development and active use of these services in Slovakia, although the overall interest of the electronic signature and electronic services has increased. We can confirm that in the field of electronic services in Slovakia lacks efficient and effective campaign and education. Electronic services in banking and public administration promotion needs to be strengthened.

We will continue in this research, with the expectation of what results we achieve in the next period, which might be strongly influenced by the fact that in Slovakia started in 2014 to issue electronic identity cards, which allows to each owner an active use of electronic signature technology. Early information on low percentage of issuance of identification cards with digital signature included indicates still low interest of Slovak citizen to use electronic services with assistance of digital signature. This fact suggests that awareness rising in this area is increasingly needed.

REFERENCES

[1] M. Yildiz, "E-government research: Reviewing the literature, limitations, and ways forward", in Government Information Quarterly 24, 2007, pp. 646–665.

[2] A. Demczuk and A. Pawłowska, "Progress toward e-government in Poland: Issues and dilemmas", in Information Polity 11, 2006, pp. 229–240, IOS Press.

[3] G. Means and D. Schneider, "Meta-capitalism: Thee-business revolution and the design of 21st century companies and markets," 2000, New York: John Wiley&Sons Inc.

[4] M. Vejačka, "Customer acceptance of electronic banking: Evidence from Slovakia," in: Journal of Applied Economic Sciences. Vol. 9, no. 3 (2014), p. 514-522.

[5] J. Bucko, "Elektronické služby v bankovníctve," (in Slovak) 1st edition, Košice: TU, EkF, 2009, p. 92.

[6] L. Révészová and D. Paľová, "ICT challenges, opportunities and choices in the knowledge society development," in: ICT as a Determinant Factor in the Development of the Public Services in the Knowledge Society. - Targu Mures : Petru Maior University, pp.13-37, 2013.

[7] P. Fiala, J. Jablonský and M. Maňas, "Modely a metody rozhodování," Praha: Nakladatelství Oeconomica , 2008.

[8] J. Bucko and M. Vejačka, "Bezpečnosť a dôvera elektronických služieb", (in Slovak), [online], 2014, Available at: <http://spreadsheets.google.com/viewform?formkey=dGk4el9nNHJPTU5iZWpBLTZRTVdYMGc6MA>.

[9] J. Tvardzík, "Systém daniarom stále hlási chyby", (in Slovak), [online], sme.sk, Bratislava, 2012, Avalilable at: <http://ekonomika.sme.sk/c/6261509/system-daniarom-stale-hlasi-chyby.html>.

A Course in Digital System Design using Unified E2LP Platform

Ivan Kastelan, Nebojsa Pjevalica, Miodrag Temerinac
University of Novi Sad, Faculty of Technical Sciences
Novi Sad, Serbia
ivan.kastelan@rt-rk.uns.ac.rs

Abstract - Computer engineering education has a strong dependence on laboratory work. Teaching students in the lab does not always consist of teaching the required course material, but also of a vast number of tutorials on different platforms. In many cases, these tutorials take significant amount of time which could be better spent doing actual course material. For that reason, E2LP project has developed a platform to be used in multiple courses, covering majority of the computer engineering curriculum. This paper presents a first step in the curriculum - a course in digital system design with laboratory exercises which are implemented on E2LP platform. The platform not only gives students hands-on experience with design and verification of digital systems, but it also provides a bridge between hardware and software because using the same platform in future courses allows students to re-use their results from previous courses and continue building their knowledge. First impression of students is positive about the experience with the platform and their feeling of usefulness of the course for their career.

I. INTRODUCTION

Embedded systems are invisible electronics and corresponding software that bring intelligence to objects, processes and devices. The main challenge in engineering education for embedded systems is a complex interdisciplinary approach which includes understanding of various systems based on different technologies and system solution optimizations.

The significance of laboratory work in electrical engineering education has been widely recognized. It provides engineering education personnel to transform passive listener's students into active learners, thus stimulating students to actively participate in the learning process [1]. Moreover, knowledge obtained through laboratory hands-on experience has proven to be more profound and more lasting.

Increased role of computer based embedded systems in various industrial applications has produced a growing need for embedded system engineers. As of today, the job market is very competitive for highly qualified electronics systems engineers [2]. Furthermore, the European Centre for the Development of Vocational Training (Cedefop) in its publication "Skills supply and demand in Europe, Medium-term forecast up to 2020" [3] predicts that increase of 10.7% in new engineering positions across all disciplines in Europe will be reached up to 2020, compared to the 2010 level. Similar projections are available for USA by the Bureau of Labor Statistics, U.S. Department of Labor in "Occupational Outlook Handbook, 2010-11 Edition" [4] and Asia given in "Employment Outlook Asia: Focus on China and India" written by Mary Anne Thompson [5]. Consequently, many technical faculties have put more emphasis on embedded systems learning by introducing a number of active learning laboratory-based courses [6].

A typical approach in the field of computer engineering education is to have a group of courses for a specific computer engineering topic. Usually, these courses comprise a set of laboratory assignments which are performed using different hardware platforms for each particular course, thus introducing high cost requirements for learning equipment. Further, the dynamics necessary to address industry needs have put teaching personnel under significant pressure to adequately design the laboratory environment for the courses, since updates of the learning environment in an embedded systems course are necessary on a regular basis to keep it relevant. Having this in mind, it is not hard to imagine technical faculties ending up with a set of different and often inconsistent laboratory setups and teaching platforms used throughout the curriculum. As a result, efficiency of laboratory work in embedded systems learning usually suffers from introduced overhead (30%) in both the time and the effort necessary to get students familiar with hardware platforms and software tools for each course. An on-going E2LP project [7]-[11] started with the goal of reducing this overhead in engineering education and providing a unified learning platform which will cover the most of the computer engineering curriculum, allowing students to concentrate on important topics in each course, rather than unnecessary tutorials on using different platforms.

The main body of the E2LP platform is the base board with FPGA connected to a wide range of interfaces – audio, video, communication, memory and user I/O. Additionally the base board can be connected to one extension board via the standardized Mezzanine interface. This paper will focus on presenting the base board of the E2LP platform and how it was applied in teaching the introductory course in digital system design at the Faculty of Technical Sciences, University of Novi Sad, during the fall 2014 semester. This course is the first in which the platform was introduced to the students and it is expected

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 317882.

to be the first in a sequence of courses using the E2LP platform in the computer engineering curriculum. Continuation of the platform usage in future courses will also allow students to feel the connection of materials learned during their studies, in a way that they will be able to reuse or extend results from the previous projects in future courses.

The rest of the paper is organized as follows: section 2 provides an overview of the E2LP platform. Section 3 discusses the theoretical instructions in the course. Section 4 gives an overview of laboratory exercises which use the E2LP platform. Section 5 summarizes the results from evaluation of the course. Finally, section 6 gives some concluding remarks.

II. E2LP Platform Overview

In order to overcome the large amount of lost time in tutorials of different platforms given to students, E2LP project was established to implement a unified platform which will be used in the complete curriculum in computer engineering studies. The Digital System Design course, which is the topic of this paper, presents the first course in the curriculum which uses the platform.

E2LP Platform consists of:

- Base Board, with Xilinx Spartan-6 FPGA,
- Extension boards based on microprocessors.

Digital System Design course uses only the Base Board, as it needs only FPGA for implementation of digital systems students design on labs.

The key building modules of the E2LP Base Board (Fig. 1, 2) are:

- Xilinx Spartan-6 FPGA,
- ARM-based control processor,
- Mezzanine connector to extension board (Xilinx FMC LPC standard),
- DDR2, flash and multimedia card memory,
- user interface (8 switches, 6 buttons, 8 LEDs, alphanumeric LCD screen),
- snapwire connector (8 pins),
- CVBS video encoder and decoder,
- video output (VGA, HDMI),
- audio sub-system,
- communication interfaces (USB, Ethernet, RS232 and Infra-red).

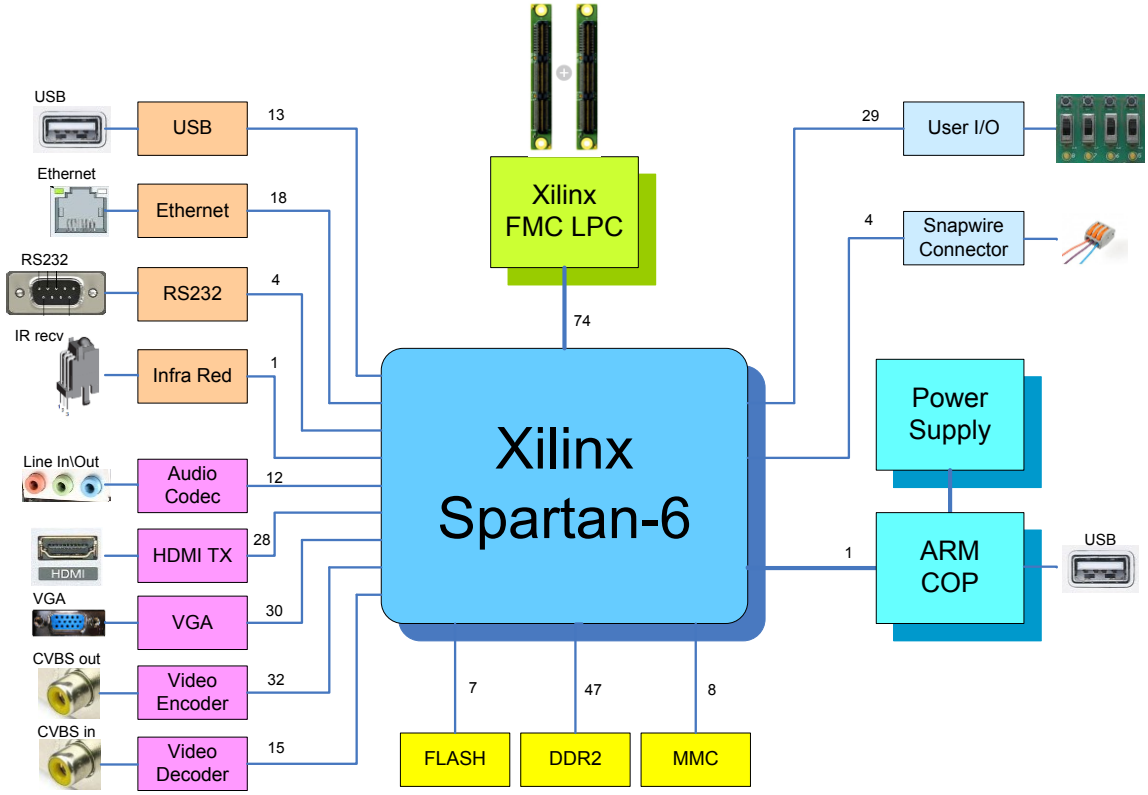


Figure 1. E2LP Base Board schematic overview

Figure 2. E2LP Base Board top view (up) and side view with some of the multimedia and communication interfaces connected to FPGA (down)

III. THEORETICAL LECTURES

Theoretical part of the course consists of introducing students to the field of digital system design, first by short introduction to digital electronics, followed by introducing the concepts of combinational circuits, sequential circuits and finite state machines. The course, since it is being offered in a computer engineering curriculum, has a goal of teaching students principles of processor design. After going through introductory topics, lectures proceed in the direction of processor design, first teaching fundamental concepts – processor architecture, memory design and data flow. Final step is teaching about the complete RISC processor design example and touching the topics of pipelining and timings inside a processor. Table 1 gives an overview of lectures each week of the course.

Since this course is the only course in digital systems for many students who are taking it, because they continue their undergraduate studies in software-oriented direction, in previous years we tried to give them a feel of as many topics in digital systems as it was possible within one semester. Some of the topics taught were: programmable logic, PAL (Programmable Array Logic), PLA (Programmable Logic Array), architectures of FPGA's, DDR (Double Data Rate) memory designs. All these topics were mentioned in at most one session per topic or several topics. The outcome of this approach, which we noticed after a few years of using it, was that students did not learn much about these topics, but had less time to spend on important topics which were the core of the course.

TABLE I. LECTURES IN DIGITAL SYSTEM DESIGN COURSE

No	**Lectures in Digital System Design course**
	Lecture Topic (per week)
1	Introduction
2	Digital logic circuits – technologies, properties
3	Boolean algebra, minimization
4	Combinational circuits – decoders, coders, multiplexers, demultiplexers, comparators
5	Arithmetic in digital systems – arithmetic circuits, number presentations, adders, ALU
6	Sequential circuits – registers, counters
7	Finite state machines, minimization
8	Information coding in digital systems
9	Memories, processor architectures
10	Control units
11	Timings in processor design
12	Review

As a trial for change, in the academic year 2014/15 we changed lectures from the previous year using the simplistic approach to teaching. We concentrated only on core topics of the course and each of them received more time in the schedule. The results of this approach, as results in section V show, are that more students learned the core topics better and their overall passing percentage in the course increased. Future generations will tell whether this was just a peak generation, but we are confident to continue using this simplistic approach in lectures.

Additionally, students were encouraged to study lecture materials by giving them optional homework tasks which required advanced understanding of material presented in the lectures. Significant group of students worked on these tasks as they were motivated by earning bonus points. The outcome was better understanding as reflected in final scores of students, presented in Section V.

Lectures were appended with auditory exercises which were held in 90 minute sessions every two weeks. There students had an opportunity to practice problems with the help of a teaching assistant. These sessions were also modified to use simplistic approach – concentrating on core material and not spending too much time on advanced and general digital systems topics.

TABLE II. AUDITORY EXERCISES IN THE COURSE

No	**Auditory Exercises in Digital System Design course**
	Exercise Topic (per session)
1	Digital logic circuits
2	Combinational circuits
3	Sequential circuits
4	Finite state machines
5	Computation structures - processors

IV. LABORATORY EXERCISES

Laboratory exercises concentrated on developing students experience in designing digital systems from scratch, simulating them and synthesizing for a real platform. For most of the students, this was the first touch with a real platform and making their system work on it produced large amount of satisfaction which further fostered learning experience.

Laboratory exercises for E2LP platform were developed from an earlier set of exercises [9] which was used on another platform. This academic year we introduced simplistic approach in the lab too, as we increased the number of tutorials and reduced the number of graded exercises. Overall, the laboratory exercise structure is given in table 3.

More detailed explanation of laboratory exercises can be found in our earlier work [9]. The changes we introduced this year, in addition to increasing the amount of tutorials, were:

- Processor design exercise was reduced from three to one, with an optional final project involving what the previous exercises 6 and 7 were,
- Each exercise was divided into two parts – homework part and problem part which was solved during the lab,

Approach to designing tasks for students in the laboratory exercise can be done in two ways:

- Students could extend a previously written design by a professor or a teaching assistant, or
- Students could design their system from scratch.

TABLE III. LABORATORY EXERCISES IN THE COURSE

No	**Laboratory Exercises in Digital System Design course**
	Exercise Topic (per session)
1	Tutorial – Digital logic circuits
2	Tutorial – Combinational circuits
3	Problem 1 – Multiplexers, adders
4	Tutorial – Sequential circuits
5	Problem 2 – Stopwatch
6	Tutorial – Finite state machines
7	Problem 3 – Car lights FSM
8	Tutorial – Complex digital systems
9	Problem 4 – Printing name on LCD
10	Tutorial – Computation structures
11	Problem 5 – Basic processor design

While the first method is good when students are learning something, our opinion is that this method leads to a significant overhead in the amount of required understanding of someone else's work in order to perform a small addition. Therefore, in each of the laboratory exercises, students were asked to design the digital system from scratch. While these systems could not be very complex, due to the limited time students have during the week, they were individual solutions and at the end of each lab, students had a finished system which was working on a real platform in real time.

E2LP platform based on a Xilinx FPGA was chosen as opposed to already available solutions because it is a custom-made solution which can easily be modified in-house to meet the needs of increasingly developing modern ways of education. Additionally, this course will teach students how to work with the platform and make their future courses on the same platform easier and faster to start and really concentrate on important topics in each of the courses. Additionally, results from this course, e.g. a designed processor, can be used as starting points in the labs in future courses. One example is the Computer System Design course where students can continue building their processor with more advanced instructions, introduce pipeline, etc.

V. RESULTS FROM THE FALL 2014 SEMESTER

The students' final grade was determined based on points earned in the following aspects of the course:

- Participation in the lectures
- Laboratory exercises
- Midterm exams
- Final exam
- Optional bonus points

Table 4 summarizes the distribution of points throughout the individual course assignments.

TABLE IV. GRADE DISTRIBUTION

%	**Grading in Digital System Design course**
	Assignment
5	Participation in lectures
20	Laboratory exercises
15	Midterm exam 1
20	Midterm exam 2
40	Final exam
0+	Optional bonus points

In previous years, students had to pass only laboratory exercises during the semester, while all exams could be repeated whole year in exam periods. Compared to earlier years, this year we introduced a final exam and made midterm exams passable only during the semester. This made rules for passing even stricter than previous years, but it also added a motivation for students to study more during the semester, which in turn led to better results as it will be shown later.

Evaluation of the course in the fall semester in academic year 2014/15 was performed in the following ways:

- Comparing students grades with earlier generations,
- At the end of each graded laboratory exercises, students filled in the Lab Feedback Questionnaire (LFQ),
- At the end of the course, students filled two questionnaires: Motivated Strategies for Learning Questionnaire (MSLQ) and the Computer System Usage Questionnaire (CSUQ).

While the analysis of MSLQ and CSUQ results is not yet available at the time of writing this paper, this section will summarize students' responses in LFQ's and their grades, as well as some comments student left in final questionnaires.

Comparing the grades with earlier generations (Table 5), we can notice a significant improvement in the percentage of students who passed the course which shows that our simplistic approach improved students' knowledge and understanding of core material. Point average improved from previous year, but did not improve significantly from 2012/13 which is expected, since the grade distribution should contain all grades in a normal distribution to be fair. This shows that our grade distribution remained fair, but more students successfully finished the course which is the main goal of education after-all. It should be noted that midterm exams and final exam were of the same difficulty as in previous years and that the rules for passing were even stricter than previous years due to the introduction of the final exam. Point averages are based on 100 points total and the whole table is based on results after the first exam period in the end of January.

TABLE V. STUDENTS' SUCCESS IN THE COURSE

	Grades over 3 generations of students		
	2014/15	2013/14	2012/13
Total students	235	229	206
Passed students	158	96	115
% Pass	67.2	41.9	55.8
Points average	63.1	52.7	61.1

Students were asked at the end of each graded exercise to fill the Lab Feedback Questionnaire (LFQ) and mark the effectiveness of the lab experiment on the scale from 1 to 10 in each of the following categories:

1. Clarity of theoretical explanations
2. Clarity of technical instructions, exercises and problems
3. Total time and efforts required
4. Ease of use of the environment – software issues
5. Ease of use of the platform
6. Assistant support – was the lab assistant helpful
7. To what extent you think you learned something valuable
8. Overall satisfaction

Fig. 3 shows their responses for the overall satisfaction. The total of 235 students attended the course and a part of them filled up to 5 LFQ's. It can be seen that the lab exercises in the course were well perceived by students in general.

Below we summarize some of the comments of students on how the course can be further improved in the future (taken from final MSLQ):

- "Better alignment of the lab assistants regarding criteria for student grading."
  (6/53 of students posted comment of that type.)
  It turned out that dispersion of the student grades and as more critical dispersion of the requirements per each grade depends on the assistant personally. Such outcome is a consequence of relatively high number of lab assistants - 10 and also a relatively high number of grades – up to 4 points per lab, plus an optional bonus point. It could be improved at least by decreasing number of grades to 3 in total for the next generation.
- "Number of lab classes should be increased."
  (5/53 of students posted comment of that type.)
  Comment placed generally, but points mostly to one lab exercise, problem 4 (LCD) with request for additional time and preparation needed for accomplishment of this, more complex lab.
- "Compliments for professor and assistants..."
  (3/53 of students posted comment of that type)
- "No complains."
  (2/53 of students posted comment of that type)
- "Tutorial lectures could be better."
  (1/53 of students posted comment of that type)
- "Literature should be better"
  (1/53 of students posted comment of that type)
- Empty comment space
  (30/53 of students left empty comment field in the survey)

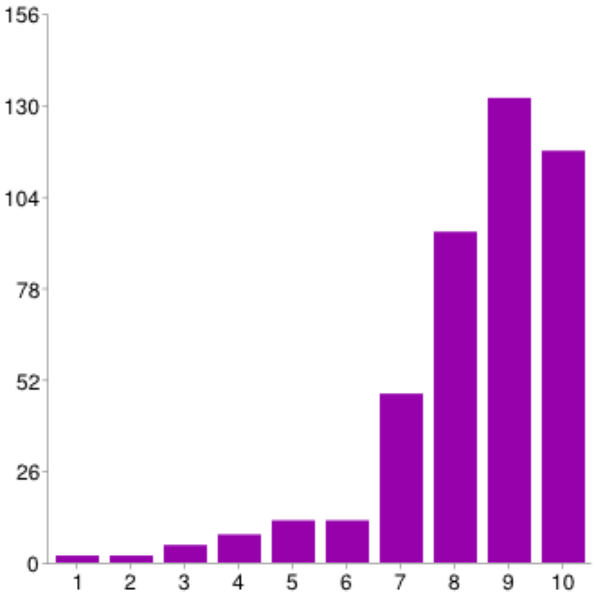


Figure 3. Overall satisfaction with the course labs

VI. CONCLUSIONS

Feedback from students showed that the changes we made in the course were well accepted. Course grades support this as the students showed better knowledge of the core material which motivates us to continue using the simplistic approach to teaching and refrain from teaching unnecessary broad material in the course.

In the future, platform will be used by the same generation of students in the courses in computer system design, signal processing, computer networks, system software and application programming courses when the education benefit of using the same platform in multiple courses will be better known. As for the Digital System Course, it will be improved by responding to most frequent student comments, e.g. by balancing the grading criteria between lab assistants. We will continue to use the simplistic approach as it proved to be a good practice.

REFERENCES

[1] K.E. Holbert, G.G. Karady: "Strategies, Challenges and Prospects for Active Learning in the Computer-Based Classroom", IEEE Transactions on Education, Volume 52, Issue 1, Feb. 2009 pp. 31 – 38

[2] R. Schneiderman: "Economy and Shortages Affect the European Job Outlook", IEEE Spectrum, March 2010, Available from (accessed on 2011-12-10) http://spectrum.ieee.org/at-work/tech-careers/economy-and-shortages-affect-the-european-job-outlook/0

[3] The European Centre for the Development of Vocational Training (Cedefop): "Skills supply and demand in Europe, Medium-term forecast up to 2020", Publications Office of the European Union, 2010, ISBN 978-92-896-0536-6, Available from (accessed on 2011-12-10) http://www.cedefop.europa.eu/en/Files/3052_en.pdf

[4] Bureau of Labor Statistics, U.S. Department of Labor, Occupational Outlook Handbook, 2010-11 Edition, Available from (accessed on 2011-12-10) http://www.bls.gov/oco/

[5] M.A. Thompson: "Employment Outlook Asia: Focus on China and India", 2011-04-20, Available from (accessed on 2011-12-10) http://www.goinglobal.com/articles/864/

[6] S. Nooshabadi, J. Garside: "Modernization of teaching in embedded systems design-an international collaborative project", IEEE Transactions on Education, Volume 49, Issue 2, May 2006 pp. 254-262

[7] I. Kastelan, M. Temerinac, N. Teslic: "E2LP: A Unified Embedded Engineering Learning Platform", IT System Conference, 2013, pp. 1-4

[8] M. Zagar, N. Frid, J. Knezovic, D. Hofman, M. Kovac, V. Sruk, H. Mlinaric: "Unified, Multiple Target, Computer Engineering Learning Platform", IEEE Global Engineering Education Conference EDUCON, 2014, pp. 926-929

[9] I. Kastelan, D. Majstorovic, M. Nikolic, J. Eremic, M. Katona: "Laboratory Exercises for Embedded Engineering Learning Platform", *Proceedings of the 35th International Convention MIPRO*, 2012, pp. 1113-1117

[10] I. Kastelan, M. Temerinac, N. Teslic: "Design of a Platform for Teaching Embedded Systems Engineering", *Proceedings of the E2LP Workshop at FedCSIS*, 2014, pp. 21-24

[11] M. Nikolic, D. Majstorovic, I. Kastelan, N. Pjevalica: "An Approach to Improvement of Students Learning of Computer System Engineering", *Proceedings of the E2LP Workshop at FedCSIS*, 2014, pp. 17-20

Improving education experience with Augmented Reality (AR)

Bojan Kraut and Jelena Jeknić
Alcyone d. o. o., Maribor, Slovenia
bojan.kraut@alcyone.si
jelena.jeknic@alcyone.si

In the modern world, information is everything. It is important that it can be accessed from anywhere and at any time. What is even more important is that information must be relevant to the user and presented in such a manner that it is easily understood.

Emerging technologies, such as augmented reality (AR) might within short period of time be widely accepted, as smart-phones are today. If this happens, it will probably change the way we perceive information and our reality. AR is not hype any more – it is a solid technology that is already used in some creative applications.

One area which might significantly benefit in the future from this technology is the education process. AR tools could guide students through learning process in enhanced way, as AR can upgrade traditional books with a digital layer. We think it will improve both, teaching and learning experience, and bring interactive dimension into the whole picture. We also predict that this new layer will encompass several senses which could speed up memorization process. Furthermore, AR learning method might raise common understanding of the learning material. Moreover, learning activities could be supervised by a mentor or automated process which might also lead to a lower school dropout rate.

Our still-in-progress ARAVET project is exploring mentioned predictions about AR learning method. We will present the results we have so far.

I. INTRODUCTION

Nowadays, information is a key to success. For decision makers it is important they have instant information access from anywhere.

On the other hand, with today's flood of information, decision making process can sometimes be overloaded. Computerized services offer centralized information management, user access control, information categorization and indexation, as well as full text search, but retrieved data are not always relevant to the query, which negatively influence decision making process. Without proper personalized presentation, content filtering and classification, decision makers are often left with excessive data that can impose a cognitive barrier for them.

As mentioned, it is important in which form the data are presented to the decision makers. Sophisticated technology, like augmented reality (AR), offers insight into data on a different level, as it combines real world with the digital one, giving users enhanced information presentation.

AR needs rich user collaborative interfaces, content filtering and real time personalization, and all these functions demand huge computing power. Rising use of cloud computing is addressing some of these issues. Therefore, in this paper, we offer a quick overview of current computing technology improvements and cloud systems that are significantly enhancing information processing. Moreover, we will glance over evolving adaptive systems, which can accommodate individual user experience and optimize learning process. As learning depends on information perception, we will also try to explain how information flow is processed inside the human mind and how do people perceive reality.

At the end, we will show our ARAVET results, which indicate that AR can be very useful in learning process as it can improve perception of received information.

II. POWER OF PROCESSING

Most computers still follow Von Neumann's computer architecture, where instructions are performed sequentially, but really fast. However, recent research in quantum and parallel computing have introduced next generation machines which process information differently and much faster than contemporary computers. Similarly to our brain, new computers are processing instructions concurrently. This kind of processing is not limited to a single processor, as instructions execution often includes multi-core systems or even more than one computer in form of computer clusters.

These days, we already have computer systems built from computers connected with each-other, forming global processing systems with thousands and thousands of cores that can process information extremely fast, which is needed for improved real-time user experience and development of Internet of things (IoT). Today, almost any electronic device can be connected to the web and it can send its sensory data to the cloud services. Moreover, these devices could also offer additional processing power, and therefore, enhance the overall performance of the system.

Moore's law [1] predicts that computer processing power will be doubled every 18 months. Reference [3]

predicts that by 2019 a 4000$ computing device could have computational capability of a human brain (20 quadrillion calculations per second). According to reference [3], by 2029 a 1000$ personal computer might be 1000 times more powerful than human brain. So, artificial intelligence might be recognized as conscious entity. The same reference also predicts that, it will be possible to use implants directly connected to the human nerve system or brain to enhance information processing. Experimental testing already proved that with use of a computer, information from human brain can be read and written directly into the brain [2]. There are also predictions saying that by the end of 2099 human brain could be completely reverse engineered and all aspects of brain functioning might be finally understood [3].

Namely, computing power could soon offer unimaginable possibilities resulting in self-adaptive global systems that could connect millions of people into collaborative environments, where user behavior could be analyzed in real-time. New technologies might offer creation of user tailored education process, where students could follow personalized curriculum adapted to their own learning skills. So, the whole education experience could be completely personalized and provided in an optimal way.

III. LEARNING

A. Nervous system and human brain

Learning is an elementary process. It starts in the womb and continues during our whole life. Senses provide us with information about our environment and nervous system transfers the electric impulses to our brain with speed up to 100 meters per second [4], where information is processed. Information processing is a base for human cognition, creating logical abstracts, reasoning, information classification and categorization.

Average human brain consists of 86 billion neurons and uses typically 100 W of power, which represents merely 20 % of the whole body energy consumption [5]. On the other hand, gorilla's brain has about 28 billion neurons, or around 33 % capacity of human brain [6]. Each neuron is capable of making around 1000 connections, representing roughly 1000 potential synapses, which mostly do the work of data storage. If we multiply each of these 100 billion neurons by approximately 1000 connections each neuron can make, we get 100 trillion data points, or about 100 terabytes of information [5].

Until recently, we believed that animals were incapable of processing information in the way we do it. However, tests on chimpanzees reveal, that they can be significantly better at certain tasks, especially those connected with motor skills and rapid movement processing [6]. Their brains are wired differently then human's, therefore, they are specialized for other purposes. Still, their overall intelligence is on the level of a three-year-old child. An experiment performed on a group of chimpanzees and human toddlers revealed, that humans, unlike other species, are capable of connecting with each other to obtain the information they need in order to solve a problem. This was something chimpanzees were not able to do. Researchers suggested that humans are born to share their thoughts [6].

Social collaboration is crucial for humans, so that they can better understand the world around them. Digital age raised this interaction to an unprecedented level. Today, almost everyone is connected to the Internet and everybody is exchanging all sorts of information. Information exchange has never been so easy and so momentous. On the other hand, modern communication often lacks emotional components.

B. Types of learning

To better understand the learning process, we are presenting various human personality types that significantly influence learning process [7].

In general, we can differentiate 7 types of learning, tightly coupled with types of human personality [7]:

- Visual (spatial) learners use pictures, images and spatial understanding, with as many colors and other visual media as possible. Good learning tools for visual learning are: spatial organization layouts with important words, mind maps, words replaced with images and extensive use of colors.
- Aural (auditory-musical) learners prefer the use of sound, rhythm and music to enhance their learning process. Very often, background recordings are used to help achieve special effects of visualization. Aural learners also use mnemonics (systems for improving and assisting the memory) and with the help of rhyme and melody they can better memorize the content. Music influences and enhances our emotions, and therefore learning.
- Verbal (linguistic) learners prefer the use of words, both in speech and writing. So, techniques involving speaking and writing such as word-based techniques of assertions and scripting help a lot this kind of learners. Reading content aloud with dramatic and varied intonation and/or role playing with others also enhances linguistic type of learning.
- Physical (kinesthetic) learners use body, hands and sense of touch. The use of physical objects helps this type of learning, too.
- Solitary (intrapersonal) type of learning is best when working alone and using self-study techniques of learning. Alignment of personal beliefs and values with goals and objectives is also helpful. Building associations and visualizations of feelings and modeling are very important, as all these techniques use the

influence of thoughts. Being creative with role playing enhances solitary type of learning.

- Social (interpersonal) learning promotes studying and learning in groups. This type of leaning is activated when a person works with others as much as possible. Role playing is very important in social learning, as it promotes interactions with others. During social learning it is recommended to extensively share thoughts with others through assertion, visualization, etc. Working in group to practice behaviors or procedures helps understand problem variations.
- Logical (mathematical) learning is based on understanding the reason behind something. This type of learning is best when extracting key points from material, making associations, especially when the content is illogical or irrational. The use of method which makes the object of learning fit into wider content, introducing skills, as well as thinking systems helps out logical learners a lot.

There might be even more types of learning that scientist haven't identified yet. Our senses and human personality significantly influence our learning process. As everyone has a different perception of the world around them, it is of a huge importance that each individual receives the information according to the nature of their own specific mind.

AR can play significant role here, as it can address different senses and personalize user experience. Therefore, it can boost performance of individuals and additionally stimulate their interests. With help of interactive technology, individual user interaction can be analyzed and then optimized to fit the needs of students.

C. Processing the information

Research done at Massachusetts institute of technology (MIT) [8] discovered that memories are stored in individual brain cells and when scientist stimulated specific brain cells, certain memory was recalled. Removing those cells resulted in loss of specific memory.

As previously mentioned, our brain's information processing is influenced by the type of person's nature. Almost all people have one dominant sense. For most of them it is vision. This means that people process visual information much better that information perceived by other senses [9]. The reason for this lies in our brain specialization.

Interestingly, smell is unusually effective at evoking memory. When smell is used to evoke a certain memory, person can remember 10-50 % more than without smell stimulation [9]. Marketers have recognized this as a very important marketing trigger, so they include it in their marketing activities to boost sales.

People who function in multi-sensory environments always perform better than those in uni-sensory environments. Their memory recall is better, and their memory lasts longer, even after 20 years from the moment of memory creation [9].

Having that in mind, students could use multi-sensory stimulation techniques to develop multi-sensory study skills and improve their memorization and thus the learning process. AR could help achieve that as it offers a multi-sensory interactive knowledge transfer environment.

According to new research, scientists discovered they can rewire our brain with special training. One type of such training can lead to partial synesthesia, where brain gets information in a multi-sensory way [9].

Synesthesia [9] is a neurological phenomenon in which stimulation of one sensory or cognitive pathway leads to automatic, involuntary experiences in a second sensory or cognitive pathway. People who report such experiences are known as synesthetes. So, simply put, synesthesia is a condition in which one type of stimulation evokes the sensation of another, as when the hearing of a sound produces the visualization of a color.

In the most common form of synesthesia, also known as grapheme, color synesthesia or color-graphemic synesthesia, letters or numbers are perceived as inherently colored. In spatial-sequence, or number synesthesia, numbers, months of the year, and/or days of the week elicit precise locations in space (e.g. 1980 may be "farther away" than 1990), or may appear as a three-dimensional map (clockwise or counterclockwise).

Only certain types of synesthesia have been evaluated by scientific research. It is very difficult to grasp this phenomenon and analyze it, as awareness and manifestation of synesthetic perception varies from person to person.

Synesthetes can remember significantly more information in comparison to average people. Special sets of exercises can also help our brain develop synesthetic functions. However, our brain would need constant practice to retain these functions, as synesthesia is not a natural condition for the majority of people. Psychological research has demonstrated that synesthetic experiences can have measurable behavioral consequences. Functional neuroimaging studies have identified differences in patterns of brain activation. Many studies [9] found out that synesthesia improves creative processes. Psychologists and neuroscientists, who studied synesthesia, have been infatuated by not only its inherent appeal, but also, by the insights it may give into cognitive and perceptual brain processes that differ between synesthetes and non-synesthetes.

AR can stimulate several senses at the same time, leading to improved learning skills (with practice – brain can be rewired to optimize performance) and memorization. With improved AR interfaces additional senses (five senses – sight, sound, taste, smell and touch) could improve cognition of information, and therefore learning process.

IV. AUGMENTED REALITY

A. Definition

Augmented reality (AR) is a live, direct or indirect view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated sensory input such as sound, video, graphics or GPS data. It is a variation of virtual environment (VE), or virtual reality, as it is more commonly called. VE technologies completely immerse a user inside a synthetic environment [10]. While immersed, the user cannot see the real world around him. In contrast, AR allows the user to see the real world, with virtual objects super-imposed upon or composited with the real world. Therefore, AR supplements reality, rather than completely replacing it. AR is related to a more general concept called mediated reality, in which a view of reality is modified by a computer [10]. As a result, the technology enhances one's current perception of reality. Virtual reality replaces the real world with a simulated one. Augmentation of visual information is presented in semantic context, coherently with environmental elements and in real-time, similarly to synchronized sports scores on TV during a match.

With the help of advanced AR technology (e.g. adding computer vision and object recognition), the information about surrounding real world of a user becomes interactive and digitally manipulated, as artificial information about the environment and its objects can be overlaid on the real world.

Nowadays, AR is often related to mobile applications, but the concept is much wider and implications are almost limitless. Some new cars cockpits are already equipped with futuristic dashboards where information is projected directly onto the windscreen, giving drivers additional view on their surroundings [11]. Furthermore, Special Forces and astronauts use hi-tech helmets, where information is projected onto their visors [11].

In a few years, we might have special contact lenses that will display all kinds of information directly to our eyes [11]. These lenses could be connected to the Internet through our mobile devices. If such technology is to be developed, it would bring unimaginable new possibilities to our everyday life. People would be in constant interaction over several interfaces with everyone and everything. The reality as we know it could blend completely with the digital world.

B. Augmented reality in learning process

In the past, information (thus knowledge), was collected in form of books and, before printing was invented, the books were hand copied by erudite people. In today's world, the information is changing and emerging so fast that information on the print materials (paper books) is often outdated the moment it is printed out. Digital information can be easily edited, and therefore constantly validated. It can also be interactive and users can always retrieve up-to-date content, use bookmarks, 3D models or social networks, where they can collaborate with other users and exchange ideas and questions about educational material.

AR brings many advantages for teachers, as well, as it helps analyze the learning process of students and identify the elements that could be improved.

The final goal is to bring highly personalized learning content to the students which will help them learn faster, easier and better. As AR offers human and content interaction, students could stay more active during their learning process. They could be offered the opportunity to improve the learning content and even become content creators by giving their positive or negative remarks. Additionally, AR presentation layers could offer better understanding and access to social knowledge. Moreover, the multi-sensory presentation might significantly improve the memorization of the content.

Development of sophisticated user interfaces could be very beneficiary for learning process, as digital content can be endless and it can be downloaded on demand in real time from the servers and presented to the students any way they want. With personalization and interactivity, automatic algorithms could monitor the progress of individual students and then tailor curriculums according to their progress, skills and understanding.

As active contributors, students could leave notes and enrich printed content. Notes could be synchronized over cloud services and all active users could retrieve the change instantly. Teachers could improve the content, as well, according to remarks, links, images and data schemes, offering their students better understanding of the learning material.

Contemporary printed materials offer solid learning content which can be read by students at anytime. However, additional digital layer could offer performance of tests and experiments related to the content. In this way, students could better understand the processes described in the learning material, as they could see different visual and audible presentations and simulations of the explained process.

For an example, diesel engine in printed material is usually showed as a scheme of engine's internal structure, however a digital layer on a tablet can present interactive model of the engine, so students could (re)move the engine body parts and observe the simulation of the ignition cycle, etc. With interactive model, students could also change initial parameters, resulting in different model behavior.

The possibilities of such learning material presentations could be of immense help in medicine, engineering, rehabilitation, pharmacy, IT, etc.

In the long term, AR might improve users' sensor experience and provide them with the way to expand their senses to electromagnetic field, infrared and ultraviolet vision, ultrasonic and subsonic hearing, stereo phonic vision, etc.

Additionally, implementation of direct memory manipulation could significantly shorten learning process, so, students could focus on better processing and connecting the information and not just on memorizing it.

The application of this futuristic technology might provide solutions for controlling phantom limbs, robots or other remote objects [12].

V. ARAVET PROJECT

Our ARAVET project (Augmented Reality Application on the Field of Vocational Education and Training) started in 2013 as a Leonardo da Vinci project. Organizations from eight different countries are involved in this project. Each organization contributes with its own specific skills and knowledge from areas of technology, education and design. ARAVET project suggests that introduction of new technologies could drastically improve existing learning techniques and users' experience.

Implementation of AR in our pilot project introduced a digital information layer into existing books. So, books were enriched with digital content, connected to special markers. This enriched view was accessible through user smart-phones or tablets. Students could access 3D models and perform pre-animated actions to additionally visualize educational material.

In scope of our project three different fields were chosen for implementation of AR in order to better demonstrate the advantages of its use.

Firstly, we have chosen textile industry and we have implemented AR technology into the learning model of a sewing machine. Students could perform different actions on the sewing machine (from changing the bobbin to inserting a thread into a needle).

Secondly, we have implemented AR to help students learn about electronics. An operation cycle of the diode was presented with AR technology. Students were able to reverse the electric current and see why it stopped working. The interior of the diode was presented with a flow of electrons, so students could see this concept in a way they would never be able to experience if they were just reading about it from the ordinary textbook.

We have also implanted AR technology in the field of informatics. The AR implementation was used in learning materials about logic concepts and electronic gates (AND, OR, XOR and NOR).

All the results of our pilot testing (data were collected through surveys and then statistically analyzed) revealed that when using AR enhanced materials students understood the content much easier and faster than when using traditional learning techniques.

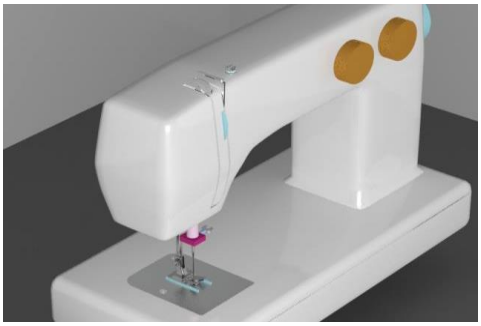


Figure 1: Sewing machine with animated actions

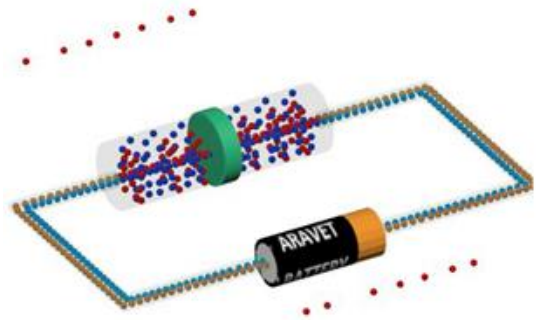


Figure 2: Animated diode behavior and electron flow

71% of the students understood the meaning of Augmented Reality in the way it was explained in the computer lab with examples from the Internet (with the use of mobile devices).

64% of the students found AR enhanced software easy to use and understood the way it functions.

Students were indecisive concerning usage of AR software. They neither agreed nor disagreed that AR software should replace the real device (e.g. sewing machine) during learning process.

62% of the students believed that AR software can help them better understand certain meanings.

59% of the students thought AR software is very interesting for using in the classroom during school lessons.

79% of the students understood very well the way that AR devices work.

55% of the students believed that AR software helped them understand specific meaning faster.

71% of the students were satisfied with the given instructions concerning AR software.

All the results confirm that AR is technology which could really improve understanding of educational material, especially in those fields, where it is important to understand the processes in detail.

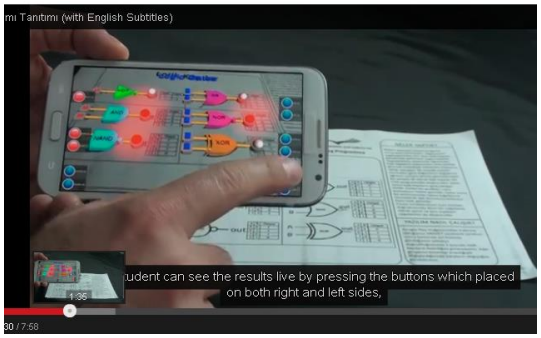


Figure 3: Virtual logical elements for testing the behavior of the printed elements

VI. CONCLUSION

In the future, the boundary between the real world and the digital one might fade away, as modern technologies offer personalized reality. In constant connection with global grid, technology offers instant access to any kind of information, no matter what kind of device user uses.

AR is a perfect example of how users can remove the boundary between the two worlds and equip the real world with digital data. Furthermore, bionic interfaces for human body might additionally improve capabilities of humans beyond anything we know today. Such systems might also provide an in depth insight into user interactions, behavior patterns and learning processes as data could be collected and stored in cloud based services for further analysis.

AR can provide interactive interfaces that enhance a process of learning and content editing in such a way, that current generation of students could learn faster, better and retain knowledge longer than before.

Evolving technology could lead to better understanding of the nature of the human mind and learning processes. This could also help improve machine learning and automated services that support learning processes.

Project ARAVET, with its limited scope, proved that digital data provided through AR interfaces can already supplement static content and improve student understanding of the processes described in the educational material. Moreover, it proved that students have positive attitude towards new technology and its use.

For now, AR is limited due to inadequate hardware. However, even this form can improve user experience. Next generation user interfaces might have the ability to present information in multi-sensory way or even connect directly to our nervous system and offer whole new range of sensors to improve human experience. With evolving next generation user interfaces AR might become commonly accepted as mobile technology is today.

REFERENCES

[1] L. G. Roberts, “Beyond Moore’s law: Internet growth trends”, Computer, Vol. 33 (1), pp. 117-119, Jan 2000.

[2] A. M. Packer, L. E. Russell, H. W. P. Dalgleish and M. Häusser, “Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo.”, Natural Methods, Vol. 12, pp. 140-146, 2015.

[3] R. Kurzweil., “The age of spiritual machines: When computers exceed human intelligence”, Penguin Group, New York, 2000.

[4] D. Kraus, “Concepts in Modern Biology”, Globe Book Company, New York, 1969.

[5] G. Brown, “The Energy of Life”, Free Press, New York, 1999.

[6] C. F. Martin, R. Bhui, P. Bossaerts, T. Matsuzawa and C. Camerer, “Chimpanzee choice rates in competitive games match equilibrium game theory predictions”, Scientific Reports, Vol. 4, Jun 2014.

[7] Blue Mango Learning, “The 7 styles of learning”, available at: http://www.learning-styles-online.com (10. 2. 2015)

[8] X. Liu et al., “Optogenetic stimulation of a hippocampal engram activates fear memory recall”, Nature, Vol. 484, pp. 381-385, Apr 2012.

[9] M. J. Mächler, “Synaesthesia and Learning”, Swiss Federal Institute of Technology, Zurich, 2009.

[10] R. T. Azuma, “Survey of Augmented Reality”, Hughes Research Laboratories, Presence, Vol. 6 (4), pp. 355-385, Aug 1997.

[11] N. Statt, “Augmented-reality contact lenses to be human-ready at CES”, CNET, available at: http://www.cnet.com/news/augmented-reality-contact-lenses-to-be-human-ready-at-ces (1. 2. 2015)

[12] N. Akpan, “How to master your third arm”, Popular Science, available at: www.popsci.com/how-master-your-third-arm (1. 2. 2015)

[13] R. Penrose, “Shadows of the Mind: A search for the missing science of consciusness”, Oxford University Press, New Yoerk, 1994.

Introduction of the automated assessment of homework assignments in a university-level programming course

Marko Poženel, Luka Fürst, and Viljan Mahnič
University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia
Email: {marko.pozenel,luka.fuerst,viljan.mahnic}@fri.uni-lj.si

***Abstract*—Modern teaching paradigms promote active student participation, encouraging teachers to adapt the teaching process to involve more practical work. In the introductory programming course at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, homework assignments contribute approximately one half to the total grade, requiring a significant investment of time and human resources in the assessment process. This problem was alleviated by the automated assessment of homework assignments. In this paper, we introduce an automated assessment system for programming assignments that includes dynamic testing of student programs, plagiarism detection, and a proper presentation of the results. We share our experience and compare the introduced system with the manual assessment approach used before.**

Keywords: automated grading, dynamic testing, programming education

I. Introduction

The ever-growing use of information technology tools in education compels us to reconsider many facets of the education process [1]. Several tasks that once had to be performed manually have been at least partially automated, such as the assessment of the knowledge absorbed by the students. In the area of programming education, the idea of automated assessment of programming assignments, the goal of which is to write a program to perform a given task, has been around for several decades [2]. In 1960, Hollingsworth [3] described probably the first program for grading programming assignments. He immediately recognized the savings in time and a possibility to teach larger group of students. We can conclude that automated assessment (and grading) of computer programs has been present since the very beginning of programming education.

The assessment of student obligations represents an important and often challenging part in the pedagogical process. Focusing on programming assignments, substantial resources are needed to manage their manual assessment, since there are many aspects related to a good program that have to be considered [4], [5]. As a result, the idea of automation or at least partial automation of the assessment process began to appear [6]. The automated assessment process is not suitable for all types of student assignments and cannot be equally effectively used in all areas. Hollingsworth [3] realized that the assignments (problems) have to be properly structured for the automated assessment to be effective. Douce et al. [7] stated that student task specifications should be defined more precisely if they are to be assessed automatically. It follows that automated assessment can be used effectively if the student's assignment is clearly defined and if its result is measurable and unambiguous and can be verified according to a strictly defined procedure [8]. One of the fields that meet all the requirements is programming.

Ala-Mutka [5] summarized the results of an internal survey on assessment practices in computer science courses. Almost three quarters of the respondents assessed the students in their courses on the basis of their practical work. While the assignments were most commonly submitted electronically, they were still predominantly assessed in a manual way. Ala-Mutka concluded that many teachers still see the possibilities of computer aided assessment to be limited to simple tasks such as multiple-choice questions at quizzes.

In theory, automated assessment of programming assignments is not very difficult to achieve. Above all, the assessment system should check the output results of student programs against the predefined correct values. However, by having access to appropriate tools and metrics, the quality and structure of programs can also be checked [9], [4]. Unfortunately, automated assessment of programs combined with hundreds of student submissions can lead to a serious problem — source code plagiarism. However, by incorporating tools for detecting plagiarism, we can also provide a greater validity of the assessment [6].

Depending on the degree of automation, assessment applications are divided into automated and semi-automated ones. Both groups use automated testing of programs. The difference is that a fully automated system tests individual components and evaluates the assignment according to the predefined criteria. It is usually used for smaller assignments. For larger assignments, it is not always possible to assess all aspects of a good programming solution. In that case, a semi-automated system is often used. In such a system, the human evaluator uses the partial results of the automated testing of less complex program parts, focuses on the assessment of demanding tasks, and makes decisions to assign correct grade. The final grade thus still depends on the human evaluator [5], [10].

Several approaches to automated program assessment can be found in literature. Typically, they are all based on two basic principles: testing the program in action, or evaluation

of its code and structure. Program assessment is based on the assumption that certain measurable attributes can be extracted from the program and can then be compared to the model or solution [5]. For an educational use, the results of the comparison also have to be backed up with the pedagogical goals of the course. The assessment of a program includes the assessment of its quality, such as correctness, complexity, reliability, coding standards, etc., and therefore involves practices and techniques that are used in software testing. Software testing can be classified into two groups according to whether the program has to be executed during evaluation or has to be evaluated statically from the program code: dynamic analysis and static analysis.

Dynamic testing techniques can be further divided into two categories: black-box testing and white-box testing. In black-box testing, a program is evaluated on the basis of its functional behavior. White-box testing comprises the analysis of the program code and of the code structure. The aim of the above two techniques is to create test cases that clearly show whether the tested program has errors or not. The test cases should have a good coverage over the program model and specifications in order to avoid misleading feedback information that can further lead to incorrect final grades [4]. When testing the students' programs using dynamic testing, we have to take into account bugs, possible malicious code inserts, and programs that could accidentally cause the collapse of the system [5].

In the introductory programming course at the Faculty of Computer and Information Science, University of Ljubljana (UL-FCIS in the following text), we recently, in 2014/15, adopted automated grading of programming assignments, mostly as a result of the growing awareness of the deficiencies of our previous assessment approach. In this paper, we present the system that we introduced for the purpose of the automatic grading of programming assignments. The system can be broadly classified into the dynamic black-box testing category. However, in one of the assignments, a part of the grade was also obtained using code analysis. In the future, we might increase the proportion of white-box testing techniques in the grading process.

The rest of this paper is structured as follows: In Section II, we briefly present the introductory programming course at UL-FCIS. In Section III, we describe our assessment system before the introduction of automated grading. In Section IV, we present our automated grading system in detail. With Section V, we conclude our paper.

II. The introductory programming course at UL-FCIS

The course 'Basic Programming' at UL-FCIS teaches the fundamentals of procedural and object-oriented programming in Java. The syllabus of the course covers control structures, arrays, classes and objects, inheritance, and fundamentals of computer graphics. The course is taught in the first semester of the first year. It comprises 15 weeks of formal lectures and 13 weeks of lab sessions. In each block of lectures (three consecutive hours per week), the lecturer introduces a programming subject and presents several basic examples to the entire body of approximately 250 students simultaneously. The lab sessions are held in small groups (15–30 students) supervised by one or two teaching assistants, depending on the size of the group [11].

In the Basic Programming course, we strive for constant improvement. In the academic year 2009/10, UL-FCIS adopted the so-called Bologna reform, which calls for a more active, student-centered way of learning with a greater emphasis on regular work. In the spirit of the Bologna reform, we introduced a system for the cooperative development of computer programs [11] to the lab sessions. In our programming course, we are also faced with a very diverse student population with varied prior knowledge. In order to animate the course and motivate students with programming skills, we introduced board game programming competitions [12]. Our last improvement was the introduction of the automated grading of homework assignments.

III. The grading system before introducing automated assessment

Throughout the history of the Basic Programming course, several grading systems have been in use. In this section, we describe the grading system between the academic years 2009/10 (i.e., immediately after the adoption of the Bologna reform) and 2013/14 (before the introduction of automated assessment), inclusive. The Bologna reform prompted us to redesign the grading system, since it recommends that approximately 50% of the student's final grade be obtained through his or her regular work during the semester, e.g., by homework assignments. Before that, the grade was determined almost exclusively by the final written exam. The homework assignments were part of the student's obligations, but they contributed at most a few bonus points to his or her grade.

During the years 2009/10–2013/14, the students were given three programming assignments per semester. For each assignment, the students had one week to finish it at home and submit it. At the next week's lab sessions, they had to implement three independent upgrades to the homework assignment. The teaching assistants then assessed the submitted homework assignments together with the upgrades. Both were assessed during the lab sessions right after the deadline for finishing the upgrades. A greater attention was devoted to the upgrades (3 upgrades $\times$ 2 points = 6 points out of 10) than to the homework assignment (4 points out of 10).

The grading system described above did have some advantages. Above all, plagiarism did not make much sense. A student who did not implement the basic assignment and did not understand the code in detail had little or no chance to implement the upgrades. Besides that, the student's work was assessed in a face-to-face manner right after the submission deadline, thus giving the student an immediate, albeit a severely time-limited feedback. (On average, we had only two minutes per student.)

Unfortunately, our old assessment system also suffered from a number of disadvantages. The greatest problem was that a lot of effort was required to prepare comprehensive and easily upgradeable homework assignments and diverse upgrades. The teaching staff for the programming course strove to prepare a set of mutually distinct homework assignments for each of the three grading weeks in a semester and for each lab session group so that each group of students could have a unique pair of a homework assignment and a set of three upgrades. The students were typically divided into 9 groups, which means that, in principle, as many as 9 distinct assignments and 27 distinct upgrades had to be prepared for *each* of the three grading weeks during the course. In practice, the number of distinct homework assignments was somewhat lower (usually, we prepared 5–6 distinct assignments per grading week), but then the number of distinct upgrade sets per assignment had to be even larger. The uniqueness of the upgrade sets was a means to to fight against academic dishonesty [13], [14], but the effort required to ensure it was tremendous. Besides that, it was often difficult to prepare a balanced set of assignments and upgrades for different groups, and the students might have had a different view of what 'balanced' is than the teaching staff.

Since the assistants could, on average, devote only two minutes to each student, the grading procedure was highly superficial: run the program once or perhaps twice, take a quick look at the program code, exchange a few words with the student, and decide on the grade within the scale of 0–10. Such a grading approach could hardly be considered optimal. Not only that it assesses the student's work on the basis of shallow, first-impression 'metrics', but it also features some typical human-factor problems such as the halo effect and disagreements between different assessors (teaching assistants) about the relative importance of individual aspects of the student's work.

IV. Introduction of automated assessment

A. Objectives

After a few years of experience, the disadvantages of our grading system have made themselves increasingly felt. Besides that, we started to participate in the organization of national programming competitions, where automated assessment of programming assignments had already been in effect. Both factors served as an incentive to introduce an automated assessment system. Thereby, we hoped that we could achieve the following goals:

1) Increase the objectivity of the assessment process.
2) Increase the amount of 'productive' time spent with the students.
3) Impart on the students the importance of creating correct programs according to a given specification, and of verifying them through a comprehensive set of tests, without actually naming concepts such as 'unit test', 'test-driven development', etc.
4) Reduce the excessive burden placed on the teaching assistants, and enable them to invest their time into the improvement of lab sessions, homework assignments (with a fewer number of distinct assignments, we can devote more time to each of them), and the grading process itself.

Whether, and to what extent, we have achieved these objectives will be discussed in Section V. Before that, we shall describe our assessment system in detail.

B. Overview of the assessment system

By the introduction of our new grading system, we abolished the upgrades and the distinct sets of assignments for each grading week. The entire body of students received the same set of assignments. This decision made it possible to increase the number of assignments to 10, approximately one per week. On the one hand, weekly assignments encourage the students to practice on a regular basis, while on the other, the relative importance of each individual assignment is lower than in our old system.

After the teaching assistants prepare an assignment, the students have to submit their solution within 9 days (between Friday and next Sunday) to the Moodle online classroom, which, at UL-FCIS, has been successfully used for several years. To facilitate the automated assessment, the names of the students' programs (Java classes) have to follow a strict naming scheme. For each assignment, a student might earn a maximum of 10 points, giving 100 possible points in total. The points received from the assignments contribute one half to the final grade; the other half is earned on the final exam.

All assignments are assessed automatically. The teaching assistants prepare the test cases, feed the students' submissions and the test cases to the system, and wait for the results. The necessary (but not sufficient) condition for a student's program to receive any points at all is that it successfully compiles. If it does, it is evaluated by a number of test cases. The proportion of successfully passed test cases determines the grade. For example, if the student's program passes 34 test cases out of 50, he or she will receive $(34/50) \times 10 = 6.8$ points. (In 2014/15, the points were actually rounded up to the nearest integer. In the future, however, we shall refrain from rounding and work with fractional points.)

For each assignment, the students were given 10 so-called *public* test cases, which were divided into classes of increasing difficulty. One of the assignments, for instance, required that the students write a program for counting the blobs (4-connected contiguous areas of ones) in a given binary (zero-one) matrix. The public test cases for this assignment were divided into 6 classes:

- In the test cases 1–4, every blob in the input matrix consisted of a single cell, and so a program that simply counts the number of ones in the matrix would pass all four tests.
- In the test cases 5–6, blobs could also take the form of horizontal or vertical lines.
- In the test case 7, blobs could also take the form of full rectangles. Still, no general flood-fill algorithm was required for this case.

- In the test case 8, every line of each blob consisted of a single consecutive sequence of ones. Even in this case, the students could do without performing flood-fill.
- In the test case 9, blobs could take an arbitrary shape. However, they were comparably small so that a recursive flood-fill algorithm would not run out of stack space.
- In the test case 10, blobs could take an arbitrary shape and could fill up the entire 1000-by-1000 matrix, which means that — owing to Java's default stack limitations — only an iterative flood-fill algorithm was guaranteed to work in this case.

The public test cases are called 'public' because every student can view them and test his or her program on them. However, the programs are graded on a separate set of 50 *hidden* test cases, which are made publicly available only after the grading process for the current assignment has finished. The hidden test cases are guaranteed to be divided into the same number of classes as the public ones, and in each class, the number of hidden test cases is proportional to the number of public ones. In the example given above, the hidden test cases 1–20 belonged to the first class, the test cases 21–30 to the second, the test cases 31–35 to the third, etc.

By dividing the test case set into different classes, we encourage the students to develop their programs in a stepwise manner; especially a beginner is expected to start with the first class (matrices with single-cell blobs in the blob-counting example), then move up to the second, etc. Despite the diversity of the public test cases, they are, of course, not guaranteed to cover all possible boundary cases. Some boundary cases might be deliberately covered only by the hidden test cases. The students are therefore warned not to rely on the public test cases exclusively; they are advised to write additional test cases on their own.

Having the test cases divided into classes of increasing difficulty serves another purpose: to make the students think about the efficiency of their programs, without actually operating with the $O(.)$ or $\Theta(.)$ notation. Although computational complexity belongs to a course on algorithms and data structures rather than to a basic programming course, the earlier the students become aware of the fact that efficiency matters, the better.

C. Input-output assessment

For a majority of the assignments, including the blob-counting one presented in Section IV-B, each public and hidden test case is divided into a *test input* and the *expected output* for that input. The program reads its input from the standard input and writes its output to the standard output, and the grading system compares, for each hidden test case, the produced output with the expected one. In the text of the programming assignment, we provide the exact format of both the input and the output, as well as the bounds of individual components of the input. For instance, we might state that the first line of the input contains an integer $a \in [1, 10^9]$, a space, and an integer $b \in [a, 10^9]$. To enable the students to focus on the programming problem rather than on the output format, the expected output is, in most assignments, composed of a single integer. We use real numbers only when absolutely necessary.

D. Assessment through test classes

In some assignments, the students have to write a set of Java classes containing a set of methods with predefined signatures. In such cases, each public and hidden test case consists of a *test class* and the corresponding expected output. The test class creates a set of objects of the classes written by the student, invokes one or more methods of them, and prints the results. The results are then compared to the corresponding expected output. Again, to be able to pass any test case at all, the student's classes have to compile successfully.

In one of the assignments of this type, the students had to write a pair of classes (`Point` and `Line`) to represent points and lines in the two-dimensional space. Each class had to contain a number of methods. Here are three of them:

- `double distance(Point p)` in the class `Point`: Returns the distance between the point `p` and `this` point.
- `Point projection(Point p)` in the class `Line`: Returns the orthogonal projection of the point `p` onto `this` line.
- `double distance(Point p)` in the class `Line`: Returns the distance between the point `p` and `this` line.

This was the only assignment in which the grade was not obtained exclusively by black-box testing. In addition to the number of correct method call results, we also considered the 'elegance' of individual methods, which contributed up to 20% of the grade. We defined (and made it known to the students) that a method is elegant if it calls other methods in a meaningful way. For example, the most elegant way to calculate the distance between a line l and a point P is to determine the orthogonal projection P' of P onto l (by calling the method `projection`) and calculate the distance between P and P' (by calling the first method `distance`). The orthogonal projection of a point P onto a line l can itself be found in an 'elegant' way: as the intersection between l and the orthogonal line passing through P (both operations had to be programmed as methods, too). Since, for this assignment, we could uniquely define a canonical graph of method calls, the elegance was determined simply by comparing the canonical graph with the graph implicitly defined by the student's program. We counted the number of edges present in the program's graph and obtained an objective and meaningful measure of elegance.

E. Assessment of computer graphics assignments

The goal of a computer graphics assignment is to write a program that produces a bitmap based on given data using the Java graphics framework. The bitmap in Fig. 1, for example, is based on an array representing the relative heights of the bars. In such assignments, each test case is composed of a test class and the corresponding bitmap file. The test class calls the student-supplied drawing method, passing to it the desired width and height of the bitmap and the underlying data for producing the bitmap. The bitmap created by that method is

then compared against the reference bitmap on a pixel-by-pixel basis. To allow for rounding-off errors, the grading system considers that the two bitmaps match at a given position (i, j) if for the pixel at (i, j) in the reference bitmap there exists a same-colored pixel within the square $(i - 1,\ j - 1)$, ..., $(i + 1,\ j + 1)$ in the program-created bitmap, and vice versa. The points contributed by individual pairs of matching pixels are inversely weighted by the total number of pixels of that color, since we want to ensure that a short thin blue line counts exactly the same as a large red-filled rectangle. To be able to compare the bitmaps on an element-by-element basis, every logical element (or set of elements) of the bitmap has to be painted with a distinct color. For example, the bitmap in Fig. 1 consists of four sets of elements: orange bar fillings, red bar borders, blue connecting line, and white background. Each set of elements contributes 25% to the grade.

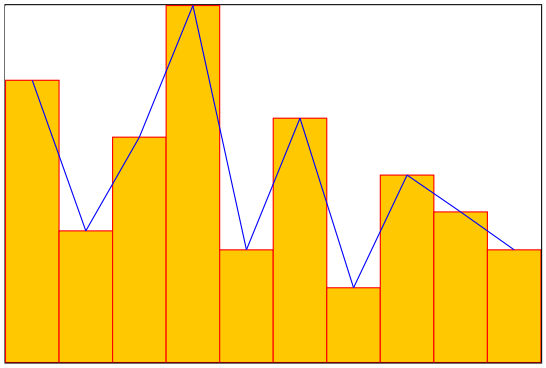


Fig. 1: The expected output for a sample assignment.

In addition to producing an output bitmap, the student might have to write additional methods for computing the dimensions or offsets of certain elements based on the bitmap size. The student might thus receive some points for understanding basic coordinate-system transformations, even if he or she fails to produce a bitmap. In the example of Fig. 1, the student could be requested to write a method to compute the height of a given bar based on the bitmap dimensions and the corresponding data value.

F. Plagiarism

To detect plagiarism, we employ the Stanford University Moss system, which quickly and fairly reliably points out potential occurrences of plagiarism even in the face of substantial code changes. Unfortunately, as many as 75 (out of 279) students were caught at least once, either as plagiarists or as overly 'generous' authors. The actual number of plagiarism occurrences was probably even higher, but in several borderline cases we decided against a punishment. One of the reasons for a high number of offenders was the mild penalty: zero points for each plagiarized assignment, with no additional penalty for repeated offenses.

G. Time and memory controls, security, etc.

For each test case, the program is given 3 seconds to produce an output. After that period, the Linux `timeout` command forcibly terminates the program and considers that test case to be failed. Memory consumption is bounded by the Java virtual machine itself. Security is ensured by running the programs under a designated non-privileged user account on a Linux machine.

H. User interface

The students obtain a command-line tool to grade their programs against a given set of test cases. This simple-to-use tool supports all three modes of testing (input-output, test classes, and bitmap-vs-bitmap). The tool presents the test results in form of an HTML report, which can be viewed in a browser. Both in the output of the command-line tool and in the HTML report, different possible test outcomes (pass, fail, timeout, exception) are represented by different colors. For each test case, the report shows both the input (or the test class) and the expected output; in the case of a mismatch, the program-produced output is shown as well. If a test case triggers an exception, the accompanying message is fully displayed. The student can thus quickly spot the test cases that he or she still has to work on. Figure 2 displays a sample run of the command-line tool and the corresponding HTML report.

V. Discussion and Conclusion

Have we fulfilled the objectives stated in Section IV-A? First, the assessment process is now undoubtedly more objective than it used to be. The computer grader is immune to the halo effect and other human-related imperfections. Second, the amount of 'productive' time spent with students has also increased: before the year 2014/15, three out of 13 lab session weeks were completely devoted to the assessment of assignments (through upgrades). Now, we are able to devote all 13 weeks to solving sample programming assignments, answering student questions, etc. The third objective was constantly being addressed through the rigorous definitions of input and output, through a diverse *and* imperfect set of public test cases (which covered a significant 'area' of the space of possible inputs but deliberately not every boundary case), by encouraging the students to write additional test cases themselves, and by the 'ruthless' automated grading system itself, in which only the number of passed test cases count. As for our fourth objective, however, we have to admit that the assistants might have had even *more* work than before. However, a large amount of time was devoted to the construction of the system itself; to the conversion of our old lab session programming assignments and more than 60 extra assignments (intended for ungraded practice) to a new format, and to the preparation of the corresponding test case sets; to the preparation of course material for three extra lab session weeks; to the renovation of the lab sessions themselves (the assistants shifted further towards the 'guide on the side' principle, while still functioning as a 'sage on the stage' when presenting the solutions to the assignments to the entire classroom); and, last but not least, to plagiarism detection and highly unpleasant email 'negotiations' with the

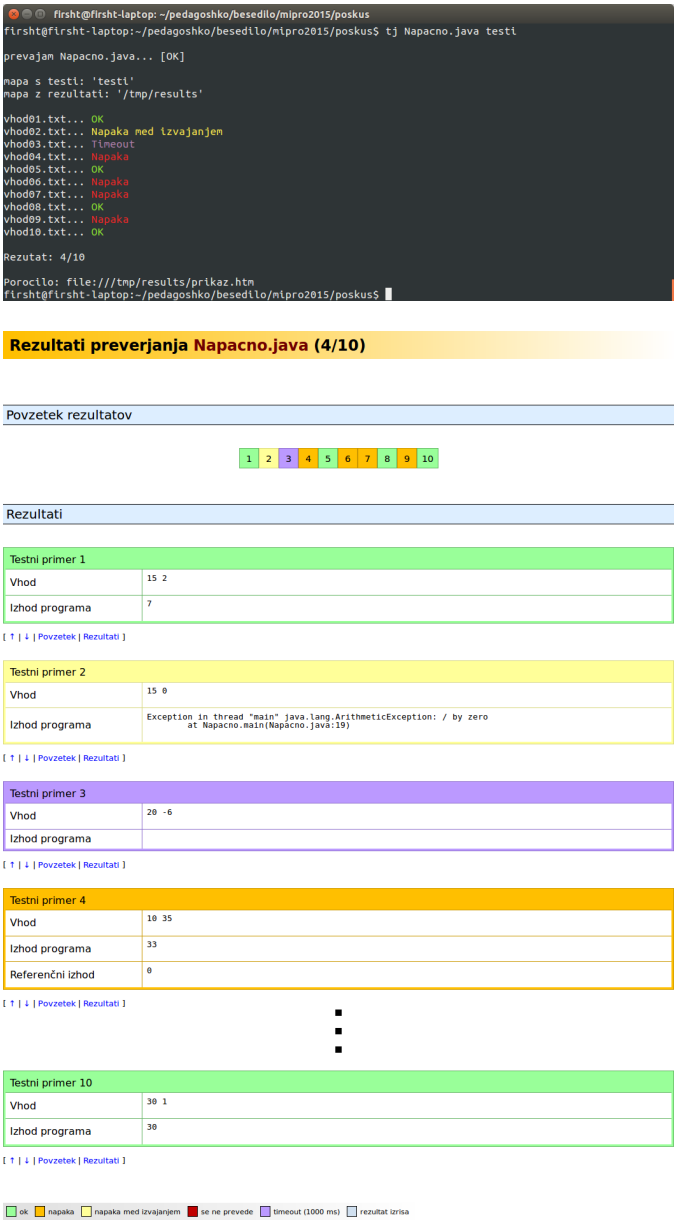


Fig. 2: A sample run of the grading program and (a part of) the corresponding HTML report.

accused offenders, most of which, nevertheless, eventually admitted their misdeed in the face of overwhelming evidence.

Unfortunately, plagiarism is definitely one of the problems brought about by our new assessment system. To cope with it, we intend to increase the penalty and to engage the designated faculty-level committee in case of repeated offenses. Another difference with the previous system is that the feedback provided to the students is limited to pass/fail for individual test cases. However, the test cases, if properly prepared, provide much more information than simple binary output. They can make the student think about different possible boundary cases and about the efficiency of his or her program. Implicitly, they guide the student in a stepwise development of the program.

While the test cases themselves cannot check adherence to coding standards or 'elegance', these concepts could, to a certain degree, be checked as well. In Section IV-D, we showed an example where 'elegance' was translated into a well-defined concept that could be easily checked. In the future, we shall investigate other ways to extend the system beyond simple pass/fail checks.

Despite certain imperfections, we consider the introduction of automated assessment to be a step forward in our pedagogical process. We have not yet conducted a survey about the students' views on automated assessment, but judging from the general anonymous remarks that we received at the end of the semester, the students seem to be satisfied with the new system. For the next year, we plan a more thorough preparation of the students to the grading system. In the lab sessions, we shall consistently use the system from the very beginning, and we shall introduce an ungraded pre-assignment to enable the students to become fully familiar with the system before the actual grading takes place.

References

[1] J. G. Doiron, "How teaching should be conducted in an {IT} era: Back to the future," *Asian Journal of Surgery*, vol. 25, no. 2, pp. 118 – 120, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1015958409601573

[2] K. A. Rahman and M. J. Nordin, "A review on the static analysis approach in the automated programming assessment systems," in *Proceedings of the national conference on programming*, vol. 7, 2007.

[3] J. Hollingsworth, "Automatic graders for programming classes," *Communications of the ACM*, vol. 3, no. 10, pp. 528–529, 1960.

[4] R. Romli, S. Sulaiman, and K. Z. Zamli, "Automatic programming assessment and test data generation - a review on its approaches," in *Information Technology (ITSim), 2010 International Symposium in*, vol. 3. IEEE, 2010, pp. 1186–1192.

[5] K. M. Ala-Mutka, "A survey of automated assessment approaches for programming assignments," *Computer science education*, vol. 15, no. 2, pp. 83–102, 2005.

[6] M. Joy, N. Griffiths, and R. Boyatt, "The boss online submission and assessment system," *J. Educ. Resour. Comput.*, vol. 5, no. 3, Sep. 2005. [Online]. Available: http://doi.acm.org/10.1145/1163405.1163407

[7] C. Douce, D. Livingstone, and J. Orwell, "Automatic test-based assessment of programming: A review," *Journal on Educational Resources in Computing (JERIC)*, vol. 5, no. 3, p. 4, 2005.

[8] V. Pieterse, "Automated assessment of programming assignments," in *Proceedings of the 3rd Computer Science Education Research Conference on Computer Science Education Research*. Open Universiteit, Heerlen, 2013, pp. 45–56.

[9] T. Wang, X. Su, P. Ma, Y. Wang, and K. Wang, "Ability-training-oriented automated assessment in introductory programming course," *Computers & Education*, vol. 56, no. 1, pp. 220 – 226, 2011, serious Games. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131510002241

[10] B. Skūpas, A. Čaplinskas, J. Augutis, E. Bareiša, G. Kulvietis, V. Marcinkevičius, D. Dzemydienė, and R. Šeinauskas, "A method for semi-automatic evaluation and testing of programming assignments," Ph.D. dissertation, Vilniaus universitetas, 2013.

[11] L. Fürst and V. Mahnič, "A cooperative development system for an interactive introductory programming course," *World transactions on engineering and technology education*, vol. 10, no. 2, pp. 122–127, 2012.

[12] ——, "Introductory programming course: motivating students with prior knowledge," *World transactions on engineering and technology education*, vol. 11, no. 4, pp. 400–405, 2013.

[13] N. C. Rowe, "Cheating in online student assessment: Beyond plagiarism," *Online Journal of Distance Learning Administration*, vol. 7, no. 2, 2004.

[14] M. R. Olt, "Ethics and distance education: Strategies for minimizing academic dishonesty in online assessment," *Online journal of distance learning administration*, vol. 5, no. 3, 2002.

Informatics Education and Requirements of Current Practice

L. Révészová
Faculty of Economics Technical University of Košice
Department of Applied Mathematics and Business Informatics, Košice, Slovakia
libusa.reveszova@tuke.sk

Abstract - Bill Gates in 1999 wrote: "Your ability to gain, process and use information decides whether you will belong to winners or losers." Since then, the meaning of his words has been constantly increasing. We often recognize an attitude that the primary issues in the development, implementation and effective application of informatics and ICT in current practice are modern technical devices, computers, hardware and software. These are surely very important part of it, however the key part consists of competent people. Investments into informatization of society and new technology are often very high. The efficiency of these investments depends on users, their motivation, interest and proper education. What are the real information competencies and students knowledge gained during their school informatics education? How can they use these knowledge and competencies? The article summarizes the results of the research of information competencies of secondary school graduates who entered the first grade at the Faculty of Economics, Technical University in Košice. We also deal with the experience in teaching and discuss the base of knowledge and methods which have been implemented into the informatics subjects at our faculty.

I. INTRODUCTION

Economists, politicians, publicists still more often talk about a new economy as a new system of organization of financial, enterprise and business activities based on intense work with information and knowledge. Knowledge plays an important role in all types of economic systems, even in the simple ones. For example: knowledge about grain growing in agrarian society, about mining and quarrying, building in industrial society, etc. (see Table 1.)

TABLE I. CIVILIZATION WAVES, ACCORDING TO [1]

Time ago	Society	Sources	Work/tools
1 million years	Primitive	Animals, seeds, fruits	Man-hunter/bow Women-counter/hands
15 000 years	Agrarian	Soil, home animals	Agricultural workers and stockbreeders/soil
300 years	Industrial	Fossil fuels, metals	Workers/machines
40 years	Information	Information	Managers/ICT

Knowledge has become even more important since the industrial revolution. The degree of including information and knowledge into economic activities is today so high that it introduces significant structural and qualitative changes in economic operations and changes the basic presumptions of getting competitive advantages.

We are currently preparing students for jobs that don't yet exist; using technologies that haven't been invented in order to solve such problems that we don't even know about them.

How technological progress in the 21st century will impact on labour market outcomes remains to be seen. Throughout history, technological progress has vastly shifted the composition of employment (see Table 1) from agriculture and the artisan shop, to manufacturing and clerking, to service and management occupations [2].

In a speech at the Harvard University in 1943 Winston Churchill observed that "*the empires of the future will be empires of the mind.*" He might also have added the battles for talent, as it is suggested in a paper from The Economist ("A survey about talent: The battle for Brainpower", in The Economist, Oct. 5th 2006). The old battles for natural resources are still with us. But they are being supplemented by new ones for talent not just among companies but also among countries. Our world is a world of brainpower. This explains why many international organizations focus on the idea of "putting people and their competence first" [3].

Human capital and the educational process of the population play a central role for the knowledge-based society development.

II. ESSENTIAL REQUIREMENTS FOR SUSTAINABLE COMPETITIVE ADVANTAGE AND LONG-RUN DEVELOPMENT

Long run development and competitive advantage are no longer to be interpreted with respect to the lower labour cost, but much more in a direct correlation with the competences, talent and the stock of intellectual capital and intangible assets.

The concept of intellectual capital is not new; in fact the economist Nassau mentions "intellectual capital" as an important factor of production in his book, published more than 150 years ago in 1836. The term "human capital" was

first introduced by Theodore Schultz (Nobel Prize winner). In 1963, Schultz was writing about the necessity of investments in education in order to increase the agricultural productivity in the USA. In 1975, Gary Becker developed a theory of human capital, which stated that the level of education and the experience of a person add up as determinant factors of their income. Economists highlighted the importance of intellectual capital as a production factor in comparison to traditional economic assets [3].

Nick Bontis defines intellectual capital of nations as "*the hidden values of individuals, enterprises, institutions, communities and regions that are the current and potential sources for wealth creation*" [4].

Intellectual capital implies human, system and market elements. Human capital refers to what people can do individually and collectively. The system component represents the knowledge in the firm which is independent from people and includes patents, contracts, databases, and information and production technology. The market component consists of the relationships between the organization and outsiders, e.g. suppliers, distributors and customers. According to this definition the components of the intellectual capital were divided into 3 parts [3], [5]:

- Human capital – can be defined as the employees characteristics that can be transformed into resources for the organization and which cannot be replaced by machines or cannot be transformed into codified knowledge.
- Organizational capital – encompasses all the intangible elements internal to the company that cannot be expressed directly on the balance sheet. This includes structural appropriateness, informal processes, quality of organizational governance or R&D productivity, tacit knowledge and brands but isn't limited to this list.
- Relational capital – can be considered to be external to the company. It comprises the outside connections that a company has; this includes customers, suppliers, partners or other types of stakeholders or organizations that are in some way relevant to the companies' activity.

The advancement in information and communication technology (ICT) has greatly enhanced the productivity of intellectual capital and intangible capital investment in the past several decades. The proliferation of software and computerized information systems (IS) as new forms of intangible assets is the most obvious change the IT revolution brought about. But more crucial, it increases the effectiveness of many other knowledge investment endeavors. For example, progress in communication technology and new media increased the reach of firms' marketing efforts. The discovering of the Internet made realization of many new business models possible, especially in the service sector. Computer networks make searching and sharing of information within and between all kinds of entities easier and faster. The use of computer software facilitated innovative work that produces knowledge assets [6].

Success of modern institutions and companies depends on engaging in a wide range of knowledge-building activities, such as designing new products and services, processes and business models, training employees, marketing brands, developing computerized assets, communicating within and outside the organization and acquiring information about markets and competitors. These activities mostly do not create any physical assets. However, they create knowledge-based resources indispensable in generating new values and not only financial returns. The nature of these activities is not very different from an investment into physical capital which both generate productive resources for the future.

The new world of work demands higher levels of expert thinking and complex communication. These skills are the keys to unlock a lifetime of learning and creative work. The set of the 21st century skills needed in current practice is focused on critical thinking, problem solving, communication, collaboration, creativity, innovation, imagination and invention. The 21st century global economy requires abilities mentioned above to continually invent new and better services and products for the global marketplace.

Market restructuring, accentuation of knowledge and increasing labour requirements carry demands on the education system. All kinds of graduates should become adequately skilled workforce. The education system should be changed simultaneously with requirements of the labour market, to adapt to these changes and react flexibly "to provide" applicable graduates.

III. INFORMATICS EDUCATION – WHAT IS NEEDED

Computer Science/Informatics is a young discipline. Its scope and focus have changed continuously over the past seven decades. The first computers ran numerical software for defense applications. In the 70´s, the advent of mainframe computers broadened the scope to include commercial applications. In parallel, large-scale circuit integration allowed exponential increases in computing power (Moore's law). In the 80´s, the convergence between IT and telecommunications opened the way for the Internet, the Web and the Information Society. In the 90´s, another very important, but less visible, revolution started with the dissemination of embedded systems technologies. More than 95 % of the chips produced today are for embedded applications. These are electronic components integrating software and hardware jointly and specifically of any kind: mobile phones, cameras, home appliances, cars, aircraft, trains, medical devices etc. In 2008, the average person used about 230 embedded chips every day [7].

Table II shows society´s educational goals throughout the ages in general (according to [8]) and Fig. 1 shows the structure and components of the skills, knowledge, and expertise students need to master to work and live successfully in the 21st century [8]. The rainbow illustrates the desired student outcomes most needed for our times including learning through additional school subjects and contemporary content themes, combined with 21st century skills. The framework adds to the additional subjects interdisciplinary 21st century themes relevant to some of the key issues and problems of our times, such as global awareness, environmental, health, financial, civic literacy. Finally, the core subjects and interdisciplinary 21st century themes are surrounded by the three sets of skills most demanded in the 21st century:

- learning and innovation skills,
- information, media, and technology skills,
- life and career skills[8].

As we think about what to teach our students today, (when there is a rapid development in Informatics) the authors in [8] recommend to consider the following questions:

- What will the world be like "x" years from now when our students have left school? Think about what life was like "x" years ago and all the changes we have seen had happened. Then imagine what will happen in the next "x" years.
- What skills will our students need to be successful in this world you have imagined "x" years from now?
- What were the conditions that made our high-performance learning experiences so powerful?
- What would learning be like if it were designed around our answers to the first three questions?

TABLE II. SOCIETY´S EDUCATIONAL GOALS THROUGHOUT THE AGES, ACCORDING TO [8]

Age	Goals for Education
Primitive	carry traditions and values forward;
Agrarian	pass on farming knowledge and traditions to the next generation; raise children in the ethnic, religious, and cultural traditions of parents and ancestors;
Industrial	learn the past knowledge of a trade, craft, or profession and pass this on to the next generation; maintain one´s own culture and values amid a diversity of traditions in urban life; connect with other cultures and geographies as communication and transportation expand;
Knowledge	quickly learn traditional knowledge in a field and apply its principles across other fields to create new knowledge and innovations; build identity from and compassion for a wide range of cultures and traditions; participate in a wide diversity of traditions and multicultural experiences; blend traditions and global citizenship into new traditions and values to pass on;

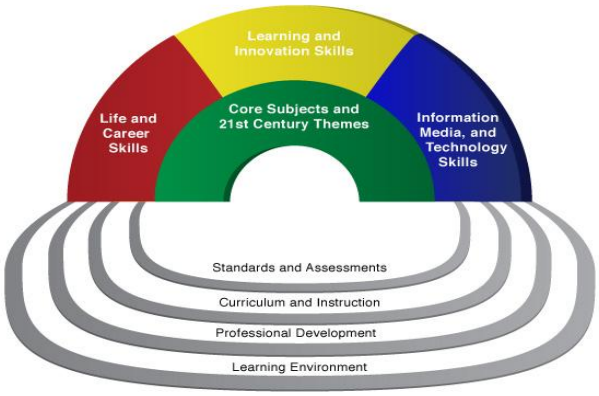


Figure 1. The 21st Century Knowledge-and-Skills Rainbow, [8]

Computer Science/Informatics subjects curricula seldom recognize the importance of system and fail to provide a holistic view of the discipline. There are the following recommendations in [7]:

- Teach students how to think in terms of systems (design processes, tools, interaction with users and physical environment).
- Teach principles rather than facts (foundations, architectures, protocols, compilers, simulation, etc.). Courses are very often descriptive and present details that can be acquired later as needed along professional life.
- Students should be prepared to deal with the constant change induced by technology and applications. They also should be kept aware of the limitations of existing theory of computing.
- Put emphasis on information and computation as universal concepts which are applicable not only to computers and provide the background for triggering critical thinking, understanding and mastering the digital world.

Informatics education in general should prepare young people for a world that doesn't yet exist, involving technologies that have not yet been invented, and that present technical and ethical challenges of which we are not yet aware [9].

Computer Science/Informatics should be a practical subject, where invention and resourcefulness are encouraged. Pupils are expected to apply the academic principles they have learned to the understanding of real-world systems, and to the creation of purposeful artifacts. This combination of principles, practice, and invention makes it an extraordinarily useful and an intensely creative subject, suffused with excitement, both visceral ("it works!") and intellectual ("that is so beautiful"). Students studying computing gain insight into computational systems of all kinds, whether or not they include computers. Computational thinking influences fields such as biology, chemistry, linguistics, psychology, economics and statistics. It allows us to solve problems, design systems and understand the power and limits of human and machine intelligence. It is a skill that empowers, and that all students should be aware of and

have some competence in. Furthermore, students who can think computationally are better able to conceptualize and understand computer-based technology, and so are better equipped to function in a modern society [9].

IV. INFORMATICS EDUCATION – REAL STATE

The evaluation of the knowledge base and ICT skills of students (entering the first year at the Faculty of Economics, Technical University in Košice) we consider as a very important step in determining the content and methods of our effective teaching. That is why in the first seminar in the subject Informatics I students (secondary school graduates) fill in a questionnaire which is focused on their general knowledge of informatics/computer science and their real ICT and problem solving skills. In this paper we present evaluation of students' responses in the period 2003 - 2013 with participation of 1755 students.

Fig. 2 presents an average percentage of respondents, who worked with the mentioned terms, programs or applications. These are needed for solving the next problems and tasks in the questionnaire.

As we can see in Fig. 2 nearly 100 % respondents are familiar with working with text editor, operating systems, using internet services and search machines. In the field of working with spreadsheet program and software for presentation also in the fields of informatics elementary terms, basics of hardware and architecture of computer, internet protocols programming languages the percentage is decreasing, in spite of the fact that these fields are a part of compulsory informatics curriculum. The surprising fact is that gradually fewer students are familiar with algorithms, basics of computer networks (e.g. topology) and databases systems.

We also want to find out how students can use their knowledge and skills to solve next problems:

1. Define/describe user requirements for a simple information system (IS) in a DVD rental where we want to install new and effective electronic IS based on modern ICT instead of "old-styled paper" IS.
2. Characterize, define, describe by your own words what does the next concept mean in general:

   2.1. system,

   2.2. information system,

   2.3. model.

Fig. 3, Fig. 4, Fig. 5 and Fig. 6 show evaluation of students´ answers and solutions. In the following graphs it can be seen the tendency of students´ knowledge development throughout the years.

In the first task we divided students' solutions according to their quality into following five levels:

- A – acceptable; it was possible to recognize functions and requirements of IS, there were structure and hierarchy between requirements, there were some graphical expressions.

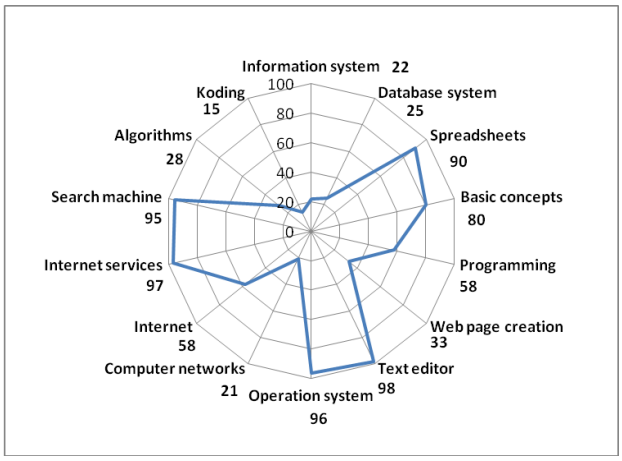


Figure 2. Average percentage – content of informatics education

- B – mostly continuous text with effort to design database tables or web site; partly incomplete.
- C – extensive continuous text with typical lack of exactness, structure, relationships; extent from ½ sheet to 2 sheets.
- D – text in extent from few lines to ½ of sheet; totally unsatisfactory.
- E – no elaboration; most of these students said: "I don´t know what to write.".

In evaluation of the second task we divided students' answers according to their quality into the levels: satisfactory, partial and unsatisfactory. In case 2.1 we monitored the appearance of two basics components – elements and relations in the system. In case 2.2 we considered a satisfactory answer to be one which enumerates at least three functions of IS (collection, saving, processing, analysis, presenting, information provision). In case 2.3 we searched connection of real object with model, purpose, simplification. The answers to these questions were evaluated individually by qualitative analysis of the students´ texts.

We have concluded from the evaluation of the questionnaire, that there is a lack of students´ ability to solve simple problems, in spite of the fact, that they worked with concepts system, IS and model every day in different forms. Despite the informatization of all aspects

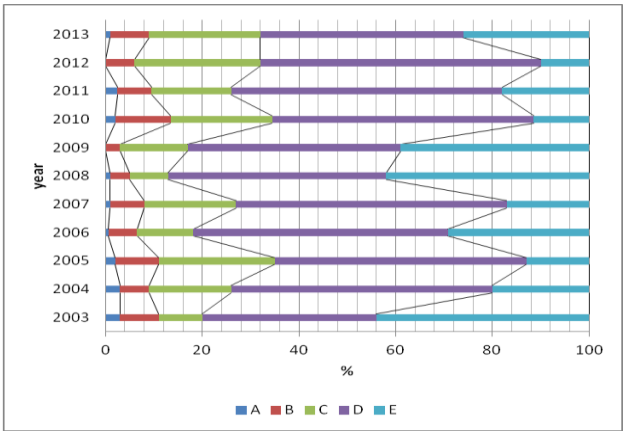


Figure 3. Comparison of evaluation IS user requirements description in percentage

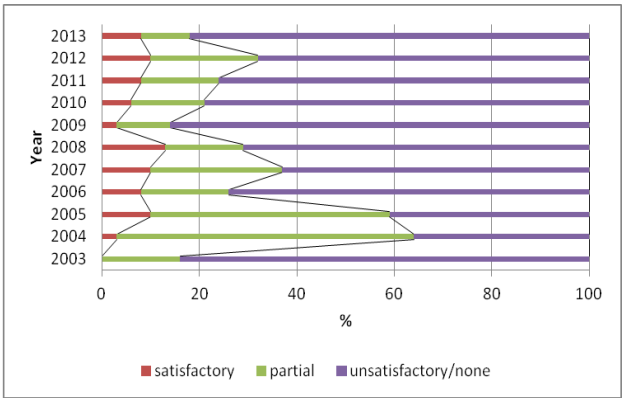


Figure 4. Comparison of evaluation of system definition in percentage

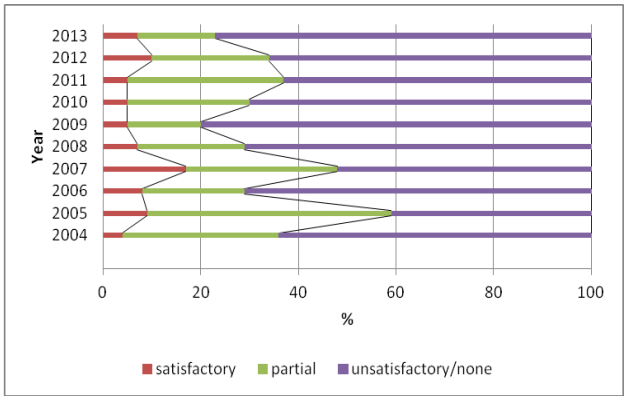


Figure 6. Comparison of evaluation of model definition in percentage

of live and development of ICT the ability of students to solve problems using informatics knowledge and ICT skills is not improving. As we can see in Fig. 2 the education is focused on handling different packages of applications´ programs. The algorithmic approach, use of knowledge for creating "something new" and interdisciplinary relations are often missing. Our findings confirm analyses, researches and facts from current practice.

According to [10] and [11] there is a visible decrease of employers´ satisfaction with the quality of graduates from all kinds of schools entering the labour market. In employers´ opinions graduates are not well prepared especially in those areas which they consider important: problem solving competencies, understanding instructions, ability to decide, will to learn new things, skills to work with information and use of computers.

The importance of problem solving skills is confirmed by PISA a PIAAC assessment. The OECD Programme for International Student Assessment (PISA) and its extension through the Survey of Adult Skills, a product of the Programme for the International Assessment of Adult Competencies (PIAAC) assessed the proficiency of pupils and adults in literacy, numeracy and problem solving in technology-rich environments. These skills are key information-processing competencies that are relevant to adults in many social contexts and work situations, and necessary for fully integrating and participating in the labour market, education and training, and social and civic

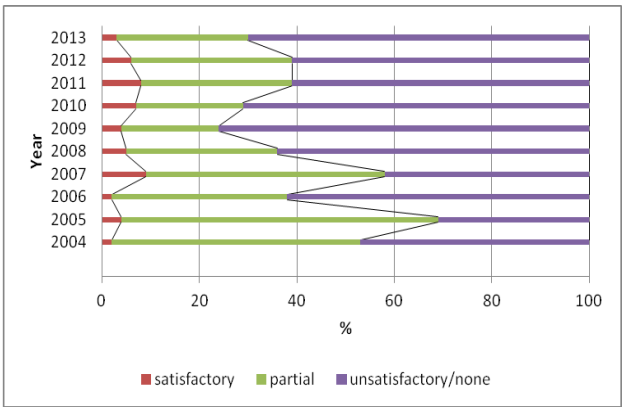


Figure 5. Comparison of evaluation of information system definition in percentage

life [12]. The analysis of results from the problem-solving assessment shows that, on average across OECD countries, only about one in five students is able to solve very straightforward problems. According to [13], [14] our students and adult population achieved a significantly lower score than the average of OECD countries.

V. INNOVATIVE WAY OF TEACHING

Considering the questionnaire results it is necessary at first to integrate level of students´ knowledge in the basics of computer science and using of basic applications in the scale and content similar to ECDL standard (first term).

To develop students' creativity, logical thinking and problem solving skills in the second term we developed an innovative manner of teaching. The students are taking part in projects of building IS in one of various types of virtual firms, companies or institutions. Except the traditional methods of teaching, as lectures and practices, we mostly used e-learning, learning-by-doing and project method. In this innovative way we are trying to transform the problem solving process into an active process of cognition. We agree that problem solution can be considered to be a very valuable skill for an individual and also for the whole society according to [15].

We assume that through designing, developing and conducting the investigative IS user requirements formulation we not only create attractive opportunities for students to employ their knowledge and experience in a complex way, neither we only provide them with productive learning environments for developing their digital, computational and social competencies, but we do more. We also facilitate the development of general problem solving skills of our students. During the investigations they were facing the problems they had never seen nor practiced before in any school subject. We also developed supplementary strategies for building modern constructionist informatics education, employ digital technology to open new directions for learning – project-based, immersion learning compatible with the Caperton and Papert's vision for education [16].

The area of creation and integration of IS is suitable to prove the importance of the above mentioned components of intellectual capital – human capital, organizational capital and relational capital. Informatics can therefore serve as a quality prerequisite for specialized subjects like

e-business, e-banking, Economics information systems [17], [18], [19].

VI. CONLUSION

Recent decades witnessed a significant increase in the share of jobs that require high levels of creative problem-solving skills. Our findings as well as international researches show that the ability of our students in problem solving area in ICT rich environment does not reach satisfactory level. It is necessary to consider, that people and their knowledge and skills, effective using of ICT are the most important components in development of our society. If universities want to prepare graduates according to current practice requirements they have to change their curricula and methods of teaching and learning.

In this new "digital world", information is available almost anywhere at almost any time, computer power is ubiquitous, communication of vast amounts of information is almost instantaneous, and storage capacities seem infinite. But these powerful capabilities only benefit those who have learned to use them effectively. Any citizen of a modern country needs the skills to use IT and its devices intelligently. All of Europe's citizens need to be educated in both digital literacy and informatics. Digital literacy covers fluency with computer tools and the Internet. Informatics covers the science behind information technology. Informatics is a distinct science, characterized by its own concepts, methods, body of knowledge and open issues. It has emerged, in a role similar to that of mathematics, as a cross-discipline field underlying today's scientific, engineering and economic progress. Informatics is a major enabler of technology innovation, the principal resource for Europe's drive to become an information society, and the key to the future of Europe's economy [20].

Informatics encourages and facilitates the process of collecting, cultivating, and organizing information digitally so that it can be better utilized by the community, businesses, organizations, and individuals. Informatics aims at improving the human experience by bridging the digital divide to bring people, information, and technology together so that technology can be put to work for solving complex problems facing humanity today.

REFERENCES

[1] B. Brestenská, "From homo sapiens to homo mobilis," In: 4th International Conference Research in Didactics of the Sciences (DidSci), Krakow, 2010. [online] Available at: http://www.statpedu.sk/files/documents/vzdelavacie_aktivity/kurikulum/brestenska_ecrice.pdf

[2] C. B. Frey and M. A. Osborne, "The future of employment: how susceptible are jobs to computerization?" 2013, [online] Available at:http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

[3] M. C. Suciu, "Intangible Assets and Intellectual Capital as Key Factors of Romania's Convergence," [online] Available at: http://econpapers.repec.org/paper/rorwpince/090109.htm

[4] N. Bontis, "Assessing knowledge assets: A review of the models used to measure intellectual capital". Working paper, Queen's Management Research Centre for Knowledge- Based Enterprises, 2004, p. 4

[5] M. C. Suciu, L. Picioruş and C. I. Imbrişcă, "Intellectual Capital, trust, cultural traits and reputation in the Romanian education system" The Electronic Journal of Knowledge Management Volume 10 Issue 3 (p223-235,) available online at www.ejkm.com ISSN 1479-4411

[6] N. X. Che, "Sectoral Structural Change in a Knowledge Economy," [online] Available at: http://mpra.ub.uni-muenchen.de/19839/1/MPRA_paper_19839.pdf

[7] J. Sifakis, "A vision for computer science – the system perspective," In Central European Journal of computer Science, 1, 2011, p. 108-116, Springer, DOI:10.2478/s13537-011-0008-y

[8] B. Trilling and Ch. Fade, "21st Century Skills: Learning for Life in Our Times," Wiley and Sons, 2009, ISBN 978-0-470-47538-6

[9] Computing at School Working Group, "Compute science: A curriculum for schools" March 2012, [online] Available at: http://www.computingatschool.org.uk

[10] M. Úlovec, "Potřeby zaměstnavatelů a připravenost absolventů škol – srovnání 2004 – 2013, Národní ústav pro vzdělávání, školské poradenské zařízení a zařízení pro další vzdělávání pedagogických pracovníků, Praha 2014, [online] Available at: http://www.infoabsolvent.cz/Temata/PublikaceAbsolventi?Stranka=9-0-93&NazevSeo=Potreby-zamestnavatelu-a-pripravenost-absolventu

[11] M. Úlovec, "Potřeby zaměstnavatelů a připravenost absolventů škol – komparační analýza, Národní ústav pro vzdělávání, školské poradenské zařízení a zařízení pro další vzdělávání pedagogických pracovníků, Praha 2014, [online] Available at: http://www.infoabsolvent.cz/Temata/PublikaceAbsolventi?Stranka=9-0-90&NazevSeo=Potreby-zamestnavatelu-a-pripravenost-absolventu

[12] OECD, Education at a Glance 2014: OECD Indicators, OECD Publishing. ISBN 978-92-64-21505-4 [online] Available at: http://dx.doi.org/10.1787/eag-2014-en

[13] PISA 2012 Results: Creative Problem Solving (Volume V): Students' Skills in Tackling Real-Life Problem, PISA, OECD ine] Publishing, Paris, [online] Available at: http://dx.doi.org/10.1787/9789264208070-en.

[14] OECD Skills Outlook 2013: First Results from the Survey of Adult Skills, [online] Available at: http://dx.doi.org/10.1787/9789264204256-en

[15] H. Brožová, M. Houška et al. "Modelování znalostí", 1. vyd. Praha: Professional Publishing, 2011. ISBN 978-80-7431-069-0

[16] Z. Galik, I. Kalaš, "Developing Digital, Computational and Social Competencies through Investigative On-Line Activities", Informatics in Education, 2009, Vol. 8, No. 2, p. 191–216

[17] M. Vejačka, "Introduction to electronic banking by elearning methods" In: MIPRO 2012, jubilee 35th International convention proceedings, May 21 - 25, 2012, Opatija, Croatia. Rijeka, p. 1406-1411. ISBN 978-953-233-069-4

[18] D. Paľová, Martin Vejačka, "On-line E-learning platform supporting education and practice of accountants in EU space" In: Mipro 2013, 36th international convention conference proceedings, May 20-24, 2013, Opatija, Croatia. - Rijeka MIPRO, 2013 P. 771-776. ISBN 978-953-233-074-8

[19] J. Bucko, M. Vejačka, "Integration of FASTER platform into education process of economic information systems" In: Confronting Contemporary Business Challenges through Management Innovation, 6th Annual Conference of the EuroMed Academy of Business, Conference Readings Book Proceedings September 23rd-24th, 2013, Estoril, Cascais, Portugal. - Estoril EuroMed Press, 2013 P. 496-507. ISBN 978-9963-711-16-1

[20] Informatics Europe & ACM Europe Working Group on Informatics Education, "Informatics education: Europe cannot afford to miss the boat" April 2013, [online] Available at: http://europe.acm.org/iereport/ACMandIEreport.pdf

Integrating Learning Management System and Faculty Information System – Service Oriented Approach

Petar Bjeljac *, Igor Zečević * and Branko Perišić *
* Faculty of Technical Sciences, Novi Sad, Serbia
pbjeljac@uns.ac.rs, igor.zecevic@uns.ac.rs, perisic@uns.ac.rs

Abstract - The current era of information technologies opens new challenges brings new approaches concerning the use of computer supported Learning Management Systems (LMS). Nowadays, the majority of higher education institutions utilize information systems for handling administrative data, while some of them have successfully introduced one or more LMSes. On the other hand, the majority of LMSes rely on an encapsulated subset of administrative data, causing data redundancy with the administrative database. If the LMS supports learning activities which need to be scored, there is a strong need for reverse flow of relevant data from LMSes back to the administrative system. The need for the development of a moderation layer becomes a must. In this paper a brief analysis of the developed architecture of a moderation layer, the state of the art and the further development direction concerning the Faculty of Technical Sciences in Novi Sad information system architecture extending to the LMS integration are presented. A service oriented approach for two-way data exchange between these systems is proposed as a flexible solution.

I. INTRODUCTION

The proliferation of IT use is causing changes in all aspects of life. Among these changes, a change to the learning process has been made by the expanding use of Learning management systems (LMS) on higher education institutions.

An LMS [1] is "a software application that automates the administration, tracking, and reporting of training events". These systems ease student-teacher communication, as well as the distribution of teaching materials. However, several problems rise due to the amount of administrative tasks teaching staff needs to execute in order to prepare this system to adequately support the teaching process.

Learning management systems contain different types of data:

- Course structure data;
- Teacher and student data;
- Group and enrollment data; and
- Grading data.

On the other hand, in most cases the faculties already have an existing system (or several systems) for handling this type of data. These systems are called Faculty Information Systems (FIS). A Faculty Information System is a single system, or a set of systems, used for executing administrative tasks at higher education institutions (student, staff, course administration etc.).

In this paper, a solution for integrating LMSes with an existing FIS, using a service oriented approach, is presented. The paper focuses on a prototype implementation, providing data exchange from the FIS to LMSes. This is initial work in the process of integrating these separate systems into one complete solution for handling administrative, as well as teaching tasks at faculty level.

The rest of the article is organized as follows. The second chapter describes the current Information System (IS) state at the Faculty of Technical Sciences (FTS) in Novi Sad, which is used as proof of concept in this research. Furthermore, the custom requirements for the integration process are analyzed. The third chapter presents existing solutions and compares them to these requirements. It also gives a description of an existing standard for integrating LMSes. In the fourth chapter, a solution, based on RESTful web services is described. Finally, in chapter five, conclusions are made and the directions for future work are outlined.

II. CURRENT STATE AT THE FTS

The Faculty of Technical Sciences in Novi Sad is the largest faculty at the University of Novi Sad, with approximately 13000 active students and 1000 members of teaching staff. It consists of 13 departments, with 87 study programs. The teaching process administration for such large institutions presents a complex task, which has been supported for years using several information systems, building a custom FIS for the FTS (Fig. 1). This IS consists of the following elements:

- The Accreditation IS – used for structuring courses and preparing study programmes for the accreditation process;
- The Human resources IS (HRIS) – used for the administration of teaching and non-teaching staff; and

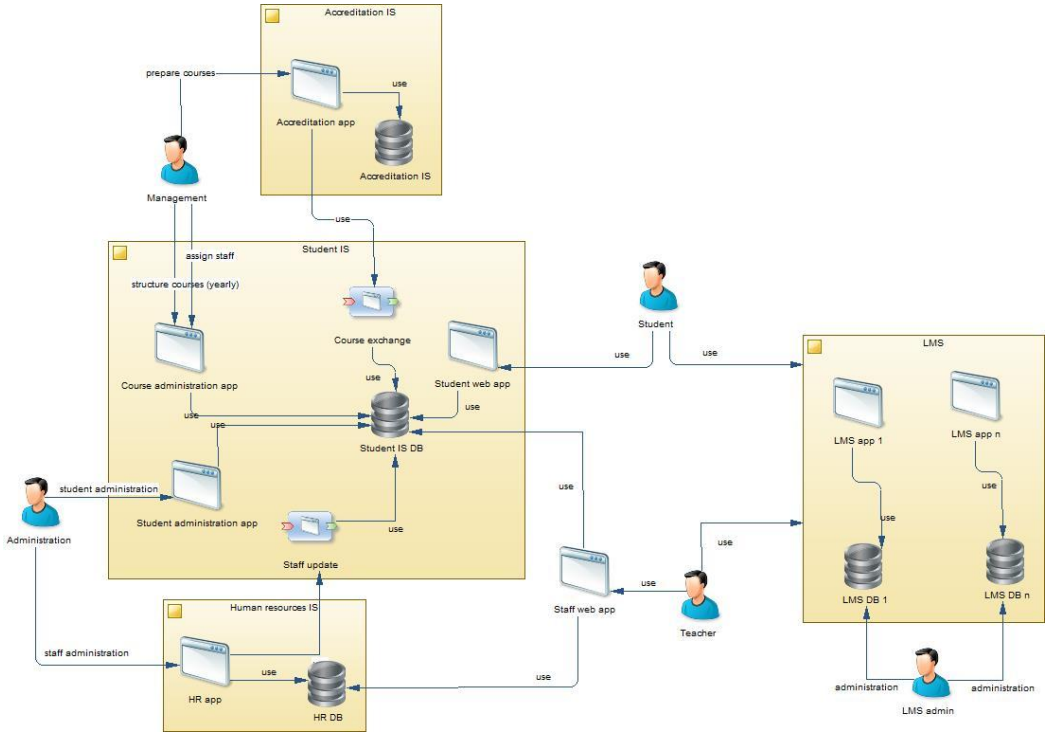


Figure 1. IS state at the FTS

- The Student IS (SIS) – used for the administration of students and their grades, as well as real-time course administration and distribution of teachers and students to courses.

In the last couple of years, several departments have made effort in adopting the use of contemporary LMSes in their teaching process. Currently, three departments use an LMS to support their teaching process: The Department of computer sciences uses Canvas LMS [2], whereas the Department of industrial engineering and management, as well as the Department of electronics use implementations of Moodle LMS [3], adapted to their specific needs. As oposed to other FIS subsystems, the LMSes have not been centrally implemented by the FTS IT service, primarily due to the fact that a compromise on using one exclusive LMS system at faculty level has not been made. These systems represent separate systems, with no level of integration with the FIS.

The following requirements for the integration process emerge:

1. A flexible solution, easily extendable to new LMSes, as only 3 out of 13 departments have an LMS, whereas other departments may decide to implement and use a different one.
2. Secure data transfer, as student and staff data is, and should remain, confidential to third parties.
3. An easily usable and understandable integration interface, in order to ease the integration process for the developers implementing the LMSes.
4. A quickly and easily implementable prototype, in order to test the level of usability of such solution for our IS.

III. EXISTING SOLUTIONS

The problem of integrating an LMS with other systems is not new. How the author states in [1] LMSes need an integration with an existing HR system in order to track employees on courses. In higher education institutions, this extends to integration with a Student administration system, in order to execute the same task with students in focus.

Most LMSes, like [2] [3] [4] [5], offer a simple way for data exchange (primarily import) using Comma Separated Value (CSV) or some other format of files. However, using this approach, files containing potentially confidential data, need to be manually imported by the administrator, exposing them to potential data theft. An even larger setback for this integration approach is the fact that different LMSes often use different file and data formats, so a separate data export script would need to be implemented for each system.

Most modern LMSes also have an Application Programing Interface (API), solving the problem of data file exposure. However, the different format (API) problem still remains.

In [6], the authors suggest an approach based on direct data integration. The LMS communicates directly with a copy of the SIS database to access data about users. This approach, although more generic, is not applicable in our case, due to the fact that the LMSes are not centrally implemented and controlled by the FTS IT service.

The same author in [7] states the disadvantage of direct data integration in the process of updating SIS data, for example importing grades back to SIS, due to the lack of authentication, as well as data validation. Here, a business logic level integration is proposed, in order to include the business-level validation of data.

The authors in [8] propose two levels of integration – data and functional integration. An interface is created for data to be exchanged between SIS and LMS. Furthermore, the functional integration of these systems, enabling the user to log in to each of these systems and directly use the functions from the other, is presented. However, no further details about the integration process are available.

A. *Learning Information Services specification*

Learning Information Services (LIS) Specification [9] is a standard created by the IMS Global Learning Consortium. LIS defines the process for exchanging information which describes people, groups, memberships, courses and outcomes within the context of learning. The specification defines six different types of exchange services:

- Person Management Services – data about students and teaching staff;
- Group Management Services – data about study groups on courses;
- Membership Management Services – data about membership for students and teaching staff – connecting them to courses and groups;
- Course Management Services – courses and course structure;
- Outcomes Management Services – grade-book data; and
- Bulk Data Exchange Management Services – bulk exchange of all data.

These services enable the integration of all LMSes supporting the LIS standard with any IS able to provide interfaces based on the specification. An integration process in full compliance with the Learning information services (LIS) specification is presented in [10].

Although this is a complete solution, based on an industry standard, the standard has not yet been supported by all LMS vendors, and is currently not considered as a solution in the case of FTS. However, this standard represents a base for our proposed solution.

IV. PROPOSED SOLUTION

Analyzing the requirements, as well as existing work in this area, we propose a prototype implementation for a data integration system for exchanging data from our FIS to the LMS. This approach is based on the LIS specification data model and service interfaces, and the implementation of these interfaces using service oriented approach.

A. *SOA, REST, JSON*

Service oriented architecture (SOA) is [11] an architectural style that supports service orientation. Services are defined [12] as course grained, discoverable software entities that exist as a single instance and interact with applications and other services through a loosely coupled (often asynchronous), message based communication mode.

Representational State Transfer (REST) is an architectural model for building distributed applications based on three distinct concepts: representation, state, and transfer [13]:

- Representation: Data or resources are encoded as representations of the data or the resource. These representations are transferred between clients and servers.
- State: All of the necessary state, needed to complete a request, must be provided with the request. The clients and servers are inherently stateless. A client cannot rely on any state stored in the server, while the server cannot rely on any state stored in the client. This does not, however, pertain to the data stored by servers or clients; only to the connection state that is needed to complete transactions.
- Transfer: The representations and the state can be transferred between client and servers.

REST is most often implemented as a combination of the Hypertext Transfer Protocol (HTTP) and TCP/IP. With this instantiation of REST, HTTP requests are used to transfer representations of resources between clients and servers.

JSON (JavaScript Object Notation) [14] is a lightweight data-interchange format, easy for humans to read and write. Compared to XML, JSON is [15] faster and uses fewer resources, and is on the other hand as widely supported in programming languages as XML.

B. *The implementation*

Based on the nature of our system, required to support several different types of LMSes, a solution based on a web service implementation on the FIS side is proposed. This approach to the integration process requires a plug-in (extension) to be implemented on the LMS side in order to access the functions. The proposed architecture for our solution is presented in Fig. 2. The system covers the complete integration process, with a Single Sign On (SSO) Service providing access to all LMS systems directly from the student/staff application, and a Data Exchange Service, responsible for handling data import/export tasks. The scope of this paper covers the implementation of Data Exchange Services allowing data export from the FIS.

The web services implemented are based on the LIS specification and consist of the following functions:

- Course export
- Groups export
- Student export
- Teacher export
- Course membership export

- Group membership export

The data model for all these web services contains two segments – input and out parameters. Here we present a more detailed description of these services and the data exchanged.

1) Course export

The course export service provides an interface for exporting data about courses and course structure to the LMS.

Input parameters:

- School year
- Semester
- Study program

Output parameters:

- Unique id – containing data about the school year, semester, study program and course
- Name – full course name
- Short name – an abbreviation for the course name

2) Group export

The group export service provides an interface for exporting data about student groups on a specific course.

Input parameters:

- Course id – the id received from the Course export service

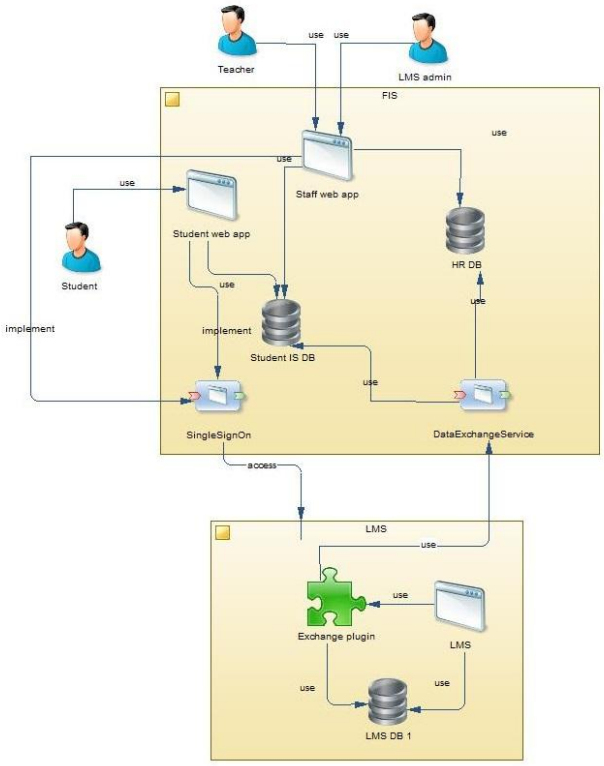


Figure 2. Proposed architecture

Output parameters:

- Unique id – containing data about the course and group
- Group name - a descriptive name for the group

3) Student export

The student export service provides an interface for exporting data about students enrolled in a specific course.

Input parameters:

- Course id – the id received from the Course export service; if empty, all active students are returned

Output parameters:

- Unique id – the id received from the Course export service; student id from the SIS
- First name
- Last name
- E-mail address

4) Teacher export

The teacher export service provides an interface for exporting data about teaching staff responsible for teaching activities on a specific course.

Input parameters:

- Course id – if empty all active teaching staff are returned

Output parameters:

- Unique id – id from the HRIS
- First name
- Last name
- E-mail address
- Office
- Phone number

5) Course membership export

The membership export service provides an interface for exporting data about student and teacher distribution enrolled on a specific course.

Input parameters:

- Course id

Output parameters:

- Student id/Teacher id

6) Group membership export

The membership export service provides an interface for exporting data about student and teacher distribution in groups on a specific course.

Input parameters:

- Course id

Output parameters:

- Student id/Teacher id
- Group id

Apart from these functions, an interface for importing grades from the LMS back to the SIS is planned to be implemented in the Data Exchange Service component of our system. However, this interface is out of the scope of this paper.

Fig. 3 presents the class model which supports these interfaces, without attribute details.

Due to REST and JSON simplicity of use and low resource requirements, a decision has been made to implement web services based on these specifications. An example of use is the call to the Teacher export service, which is made as a simple HTTP GET call to an address in the following format:

```
[serveraddress]/teachers/2014.1.L.1.1.SRP.SE1.13
.SE0011
```

where `2014.1.L.1.1.SRP.SE1.13.SE0011` is the courseId received from the Course export service.

The responce received is in JSON format:

```
{"teachers":[{"teacherId":"N361","firstName":"Br
anko","lastName":"Perišić","email":"perisic@uns.
ac.rs","office":null,"phoneNum":"0214852244"},{"
teacherId":"N1649","firstName":"Petar","lastName
":"Bjeljac","email":"pbjeljac@uns.ac.rs","office
":null,"phoneNum":"0214852246"}]}
```

and can be imported into the LMS.

V. Conclusion

The proliferation of learning management systems use in contemporary educational process opens several challenges to the institutional integrated information system development. However, a large drawback in the adoption of LMSes by institutions and teachers is the amount of redundant administrative work needed for their utilization. This is due to the unavoidable heterogeneity and the lack of integration support mechanisms being the common situations that have to be addressed.

The solution based on RESTful web services, with JSON as the data exchange format, presented in this paper, is easy to implement and use, and represents an open solution that may be extended and improved in a simple manner. Furthermore, the fact that we use a data model similar to the one in LIS specification enables simpler implementation of plug-ins needed for the exchange of data and an easy understanding of all the concepts used. The initial tests were subjected on four courses, on a subset of approx. 300 students and five members of teaching staff. The approach has proved to satisfy all specific requirements at the Faculty of Technical Sciences.

Future work directions are: the implementation of services needed for the import of grade-books back to the SIS; the implementation of an authentication and authorization mechanism for these services; the implementation of a Single Sign On (SSO) solution, enabling users to login to one system (SIS), and use functions from other systems (LMS).

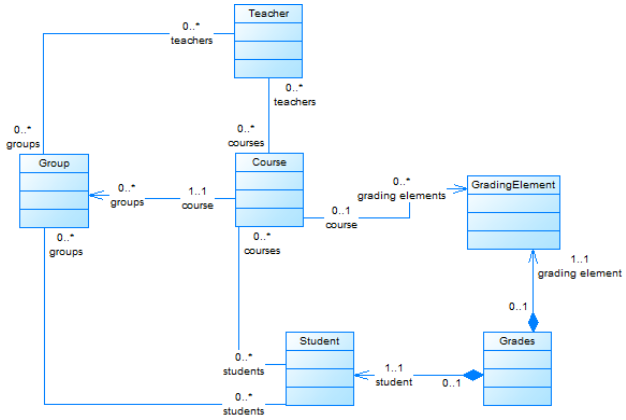


Figure 3. Exchanged data model

References

[1] R. K. Ellis, "Field guide to learning management systems." ASTD Learning Circuits (2009).

[2] Canvas LMS http://www.instructure.com/, consulted February 01, 2015

[3] Moodle LMS https://moodle.org/, consulted February 01, 2015

[4] Sakai LMS https://sakaiproject.org/, consulted February 01, 2015

[5] Blackboard LMS http://www.blackboard.com/Platforms/Learn/Products/Blackboard-Learn.aspx, consulted February 01, 2015

[6] B. Abu Shawar, "Utilizing AOU'VLE with other computerized systems."International Journal of computer Science and Security (IJCSS) 1.4 (2007): 13-24.

[7] J. Al-Sadi, B. Abu Shawar, and T. H. Sarie. "Quality Assurance Procedures: New Enhancements to the Learning Management System at AOU." CSREA EEE. 2006.

[8] J. Molins and J. Busquiel. "SIGMA AIE OFFERS TO FACULTAT PERE TARRÉS THE MOODLE LEARNING PLATFORM, INTEGRATED WITH ACADEMIC MANAGEMENT SYSTEM FOR TEACHERS IN SOFTWARE AS A SERVICE (SAAS)." INTED2014 Proceedings (2014): 5760-5765.

[9] Learning Information Services http://www.imsglobal.org/lis/, consulted February 01, 2015

[10] INTEGRATING YOUR LEARNING MANAGEMENT AND STUDENT ADMINISTRATION SYSTEMS, http://www.oracle.com/us/industries/education-and-research/056908.pdf, consulted February 01, 2015

[11] http://www.opengroup.org/soa/source-book/soa/soa.htm#soa_definition, consulted February 01, 2015

[12] Brown, Alan, Simon Johnston, and Kevin Kelly. "Using service-oriented architecture and component-based development to build web service applications." Rational Software Corporation (2002)

[13] M. Rouached, S. Baccar, and M. Abid. "RESTful sensor web enablement services for wireless sensor networks." Services (SERVICES), 2012 IEEE Eighth World Congress on. IEEE, 2012.

[14] JSON http://www.json.org/, consulted February 01, 2015

[15] N. Nurseitov, M. Paulson, R. Reynolds, and C. Izurieta. "Comparison of JSON and XML Data Interchange Formats: A Case Study." Caine 2009 (2009): 157-162.

Evaluating the effect of the automatic assessment introduction in the ADS1 course practical work

I. Rožanc and J. Mihelič
Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia
{igor.rozanc, jure.mihelic}@fri.uni-lj.si

Abstract - The introduction of modern teaching and assessing approaches importantly enhance the quality of teaching. Apart beneficial effects such as effective knowledge presentation, better communication and easier practical work execution, changes include risk to ruin good results of well-established teaching routine.

This paper is motivated by recent changes of the content, structure and assessing method of practical work of the Algorithms and Data Structures 1 course, which is taught at the FCIS at the University of Ljubljana. The most critical part was the introduction of automatic assessing of practical work as it considerably elevated the number of shorter tasks to be handed in on time by students.

In our work the effect of such approach is objectively evaluated. First, the previous and current assessment data is used to examine the number and distribution of results for students involved in practical work. Second, the current grades are analyzed to check the correlation with previous grades obtained manually. Next, the results of a student questionnaire are presented, which confirm the successful introduction of new approach. Finally, our experience is summarized to support similar initiatives elsewhere.

I. INTRODUCTION

In Slovenian higher education the principles of Bologna reform [1] were adopted recently. Its implementation caused important changes in the structure of university programs such as uniform length of study, course selectiveness and on credits based promotion. Additionally a new teaching style was implemented to enhance efficiency of study with the emphasis on practical student work, which must be concurrently evaluated and later incorporated into final course grade.

In practice we deal with requirement to upgrade the practical work (i.e. exercises) with more student-assistant interaction, more student tasks of different types and more task assessments. The amount of assistant work is growing towards uncontrollable level, especially in cases of the courses with large number of students.

The introduction of modern (automated) teaching and assessing approaches can be of a great aid. Among well-known beneficial effects such as modern knowledge presentation and upgraded interaction with new ways of communication they especially ease exercises execution by supporting automatic submission and assessment with appropriate tools. However, the elevated use of automatic tools has pitfalls, too. Important aspect of assistant's work is now the preparation of interactive materials and fraud detection instead of face-to-face communication between student and assistant. Additionally the implementation of new approaches can ruin the beneficial effects of a well-established teaching routine.

The optimal relation between the automated and non-automated approaches is obviously different from case to case. In this paper we present the experience gained at the Faculty of Computer and Information Science (FCIS) of the University of Ljubljana on this aspect. It was motivated by recent changes of the content, structure and assessing method of the Algorithms and Data Structures 1 (ADS1) course, which is taught at our undergraduate professional program of computer and information science. Our aim is to analyze available data to confirm expected advantages and detect issues of new approach. This experience can provide valuable advice for university teachers with similar mission, as in case of research works with similar motivation [2, 3].

ADS1 is one of basic programming courses. In academic year 2010/11 it was performed for the first time in the Bologna compliant way, but with a moderate involvement of automatic tools only [4]. At first the students had to solve a few big programming projects during the laboratory work, and they had to present (defend) the solutions to the assistant during a face-to-face meeting. Then and there all the issues were discussed and a grade was determined.

In academic year 2014/15 a new lecturer started to teach the course and consequently the changes of content and organization were introduced. The most important change was the introduction of the automatic assessment of exercises. Students now solve much more tasks of different types, and the main part of solution delivery and assessment is automated.

We are able to collect enough past and current data to objectively evaluate the effect of this change. Thus we focus our work on statistical analysis of this data with the aim to discover which approach yields better outcome. Additionally, the results of an anonymous student questionnaire are analyzed to add an alternative view of these topics.

In the second section the ADS1 course is presented with focus on its content, structure and assessment changes. The third section describes the analysis background with data and statistical analysis selection. In the fourth section the actual analysis results are presented in detail and discussed. Finally the paper is concluded with a main findings and a proposal for future work.

II. THE ADS1 COURSE DESCRIPTION

The ADS1 course contains the first part of the classical ADS content as defined in the IEEE curricula for the Computer Engineering [5]. As such it is aimed at teaching selected chapters of programming concepts and data structures, while its successor course ADS2 appends the remaining part of the ADS content.

The ADS1 is taught in the first semester of the second year of study. It is part of undergraduate professional program, which has intentionally a large number of free choice courses to enhance the adaptation of student curricula to his or her wishes [6, 7]. The ADS1 course is formally a free choice course, but in practice all students chose to study it as it is a necessary prerequisite for most of the free choice courses students choose in future.

The implemented changes aim to enhance the course performance in an effective way. Thus we preserved good solutions and upgraded the others. E.g. previous and current course's teaching style remained similar, but major changes are in course's topics (i.e. contents), it's performing in practice (i.e. structure) and especially in the implementation of exercises grading (i.e. assessment).

A. The content differences

The contents of the previous ADS1 course was aimed on the measurement and analysis of algorithm complexity, introduction of some programing paradigms (e.g. divide and conquer) and (especially) on detailed presentation of the data structures (ADT, simple and complex data types, diverse implementations of data structures). Important emphasis was given to the actual implementation of data structures in Java. A majority of tasks at the exercises were of programming nature, too.

With new teacher the division of topics between ADS1 and ADS2 courses changed. Some additional topics emerged in the ADS1 course (i.e. arithmetic algorithms, sorting algorithms, dynamic and greedy algorithms), and some data structures are now covered in ADS2. The lectures are now presented at a higher level of abstraction using pseudo code. The actual programming tasks are now present at some exercises only, and a lot of non-programming tasks are given to students, too.

B. The structure differences

The main organizational differences were introduced in the practical execution of the course, especially at the level of the exercises. Previously the main part of each the ADS1 exercise was the presentation of programming examples, which were solved and explained in detail by assistants. The student participation was obligatory and they had to be involved in discussion. The course was (and is still) efficiently supported by Moodle [8].

The main student's obligation was to solve four (or five) big programming projects, which had to be finished in two weeks and then correctly presented to the assistant on time (sharp deadline). Additionally just before the end of semester the students had to participate at a theoretical quiz. The last (and the most difficult) obligation for student was written final exam. The oral exams were performed in exceptional cases only.

Due to last year's changes the exercises are now performed in a different way. Only the initial part of each exercise is devoted to example presentation. In remaining time the students solve smaller programming tasks alone. If they don't complete the solution on site, they have to finish it at home. The solution has to be delivered before next exercise using Moodle. The big programming projects are preserved as well, but now the students solve only three middle size projects in two weeks. The solutions of projects are automatically assessed by a custom made tool. Instead of one big quiz many small automatically assessed Moodle quizzes are introduced. Each is devoted to a single lecture topic and contains 5 to 10 questions. As before the students have to pass a written exam, and the oral exam is performed rarely.

C. The assessing differences

In both previous and current approach the students earn up to 50 points before the final exam (at exercises) and up to 50 points at the final exam. To participate at the final exam they have to earn at least 25 points at exercises. 25 points are required to pass the final exam, too. Thus 50 points is needed for positive grade and the final grade is calculated from the sum of all points.

Previously (before academic year 2014/15) the total of exercises points was calculated as a sum of up to 40 points for programming projects (4 times 10 points) and up to 10 points for quiz. Currently the students earn up to 10 points with 12 small programming tasks, up to 30 points with three programming projects (3 times 10 points) and up to 10 points with 18 short quizzes. The final exam brings the same amount of points as before, but before it consisted of four programming exercises. Now the final exam consists of 6 shorter tasks.

The main difference is the use of automatic assessment. Before the Moodle was used to hand in the required solutions, but the assessment was performed manually by assistants. Now the programming projects use a special tool to hand in and assess solutions, and the quizzes are automatically assessed by Moodle. Only the small tasks and the exams are still checked manually.

D. The automatic assessment

A special tool was developed for the automatic assessment of programming projects written in Java. There are well-known tools (i.e. Marmoset [9]) and initiatives to automate part of assessment process [10], but we preferred to build a customized tool to exactly meet our requirements. Here a short description of functionality without technical details is given only.

The tool runs on a dedicated UNIX server, where all students and assistants have their own account. First each programming project has to be carefully configured by specifying tests (solutions for various inputs and points for of correct answers), the deadline and the time between two admissions. All programs use the precisely defined standard input and output format only. The students have to follow the instructions to correctly conduct the automatic assessment.

The students copy solutions for an active project to their account directory as one java file. Then they run a

script, which compiles the sources and executes several prepared tests. Each test can contain more inputs, and each result of test execution has to syntactically precisely match the predefined output. Successful execution is awarded with a predefined amount of points, and the points for all tests are summarized into final score. The complexity of tests usually gradually increases. It is possible to test different characteristics, e.g. the test can measure the execution time: if the calculation takes too much time, it fails.

After the execution of all tests the student receives the results on-line. Additionally, the outputs of all tests are in the predefined form placed into a dedicated directory. Students usually compile solutions frequently with the aim to concurrently solve issues and meet the desired grade. When the project assessment is over, a special plagiarism check (i.e. Moss [11]) is performed on all solutions. Only after manual control of all discovered similar solutions the grades become officially valid.

The automatic assessment of quizzes is conducted by Moodle. This system has some limitations, but it suffices for our needs, if we carefully prepare the questions.

III. ANALYSIS BACKGROUND

To make an objective comparison of both approaches we have to analyze the effect of introduced changes. Thus, we carefully select suitable data for analysis, prepare it to be correctly used and select appropriate statistical analysis to be performed on data.

A. The data selection

Previous ADS1 teaching performance is analyzed on data from four consecutive academic years: 2010/11, 2011/12, 2012/13 and 2013/14. The current ADS1 performance analysis use data from the academic year 2014/15 only. Four groups of data are used for analysis:

a) The course selection data is used to determine the exact number of ADS1 students per academic year;
b) The assessment results of laboratory work contain detailed data on points earned by student for all programming projects, tasks and quizzes. Only the sum of all points per student (namely **exercise points**) is used to determine the number and distribution of exercises results per academic year;
c) The examination results contain exact student's points for each task on each exam. However only the **final grades** for 14 exam terms are used to analyze the number and distribution of the examination results per academic year;
d) The data from anonymous ADS1 questionnaire are used to demonstrate student opinion on current course execution. The questionnaire was performed in January 2015. 31 students answered on 19 questions (with Likert [12] scale answers) and on one free-form answer.

B. The data preparation

Most data were obtained from the official faculty student information system StudIS [13], some from assistant's personal archive and from a specially prepared questionnaire. All collected data can be considered accurate and complete. Unfortunately, certain sets of data are not consistent and they must be manipulated to be properly compared.

a) As we use aggregated values only, the course selection data don't need any preparation.

b) This is not the case for the exercise data. Because of the unreliable data for unsuccessful students we use cumulative results for the students who collected enough laboratory work points to be able to participate at the final exam only. The number of unsuccessful students is still shown. As the required number of points for successful laboratory work changed over years, different limits were used for different academic years: 10 points in 2010/11 and 2011/12, 20 points in 2012/13 and 25 points in 2013/14 and 2014/15.

c) The examination data has two issues: (i) the unequal distribution of positive and negative grades and (ii) missing data for current academic year as only two terms are concluded by now.

   (i) The use of only one negative grade (5) in examination causes unequal distribution between positive (normal distribution) and negative (only one value) grades. This error propagates into inaccurate statistical results (i.e. mean value, standard deviation). By analyzing positive grades only a lot of important information in lost, so we act differently. Our solution is to change all values 5 into values from the range 5 to 1 in the same distribution as positive grades from 6 to 10 are. Such data preparation is artificial, but still much closer to the real student results.

   (ii) The successful students pass the exam on 1st or 2nd term, while the 3rd term is mostly reserved for bad students. As this year's 3rd ADS1 term will be in September 2015, we compared only the results for first two terms in each academic year.

d) The questionnaire data contains many questions about lectures, which are omitted in analysis. Thus we analyzed two questions on exercises execution in general, three on quizzes, three on small programing tasks, two on programming projects and one on automatic assessment. Additionally, the most notable free form answers are taken into account only.

C. The statistic analysis selection

After the careful selection and preparation of good quality data the statistical analysis are selected.

First, many general characteristics may be recognized by applying simple statistical analysis such as sample count, percentage, mean values and standard deviation values calculation. We perform such analysis on exercise data, examination data and questionnaire data.

Second, to statistically prove approximate equality the right statistic test has to be selected. As we have approximately normal distribution of exercise and examination data we could choose t-test, but z-test is a better choice as standard deviation can be computed. In our case we selected **large sample z-test for the difference between two mean values** [12]. To correctly perform z-test the necessary conditions must be checked first: the test variables must be independent, sample large enough and normally distributed. Histograms and normal distribution plots are used to show if the data is approximately compliant with normal distribution. Next, suitable hypothesis are formulated and tested. Technically we define an original hypothesis H_1 and an alternate null hypothesis H_0 for both cases, which we try to prove false during the process. Using this test we try to confirm with statistical significant probability (usually 97%) that the mean number of exercise points and grade of the previous and current course are related as hypothesis states.

Finally in case of non-structured data (as free-form opinion in questionnaire) a non-formal analysis is needed.

IV. ANALYSIS RESULTS AND DISCUSSION

A. *The simple analysis of the exercise data*

Results in Table I show the basic characteristics of the exercises data by academic year: the number of students, who selected the ADS1 course, the number (and percentage) of those, who succeeded to collect enough points and passed the exercises, the mean (average) number of exercise points and the standard deviation of those points. The first four (five) columns contain data for previous (manual) approach, while the last column presents data for current (automatized) approach.

TABLE I. THE BASIC CHARACTERISTICS OF EXERCISE DATA

	2010/ 11	**2011/ 12**	**2012/ 13**	**2013/ 14**	**2010 - 2013**	**2014/ 15**
Total number of ADS1 students	217	141	132	142	632	129
Number of students with enough points	104	84	87	111	386	108
% of students with enough points	47,93	59,91	65,91	78,17	61,08	83,72
Mean number of exercise points (μ)	27,05	25,98	28,52	30,75	28,21	34,06
Standard deviation of exercise points (σ)	7,79	8,29	3,94	4,40	6,61	7.45

The data in first two columns are unreliable as they contain "old" non-Bologna students, who were enrolled for ADS1 exercises, but who did not participate actually. In fact only the last two columns are directly comparable as in last two years the same rules were applied.

Still two clear tendencies may be recognized from the table: the group of successful students rapidly grows and the average number of points notably grows (in the last three years) as well. This demonstrates the success of new approach, thus a larger number of smaller tasks and automatized solution handling ease the work for students. The lower mean values of exercises points before 2013/14 are a consequence of different rules applied, however this is not the case for last two years.

B. *The distribution of the exercises points*

To gain more insight, the distribution of exercise points is analyzed. The Figure 1 presents the histogram of exercise points for five consecutive academic years (2010-2014). Considering the issues explained earlier we focus on the data for last two years (green and red line on graph), which prove similar characteristics.

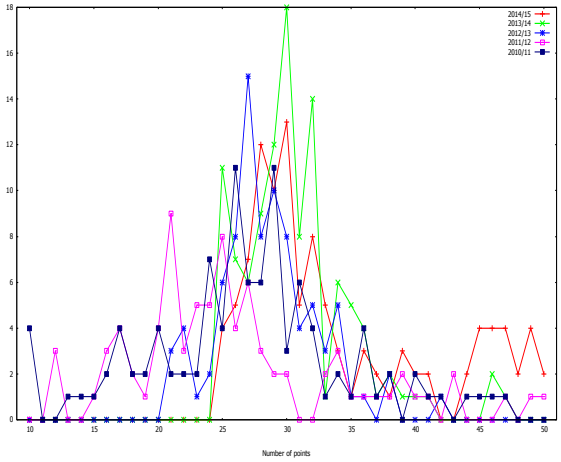


Figure 1. The histogram of the exercise points for academic years 2010/11 – 2014/15

Students are very motivated to collect enough points to pass the limit of 25 points, but to achieve more is obviously too demanding for most of them. There is a smaller group of excellent students with almost all points, thus the curve has a main peak in the middle and another one at the right edge. The data is not entirely normally distributed, but it is close enough as demonstrated in distribution plot (containing data for last two years) on the Figure 2.

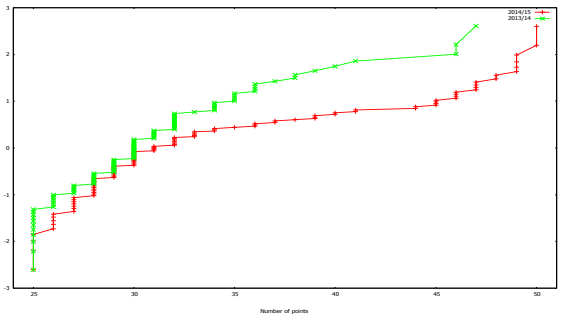


Figure 2. The distribution plot of the exercise points for academic years 2013/14 and 2014/15

C. *The hypotesis testing for the exercise points*

By hypothesis testing (using z-test) we can be even more accurate. All the necessary conditions are valid: variables are independent (as we compare results for different students in different academic years), we analyze much more samples than minimum 30 and our data is approximately compliant with normal distribution.

Let our exercise hypothesis state that the mean values of exercise points are now higher than previously. More precisely the difference between mean value of previous and current exercise points is 2 or more, while null hypothesis claims the opposite. If we define previous data as X and current data as Y then the hypothesis are (1):

$$H_{P1}: \mu_{Py} - \mu_{Px} > 2 \quad (1)$$
$$H_{P0}: \mu_{Py} - \mu_{Px} \leq 2$$

The mean values (averages) μ and standard deviations σ shown in Table 1 enable us to calculate **z-score** and **P-values** for H_{P1} and H_{P0} hypothesis. Table II summarizes the final results on the data for academic year 2013/14 and for all years together. The result for 2013/14 (111 totally compatible samples) is the correct one, while the other result (on bigger sample) requires careful interpretation.

TABLE II. THE EXERCISE HYPOTHESIS RESULTS

	z-score	P-value	H_{P0} rejected?
2013/14	**-3,14**	**0,0008**	**Y**
2010-2013	-11.44	<0,0001	Y

The P-value is in both cases very small and strongly rejects the null hypothesis, thus making our hypothesis confidently plausible. If we set the required confidence limit to usual 3%, the same can be claimed even for difference of 2.5.

D. The simple analysis of examination data

To demonstrate the basic characteristics of examination data a similar analysis was performed as in case A. Results in Table III demonstrate the number of exam terms, the number of all exams, the number and percentage of exams with positive grade, the mean positive grade, the mean grade and its standard deviation for same range of academic years. The data was previously manipulated as described before.

TABLE III. THE BASIC CHARACTERISTICS OF EXAMINATION DATA

	2010/ 11	2011/ 12	2012/ 13	2013/ 14	2010 – 2013	2014/ 15
Number of exam terms	3	3	3	3	12	2
Total number of ADS1 exams	172	120	130	160	582	119
Number of exams with positive grade	88	85	91	109	373	74
% of exams with positive grade	51,16	70,83	70,00	68,12	64,09	62,18
Mean positive final grade	6,96	7,03	6,97	7,31	7,01	7,51
Mean final grade (μ)	5,91	6,26	6,00	5,83	5,92	6,60
Standard deviation of final grade (σ)	1,88	1,85	1,75	2,07	1,85	2,25

All examination data can be considered valid as they reflect the result of similar final exam despite issues in the exercise data. The exercises points are evaluated as a part of total exam points, but only when the student collects enough points for correct exam task solutions.

The effect of unsuccessful "old" non-Bologna students may be recognized in pass rate for academic year 2010/11, but latter the pass rate becomes very stable till current year, when it suddenly drops. In light of exercise results this is a surprise, especially because of the missing third term data. Obviously the sample for current year is too small and actually reflects the effect of one exam term with too difficult tasks.

On the other hand the mean positive grades are very much synchronized, and, as expected, with clear growing tendency in last two years. More fluctuation may be recognized for mean value of all grades, but still the mean value for current data seems to be considerably better. Our further analyses are focused on all final grades.

E. The distribution of the final grades

In Figure 3 is the histogram of final grades for each academic year between 2010/11 and 2014/15. To be objective the data for the first and second exam terms are analyzed only.

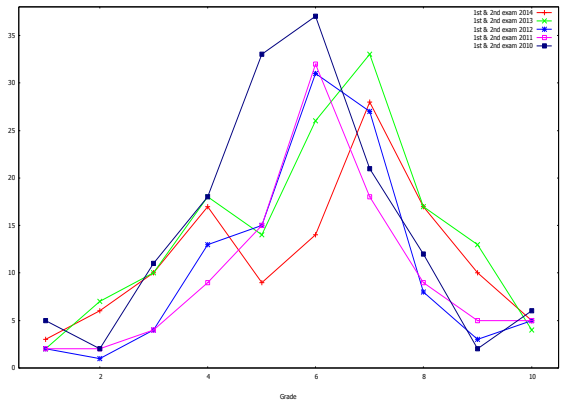


Figure 3. The histogram of the final grades for academic years 2010/11 – 2014/15

All histograms show approximately normal distribution with central peak moved to the right. Due to high average grade and previously described manipulation of negative grades the histograms for the last two years demonstrate even two peaks. The distribution plots of final grades for 1st exam term in years 2013/14, 2014/15 and for years from 2010 till 2013 are presented in Figure 4.

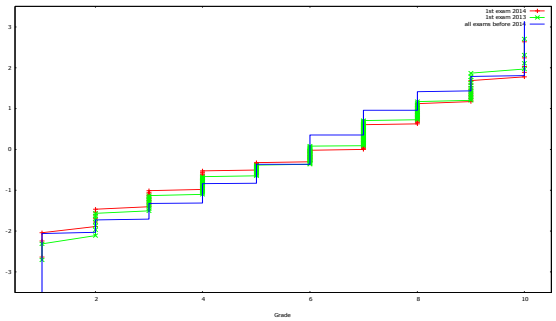


Figure 4. The final grades distribution plot for academic years 2013/14 and 2014/15

F. The hypotesis testing for the final grades

The hypothesis testing is applied to prove the final grades for current (automated) approach are considerably higher. The final grades hypothesis states the difference of mean final grades for previous and current data differ for at least 0.5. For previous data X and current data Y we define the following original and null hypothesis (2):

$$H_{G1}: \mu_{Gy} - \mu_{Gx} > 0.5 \quad (2)$$
$$H_{G0}: \mu_{Gy} - \mu_{Gx} \leq 0.5$$

The hypothesis testing follows the same steps as in case C. The results are presented in Table IV.

TABLE IV. THE FINAL GRADES HYPOTHESIS RESULTS

	z-score	P-value	H_{G0} rejected?
2013/14	-1,65	0,0495	N (3%), Y (5%)
2010-2013	**-2.34**	**0,0096**	**Y**

The hypothesis is strongly plausible as we consider all data to be consistent. If we use just the data from academic year 2013/14, the hypothesis may be confirmed with a lower confidence rate (95%) only.

G. *The questionnaire analysis*

To complete the picture the analysis of questionnaire results is performed. From 19 questions with answers in Likert [12] range from 1 (bad) to 5 (excellent) only 11 were considered relevant as they express opinions on (i) exercises in general, (ii) quizzes, (iii) programming tasks, (iv) programming projects and (v) automatized assessment. For each topic the questions were aimed on the quantity, the difficulty and the expected effect. The basic results are presented in Table V. The histogram with distribution of answers is shown in Figure 5.

TABLE V. THE BASIC QUESTIONNAIRE RESULTS

	Exercises	Quizzes	Tasks	Project	Automatic assessment
Number of answers	62	93	92	62	31
Average grade	4,19	3,48	3,46	3,38	4,35

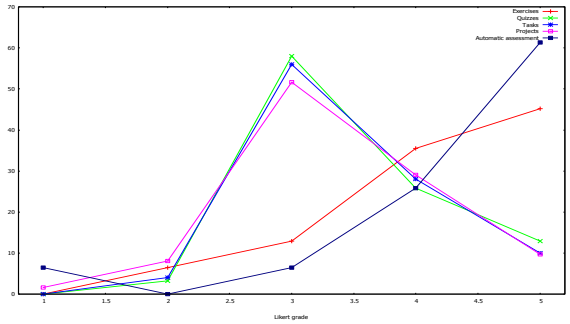


Figure 5. The histogram of the questionnaire answers

The results demonstrate that students like the new exercise approach and especially automatic assessment, while the results on different types of tasks are slightly positive only. Similar conclusion can be recognized from relevant free-form answers: the practical work was very well performed and at right level of difficulty, automatic assessment of programming projects is great. Some students like quizzes and programming tasks as they force them to study regularly, but on the other hand many of them complained over too much work with them.

Only one student would like to have more programming oriented lectures and exercises (as before). Still not much can be claimed as we don't have similar questionnaire results from previous years and current ADS1 students don't have experience with such course performance to be able to objectively judge.

V. CONCLUSION

After the careful analysis of different types of relevant data we established the following conclussions:

First, in many aspects (e.g. the distribution of exercise and examination data) the effects of previous (especially last year's) and current approach are totally comparable.

Second, due to considerable experience with ADS1 teaching, effort invested and accurate data collected we are able to confidently show that some results (e.g. the exercises points and final grades) are better than before.

Third, the automatic assessment works well and it is praised by students. It causes different structure of tasks and projects, which at the end enable higher grades.

Finally, the relation between grades earned and actual knowledge is not clear. Some incompatible results (e.g. low current exam pass rate) point out this question. Unfortunately, the actual knowledge is difficult to directly measure and we don't have better data to use. In this view the previous approach was better as the actual discussion with students enabled a better understanding of their issues.

Apart these results this paper presents a demonstration on how to use suitable statistical analysis to point out selected characteristics. In future we plan to perform similar analysis for other courses, which are in the process of modification.

REFERENCES

[1] "The Bologna Declaration" – Joint declaration of the European Ministers of Education, 19 June 1999, http://www.bologna-bergen2005.no/Docs/00-Main_doc/990719BOLOGNA_DECLARATION. PDF

[2] P. O. Duarte, M. B. Raposo, and H. B. Alves, "Using a Satisfaction Index to Compare Students' Satisfaction During and After Higher Education Service Consumption", Tertiary Education and Management, Volume 18, Number 1, pp. 17-40, 2012.

[3] V. Ivančević, M. Knežević, M. Čeliković, S. Aleksić, I. Luković. "Database courses: Curriculum changes and student results" in Proceedings of the 6th PSU-UNS International Conference on Engineering and Technology, pp.1-5, Novi Sad, Serbia, 2013.

[4] I. Rožanc, "One year later: the effect of the Bologna reform on Algorithms and Data Structures Course Teaching" in Proceedings of the WSEAS International Conference on Recent Researches in educational technologies, pp. 221-226, Korfu, Greece, 2011.

[5] Joint Task Force on Computing Curricula, "Software Engineering 2004: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering", IEEE CS and ACM, 2004, http://sites.computer.org/ccse

[6] "Študijski program Računalništvo in informatika (Dodiplomski – visokošolski strokovni)", B. Slivnik et. al., Faculty of Computer and Information Science, University of Ljubljana, 2008.

[7] I. Rožanc and B. Slivnik, "Analysis of elective courses selection in post-Bologna programmes" in Proceedings of the MIPRO 2014, pp. 862-867, Opatija, Croatia, 2014.

[8] *Moodle*, http://docs.moodle.org/

[9] *Marmoset*, http://marmoset.cs.umd.edu/

[10] L. Fürst, V. Mahnič, "A cooperative development system for an interactive introductory programming course". World transactions on engineering and technology education, Volume 10, Number 4, pp. 1557-1584, 2012.

[11] *Moos,* http:// theory.stanford.edu/~aiken/moss/

[12] William Navidi, *Statistics for Engineers and Scietists, Third Edition*, McGraw-Hill International Edition, 2011.

[13] *StudIS*, http://studij.fri.uni-lj.si/

Conceptual Design, Development and Analyses of "E-Learn SEEU" Model and Framework

Bekim Fetaji*, Majlinda Fetaji*, Orhan Rakipi* and Mirlinda Ebibi*
* South East European University, Contemporary Sciences and Technologies, Ilindenska bb, 1200 Tetovo, Macedonia
** International Balkan University, Information Technology, Skopje, Macedonia
b.fetaji@seeu.edu.mk, m.fetaji@seeu.edu.mk, or13562@seeu.edu.mk, m.ebibi@ibu.edu.mk

Abstract— **The research study is focused on the exploration and analyses of development methodologies of e-learning software and proposes a model and a framework for approaching such development, while applied it on a particular practical application E-Learn SEEU were has been implemented and tested the proposed model as well as evaluated with the users. Analyses of the current methodologies and best practices in e-learning, comparison of e-learning with m-learning as well as analyses of development of e-learning software has been realized. Also developed a software solution to test the proposed model and apply the framework as well as tested and evaluated it with the users using questionnaire. In conclusions insights are provided and recommendations are stated based on the study results realized.**

Keywords—e-learning, m-learning, e-learning software, model and framework for e-learning, software engineering reliability

I. INTRODUCTION

Today many technological innovations are redefining higher education, where students have to read many books, publications and articles during their studies. Teachers have to provide a number of lists of books and publications in many fields for their students. As an alternative teaching method, universities have adopted E-learning courses. By development of new pedagogical and technological methods of teaching and computer based devices, e-learning has rapidly emerged over the past few decades as an effective alternative to traditional education. E-learning models are attempts to develop frameworks to address the concerns of the learner and challenges presented by the technology so that online learning can take place in effective way [1]. Using web technology teachers can make their tasks more effectively. Instead of distributing materials separately to each of the students, a teacher can host it on the web so that students can download it when they need. According to [2], e-learning is also a very learner-centered form of education, and given its inherent need to archive teaching material, can be a useful platform for storing and sharing information. Teachers want to create course websites in a flexible way and quickly update and publish already existing Word, Excel or PDF documents. A Web-based teaching tool that enables to do such teaching tasks as distributing documents and exercises on the subject being taught will improve the teaching quality. Therefore, our purpose with this study was to explore the ability of an e-learning tool, for teachers and students for managing academic literature, and providing subject lists with literature for users.

II. LITERATURE REVIEW

There's a broad literature on e-learning published from around 2000. Different authors use different terminologies including online learning, Internet learning, distributed learning, networked learning, tele-learning, virtual learning, computer-assisted learning, web based-learning and distance learning instead of e-learning [3] [4] [5], which makes it difficult to develop a general definition for e-learning. So, there is no clear and explicit definition of the concept of e-Learning. Sigala [6] portrayed the Internet as an education platform for students, enabling them to receive educational materials and interact with teachers and peers in ways that previoulsy may have been impossible. According to [6], Internet is not the only method for delivering e-learning; in his definition includes several other methods such as TV, CD-ROM, utilized to provide distance education. Anderson [3] in his book used the term online learning throughout and he defines online learning as "[t]he use of the Internet to access learning materials; to interact with the content, instructor, and other learners; and to obtain support during the learning process, in order to acquire knowledge, to construct personal meaning, and to grow from the learning experience"; Itmazi [4] defines e-learning as "the use of new multimedia technologies and the Internet to improve the quality of learning by faciliating access to resources and services as well as remote exchanges and collaboration."

According to [5], the definition of e-learning comprises a lot more than online learning,virtual learning, distributed learning, networked or web-based learning. As the letter "e" in e-learning stands for the word "electronic", e-learning incorporates all educational activities that are carried out by individuals or groups working online or offline, and synchronously or asynchronously via networked or standalone computers and other electronic devices. Naidu [5], various types of e-learning activities down into the following modalities:

Fig. 1. E-learning modalities

2.1 Individualized self-paced, based on Naidu [5]

2.1.1 Online

In this type of learning an individual learner accesses learning materials online

2.1.2 Offline

In this type of learning an individual learner accesses learning materials offline (without Internet)

2.2 Group Based

2.2.1 Synchronous

In this type of learning a group of learners work together in real time via Intranet or Internet. Syncrhonous e-Learning is a method of e-Learning where interactions between the learners and instructors happen in real time [8]. It is commonly supported by media such as videoconferencing and chat, teleconferencing etc. The advantage of this type of learning is that it is independent of the geographical location and distance, and for learners online learning knows no time zones [3];

2.2.2 Asynchronous

In this type of learning a group work together through Intranet or Internet, but there is a time delay during exchanges among groups. Asycnchronous e-Learning is a method of e-Learning where interactions between learners and instructors happens independent of time and space. So, learners are in situation to interact with course materials and with instructors at a time of their choosing [8]. According to [10], the most common type of e-learning being used by different universities is asynchronous e-learning.

Research study by [7] on what university teachers think e-learning is good for in their teaching, where e-learning was conceived of as a medium, from the analysis found that e-learning as a medium can provide information to students, as a medium support 'occasional' online communication, as a medium help for engaging students in online discussions and as a medium support knowledge-buidling tasks. To ensure its success, e-Learning depends on its delivery method and content. And for this reason, e-Learning modules have to be interesting, interactive and informative in order to be effective [8]. The research discussed by [9] demonstrates that asynchronous and synchronous e-learning complement each other and support different purposes. The advance of e-learning has been influenced by numerous software programs, such as Moodle, Blackboard system, and others. So, students are enabled to be in constant contact with their professors virtually. A MOOC (Massive Open Online Course) is a course of study made available over the Internet (usually through videos and other materials) without charge to a very large number of people. According to [11], over the last few years, the MOOCs phenomenon has become widely acknowledged as crucial for freely accessible courses provided by international institutes for informal as well formal education.

III. ANALYSES E-LEARNING COMPARED TO M-LEARNING

Mobile learning or m-learning is an intersection of online learning and mobile computing [12]. That is an electronic learning (e-learning) accessed by mobile devices. Mobile Learning (m-Learning) is a new stage of the progress of distance learning (d-Learning) and electronic learning (e-Learning). According to [13] m-Learning is a new stage of e-learning that makes possible the ability to learn everywhere at any time with this enabling the well known paradigm "any time any location". M-learning is achieved by the use of mobile and portable devices such as PDA, cell phones, portable computers and Tablet PC [13]. One of the expected outcomes of m-learning is to minimize the dimensional constraint of time and location. In general it promised flexibility, convenience and the ability to learn at any time and any place and at learner own place.

IV. RESEARCH METHODOLOGY AND HYPOTHESES

The main goals of this research study are:

1. Analyze existing e-learning software systems.
2. Focus on issues that relate to the users' requirements and to establish them as structural elements.
3. Validate, test, evaluate software to ensure the proper fun ctioning of the software.
4. Develop a model to integrate, to engineer systems thinki ng, learning processes using the system and adequate doc umentation.
5. Utilize and use techniques to increase reliability during software development and describe the challenges for ac hieving these goals.

The main hypotheses addressed in this research study are as follows:

1. The use of the methodologies and best practices involved in development of e-learning
   software can be used to increase the "level of learning" (det ermined by "knowledge transfer" and "learning outcome") by using software.

2. Proposed Framework of recommendations and guidelines for learning systems can increase "level of learning"

V. ANALYSIS OF ISSUES IN E-LEARNING SOFTWARE

In order to establish and implement their e-learning system, institutions must carefully assess the factors that could affect the systems usability within the structure, because new systems can fail if the end users do not accept to use them. There are a lot of research studies in the area of e-learning but only small number of them addresses the critical area and issues in the field [15]. According to the experience of [16], implementation of e-Learning systems still facing a number of challenges, such as students IT skills, performance levels, cost, availability of resources, etc. [17] pointed out that "in general, like any information systems, user acceptance and usage are important primary measures of system success". Therefore, students involvement and acceptance of e-learning must be considered, as the major step toward implementing and developing a successful e-learning system [18]. E-learning system designers must decide which of the multiple available features should be implemented to assist specific target groups and maintain their satisfaction level [19].

VI. PROPOSED MODEL FOR DEVELOPING E-LEARNING SOFTWARE

Due to the large number of issues and challenges facing the E-learning, there is a need of new model, in order to explain and clarify the basic phases that study specific problems that may occurs at each phase.

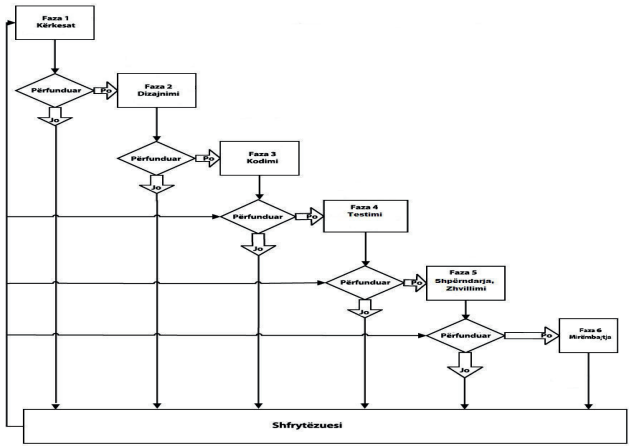


Fig 2. Proposed model of e-learning

In software engineering, the Systems Development Life Cycle (SDLC) concept underpins many kinds of software development methodologies, which can form the framework for planning and controlling the creation of an specified system [20]. There are many and different definitions and models for SDLC.

According to [23] , the systems developers must first identify the optimal requirements and definitions so that applications be as powerful and reliable. The proposed model is based on Waterfall model and includes six components (phases) of the process of developing reliable software for e-learning.

This model is based on strong management and proven tools, techniques and methodologies that are characterized by the key elements of the model. After each stage is defined, development goes on the next stages, so firstly one phase has to be completed before moving onto the next phase. After the requirement analysis has been done, the next phase was the design and coding using laravel framework [24] and PHP. After the coding, the following phases are testing, deployment and maintenance.

VII. CONCEPTUAL DEVELOPMENT OF "*E-Learn SEEU"*

e-learn SEEU developed as software for teachers and students, in order to create course websites in a flexible way and quickly update and publish already existing Word, Excel, PDF documents or any other format.

Search is the same as in Google; you can search either by author or subject and you can find uploaded materials that you care for only a few seconds, if available.

With a few more clicks, you can have your own account, create your own subjects where you and other registered users can personalize their profile and can upload materials. The search options is available to all visitors; create, upload and delete only to registered users.

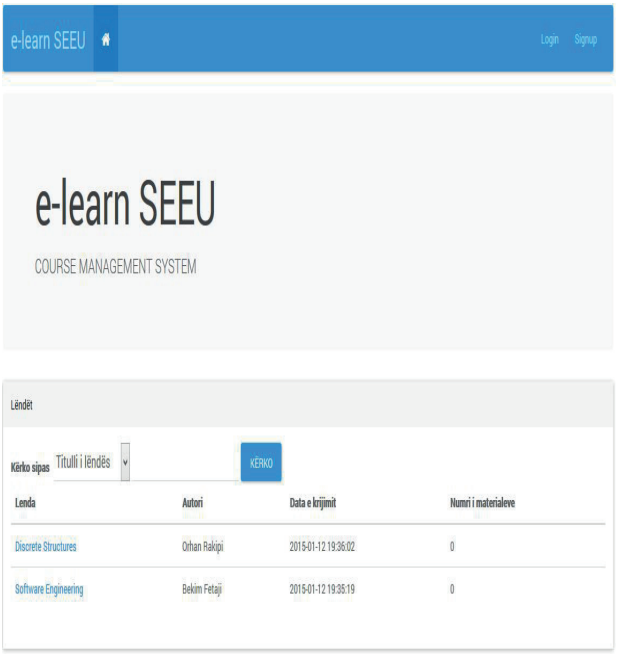


Fig 3. e-Learn SEEU

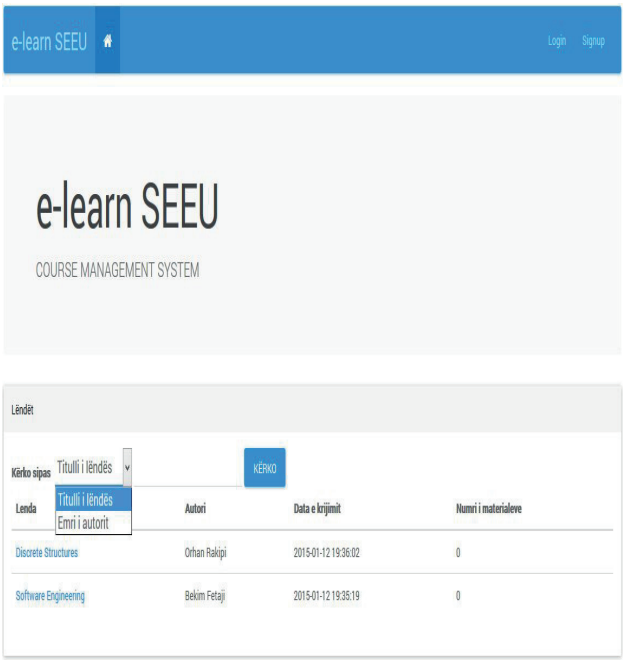


Fig 4. Search option, either by subject or author

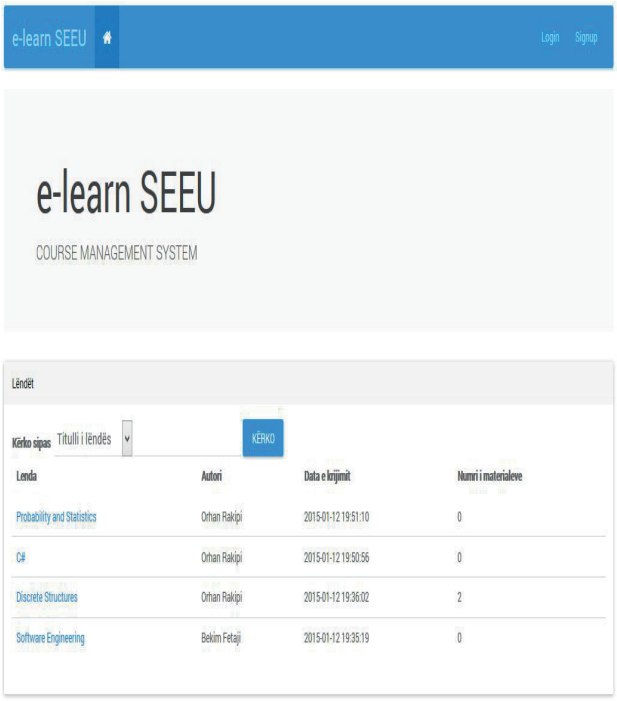
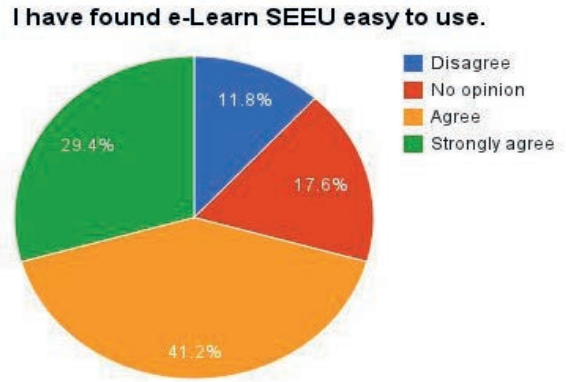
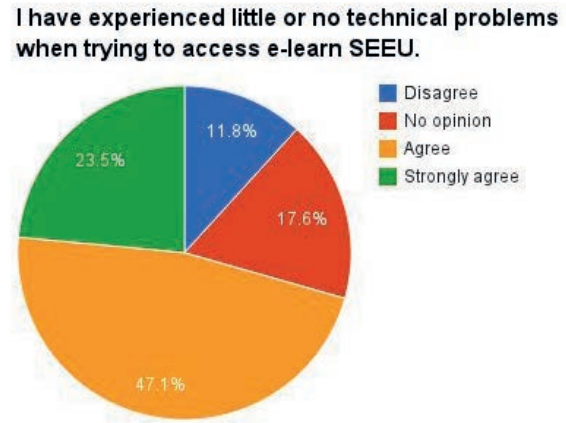


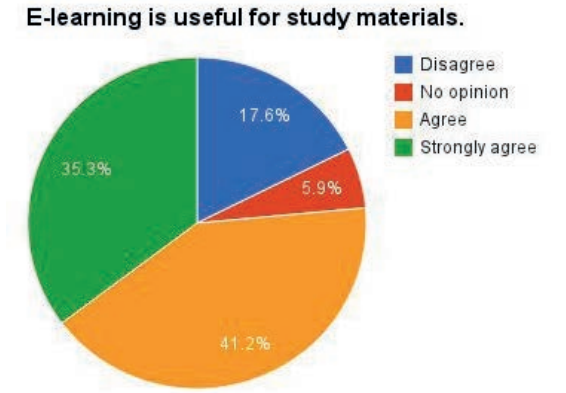
Fig 5. After creation of subjects and uploaded materials (date of creation and number of materials is shown)

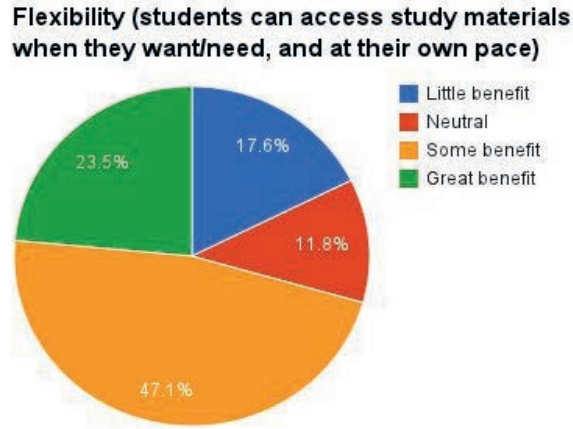
VIII. RESULTS AND DATA ANALYSIS

There were 17 responses from the participants in the questionnaire that had 8 questions and an option for additional comments and addressed the main identified issues regarding e-learning and the proposed model e-learn SEEU. Indicate the benefit level characterizing each one of the following qualities related to e-learn SEEU web application.









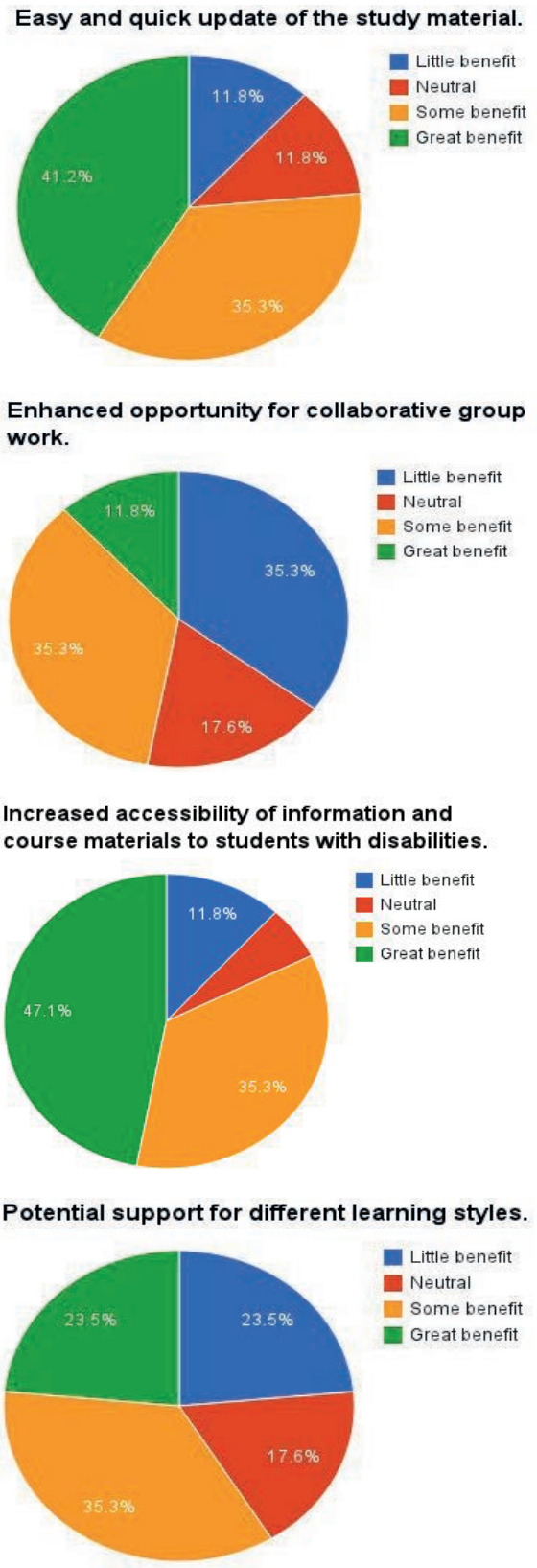


Fig 6. User Feedback Analyses

IX. PROPOSED FRAMEWORK

Laravel is a web application, framework with expressive and elegant syntax, that tries to ease the development by simplifying the common tasks used in the majority of web projects, such as authentication, routing, caching and sessions. It's accessible and powerful, providing tools needed for large applications. Inversion of control container, migration system, tightly integrated unit testing support, are some of the tools needed to build any application with which we are tasked [25]. Modern PHP frameworks, like Laravel, follow the Model-View-Controller (MVC). The MVC design pattern is very popular in the web development space, composed from the three components to the architectural pattern, which are the model, view and the controller [26].

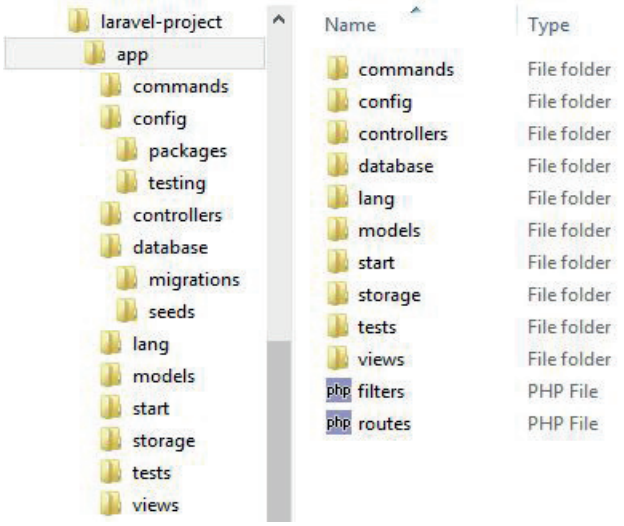


Fig 7. Laravel project directory structure

In order to implement e-learn SEEU, we have used Laravel framework with PHP, where the main pages are as follows:

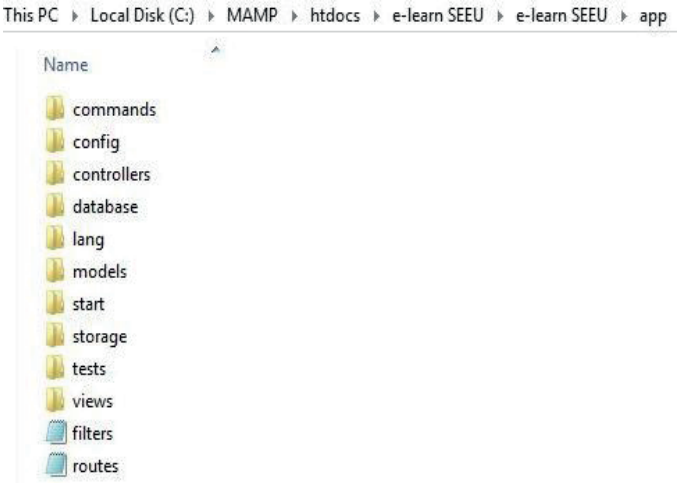


Fig 8. E-learn SEEU project directory structure

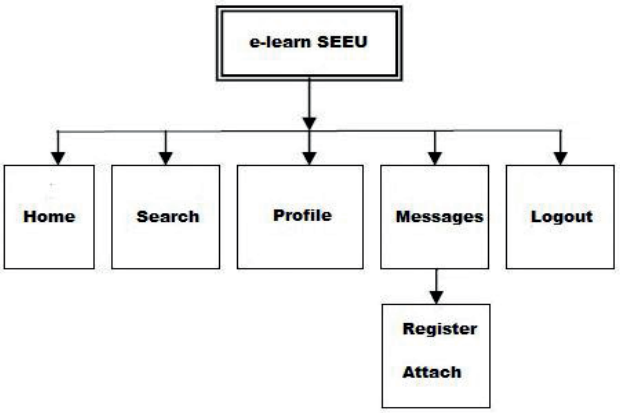


Fig 9. Overview of the e-Learn SEEU system

The version of the e-learn SEEU contains the following components :

- Home – move to the index page
- Search – any visitor can search either by subject or author
- Profile – any visitor can create his profile
- Messages – for any successful or unsuccesful registration, edit, or attachment message appears
- Logout – logout from the system

X. CONCLUSION

Based on the literature review, as the most significant steps toward implementing and developing a successful e-learning system are the students involvement and acceptance of e-learning, deciding which of the multiple features should be implemented, in order to assisst specific groups and maintain their satisfaction. Also, identifying the optimal requirements and definitions is considered as a major step in order application to be powerful and reliable.

We have reviewed the literature and developed an e-learn SEEU model that can faciliate interaction between student and lecturers virtually. The primary objective is premised on flexibility, benefits, easy to use, access and quick update materials through adaptation of different technologies. The participants in the questionnaire as an instrument to evaluate the system, have found e-learn SEEU to be a beneficial model and convenient platform. An effecient way to improve the system according them should be integration of the different modules, such as possibility for creating the online quizzes and taking them, audio and video lectures, etc, that should be considered as future work. Also they agree that e-learning can have great benefit for students with disabilities by increasing accessibility of information and course materials to them.

Based on the feedback and evaluations we can conclude that this study has indications to have positive effect on the improvement of the design concept e-learning software, as well as serve as a good reference for the further studies on the development of e-learning software.

REFERENCES

[1] Zhang, B.D.; Zhao, J.L.; Zhou, L.; Nunamaker, J.F.(2004). Can E-learning replace classroom learning In Communications of the ACM, 47(5),(pp.75-79).
[2] Engelbrecht, E. (2003) A Look at E-learning Models: Investigating their Value for Developing an E-learning Strategy, *Progressio,* 25, 2, 38-47
[3] Anderson, T. The Theory and Practice of Online Learning. AU Press, 2008. Available from:https://facultycommons.macewan.ca/wp-content/uploads/The-Theory-and-Practice-of-Online-Learning-by-Terry-Anderson.pdf
[4] Itmazi, J. A. eLearning. online. Available from: http://elearning. ppu.edu/file.php/1/eLearning/elearning.pdf.
[5] Naidu, S. E-Learning: A guide of Principles, Procedures and Practices, 2nd. ed ed. Commonwealth Educational Media Centre for Asia, 2006.
[6] Sigala, M., The Evolution of Internet Pedagogy: Benefits for Tourism and Hospitality Education, *Journal of Hospitality, leisure, Sport and Tourism Education*, Vol.1, No.2, 2002, pp. 29 – 45.
[7] Gonzalez, C. What do university teachers think elearning is good for in their teaching? Studies in Higher Education 35, 1 (February 2010), 61-78.
[8] Lado, A. Asynchronous e-Learning. Online. Available from: http://www.rau.ro/websites/e-society/lucrari/adrian%20lado.pdf.
[9] Hrastinski, S. Asynchronous and Synchronous e-Learning. Online, 2008. Available from: http://net.educause.edu/ir/library/pdf/EQM0848.pdf.
[10] Technologies, T. L. Asynchronous Learning. Online. Available from http://www.timelesslearntech.com/asynchronous-learning.php.
[11] Brown, S. (2013). Back to the future with MOOCs?. *ICICTE 2013 Proceedings,* pp. 237-246.
[12] Yordanova, K., 2007. Mobile learning and integration of advanced technologies in education, in B. Rachev, A. Smrikarov, & D. Dimov eds. 2007. *Proceedings of ACM International Conference on Computer Systems and Technologies*, vol. 285, Sofia, NY: AMC. Available at: http://dl.acm.org/citation.cfm?id=1330695
[13] Georgiev, T., Georgieva, E. and Smrikarov, A. (2004). Me-Learning: A new stage of E-Learning. Proceedings International conference on Computer Systems and Technologies, CompSysTech'2004, IV.28, 1-5.
[14] Allen, M.W. (2011) Michael Allen's 2012 E-Learning Annual. Pfeiffer, San Francisco.
[15] Hassan M. Salem, 2005"Critical success factors for e-learning acceptance: Confirmatory factor models" Computers&Education, vol.49 page396–413,
[16] O'Donoghue, J., Singh, G. and Handy, D. (2003) Higher Education—IT as a Catalyst for Change. *On the Horizon*, 11, 23-28. Available from http://dx.doi.org/10.1108/10748120310500099
[17] Saade, R., Nebebe, F. & Tan, w. (2007). "Viability of the "Technology Acceptance Model" in Multimedia Learning Environments: A Comparative Study". Interdisciplinary Journal of Knowledge and Learning Objects, 3, pp.175-184.
[18] Butorac, M., Nebic, Z & Nemcanin, D. (2001). "Blended E-learning in Higher Education: Research on Students' Perspective". Issues in Information Science and information Technology, 8, pp.409-429.
[19] Kastner, Margit, Stangl, Brigitte.. (Dis-) Satisfiers for e-Learning User Interfaces. In Proceedings of the International Conference on Information Systems (ICIS 2011), Hrsg. ICIS,1-19. Shanghai, China: AIS Association for Information Systems (2011)
[20] Kevin Roebuck, Systems Development Life Cycle (SDLC): High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Published by Emereo Pty Limited, 2011
[21] Bill Shackelford. Project Managing, E-learning. 2002 by the American Society for Training & Development. Reprinted 2005. Third printing.
[22] SDLC Waterfal Model. Available from http://www.tutorialspoint.com/sdlc/pdf/sdlc_waterfall_model.pdf
[23] Booch, G. (1994). Object-Oriented Analysis and Design with Applications (Vol. 2nd Ed.). Menlo Park, CA: Addison-Wesley.
[24] The PHP Framework For Web Artisans. http://laravel.com/
[25] Intoduction to Laravel [online]. Available from http://laravel.com/docs/4.2/introduction, Accessed: 15 January 2015.
[26] Architecture of Laravel Applications [online]. Available from http://laravelbook.com/laravel-architecture/, Accessed: 15 January 2015.

Hands-on System Programming Labs & Projects Based on E2LP Platform

B. Kordic*, V. Marinkovic*, M. Popovic* and I. Kastelan*
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
branislav.kordic@rt-rk.com

Abstract - This paper presents a set of laboratory exercises and student's projects developed for an E2LP embedded learning platform. The E2LP is educational and research embedded platform composed of base and daughter boards based on FPGA and ARM, respectively. The exercises and student's projects cover different topics in embedded and real-time system developments. The main goals of the exercises is to introduce writing Android kernel modules and devices, and multithreaded user space applications using cross-compiling tool chain for embedded platforms and systems. In course of their projects, students implement real-time support for various virtual and hardware peripherals such as RS-232, VGA, LCD, random generator number, encryption device, and etc. The exercises and projects are basic functional building blocks of a global project which provide simple text/chat communication between two E2LP boards. The achieved results are positive and promising.

I. INTRODUCTION

Industry needs for educated, high-skills, well-trained and experienced embedded engineers are constantly growing, which put high bars on a relevant university faculties. In general, active university teachers involved in different areas of a computer engineering has difficult task to fulfil this demand due to rapid technology changes as well as learning process changes. Teaching embedded systems and real-time system programming is always challenging and not easy task because beginners, for the very first time, face with bare hardware, complex and not so user-friendly tools, and minimal system software support. As a result of efforts to make embedded development more suitable and more productive for the beginners and teachers, in the recent past different case studies [1] and cognitive methodologies [2] were introduced to improve that process.

Laboratory exercises and projects are very important part in a student's education. It is essential because they provide a chance for students to tryout and evaluate their theoretical knowledge, and thus upgrade its practical experience. No doubts, maintaining laboratory has its own benefits and drawbacks [3]. Nowadays, in a different educational fields, a remote and e-learning technologies are widely accepted and are in common use, as well as a new technologies like augmented reality [4]. The remote learning laboratory for embedded learning was also introduced. The remote laboratory is equipped with different kind of embedded boards and measurement equipment tools which are monitored locally and controlled over the Interned within a Web browser [3] [5], which is a significant improvement in the area of embedded engineering education.

Maintaining a course laboratory syllabus aligned with industry needs and keeping it attractive and usable from student's point of view is difficult from many reasons: (i) a set of real-time system programming and embedded development exercises highly depend on hardware specification of underlying platform, and its migration to other platform are very time consuming, (ii) various embedded boards based on ARM, MIPS, FPGA, advanced microcontrollers and etc. are available on the market, but none of them, for a reasonable price, can't provide comprehensive set of available peripherals, and (iii) a system user interface is more likely implemented as a minimalistic interactive command line driven tool, which today is, unattractive for students and beginners which want to enter in a world of embedded systems.

A few years ago the E2LP project was introduced bold and ambitious project with its goals to create a unified, comprehensive, and viable embedded engineering learning platform for student's use [5]. The E2LP platform provides FPGA based base-board and two extension-boards equipped with ARM-v7 and DSP processors. Advantage of the E2LP board compared to other embedded platform on the market are manifold because it covers a very wide range of disciplines which embedded systems engineers should learn about – hardware design, digital systems, computer systems, signal processing, networking and communications, system and application software [5].

Professors and assistants at the University of Novi Sad, Faculty of Technical Sciences in the academic year 2014/2015 for a first time use the E2LP base and DSP based extension boards as the dominant learning platforms for a Real-time systems 2 course. After initial investment of significant time and effort to adapt existing and develop new exercises, which are not part of startup kit, for running on the E2LP boards, and if we take into account that we did not exceeded its full potential, seems that using the E2LP boards in long term will really pay off.

The rest of the paper is organized as follows: section one briefly introduce the E2LP architecture and architecture of DSP based extension board, chapter three in details describe student's exercises and projects, and section four conclude the paper with the final remarks and future work.

II. E2LP Platform Architecture Overview

A. The E2LP Architecure

As already introduced in the previous section, the acronym E2LP stands for *Embedded Engineering Learning Platform*. In its specification, the E2LP platform include a FPGA base board, a microcontroller extensions board based on ARM-v7 series, and a DSP extension board based on Marvell ARMADA 1500 processor [3]. The both extensions boards are also called the daughter boards. At the moment when exercises for a Real-time systems 2 course was developed, the DSP daughter board was fully functional and the microcontroller daughter board is still work in progress. Fig. 1 show the E2LP base-board and DSP based daughter board in its laboratory casing. In the Fig 1 is marked the DSP extension board.

B. The E2LP Base Board Overview

The E2LP base board is a backbone of the E2LP system. It is equipped with various peripherals which are intended to be used in Digital Systems Design Principles, Real-time systems, Operating systems, Digital signal processing and Network programming courses. The available peripherals covers wide range of functionality starting from a set of ordinary LE diodes and push buttons to a complex peripherals like an Ethernet. The comprehensive list of available peripherals on the E2LP base board is given in the Table I.

The core of E2LP base board is Xilinx Spartan-6 FPGA to which are connected all available peripherals. The FPGA provide more flexibility in a case that specific peripherals need to interconnect on a custom way. Of course, the nature of the FPGA also provide us the ability to implement additional functional components which are not already available such as pseudo-random generators, timers, communication protocols, signal processing filters, and etc. The FPGA keep the main role in the E2LP because it is used to handle most of signals between daughter boards and peripherals on the base board.

Concerning a software support, on the E2LP base-board are available (i) a mandatory house-keeping software which held the board in fully operational state, (ii) and a custom FPGA firmware. Depending of course

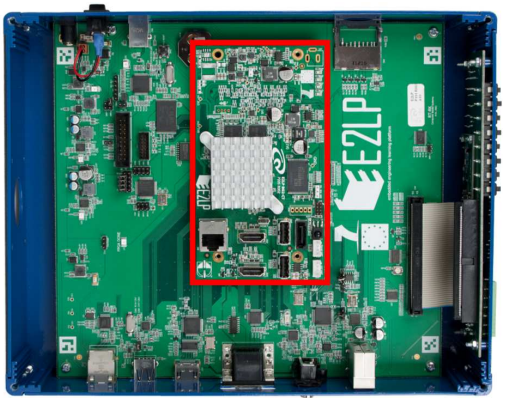


Figure 1. The E2LP FPGA Base and DSP Extension Boards in Laboratory Casing

TABLE I. The List of Peripherlas on The E2LP Base Board

Peripheral Class	Peripherals
High-speed communication	Ethernet, USB
Low-speed communication	RS-232, I2C, SPI, Infra-Red
User I/O	GPIO pins, LEDs, DIP Switches, Push Buttons, Snap-Wire Connector, LCD 16x2 Character Display
Audio/Video	Audio Codec, VGA, CVBS, HDMI
Mass Storage	SD-Card

the FPGA custom firmware may be different in order to suite the course specific needs.

C. The E2LP DSP Based Extension Board Overview

The E2LP DSP based extension board relay on a Marvell's ARMADA 1500 processor. The ARMADA 1500 is high-end dual-core ARM v7-compatible processor intended for use in cut-edge multimedia devices such as a Set-Top-Boxes, smart Digital TVs, tablets, and etc. With Marvell's QDeo video post processing technology the ARMADA 1500 processor is the central component of the Google TV 2.0 platform. In addition to the QDeo technology, it also integrates a 3D graphic core and audio post-processor for advanced video and audio processing algorithms.

The DSP based extension board is connected to the E2LP base board via Mezzanine connection. Using the Mezzanine connection and GPIO pins available on the extension board some of peripherals on the E2LP base board are also available on the extension board. For instance, from the extension board may be directly accessed to the LEDs, DIP switches, push buttons and RS-232. Peripherals on the DSP based extension board are mainly intended for exterior usage, USB, HDMI, LAN interfaces, and for debugging purpose, dedicated RS-232 interface.

The extension board based on ARMADA 1500 is suitable for implementing laboratory exercises and student's projects in field of 1D and 2D signal processing, low level and algorithm optimizations, computer networks, real-time systems, and different areas of application development. The application development is possible due the Android operating system. One of the most attractive and exciting regarding the DSP based extension board is the fact that on-top of its hardware is running a custom version of Android *Jelly Been* OS. The custom version of Android JB OS is ported in order to support hardware features available on the E2LP boards. Further, the Android itself, as the fastest growing operating system on the market today, provide to teachers and students a chance to develop different kind of application and system software from JAVA and WEB based Graphic User Interface (GUI) applications to low level Linux kernel and device driver development.

III. System Programming Course Overview

In the recent past a set of laboratory exercises in field of a digital systems design for a FPGA based board like

the E2LP base board was introduced [6]. Further, the presented results based on student's questioners tell us that even more changeling exercises could be attractive to the students. Based on lessons learned, the exercises and student's projects developed for the Real-time systems 2 course try to comprise the core knowledge about real-time and embedded system programming with the challenging and attractive problems which may be interesting to the students and beginners in the field of embedded systems.

During the course, both the E2LP base and DSP based extension boards are used. The E2LP (FPGA) board is used to handle low level signal communication between peripherals on the E2LP base board and software modules on the DSP based extension board. In the Fig. 2 E2LP software stack layers architecture is given. The house-keeping software ships with the E2LP base board and it is not relevant for the course. The FPGA custom software is specially tailored to suit the needs of the course. Due to the limitation that only a few peripherals located on the E2LP base board are interconnected with the E2LP extension board through the Mezzanine connection, a simple communication protocol which overcome this drawback is developed. In its essence, the communication protocol provide control of all peripherals available on the E2LP base board using the GPIO pins from the E2LP extension board side. More information about communication protocol, how it work, and which peripherals are supported is available in [7]. In the Fig. 2 the Android kernel and user space which is running on the E2LP daughter board are depicted – daughter board based on Marvell ARMADA 1500 processor. All exercises and projects are performed on the daughter board.

A. Laboratory Prerequisites

From a prerequisite's point of view, the laboratory hardware equipment and software needed for class lecturing are quite ordinary, and it follows:

1. The E2LP kit (may share among the students)
2. A PC running any Linux distribution
3. A cross-compile toolchain and an Android Debug Bridge (ADB) tool

The E2LP kit include the FPGA base and DSP extension boards. One E2LP kit may be shared among the students during the exercises. The cross-compile toolchain and ADB tool is necessary to compile a source code for ARM architecture and to download binaries to the target (Marvell's ARMADA 1500) platform, respectively.

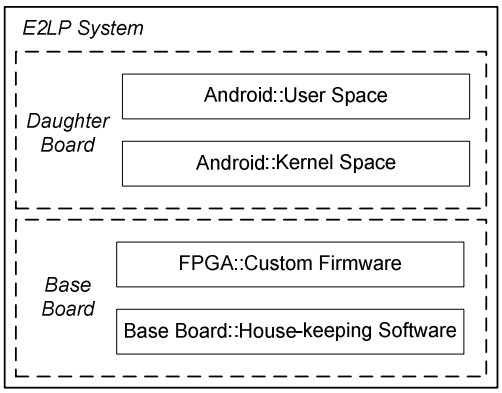


Figure 2. The E2LP Software Stack Layers Architecture

B. The Exercises Overview

The exercises are divided in a two subclasses. The exercises from both subleases are in according to the taxonomy described in [2]. The first subclass is about user space programming. It deals with common multithreading programming topics. The main goals of the exercises which covers user space programing are following:

1. Basics about using cross-compiler toolchain and embedded tools
2. Introduction to the multithreading concepts using POSIX standard (*pthread* library)
3. Thread synchronization and data-race problems
4. Introduction to mutual exclusion objects and semaphores
5. The consumer-producer problem implementation

The second subclass of the exercises cover a Linux kernel and device driver development. As the exercises regarding user space development, the exercises about the Linux kernel development include demonstrative exercises (tutorials) and exercises with specific problems which students have to solve. The Linux kernel development exercises demonstrate how kernel modules (drivers) should write and how it works in real operating systems. The following topics are covered:

1. Write a simple and complex (virtual) device drivers using referent examples
2. Develop a simple GPIO driver to handle LEDs and push buttons on the E2LP base board

C. The Projects Overview

In general, each project consists from its user space and kernel space part. In the kernel space part functionality which students have to implement is defined. Not necessary, but in the common case, the students have to develop a Linux device driver for a specific peripheral, i.e. to provide real-time system handler for a particular peripherals on the E2LP base or daughter board. The exceptions are virtual devices and devices implemented in the FPGA such are encryption device and random generator number device. The user space part is used to develop a simple test application which should implement a relevant use cases in order to tests driver's basic functionality. For instance, in the case of a RS-232 (UART) driver, the user space application is used to test send and receive functionality. Based on the student's affinity, it also may choose project which is entirely done in the Linux user space. Also, depending on a projects difficulty and complexity, the students may be split into a groups of two or three students. In Table II is given a basic set of the projects with functionality which should be implemented.

The projects described in the Table II defines a set of basic functional blocks which are used to compose a simple text/chat application on top of the E2LP platform.

TABLE II. PROJECT'S BRIEF DESCRIPTION

Linux device driver	Functionality
RS-232	Implement read/write commands using *pooling* mechanism.
RS-232	Implement read/write commands using *inerrupts* mechanism.
LCD	Implement read/write and reset commands. Enable text rotating.
VGA, HDMI, CVBS	Implement resolution change, text and background update, font size and font color change.
Encryption	Implement encryption functionality using random generator number in FPGA. Provide random generator's seed initialization.

The simple texting application is used to demonstrate sending and receiving a short text message between two E2LP platforms. In application composition VGA or other video output interface like HDMI or CVBS are involved, as well as RS-232 and encryption module for communication between two boards module. The LCD could be used to show communication protocol's commands which is developed in user space application. In the Table II are not shown the projects regarding communication protocol development. As mentioned before, if students are interested in user space applications they may work on a communication protocol development. It is mandatory that the protocol relay on the RS-232 device driver.

The idea behind the student's projects integration in a simple small scale global project is to demonstrate integration process which is inevitable part of any industrial project. At his phase, the students have to modify and fix an issues raised in the integration phase. The integration process is done by the course instructor.

IV. CONLUSION

In this paper we presented a set of student's laboratory exercises and projects intended to be used on the E2LP base and DSP based extension boards. The exercises and projects are mainly developed for a course about real-time system software, and embedded systems, but it may be used as solid background for other kind of courses and trainings. During semester, in the Linux based operating system students learn how to develop and integrate (i) user space applications, (ii) write multithreaded applications, (iii) develop kernel module drivers and devices, and (iv) write drivers for real peripherals using *pool* and *interrupt* mechanism. Students also learn how to use cross-compiler toolchain and embedded debugging tools which are common in embedded systems. At the end of semester, the output is a simple text/chat application which is composed of student's projects. The achieved results, based on a student's feedback, are positive and promising.

Test and verification process plays important role in the education of future engineers. The presented set of laboratory exercises and projects do not cover any of test and validation frameworks such as CUnit, CppUnit, JUnit, or other kind of frameworks which are using script languages. In order to improve this area of education, in the future it is necessary to make one of the frameworks available for running on the E2LP DSP extension board.

Recently various embedded boards are equipped with a couple of cores and plenty of operational memory, which provide to end user more space to develop and investigate more complex and time consuming tasks like image processing or non-trivial mathematical and chemical simulations. Along with supporting the test and verification frameworks, the goal in further work on the E2LP boards is to provide to students one of the powerful frameworks for parallel processing such as Intel's Threading Building Block and Cilk, and OpenMP.

No doubt, the main advantage of E2LP board is the DSP based extension board with Android OS support. This provide ability to teachers to develop a different classes of exercises targeting topics in communications networks and protocols, real-time and non-real system software, as well as JAVA, GUI, and WEB based application software. In the future, good compromise between the available E2LP boards and software development is to combine already developed exercises for the Real-time system course 2 and define new exercises, in the area of GUI development, which can be used to control the underlining hardware. This approach provide to students comprehensive view on computer stack architecture in one modern operating system like the Android.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No 317882.

REFERENCES

[1] I. Kastelan, M. Katona, D. Majstorovic, E. Neborovski: "Students perspective of the embedded engineering learning platform - A case study in digital design", Proceedings of the 34th International Convention MIPRO, 2011, pp. 1178-1182

[2] I. Kastelan, M. Barak, V. Sruk, M. Anastassova, M. Temerinac: "An approach to the evaluation of embedded engineering study programs", 36th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO, 2013, pp. 742 - 747

[3] Temerinac M., Kastelan I., Skala K., Medved Rogina B., Reindl L, Souvestre L. et al. "E2LP: A unified embedded engineering learning platform" Euromicro Conference on Digital System Design DSD, 2013, pp. 266 – 271

[4] Jorge R. Lopez Benito, Enara Artetxe Gonzalez, Margarita Anastassova, Florent Souvestre: "Engaging computer engineering students with augmented reality software for laboratory exercises", IEEE Frontiers in Education Conference FIE 2014, Madrid, Spain, 22-25 October 2014

[5] Ivan Kastelan, Miodrag Temerinac, Nikola Teslic: "E2LP: An embedded engineering learning platform", IT Systems Conference 2013, Bol, Croatia, 12-14 September 2013

[6] I. Kastelan, D. Majstorovic, M. Nikolic, J. Eremic, M. Katona: "Laboratory Exercises for Embedded Engineering Learning Platform", Proceedings of the 35th International Convention MIPRO, 2012, pp. 1113-1117

[7] V. Marinkovic, B.Kordic, M. Popovic, I. Kastelan: "Supporting System Programming Labs & Projects in FPGA on E2LP Platform", unpublished

Wiki-based Books as Alternatives to Printed University Publications

M. Sakal*, P. Tumbas*, P. Matkovic* and V. Pavlicevic*
* University of Novi Sad, Faculty of Economics, Subotica, Serbia
marton@ef.uns.ac.rs

Abstract - Universal dissemination of digital technology has produced a convincing alternative to printed university textbooks, workbooks and handbooks, changing the approach to their production, distribution and use, whilst irreversibly transforming the reading habits and means of knowledge acquisition of the student population made of members of the Net generation. Being continuously and ubiquitously exposed to digital and network technologies while growing up, they effortlessly and instinctively developed ICT skills at an admirable level. Multitasking, being on-line and use of collaborative environments come naturally to current generations of students, which is an essential prerequisite for the introduction and use of Wiki-systems as an alternative to printed university literature. Positive and negative aspects of such systems are the subject of analysis in this paper, as well as the proposal of a possible scenario for developing Wiki learning materials in a particular environment.

I. INTRODUCTION

Despite the evident increase in sales of e-books (Forrester Research predicts that revenues from e-book sales will amount to 3 billion dollars in 2015 [1]), research results suggest that printing business is not losing momentum: more than 2 billion copies of books, 350 million copies of magazines and 24 billion of newspapers are printed in the United States alone, which consumes around 68 million tons of paper [1].

Business aspects are present in the university textbook publishing as well. The key motif is the potential profit, divided amongst a great number of participants in the chain from authors to students: editors, publishers, printers, distributors, persons involved in the development of study programmes, etc. [2]

Although, according to some sources, electronic textbooks are deemed more cost-effective – given that they do not incur printing costs, distribution costs are reduced to a reasonable minimum, there are no holding costs associated with (obsolete) inventory, and the fact that they can be printed in full or partially if necessary [2] – opposite opinions have emerged. They point out some particular costs that do not exist with printed textbooks: hardware costs (e-readers), costs of software updates, potential costs of replacing lost hardware, etc. [3]

Even if there are reservations about cost effectiveness of e-books, it seems that there is no doubt about their following advantages: digital textbooks contain up-to-date, current information presented using text, graphics, videos and other formats; they enable quick search, text highlighting, electronic comments, incorporate easily searchable glossaries and other tools that support reading and making notes. In addition to that, students can take many books with them, owing to the portability of e-readers, which was not the case with printed books. However, issues associated with e-books have emerged, such as readers' preference of printed textbooks over e-books, legal consequences of using and sharing materials in electronic formats etc. [2]

In context of the debate between printed and electronic books, Wiki-textbooks present themselves as sort of a hybrid. Developed using Wiki technology, they combine attributes of both printed and electronic textbooks, as well as classic Wikis, with the aim to make the most of their advantages and overcome their limitations.

Wiki-textbooks, or more specifically, students' participation in their development, can also be observed in the context of acquiring Web 2.0 literacy and knowledge management competencies. Collaboration in finding existing knowledge, creating and/or acquiring new knowledge, storing, retrieving, applying and re-use of knowledge [4] is a necessity that a knowledge-based economy imposes on modern organizations. Realization of these activities is virtually impossible without Web 2.0 technologies [5]. Therefore, Web 2.0 literacy and knowledge management operations are competencies that students need to gain throughout the course of their studies, in order to suffice in skills and knowledge required for applying for knowledge-based jobs immediately after receiving their degrees.

Considering that the traditional way of learning, based mostly on acquisition of knowledge from conventional printed (but also electronic) textbooks, workbooks and handbooks, does not fully facilitate acquisition of aforementioned competencies, this paper builds on a presumption that they can be significantly improved thru inclusion of students into collaborative development of Wiki-based university content that would become an alternative to printed and electronic university publications.

Reasonableness of this assumption is explored thru an overview of learning ways of the Net generations (to which most of the current students belong), examination of possibilities of Wikis as e-learning tools, contrasting positive and negative experiences of introducing Wikis in university education and an outline of selected segments of research that the authors had carried out on students at the Faculty of Economics in Subotica, which was focused

on their readiness to participate in the development of a Wiki system.

II. NET GENERATION

Digital and network technologies have had a determining impact on the current generation of students throughout their childhoods. Hence, they have effortlessly and instinctively developed ICT skills at an admirable level [6]. Networking, multitasking, interactivity and use of collaborative environments have become a part of their everyday activities [7, 8, 9, 10], which is reflected in terms widely used to describe this generation: the Net generation, millenians, www generation, digital generation, N-generation, generation E, cyberkids, Dot-com generation, Next generation etc. [11, 12, 13, 14, 15, 16]. However, besides the ICT skills, their distinctive characteristics involve individuality and narcissism [17], as well as the yearning to be entertained in every situation [18].

Sophisticated use of digital technology and a significantly different approach to information and media has led to changes in ways that members of the Net generation think and learn. At the same time, expectations towards the education system have undergone changes [19]. As far as in 2001, Prensky [7] had determined that there are divergences between the Net generation's way of learning and one imposed by the existing educational system: changes in the Net generation's way of learning are not just incremental, but there is an evident discontinuity in the way information is processed. Present generation's way of thinking is fundamentally different than one of their predecessors. Members of the Net generation are also called 'digital natives', while professors – who were not born in the digital era, but became fascinated by it at some point of their lives and accepted many of its benefits – are referred to as 'digital immigrants' [7]. The latter, speaking a language learnt in the pre-digital era, in fact "... are struggling to teach a population that speaks an entirely new language" [7].

The idea of substituting the traditional, industrial model of education with collaborative learning has emerged as a solution to a growing gap between 'digital natives' and 'digital immigrants', for which "we need an entirely new modus operandi for how the subject matter, course materials, texts, written and spoken word, and other media (the content of higher education) are created" [20]. Considerations on new ways of teaching and learning presented in Figure 1. speak in favour of aforementioned assumptions.

Traditional learning	**New learning**
Teacher-centered instruction	Student-centered learning
Single-sense stimulation	Multisensory stimulation
Single-path progression	Multipath progression
Single media	Multimedia
Isolated work	Collaborative work
Information delivery	Information exchange
Passive learning	Active/inquiry-based learning
Factual, knowledge-based learning	Critical thinking/informed decisions
Reactive response	Proactive/planned action
Isolated artificial context	Authentic, real world context

Figure 1. New ways to teach and learn [9]

III. WIKI SYSTEMS

Wikis are one of the most popular Web 2.0 applications. Of several key concepts that have led to the explosive increase in popularity of Web 2.0 applications [21], Wikis mostly rely on Wisdom of crowds and Architecture of Participation. Wisdom of Crowds, as one of the manifestation forms of the Harnessing the Power of the Crowds principle [21], "is the basic idea of Web 2.0-style thinking, starting from the viewpoint that the solution to the problem, proposed collectively but independently by individuals comprising a large group, the so-called 'crowd', is better in quality than a solution offered by the most intelligent group member" [22]. Architecture of Participation reflects the idea that a Web 2.0 based service should be easy to use. Such characteristics eliminate obstacles to accessing and using the service, encourage and ease mass participation in its usage. Furthermore, mass participation contributes to the quality of a Web 2.0 based service, owing to the effect of principles of wisdom of crowds.

In practice, these ideas are implemented in a way that Wiki users interactively, collaboratively and incrementally edit existing and add new content to the Wiki system, thus constantly increasing knowledge repositories under permanent community based peer revision.

Eide and Eide [23], who studied the potential of applying Wikis in education, have pointed out that they promote critical, analytical, creative, intuitive and associative thinking and represent a powerful medium for accessing and sharing quality information.

Kokkinaki [24] highlights the potential of wikis reflected in the fact that it gives the student, as an individual, the opportunity to give their way of learning a specific identity, while improving their teamwork skills at the same time [22]. Karasavvidis [25] believes that, for several reasons listed below, Wikis have the greatest potential of all Web 2.0 tools:

- A regular web browser is sufficient for their use;
- Level of skill necessary to create and modify Wiki content is relatively low;
- Wikis facilitate individual work, as well as collaboration, communication and evaluation;
- Wikis support different types of educational activities, such as: collective learning, collective generation of databases, improvement of interaction among students, on-line classes and assessment, collaborative generation of materials, creation of Wiki textbooks, etc.

Wikis can also be used for brainstorming, group projects, meeting, support, creating various lists, (e.g. glossary of terms), preparing answers to FAQ, building repositories of relevant documents, assembling link collections (social bookmarking), building a group portfolio [26], posting homework/course material/study guides, peer review and feedback [9] etc. In addition to that, Wiki does provide students with unconstrained access in terms of place and time, allowing each individual to choose the work pace that suits those best [5].

Hypothetically, adoption of Wiki-technologies should stimulate positive qualitative changes in the education system. However, these expectations have not been completely realized (or not even to a desirable extent) [5], which can be concluded from the results of several studies that will be referenced in the following section.

IV. University Experience in the Application of Wiki Systems

The common idea in all examined studies is that development of Wiki-textbooks necessitates significant efforts of both teachers and students [2, 25, 27, 28, 29, 30]. These studies pointed out several positive aspects of Wiki-textbooks:

- Students are better motivated for creating quality content due to the awareness that their work will be presented to the public (even just within the group);
- Process of developing wiki textbooks stimulates group work, collaboration, knowledge sharing;
- Assigning credits (ETCS) for correcting other' errors motivates students to read those parts of Wiki-textbooks they normally would skip;
- Development of Wiki-textbooks results in better familiarization with the learning materials;
- Teachers work more closely with students and the group as a whole;
- Students become aware of the responsibility associated with the decisions what to include and what to exclude from Wiki-textbooks;
- Students who participate in the development of Wiki-textbooks achieve better grades in their exams.

However, the mentioned studies [2, 25, 27, 28, 29, 30] point more to the negative aspects, that is, problems associated with Wiki-textbooks. Of all the mentioned problems, the following particularly stand out:

- Challenging coordination of implementation activities;
- Failure to complete activities within the set timeframe;
- Students' concerns about fairness of the system used for evaluating (assigning credit) of efforts put into developing parts of Wiki-textbooks.
- Frequent frustrations caused by the lack of skills necessary for using Wiki-systems.
- Students' general need to protect (so to say, copyright) their ideas, i.e. parts of Wiki-textbooks they worked on, combined with the simultaneous disregard for their colleagues contribution;
- Anxiety due to public availability of texts entered in the Wiki-textbook (this was identified as a motivational factor as well);
- Contrary to the expectations, altruism often did not develop among students (or at least not to a desirable level);
- A relatively small number of students provide significant contribution to the development of Wiki-textbooks, while contribution of most students is negligible;
- Student prefer lower grades over the obligation to use a Wiki-system and participate in on-line collaboration;
- Students prefer individual work over on-line collaboration;
- Students' satisfaction from using Wiki system and collaboration in the content development is not at a desirable level;
- Students consider that they would achieve better final grades if they work individually;
- After a certain amount of time, there were no new contributions to Wiki;
- Students regard development of Wiki-textbooks as an excessively demanding task, both in terms of time and effort to be put into writing, entering content, reading, etc.;
- Plagiarism: copy-paste has become a regular method for 'loading' Wikis;
- Students complain over limited possibilities for communication (discussions) thru Wiki systems;
- Apart from live discussions in the classroom, online discussions within the wiki system are scarce;
- Wiki-textbook development project encourages competition more than it encourages collaboration;
- Students are sceptical about their understanding of certain concept;
- Students are reluctant to contribute and make changes to the text entered by their colleagues, although they have constructive ideas;
- Unbalanced 'depth' of content: superficially written texts were mainly a result of a copy-paste strategy and considerably greater in length than texts that were more devoted to the essence of the issue and required more work;

- Students are used to learning from authoritative texts that mostly reflect dominant theories within a discipline, authored by leading experts in their respective fields; texts originated from collaboration on developing Wiki textbooks were not deemed authoritative by students;
- Students were worried that every mistake, vagueness or inconsistency in their texts might adversely affect their final grade.

As a result of the analysis of identified positive and negative experiences related to development of Wiki-textbooks presented in the literature, the following suggestions and recommendations were proposed [2, 25, 27, 28, 29, 30]:

- Organize an in-depth training for using Wiki-systems and give students more time to adapt to a new medium;
- Define explicit rules on participation, convey them to students in a clear manner and insist on the netiquette;
- Instructor should convince students that revision is a natural, inevitable part of creating Wiki-content, and that it does not imply that the original text was wrong;
- Consider giving a list of chapters and titles for the future Wiki-textbook before the development commences.

V. RESEARCH RESULTS

On the basis of mentioned experiences in the introduction of Wiki systems into university education, with specific traits of the Net generation in mind [31], the authors have conducted a research on students' willingness to participate in the development of a Wiki system at the Faculty of Economics in Subotica [5]. A summary of research results is given in the following text.

A questionnaire was sent to the addressees of 216 third-year and fourth-year undergraduate students, selected by random sampling. 75% of students have responded to the request to participate in the survey. 90 of the respondents were male (55.56%), while 72 (44.44%) were female. Structure of respondents according to the grade point average criterion conforms to the normal distribution: 9.88% students had a grade point average between 6.00 and 6.99, 39.51% between 7.00 and 7.99, 38.27% between 8.00 and 8.99, while 12.35% respondents had a grade point average of 9.00 or more. 100% of respondents stated that they use computers and internet. 81.25% of respondents stated that they have the ability to access the Internet using their smartphones, which goes in favour of the assumption that students are able to access the (future) student Wiki virtually anytime and anywhere.

All respondents (100%) stated that they use Wikipedia. The greatest portion (60.49%) of students use it on a weekly basis. However, there are significantly less active Wikipedia users: only 12.35% of respondents had participated in creation or modification of its content. In spite of relatively frequent use of Wikipedia, one third of respondents do not use it when preparing for exams. It can be said that students who use Wikipedia for preparing exams do not have a clear vision on its potential benefits for this purpose: 49.05% of respondents have stated that Wikipedia is useful for exam preparation, i.e. 50.95% disagreed with that claim (10% = "I agree completely", 0% = "I disagree completely"). Moreover, less than 10% of respondents have stated that they do not verify information obtained from Wikipedia at all, 62.96% verify them sometimes, while 14.81% do it often.

Students have rated the idea to develop a student Wiki on a certain topic within lectures and exercises thru student-teacher cooperation with an average score of 4.40 (1=lowest score; 5=best score).

Proposals that the content of student Wiki be a part of the exam material and that activities on building the student Wiki be additionally valued in the assessment of the final grade were scored similarly high: the average scores were 4.25 and 4.16, respectively. The idea that each student would be obliged to participate in the development of the student Wiki was not accepted that well; it was scored with an average grade of 2.77. However, when students were offered a choice between participation in the development of a student Wiki for a higher final grade and non-participation in the development, but with a lower grade, the results clearly suggested that most students would prefer to take part in the Wiki development. To be specific, only less than a 5% of surveyed students have said that they would accept a lower grade not to participate in the development of a student Wiki. Just less than 15% of students were indecisive, while almost 70% stated their readiness to participate in the development of the Student Wiki for a higher grade. The category 'other', which included 11.11% responses, contained answers such as: "It depends on other responsibilities", "It depends on the topic that Wiki would cover", "Depends on my free time", "It depends on the professor I would have to work with", etc.

Idea to make the student Wiki accessible to public received an average grade of 2.77, while the idea to make it accessible only thru the Faculty's intranet with user authentication received somewhat better grade of 3.79 (1=lowest score; 5=best score).

When it comes to reviewing the content of the student Wiki, surveyed students have unanimously expressed their wish that professors be involved as active reviewers of the content developed by students (average grade of 4.37), while the idea that selected students help professors in reviewing the content for bonus points received somewhat lower average score of 4.03. The idea that the review would not be done by professors, but be assigned to students instead, was not accepted well and was scored with an average of 2.66. Defining and applying strict rules for referencing sources in the material uploaded to the student Wiki was received a score of 4.17.

When asked which they would prefer – that the content uploaded to student Wiki be signed with their real names or pseudonyms (professors would know the real identities), 38.27% of students answered that they would prefer pseudonyms, while 56.68% answered that they would prefer that their real names be clearly stated. The

remaining respondents did not express their attitudes toward this. Pivoting of collected data showed that students with lower academic performance throughout their studies would prefer anonymity, while the option to have their names clearly stated was more frequently chosen by students with a greater grade point average, as was expected. It was similar with teamwork: students with a lower grade point average were readier to accept teamwork that those with a high grade point average.

VI. Proposition of a Possible Scenario for Developing a Wiki-textbook

Based on results from selected research on the application of Wiki systems in university education, with regard for specific student preferences related to their participation in the development of a Wiki system and, of course, characteristic traits of the Net generation, the authors have defined a set of assumption to guide the development of Wiki-textbooks in several courses within the Business Informatics study programme at the Faculty of Economics in Subotica. A brief overview of these assumptions is given below:

- Wiki system would consist of two parts: a core textbook and content created by students. The core textbook would be a source of 'authoritative' texts which would provide students with information on dominant theories and ideas within a particular discipline. It would be cataloged and categorized by a relevant institution and could serve as a proof of compliance with requirements for the appointment of professors or associates to certain university positions.
- Students' proposals on possible improvements or additions to the textbooks should be much welcomed. Bearing in mind the fact that errors in the content of a core textbook can go unnoticed even by reviewers and that the data in the textbook may become obsolete, students would be able to point to possible faults in the textbook, or suggest new sources of up-to-date or more detailed data. Students would be encouraged to point to the parts of the textbook which they think are not written clearly enough or lack in content. It would also be considered a great contribution if students could find sources with different perspectives of particular topics covered by the textbooks, since that would help to reduce the subjectivity of content to the least extent.
- It would be assumed that use of Web 2.0 applications and services (including Wikis) in the context of education differs to their use for personal reasons. Therefore, students would be acquainted with the techniques for using Wiki systems in detail, thru an extensive training for all Wiki-related activities: writing, editing content review, etc. Using prepared examples, students would be comprehensively trained in the use of the Wiki system, rules, spelling, grammar, formatting, desirable and unacceptable writing styles, netiquette, etc.
- Traditional seminar papers would be replaced with activities on creating, i.e. supplementing the present Wiki content. Structure of the content which would be added to the Wiki would be predetermined, but open for all students' suggestions, amendments and contributions.
- Topics on which students should enrich the Wiki system should be similar in terms of complexity. Each student would be assigned a particular topic, while teamwork and collaboration would be achieved thru organized checkpoint and brainstorming meetings where students would be able to make their comments, express criticism and commendations for the work of their peers and the group as a whole. This would involve students in directing the development of the Wiki system, but also encourage them, or even compel them to read content posted by their peers.
- Wiki content would be checked for plagiarism by use of specialized web services. Faculty of Economics in Subotica has been using Ephorus successfully for a number of years. Special attention would be devoted to referencing used sources of information and constant review by professors.
- The Wiki would be one of a closed type, accessible to registered users, which would help reduce students' anxiety over the public availability of their texts.
- Students would contribute to the Wiki under their real names. Each student should be awarded a grade for their work on the Wiki, just like with seminar papers. Wiki content created by students would not become a part of the exam material, which would be limited to the content of the core textbook.
- At the end of each semester, the core textbook would be supplemented with the best sections of the students' part of Wiki, of course, with proper author credits.
- Comparison of students' contribution to development of the Wiki system and their final grades would be presented at the final checkpoint meeting, where every student would have the opportunity to share their opinion.

VII. Conclusion

Wiki-textbooks still cannot be considered 'mature', but that should not be an obstacle to considering (and applying) them as an alternative to the traditional and electronic textbooks, in the context of creating new, different relations between stakeholders of the education system [2]. Wiki-textbooks put more emphasis on interaction among students (as opposed to traditional lectures), P2P collaboration, study programmes with a focus on students thru 'production' of knowledge in which students take part as well, by using (mostly free) digital resources, as well as traditional ones [5]. In such scenario, the primary goals of learning change as well: memorizing

and attaining factual knowledge give way to critical thinking, decision making, synthesis and integration of knowledge and skills [3]. This can be considered as an arising, new educational culture, which enables students to reach competencies required by the knowledge-based economy.

Based on educational potential of Wikis, specific traits of the Net generation to which most of current students belong, positive and negative experiences in the application of Wiki-systems in university education and specific attitudes of students at the Faculty of Economics in Subotica towards the development of a Wiki-system, the paper presented a concise initial proposal of a possible scenario for developing a Wiki- textbook within several courses in the Business Informatics study programme. This briefly presented possible scenario for developing Wiki textbooks is a starting point for further in-depth discussions and research, prior to its experimental implementation.

REFERENCES

[1] S. Ludwig, "Print is dying: E-readers start slaughtering print book sales," 10 November 2011. [Online]. Available: http://venturebeat.com/2011/11/10/print-is-dying-e-readers-start-slaughtering-print-book-sales-infographic/. [Accessed 1 July 2012].

[2] G. Ravid, K. M. Yorad and R. Sheizaf, "Wikibooks in higher education: Empowerment through online distributed collaboration," Computers in Human Behavior, pp. 1913-1928, 2008.

[3] M.-F. G. Lin, S. Sajjapanroj and C. J. Bonk, "Wikibooks and Wikibookians: Loosely Coupled Community or a Choice for Future Textbooks?," IEEE Transactions of Learning Technologies, vol. 4, no. 4, pp. 327-339, October-December 2011.

[4] A. K. Pervaiz and C. D. Shepherd, "Innovation management. Context, strategies, systems and processes", Harlow: Financial Times Pretince Hall, 2010.

[5] D. Bobera, M. Sakal, P. Tumbas and P. Matković, "Innovation Opportunities Offered By Wikis In Higher Education Courses: Implementation Modelling," Croatian Journal of Education, in press.

[6] C. Dede, "Planning for Neomillennial Learning Styles: Implications for Investments in Technology and Faculty," in Educating the Net Generation, D. G. Oblinger and J. L. Oblinger, Eds., EDUCASE, pp. 15.1-15.22, 2005.

[7] M. Prensky, "Digital natives, digital immigrants," 5 October 2001. [Online]. Available: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf. [Accessed 10 May 2012].

[8] J. Frand, "The information mindset: Changes in students and implications for higher education," EDUCASE Review, vol. 35, no. 5, pp. 15-24, September/October 2000.

[9] J. Siron, Educating Today's Students Using Social Networking Technologies, University of St. Francis, The Center for Instructional Delivery, 2008.

[10] S. E. Metros, "Educase," 2008. [Online]. Available: http://net.educause.edu/ir/library/pdf/MWR08046.pdf. [Accessed 17 June 2012].

[11] R. Rettie, "Net generation culture," Journal of Electronic Commerce Research, vol. 3, no. 4, pp. 254-264, 2002.

[12] D. Buckingham, "Is there a digital generation?," 2006. [Online]. Available: http://www.childrenyouthandmedia.org/cscym/attachments/074_Buckingham%20Digi%20Gen2.pdf. [Accessed 20 Jun 2012].

[13] H. Green and C. Hannon, "Their space - Education for a digital generation," 2007. [Online]. Available: http://www.demos.co.uk/files/Their%20space%20-%20web.pdf. [Accessed 2 August 2012].

[14] K. Moore, "Managing and Perhaps Leading The Millennial Generation," 24 January 2008. [Online]. Available: http://www.mcgill.ca/files/medicinefacdev/mermoore.pdf. [Accessed 15 Jun 2012].

[15] J. Sandars and C. Morrison, "What is the Net Generation? The challenge for future medical education.," March 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17701615. [Accessed 1 June 2012].

[16] M. Bullen, T. Morgan and A. Qayyum, "Digital Learners in Higher Education: Generation is Not the Issue," Canadian Journal of Learning and Technology, vol. 37, no. 1, pp. 1-24, 2011.

[17] E. Hubschmid, "Generation Y in Switzerland and Russia: the challange for cross-national employer branding," 25 April 2012. [Online]. Available: http://www.iop.unibe.ch/userfiles/file/Lehre/IHRM/Gen%20Y%20in%20RUnCH.pdf. [Accessed 5 August 2012].

[18] B. D'Netto, "Generation Y: Human Resource Management Implications," in Proceedings of Annual Paris Business and Social Sciences Conference, Paris, 2012.

[19] C. Jones, "Networked Learning, Stepping Beyond the Net Generation and Digital Natives," in Exploring the Theory, Pedagogy and Practice of Networked Learning, L. Dirckinck-Homfeld, V. Hodgson and D. McConnell, Eds., Springer Science+Business Media, 2012, pp. 27-41.

[20] D. Tapscott and A. Williams, "Innovating the 21st century university: It's time.," 4 February 2010. [Online]. Available: http://www.educause.edu/ero/article/innovating-21st-century-university-it%E2%80%99s-time. [Accessed 5 Jun 2012].

[21] P. Anderson, "What is Web 2.0? Ideas, technologies and implications for education", February 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.9995&rep=rep1&type=pdf. [Accessed 2 September 2010].

[22] M. Sakal, P. Matković and P. Tumbas, "Willingness to Adopt and Apply Web 2.0 Technologies in Secondary Education – Case of Vojvodina," Croatian Journal of Education, vol. 14, no. 4, pp. 743-770, 2012.

[23] F. Eide and B. Eide "Brain of the Blogger", March 2005. [Online]. Available: http://eideneurolearningblog.blogspot.com/2005/03/brain-of-blogger.html. [Accessed 28 September 2010].

[24] D. Kokkinaki, "The potential use of Wikis as a tool that supports collaborative learning in the context of Higher Education" in Proceedings of m-ICTE2009 V International Conference on Multimedia and ICT in Education, Badajoz, 2009.

[25] I. Karasavvidis, "Wiki uses in higher education: exploring barriers to successful implementation," Interactive Learning Environments, vol. 18, no. 3, pp. 219-231, 2010.

[26] University of Delaware, "Wikis in Higher Education," University of Delaware, Delaware, 2008.

[27] S. Wheeler and D. Wheeler, "Using Wikis to promote quality learning in teacher training," Learning, Media and Technology, vol. 34, no. 1, pp. 1-10, March 2009.

[28] M. Cole, "Using Wiki technology to support student engagement: Lessons from the trenches," Computers & Education, no. 52, pp. 141-146, 2009.

[29] W. W. Ma and A. H. Yuen, "News writing using Wiki: impacts on learning experience of student journalists," Educational Media International, vol. 45, no. 4, pp. 295-309, 2008.

[30] I. Elgort, A. G. Smith and J. Toland, "Is Wiki an effective platform for group course work?," Australasian Journal of Educational Technology, vol. 24, no. 2, pp. 195-210, 2008.

[31] M. Sakal, A. Slavić and Gy. Szretykó, „The Y-Generation Through the Prism of HRM," Herald of Omsk University, vol. XI, no. 2, pp. 6-12, 2013.

Analysis of learning by doing in the creation of instructional video in teaching optics

A. Wirth* and R. Repnik*
* Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
alex.wirth@student.um.si, robert.repnik@um.si

Abstract – It is well known that videos in education have some advantages in specific areas. We have made the experiment to see the effect of video in physics education. Our goal was to see if short video clips can improve understanding of physics – optics. In our experiment we wanted to test the principle "*learning by doing*" with the specific emphasis on creativity. Students were divided into two separate groups with the same assignment. First group was to prepare a video without any additional instruction from lecturer but they were not limited by any other means. The second group was further divided into two teams. Both teams got the same instruction how to create educational video clips. The difference between these two teams was that one team was given the detailed analysis of previously created bad educational video on the same topic while the other did not. Since creativity cannot be though and has to come from the individuals own inspiration we wanted to know if too many instructions actually influence the outcome of the students' performance.

I. INTRODUCTION

Instructional video is relatively new in education although it was first used in World War II. In distance education we have years of experience with the e-learning materials production. Only in the last few years recordings of lectures on the web becomes feasible [1]. Never the less e-materials are undoubtedly on the raise, enhancing many of their advantages and minimizing disadvantages (like burdening the eyes). But there is still an open problem of reliability of information if they are not reviewed [2]. Furthermore e-learning materials are useful addition to the educational process. If used appropriately they can help improve students' understanding the basic concepts of natural sciences [3].

At the Faculty of Natural Sciences and Mathematics all students of special didactics study programs have ICT course. This course prepares them for the various application of ICT in education.

In the past research it was shown that the use of instructional video brings only minor improvements in students' knowledge but its other value is tremendous [1][4]. They proved that the instructional video is very good for skill training [1]. But making of the instructional video is very time consuming. If there are no evaluation of the additional workload, teachers will not prepare any additional instructional videos [5]. This fact is neglected in our educational institutions and is actually quite a shame because e-learning materials are interesting addition to the learning process, in particular for natural sciences [6].

We have to emphasise that instructional videos cannot substitute lectures and should only be used as a secondary source of knowledge. Not only instructional videos but all e-learning materials have some drawbacks. If they are (too) good they lower the classroom attendance [5]. Anyway, the student population is not entirely uniform and it is shown many times that older the students are less they are familiar with the technology and more they are conservative in their preferences [7]. But new generations are more ICT literate and they will demand more, high quality instructional videos and e-learning materials in general.

II. INSTRUCTIONS FOR MAKING A INSTRUCTIONAL VIDEO

With optical set from a seller nT-BROG students were to prepare instructional videos that will help understand optics. Video is not meant to substitute the experiment, but should be consider as its supplement. Video is meant to help learners to refresh their knowledge gained somewhere else (in the school at physics lessons or physic courses – summer schools, etc.)

Though instructional video looks simple on the design paper its actual production requires many skills. Light consideration is the one that cannot be dismissed in video recording for good visibility and reflection issues. . Viewers of the experiment need to get good impression of the surroundings of the experiment and scale. To prevent video shaking a tripod should be used. These are only technical consideration in the video recording. The other considerations are also highly important – aesthetic aspect. No matter what kind of learning materials is prepared it needs to be aesthetically pleased and didactically suitable. [1] [2]. To be didactically suitable the physics experiment must be easy to understand with the clear intention for the viewer. Focus of the video is experiment and not the preparation of the experiment even if it may be considered highly important for teachers who want to prepare the same experiment. Though sound is not important in the video of the experiment it is not uncommon to prepare suitable sound background or audio narration if this is feasible. For the real universal video learning materials (hearing impaired people too) all elements should be visual. Textual of image explanations are therefore necessary in such educational video. If video is further processes it is possible to include subtitles. From the users perspective subtitles can be turned on and off and even

switched between different languages. This would enable to produce multilingual learning materials. Any text, schemas and images used in video should be verified by the experts who know the experiments thoroughly.

Students get the following assignments for the video production:

- Concave mirror,
- Convex mirror,
- Mirror 1,
- Lens 1,
- Lens 2,
- Lens 3,
- Lens 5,
- Lens 6,
- Lens 7,
- Lens 8,
- Refraction of light,
- Reflection of light,
- Model of a healthy eye,
- Model of short-sighted eye (+correction),
- Model of far-sighted eye (+correction).

Student also receive additional instructions that video should be interesting for the learners and that all exposed parts of the body that are recorded on the video should have not jewellery or tattoos. Audio background should match the scene of the video, be neutral to the viewer, and obay the copyright restrictions.

III. METODOLOGY

In our case we prepare single case study. We analysed videos that were made by the students of previously described groups. In the analysis we exposed the good and the bad issues in the instructional videos. But to make things more interesting we compare our result with the instructional video made from the interdisciplinary group of students from physics and computer science who have different instruction but the same outcome. Their goal was to made instructional video for students competition and that was all they were instructed. The topics, approach and implementation was entirely up to them.

IV. INSTRUCTIONAL VIDEO

First group that had no instructions how to make a video generally made bad instructional videos. Their videos were shaky, experiment was not in centre of video, and hand that moved the lenses and mirrors is considered the nuisance (fig. 1 and 2). Though students today have sufficient ICT knowledge and are considered highly innovative their video were dull and provide no viewers appealing.

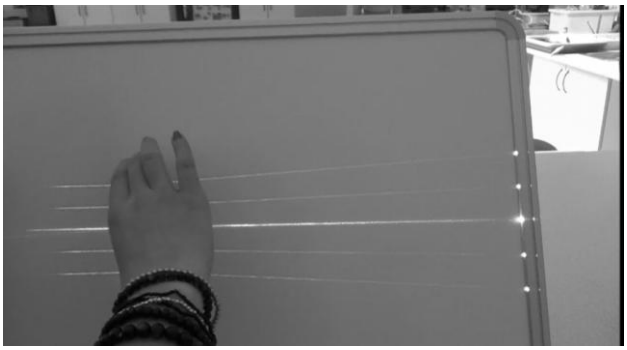


Fig. 1: Example of a bad instructional video.

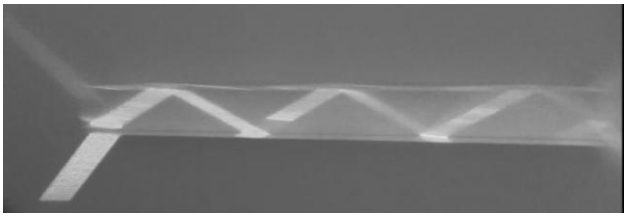


Fig. 2: Blurry image due to quick movement of a camera.

The second group got more instructions how to prepare instructional video. But for this experiment we divide them into two subgroups. First subgroup did not see videos (from other groups) and receive analytical view what is considered good and bad instructional video, while the second subgroup does. As expected the subgroup that has instruction and sees video from other groups made the best instructional video. Room in which they filmed videos was darker than room from the first subgroup. Darker room exposed the contrast between laser and white blackboard in the background but the scale is still visible. They used a scale and showed the point where the rays meet (focus) but they did not show that the lens is at the base of a scale (fig. 3). Unfortunately the reflection of cameraman is seen on the video (fig. 4).

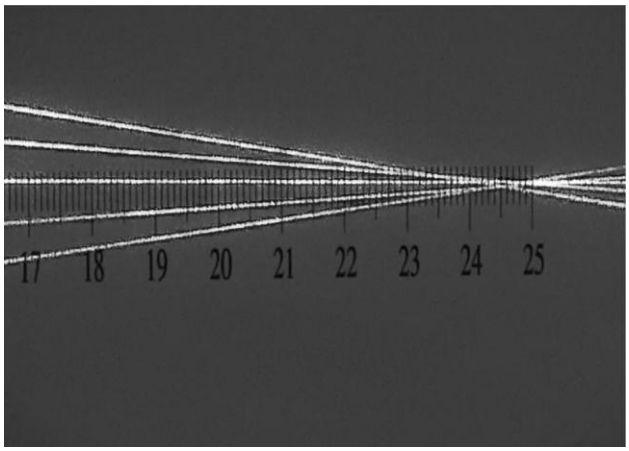


Fig. 3: Scale was used. Images tend to be black and white. Room is darker but you can still see the scale – that way contrast between background and laser is better. Sadly it was not shown that lens is at the base of a scale. From this image you cannot identify the focal length of a lens.

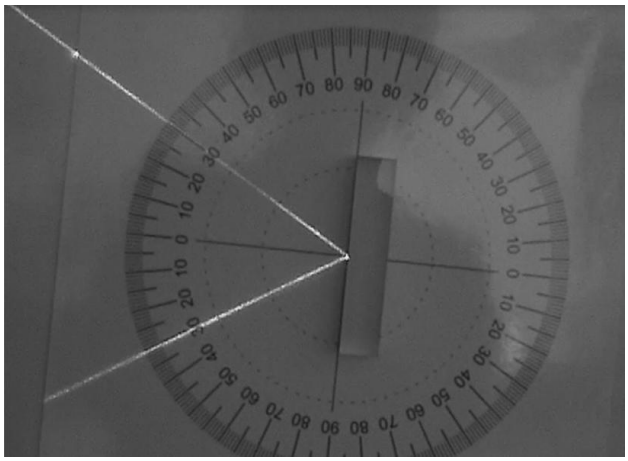


Fig. 4: Entry angle and reflective angle are nicely visible but there should be no shadow of camera operator.

The best videos made the team from the second group that saw bad videos from the first group. They got the chance to see and correct all the mistakes of the first group. It turn out that they took their chance to improve flaws of the first group. They also made better videos then the other team of the second group. For example they showed that the centre of a lens was at the base of the scale and enable viewer to determine focal length of a lens. Table 1 shows us bad and good activity in instructional video production.

Table 1: Differences between good and bad doings when making an instructional video.

BAD	**GOOD**
shaky videos	use a tripod
recording room in which you are making videos	experiment in centre of a video
kitschy transitions	neutral music
to dark or to bright videos	there should be just enough light that scale is visible but not too much so that there is a good contrast between laser and background
video substituting the experiment	appealing videos so that students are motivated to watch it at home – that way they refresh their acquired knowledge

V. INTERDISCIPLINARY TEAM

We also analysed one older video which was made by five students. Two of them were physicists, the other three were computer science teachers. They made very good instructional video on optics. Not only did they use video but they included animations (fig 5). Animations are unrealistic but they are very pure and minimalistic. In real world video it is often very difficult to expose everything but animations can be prepared in sterile environment and easily achieve that. In video they included nice neutral music. The transitions between inscriptions, animations and experiments are not too aggressive; and experiments were all in the centre of the video.

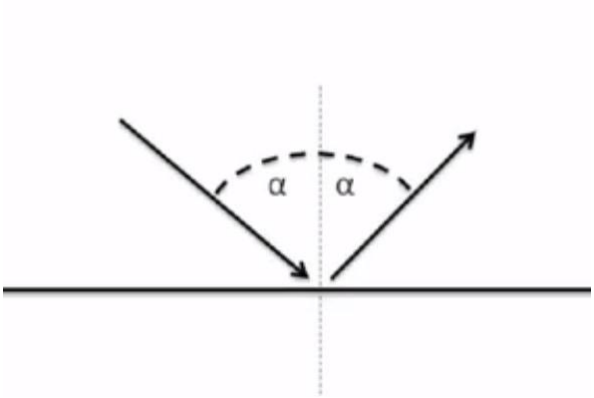


Fig 5: Use of animations in instructional video.

Though they were not given any instructions for their work they prepare good instruction video. This may contradicts our findings that students will make better instructional video if they are given instructions. But we must not forget few vital things they had. They knew that their product is going to be for the competition, and they were highly motivated because they know the prize they can achieve in the competition. In general such high motivation is hard to achieve between students and it is also true that students know they can get different grades for their work. Some of them see no need to score the highest grade. This does not contradict with the findings that not all students are suitable to participate in different projects.

VI. CONCLUSION

In this article we presented the results of research made between students of physics. Now it is evident that instructions do not kill the inventiveness. We showed that student would make better instructional videos when they are provided with good end extensive instructions. But there are some experiences that come only with time and practice. One who has more of them will make better instructional videos. And at the end we would really like to emphasize the effect of motivation on the quality of the product.

REFERENCES

[1] M. Krašna and T. Bratina, Educational value of instructional video, University of Maribor, Faculty of education.

[2] R. Repnik and M. Krašna, WEB 2.0 in education: Do we really always need it?, University of Maribor, Facilty of natural sciences and mathematics, Faculty of education, Faculty of arts.

[3] R. Repnik and V. Grubelnik, E-learning materials for 3rd grade of primary school – physics, University of Maribor, Facilty of natural sciences and mathematics, Faculty of education, Faculty of electrical engineering and computer science.

[4] T. Bratina and M. Krašna, Analysis of instructional video effectiveness for learning SPSS, University of Maribor, Faculty of education, Faculty of arts.

[5] M. Krašna and T. Bratina, Video learning for better students' performance, University of Maribor, Faculty of arts, Faculty of education.

[6] R. Repnik, M. Ambrožič, M. Krašna and V. Grubelnik, Teachers' experiences and recommendations about using E-learning materials in primary school, University of Maribor, Facilty of natural sciences and mathematics, Faculty of education, Faculty of arts, Faculty of electrical engineering and computer science.

[7] M. Krašna, M. Duh and T. Bratina, E-learning next step – learning materials for students, University of Maribor, Faculty of Arts.

Examples of Best Practice of University-Industry Collaboration in the Domain of Electro-CAD Application by using EPLAN P8 and E-learning Systems

D.Lukac*
* Rheinische Fachhochschule Köln gGmbH,
University of Applied Sciecens, Köln, Germany
lukac@rfh-koeln.de

Abstract - Requirements of the industry are reflected into the curricula of the universities oriented to application oriented sciences. Nevertheless, a coordination of the teaching contents to the industrial requirements often does not work efficiently enough because, on the one hand, the dynamism of the changes of the useful teaching contents becomes sluggish at the universities compared with the industry, on the other hand it often does not come to the desired long-term collaboration between universities and industrial companies. In this work different collaboration examples are presented. Also the state of the art of a successful collaboration and the further development of certification program between university and a next-door international enterprise in Cologne, Germany, in the domain of the Electro-CAD application are presented. Special look is give on the technical part of data exchange management. Some examples of such collaboration in the everyday practice at the Cologne University of Applied Sciences by use of the E-learning systems ILIAS and EPLAN Electric P8 software are presented.

I. INTRODUCTION

Who has influenced first whom, one theorist a practical person or vice versa, remains an unanswered philosophical question. However, one thing is sure, theory needs a useful - meant is for the industry useful - practice. This is nothing new; however, till these days this so called university-industry education often does not work in fact. There are many companies which have recognized these problems and out of it, have tailor-made education-oriented continuing education offers for educational institutions. Those especially in regard of different hardware and software solutions like companies as Festo [1], Christiani [2] or Lybold didactics [3] do that. Such systems are surely advantageous for special or university education, but they have a great disadvantage. It is, that without suitable training and permanent continuing education of students and above all of lecturers who have a task to transmit the knowledge to students, such systems becomes very fast outdated provided knowledge. Not out of dated but an up-to date education is asked, which is oriented to the developments of the industry. Consequently the questions arise, how such educational purpose can be reached, and above how this objective can be financed and technically developed in the long-term and which requirements have to be taken into consideration behind the sole readiness to collaborate?

II. BEST PRACTICE EXAMPLES OF UNIVERSITY-INDUSTRY COLLABORATIONS

A. *Technical University in Berlin and Deutsche Telekom*

There are many examples of collaboration between colleges or universities and industrial companies. One interesting example is collaboration between Technical University in Berlin and Deutsche Telekom in the area of innovations in ITC in form of so-called "Deutsche Telekom Laboratories" [4].

Figure 1. Deutsche Telekom Laboratories

Innovations in the sphere of information and communication change the world and are the base for a sustainable development within the national economy. They can be inspired by customers or by research, to solve problems basically better. To form promising innovations, is a job which must be solved by the science as well as by the enterprises. The connection of both worlds has chosen Deutsche Telekom as a beginning of their promising objective. Together with the Technical University in Berlin it has founded the Deutsche Telekom Laboratories on the campus of the Technical University in Berlin. The

mission is: to set up indicative research of innovative information technologies close to market and telecommunication technologies. Top researchers realize together with experts of Telekom they visions to the creation of the information technology and telecommunications from tomorrow. In Telekom Laboratories the know-how of Deutsche Telekom with the potentials by top researchers from all over the world is bundled up. Telekom Laboratories form the crystallisation point of state-of-the-art knowledge to disruptive, meant is revolutionary, technologies of the future.

B. *Technical University and SAP*

Another interesting example is the cooperation of the Technical University in Darmstadt and the company SAP, in the form of the so-called "SAP Research Centre" also called “SAP Living Lab” [5].

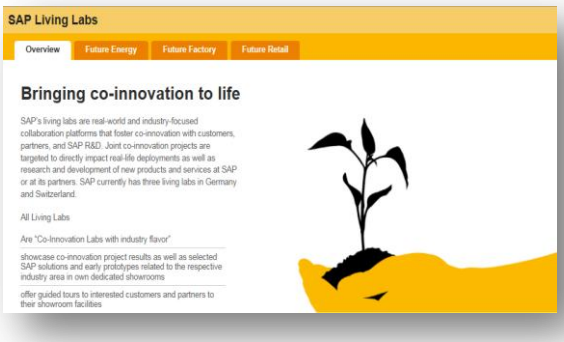


Figure 2. SAP Living Lab

“SAP Research”, the SAP own research department with the head office in small German town Walldorf and scientist of Technical University in Darmstadt co-operate successfully on several research fields since many years as for example "knowledge work in highly dynamic working surroundings", "interaction models for improved use friendliness" and "new methods for the software engineering". These areas are observed in their interlinking potentials and are from special strategical interest for SAP, because they have a strong influence on the use friendliness of the future products. Thus the intelligent software environments are developed for the business application which open varied or new services to the user. According to [6] "The cooperation is another example of the close collaboration between the Technical University in Darmstadt and an extremely innovative enterprise". The cooperation also encloses a common doctoral candidate's program: Together with SAP scientists research will do each year up to 14 doctoral candidates in subjects like "creation of future working environments". In cooperation with more than 30 leading universities SAP Research investigates worldwide new technologies and examines this for their potential for innovative software solutions or commercial models. With it the pioneering role SAP should become protected by applied technology research and be further developed. Besides, the new location the SAP Research strategy underlines the collaborative research with the aim to strength the European IT-research scene which should be opened even further for new partners. In common projects with customers, technology partners and universities the bases for future commercial uses and their infrastructures are investigated and developed. Besides, German SAP Research pursues two other partners in Brisbane (Australia) and Pretoria (South Africa), in addition also three SAP Research Centers in Palo Alto (USA), Montréal (Canada) and Sophia Antipoli (France).

C. *University of Heidelberg and Nikon LLC*

Also is to be mentioned the collaboration between University of Heidelberg with the company Nikon LLC in the form of so-called "Nikon Imaging Center (NIC)".

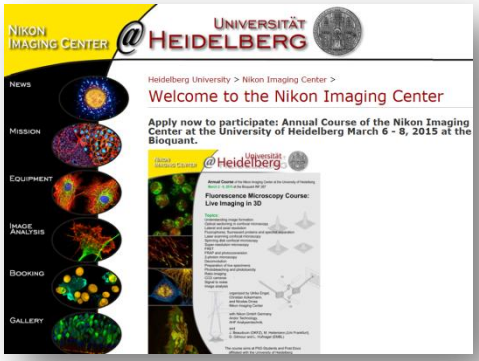


Figure 3. Nikon Imaging Center (NIC).

Nikon Imaging centre is a light-microscopic equipment of the university at BioQuant, the Center for “Quantitative Analysis of Molecular and Cellular Biosystems” [12] in cooperation with the Nikon LLC, Nikon Instruments Europe B.V. as well as other internationally operating enterprises. The life sciences and the medicine of the university receive with the NIC access the newest microscopic procedures and instruments [7] .

D. *University in Passau and Siemens AG*

There are not so many examples in the area of the CAD / CAE cooperation between universities and industrial partners. Nevertheless the cooperation between Siemens AG and University in Passau is to be mentioned as one example. Thus company Siemens supports new seminar offer with CAD software.

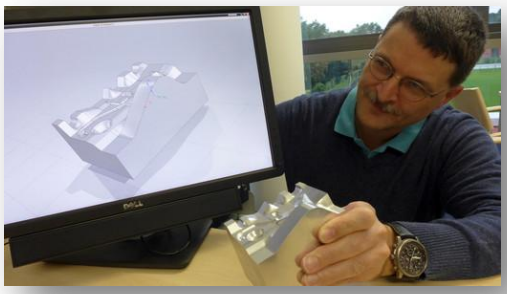


Figure 4. Nikon Imaging Center (NIC) at University in Passau .

Students of the informatics courses as well as mobile and embedded systems courses can visit since winter semester in 2013/14 for the first time a seminar to the subject CAD and learn to work with the professional CAD software NX. Cooperation with Siemens makes this offer possibly [8].

E. Rheinische Fachhochschule Köln and EPLAN Software and Services GmbH & Co.KG

Another successful example in the area of CAD/CAE about which already 5 years ago has been reported at the MIPRO conference and which is still successfully functioning, is cooperation between EPLAN Software and Services GmbH & Co.KG and Rheinische Fachhochschule Köln – University of Applied Sciences [9]. To develop and keep functional such wide and successful cooperation, are different success factors determinative. Among the rest, these factors are: functioning data interfaces (in this case based on E-Learning system ILIAS) to the industry tailor-made education offers at schools and colleges, permanent continuing education of the lecturers and development of suitable practice exercises and study course plans for schools and universities [10]. It means, in order to come along with the development of the new versions of the software, it is necessary to continuously develop examination questions, either in English language for international certifications and also in German language for national ECS examinations. Export function of the data exchenge file is presented in following figure.

Figure 5. Interface for creation of XML data exchange file

This enormous and complex task can only be efficiently solved, if collaboration partners, Rheinische Fachhochschule and EPLAN Company agree to work simulations and to develop clear interfaces for data exchanges. In the case of definitions of the examinations questions, for data exchange is XML file defined. One example of the XML file is presented in following.

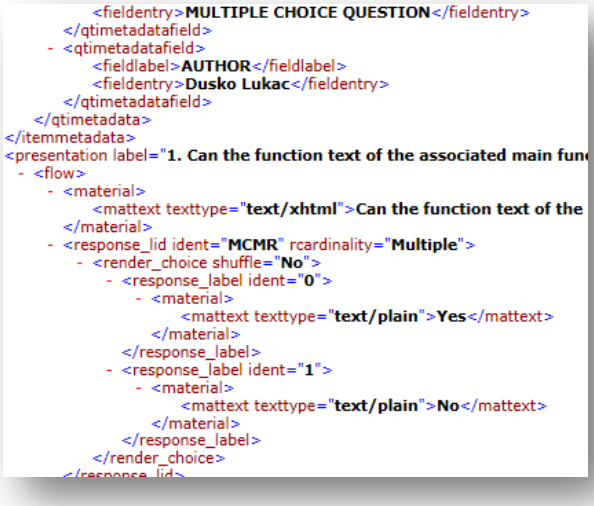


Figure 6. Example of edited XML file used for data exchange

This file allows definitions of the questions type as e.g. multiple choice with single or multiple answers, free text questions, java applet combined questions and answers. On this way, efficient work is possible and technical fundament for well-organized development of the whole program is achievable. The online-examination portal for EPLAN Certified Student (ECS) exam is presented in following:

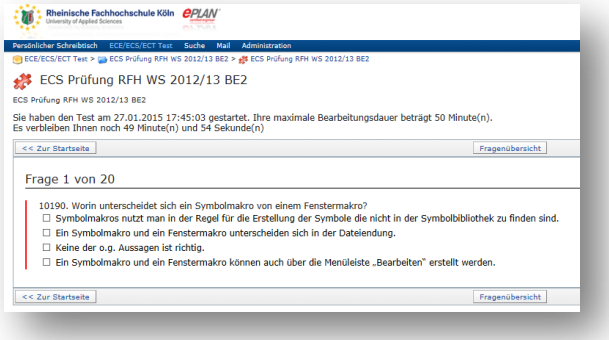


Figure 7. Online examination portal based on ILIAS Learning Management System

III. UNIVERSITY-INDUSTRY COLLABORATIONS IN DER AREA OF CAD/CAE

A. Eplan Certified Student – Examples based on platform 2.4

As already mentioned, a successful education oriented to industry at colleges requires a structured and planning oriented education, which takes into consideration the newest technical developments. Thus Rheinische Fachhochschule Köln – University of Applied Sciences uses currently the newest version of the software EPLAN Electric P8, not only to provide the basis knowledge on the new platform, but also new functions from the version 2.4. In the following some of the new functions used in everyday educational practice are presented. So, for

example, the simplistic representation of the PLC structures with one-pole functions under 2.4 makes the documentation and management more clearly and with it much easier. Also the management of the, so called “part projects”, increases the overview of the whole project. From the project, relevant parts can be extracted, worked on individually and bring together again later in the whole project. With it an entire and highly qualitative documentation is generated and users can better concentrate upon his/her area of responsibility. The platform EPLAN supports numerous norms as for example the IEC 81346; now the marking of project sides according to IEC 61355 is possible. With this kind of the marking an object characteristic is used for the definition of the project structure. Besides, with the object sign a document, it means a project sides, is assigned to certain object (function, place or product). This marking can be also used in connection with an interface to PLM/PDM or ERP systems. Now also other devices like converter, engines or valves with bus connections can be a component of a PLC bus system. Devices with several, different bus systems are better supported with it. Also additional PLC manufacturers were taken up for the bidirectional data exchange - for example ABB., Mitsubishi Electric and 3S. [11]

B. Example 1 – Device oriented projecting

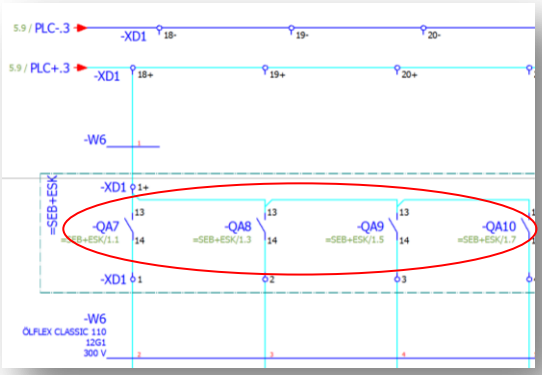


Figure 8. NOC placed device oriented

The present example concerns so-called device oriented projecting of the motor protection switch contacts -QA7 to- QA10. This means that the so called normally open contacts (NOC) are not to be placed out of symbol library, but there are dependent on company product (article) observable in the device navigator. This functionality becomes clearer to understand if live presented.

C. Example 2 – Enlarged data exchange with Siemens SIMATIC STEP 7

For the new EPLAN version 2.4. the PLC interface was extended for the data exchange between EPLAN Electric P8 and the configuration PLC software " Siemens SIMATIC STEP 7". Now for PLC cards different start addresses and addressing ranges can be defined. For this purpose both following propreties are available in the PLC boxes in the switch plan:

1. Addressing range 2 (ID 20299)
2. Start address 2 of the PLC card (ID 20255)

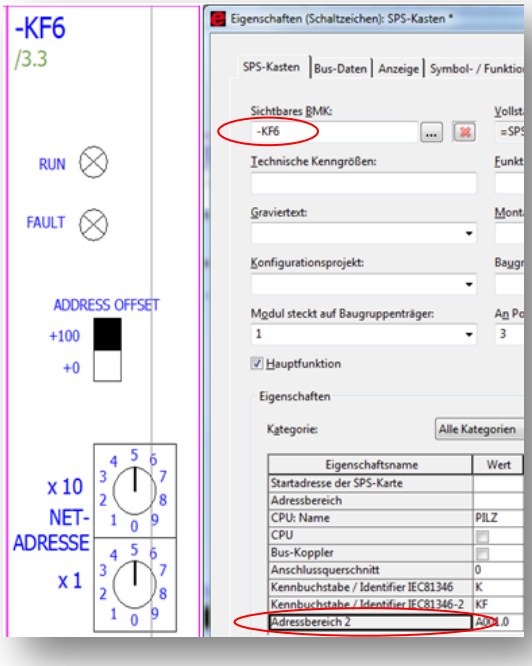


Figure 9. New possibilities for definiton of start addresses and addressing ranges for PLCs

With it the addressing range 2 and start address 2 of the SPS map are evaluated. The control box “Separate addressing area for inputs and outputs” must be deactivated. In that case for the PLC cards which have inputs as well as outputs with the help of the property “Addressing range 2” one can fix a separate addressing range for the outputs. Every PLC connection must have an unequivocal name, i.e. if the address "A001.0" exists, the address "E001.0" cannot be awarded any more. With the new function a common addressing range is used for PLC in and outputs. In this case the addresses are distinguished, in addition, by the prefix, i.e. within a PLC addresses with the same byte and bit address may seem, e.g., "A001.0" and "E001.0".

IV. CONLUSION

Paper gives examples and an overview of best practices collaboration between the university and industrial companies in Germany. Examples of the best practice collaboration in the area of CAD/CAE applications are comparatively seldom in regard of wide use of the CAD/CAE software in developed industrial countries. On the one side the reasons are to be find in the undeveloped trials to start such collaborations, on the other side, long-term collaborations between industrial companies and universities are threatened by slow and inertially development up to date course at universities and the lack of efficient working interfaces, either

technical but also managerial. In this paper an example of functioning technical data exchange is presented as well as examples of day by day examination tasks based on the newest release of EPLAN P8 software. Also one important reason of failure of such collaborations is lack of access to the newest industrial development by universities, and on the other side, lack of financial support to update the equipment and software at universities in order to come along with the industrial development. Industrial companies on the other side, search for same working attitudes which are not always given at universities, because of the different nature of the work. Therefore, industrial companies must make concessions in regard of financial support of universities as a part of their long-term marketing strategy as well as possible acceptance of the different working attitudes, on the other side, universities must offer potential parsons which are able and compliant to sustainable work on the maintenance and further development of the joint educational programs between the mentioned parties.

ACKNOWLEDGMENT

The development for this paper was supported by the Rheinische Fachhochschule Köln and EPLAN Software and Services Company. In developing the ideas presented here, I have received helpful input from training team of the EPLAN Company. I also thank reviewers for their feedback.

REFERENCES

[1] Festo Didactic "Ihr Partner für Qualifizierung und Consulting", avaliable at <http://www.festo.com/cms/de_de/Festo-Didactic.htm>, 30.01.2015, 14:43h

[2] Christiani "Didaktik & Personal", avaliable at http://www.christiani.de/ Didak-tik-&-Personal.html, 30.02.2015, 15:33h

[3] LD Didactic "Products and Solutions", avaliable at < http://www.ld-didactic.de/ produkte-loesungen.html >, 30.01.2015, 17:28h

[4] Deutsche Telekom "Telkom Innovation Laborateries", avaliable at http://www.laboratories.telekom.com/public/Deutsch/Pages/default.aspx, 01.02.2015, 13:33h

[5] SAP "Bringing co-innovation to life",avaliable at http://www.sap.com/corporate-en/about/our-company/innovation/living-labs/index .html, 01.02.2015, 15:04h

[6] SAP News Center „TU Darmstadt und SAP eröffnen campusnahes Forschungszentrum", avaliable at „http://de.news-sap.com/2006/03/28/tu-darmstadt-und-sap-eroffnen-campusnahes-forschung-zentrum/, 01.02.2015, 20:21h

[7] Nikon "Nikon Imamging Center", avaliable at <http://www.nic.uni-hd.de/>, 01.02.2015, 22:05h

[8] K.. Jordan „Siemens unterstützt neues Seminarangebot mit CAD-Software", avaliable at <http://www.uni-passau.de/wissens-transfer/aktuelles/meldung/detail/siemens-unterstuetzt-neues-seminarangebot-mit-cad-software/>, 02.02.2015, 12:06h

[9] D. Lukac "University as third party CAE/CAD certifier for the industry: A case of on open source LMS based certification as EPLAN Certified Engineer", MIPRO, 2010 Proceedings of the 33rd International Convention , 24-28 May 2010, pp. 903 – 907, IEEExPlore

[10] D. Lukac "New dimensions in the CAE/CAD standardisation and certification process in the industrial and the educational sector", Int. J. of Knowledge and Learning, 2011 Vol.7, No.1/2, pp.145 – 155

[11] EPLAN " Neue EPLAN Plattform 2.4", avaliable at <http://www.eplan.de/de/unternehmen/news/alle-news/view/ article/zum-download-neue-eplan-plattform-24/>,02.02.2015, 14:28h

[12] University of Heidelberg, "BioQuant", avaliable at <http://www.bioquant.uni-heidelberg.de>,03.02.2015,10:15h

Self-Paced Computer Aided Learning of Music Instruments

Mladen Konecki
Faculty of Organization and Informatics, Varaždin, Croatia
mladen.konecki@foi.hr

Abstract - Many people want to learn how to play a musical instrument. Modern business life however makes hard for people to make time for scheduled musical lessons. People tend to find time for this kind of activities in various periods of their working day. This implies that some kind of self-paced learning system would be well-suited for this kind of learning. Self-paced learning systems would have to be able to track progress of its users and also adapt its courses depending on one's abilities and the speed of progress. In this paper this kind of system is described and discussed. The question that needs to be answered is whether people in fact have real interest for this kind of system and what aspects of this kind of system they find important. This paper presents research results that give answers to these questions. Presented results are discussed and directions for future research are also stated and elaborated.

Keywords **- musical instruments, self-paced learning, automated and adaptive learning system**

I. INTRODUCTION

A. The impact of music and music education

Music is an integral part of everyday life. Most of people like to listen some kind of music. There are reports which have shown that music can have beneficial effects on people in many different ways. For example, listening to music can enhance one's spatial-temporal reasoning [1]. It also can have positive effect on listeners' mood [2].

But there are even more research papers which have shown that music training has positive effect on people. In [3][4] is shown that music lessons can enhance one's IQ. Music training can also enhance one's auditory skills [5]. Also, short-term music training can enhance listener's verbal intelligence and executive function [6]. Another research has shown that brain structures differ between musicians and non-musicians [7]. Musicians practice various complex motor, auditory and multimodal skills and [8] reports that children who had music training have long term enhancement of visual-spatial, verbal and mathematical performance.

B. Statistical data about music education

Statistical data shows that in recent years more and more people tend to learn how to play musical instruments [9]. In the U.S., in 2002 there were 15,74 million musical players and in 2010 that number has raised to 18,08 millions. International Music Products Association conducts a telephone surveys every 3 years [10]. Their research data shows that in 2009 there was a musical instrument player in 58% of households in the U.S., compared to 1997 when this number was 38%. The percentage of households with two or more players was 34% in 1997 and in 2009 it was 43% [10]. When considering how many people have taken music lessons during their lifetime in the U.S., it can be seen that this number has decreased from 47% to 35,6% from year 1982 to year 2012 [11]. This fact raises a question how it is possible that there are more musicians and less people that are taking formal music lessons. The most probable answer to this question is that today there are more ways to learn how to play musical instrument than before. In the year 2000, 25% of musicians have taken their first music lessons at school. In the year 2006, this percentage dropped to 16% and then in 2009 came back to 22%. The number of people that have taught themselves how to play musical instrument in the year 2000 was 14%, in the year 2006 was 26% and in the year 2009 decreased to 20% [10]. Today, people can search the Internet and find lessons and learn how to play musical instrument on their own. In the year 2003, 20% of people used Internet and other alternative sources to learn how to play musical instrument, but in the year 2009 this percentage was increased to 30% [10] which was almost third of all music players. There is so much knowledge on one's disposal today and people can learn things much easier than before. Survey conducted by Australian Nexus Survey showed that in the year 2001, 15% of music players became interested in music on their own while in the year 2007 this percentage has increased to 21% [10]. In the U.K. 43% of music players stated that they learned to play musical instrument on their own. Music Council of Australia conducted survey in 2001 and found that 75% of people that don't play musical instrument wanted to learn how to play one and 49% of them still wanted to learn [12]. These data show that there is a big interest for music education. Today, various technology can be used to find new ways of teaching people how to play musical instruments and that is the trend of the future.

C. Paper overview

In this paper, a short overview of computer educational systems that exist in the domain of music education is given. A short overview of proposed model that is adaptive to learners' skills and needs and that enables self-paced learning is also presented. Conducted survey among the students with the goal to conclude whether people have real interest in this kind of system is

also presented and elaborated. In the conclusion some final remarks about presented topic are made and pointers for future research are given.

II. RELATED WORK

A. Computers in music education

There are many applications of computer technology in music education. According to [13], there are several main categories where people can apply computers in music education:

- teaching fundaments of music
- teaching of musical performance skills
- performing of various analysis of music
- teaching of musical composition skills

There are numerous representatives for each of given categories. Of course, computer programs in these categories are partially overlapping. In [13] several created systems for each category are stated. Some of the most popular are:

- Hyperscore [14] which teaches children to make music and rhythmic patterns by hand drawing.
- Continuator [15] which teaches children music concepts through simple computer games.
- Music Ace 2 [16] which consists of 24 short games and enables its users to follow their progress and earn medals.
- MiBAC Music Lessons 2 [17] which can help in acquiring ear training and sight-singing skills as well as harmony concepts.
- Practica-Musica [18] which has 225 lessons that teach basic music playing concepts.
- GNU Solfege [19] which is free ear training program that provides customization of existing exercises.

B. Computers as tutoring systems

The category that is the most important for this topic is teaching of music performance skills. To help people in acquiring playing skills these systems must be able to capture players' input, to analyze captured data and to give proper feedback. Some of the most popular educational systems in this domain that can be found are:

- i-Maestro [20] is computer system that enables its user to learn bowed musical instruments, especially violin. It has many modules and it enables collaboration between professor and students in the class, while learning music theory and making music sheets. Based on input audio and movement sensors it can give its users feedback on their playing accuracy, body posture and gesture accuracy although interpretation of results is done by professor. It's however not fully automated. System can also generate different exercises.
- Violin Tutor [21] is a software that helps beginners that are just starting to play violin. It captures audio data over microphone and transcribes that signal into music notes. On the interface learners can see interactive music notes, 2D violin fingerboard where learners can see what they need to play and 3D animation that shows to the learners how to play. By audio frequency analysis, system can show its learners weather they played correctly or not and what was meant to be played.
- pianoFORTE [22] is a system that can capture users' piano playing over MIDI interface. It tracks several aspects of playing: dynamics, tempo, articulation and synchronization of music events. After playing, system shows to users in graphical form what was played correctly and where mistakes were made. This is useful since new players usually have problems with self-evaluation.
- PianoTutor [23] has some main features of intelligent tutoring system. It captures users' piano playing over MIDI interface, it can detect various mistakes in playing (in pitch, dynamics, notes length) and based on the type of mistake it gives proper exercises to help users to learn how to correct those mistakes. It has around 100 lessons. Order of the lessons is determined by users' level of skill.
- IMUTUS [24] is a project that is aimed at the development of an open platform for training users to play by using recorder. Practicing environment was created where developed system could record users' playing over the microphone and then analyze and evaluate users' performance. It points out where mistakes were made and helps users to focus on those parts of the score.
- VEMUS [25] project is based on IMUTUS project but its use is expanded on brass instruments, such as saxophone, flute and clarinet. Its goal is to make intelligent tutoring environment for practicing of playing alone, where system gives proper feedback based on the type of mistakes that are made. Its main goal is to make a virtual music school.

There are also many commercial products on the market [26]. Most of them are made for learning of piano playing since almost third of all music students play piano [10]. These products usually have simple features which can capture users' piano playing over MIDI interface and give feedback on what was played correctly. Lessons are always the same for all users.

There are also many educational games [27]. They are mostly fun to play with, but some of them have educational features. For example, "Rocksmith" is a computer game in which people can plug their real electric guitar (or electric bass) and learn how to play real instruments. There are also games like "Rock Band" in which people can play drums with playing mechanics that is very similar to real drumming.

C. Motivation

Educational goals and educational purpose of computer aided systems in music education have been mentioned and established, but another very important factor for music education is also motivation. Without proper motivation in the learning process, students will not learn much and will struggle while learning. On the other hand, motivated students are learning faster and better. Csikszentmihalyi directly talks about education and motivation and brings these two concepts together in his motivational theory called "The flow theory" [28]. He defines flow state as a state that's optimal for efficient learning. He defines 9 dimensions that are necessary in order to achieve the flow state [28]:

1. Clear goals in every moment - to have clear goals, there has to be instructional model in computer system that will guide users and tell them what to do next, based on their previous results.
2. Clear and immediate feedback - there has to be a proper feedback that shows what was played good and where mistakes were made. Most of computer systems described have this component.
3. Balance between the perceived challenges and perceived skills - lessons must be challenging so learner is not bored but they also shouldn't be too difficult because learner might get discouraged. Based on playing results, computer system should chose harder or easier lessons for the learner.
4. Merging of action and awareness - it's a state in which learner is focused while achieving his goal and he is not aware of that. Computer systems are suitable for achieving this state. By making learning a fun experience, learner is perceiving that he's having fun and not that he is learning. For example, by adding backing tracks while playing lessons, learner might have a feeling that he is in a band and that should make the experience more fun.
5. Freedom from distractions - with computer screen and headphones, computer systems can minimize distractions from surroundings and make learner fully focused on his actions.
6. No fear from failure - sometimes people can feel anxiety when they have to learn something they are not good at. When learning on a computer system, nobody is listening except the "computer", so this kind of learning can lower this anxiety.
7. Loss of self-consciousness - elements of computer games can make people forget about themselves as they become fully emerged in situation that they are in. As was mentioned before, computer systems can make learners fully focused on activities they are doing and with background music learners can have feeling that they are performing in a band, that they are part of a successful music group, etc.
8. Time distortion - when fully focused, learners can easily lose track of time and that is good if that happens in the learning process because learners are in that way willing to spend more time learning. This usually happens when people play computer games and when they are having fun. By making learning process game-like and fun, learners are encouraged to spend more time learning and by doing that they learn more.
9. Autotelic or self-rewarding experience - when activity is so fun and good that by just doing that activity, learner feels rewarded, then learner is experiencing autotelic state. In this state learner can learn in optimal way.

As can be seen, motivation is very important factor in the learning process. By having these dimensions in mind, a successful computer learning system that will be able to optimize learning experience can be developed.

D. Proposed model

Existing intelligent tutoring systems are based on capturing users playing and giving feedback based on the type of mistakes that were made. Database of lessons are mostly static and all users get the same lessons. Fig. 1. [29] shows reasoning mechanics that was developed and used in previously developed computer systems. Pointing out where mistakes were made during instrument playing is one of the main dimensions in Csikszentmihalyis model. Most of developed systems have this feature but usually goals are not clear and specially balance between perceived challenges and skills is not managed in any way. Usually, there are lessons that are static (always the same) and there is only one path through the knowledge base and lessons.

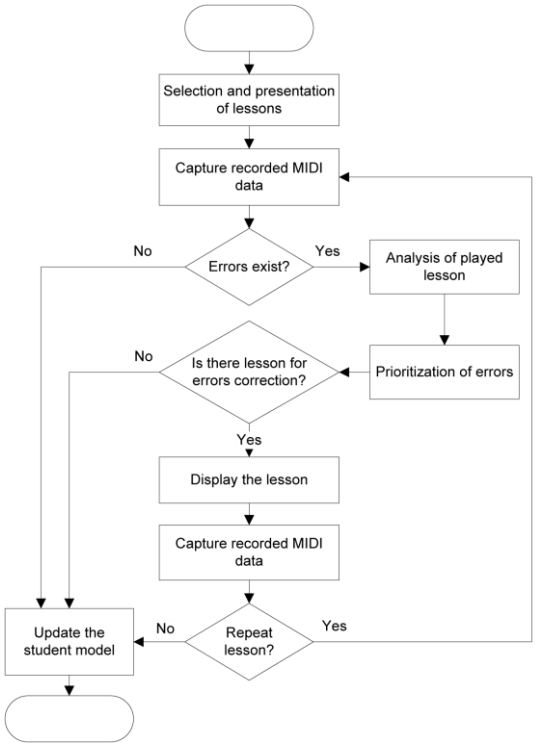


Figure 1. Model of reasoning mechanics in previously developed computer systems

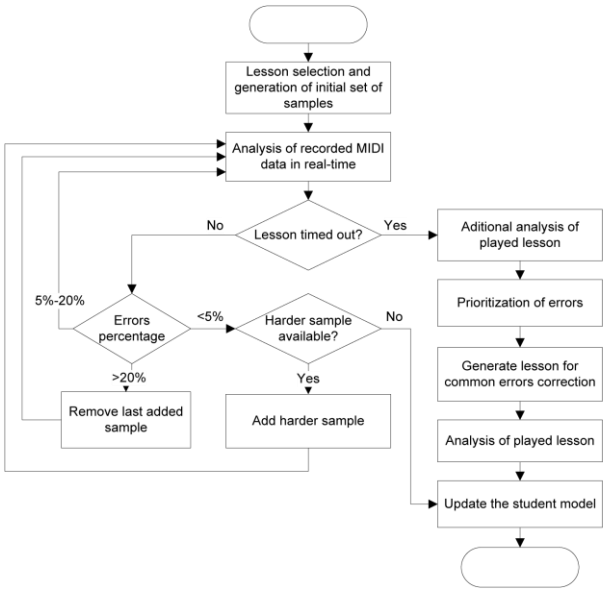


Figure 2. Proposed model of reasoning mechanics in real-time

In order to provide somewhat improved solution a new model which is shown in Fig. 2 is proposed and it enables real-time feedback and real-time adaptation to learners' skills [29]. Music lessons are divided into smaller patterns and based on percentage of correctly played patters, lessons are adapting in real-time. If percentage of correctly played patters is high, then new harder patterns are added into lessons. If that percentage is lower than expected, then harder patterns are removed and easier patterns are included. Also, after every particular lesson, types of mistakes are analyzed. Based on the highest amount of specific types of mistakes, additional lessons are generated. These lessons are always a bit different, because they are generated based on patterns from the knowledge base and types of mistakes that were made. Finally, path through knowledge base is not static but dynamic. Players that will have higher percentage in pattern playing will go faster through the knowledge base while those that learn slower will have a slower pace. Each lesson has a clear goal and that is to learn patterns that are presented. To additionally motivate learners, musical backing tracks are used to make playing experience even more interesting.

III. RESEARCH RESULTS

A. Motivation

Statistical data that was presented in the introduction of this paper shows that high number of people want to learn how to play music instrument. Important question was how these people perceive the concepts that have been presented in proposed new model. One of the important steps in creating adaptive learning systems is to determine users' needs but generally not much evaluation in the domain of intelligent tutoring systems can be found. For example, Violin Tutor has been evaluated by 12 users [21]. PianoFORTE was evaluated but evaluation results have never been published [22]. PianoTutor was evaluated by 18 users [23], where 10 finished evaluation process. i-Maestro has many modules for technology-enhanced learning but evaluation results were never published [20]. IMUTUS project was tested on 12 users, 6 in IMUTUS group and 6 in control group [24]. Although these studies gave positive feedback, studies on a larger scale are needed to get statistically more significant results.

B. Questionnaire results

We conducted a survey to determine perception of potential users of our proposed model and its features. Survey was conducted among 126 students (n=126). Answers were based on Likert scale, where 1 means "strongly disagree", 2 "disagree", 3 "neutral", 4 "agree" and 5 "strongly agree". Results of conducted research are shown in Table I.

TABLE I. RESULTS OF QUESTIONNAIRE RESEARCH

No.	Questionnaire item	Mean	Std. dev.
Q1	I want to learn how to play a musical instrument.	3,72	1,25
Q2	In my daily schedule I don't have time to attend music school.	3,89	1,25
Q3	I would like to have possibility to learn how to play a musical instrument with self-paced tempo and schedule.	4,22	0,97
Q4	I would like to be able to learn how to play a musical instrument with the help of a computer.	3,90	1,19
Q5	I would like to get feedback of my learning progress while learning to play a musical instrument with a computer.	4,12	1,14
Q6	I want music lessons to adapt to my knowledge and my tempo of progress.	4,29	0,96
Q7	I think it's important that the content of music lessons adapt individually to every learner.	4,24	1,01
Q8	I think it's important that the order of music lessons adapt individually to every learner.	4,28	0,92
Q9	I consider that automated learning of playing a musical instrument by using computer is important.	3,72	1,02
Q10	I would recommend automated learning system of playing a musical instrument that adapts to learners skills and needs to my colleagues.	3,96	1,03
Q11	Automated and adaptive way of learning how to play a musical instrument would make learning process easier.	4,07	0,96
Q12	Automated and adaptive way of learning how to play a musical instrument would make learning process more interesting.	4,20	0,97

The research results have shown that most of the students would like to learn how to play a musical instrument (58%). This data is consistent with the statistical data presented in the introduction about the number of people that still want to learn how to play. Most of the students perceive that they don't have enough time in their daily schedules to attend music lessons. But as it was mentioned before, there are alternatives. More and more people learn how to play a musical instrument on their own. But that is highly unstructured and unsystematic way of learning. Research results have shown that people would like to learn how to play a musical instrument with the help of computer, considering their own tempo and schedule. Feedback has been

reported as very important to everyone. But aspect that was perceived as even more important is to have adaptive way of learning where lessons and their content adapt to students' knowledge and skills (Q6, Q7, Q8). Students perceive that this kind of learning would be both easier and more interesting and that they would recommend this kind of learning to their colleagues.

IV. CONCLUSION

Music is an integral part of modern society and music education is important for many people. More and more technology is becoming an integral part of all educational systems. There are a few intelligent tutoring systems that are designed to learn their users how to play a musical instrument and provide users with feedback about their playing. Adaptability to users' needs, skills and their knowledge is an integral part of those systems. The new proposed model extends existing systems with real-time feedback mechanism and generated lessons that help users to correct their errors in playing. Research results have shown that those capabilities are very important to students. With higher adaptability, lessons are more suited for each individual and that optimizes educational process. Students perceive that with this kind of system, learning process would be easier and more interesting. Both motivational and educational goals would be reached. Development of adaptive intelligent tutoring system based on new proposed model that will teach students how to play drums will be a part of future research. This system will then be compared to other alternative ways of learning how to play drums in order to conclude whether its additional adaptability has any positive effects on achieving educational goals and whether students will be more motivated by using this system.

REFERENCES

[1] L. Hetland, "Listening to music enhances spatial-temporal reasoning: Evidence for the 'Mozart effect.'", The Journal of Aesthetic Education, vol. 34, no. 3/4, 2000, pp. 105-148.

[2] W. F. Thompson, E. G. Schellenberg, and G. Husain, "Arousal, mood and the Mozart effect", Psychological science, vol. 12, no. 3, 2001, pp. 248-251.

[3] E. G. Schellenberg, "Music lessons enhance IQ", Psychological Science, vol. 15, no. 8, 2004, pp. 511-514.

[4] E. G. Schellenberg, "Long-term positice associations between music lessons and IQ", Journal of Educational Psychology, vol. 98, no. 2, 2006, pp. 457-468.

[5] N. Kraus, and B. Chandrasekaran, "Music training for the development of auditory skills", Nature Reviews Neuroscience, vol. 11, no. 8, 2010, pp. 599-605.

[6] S. Moreno, E. Bialystok, R. Barac, E. G. Schellenberg, N. J. Cepada, and T. Chau, "Short-term music training enhances verbal intelligence and executive function", Psychological science, vol. 22, no. 11, 2011., pp. 1425-1433.

[7] C. Gaser, and G. Schlaug, "Brain structures differ between musicians and non-musicians", The Journal of Neuroscience, vol. 23, no. 27, 2003, pp. 9240-9245.

[8] G. Schlaug, A. Norton, K. Overy, and E. Winner, "Effects of music training on the child's brain and cognitive development", Annals of the New York Academy of Sciences, vol. 1060, no. 1, 2005, pp. 219-230.

[9] Number of people playing musical instruments in the U.S. from 2002 to 2010, 2011. [Online] Available: http://www.statista.com/statistics/192834/people-playing-a-musical-instrument-in-the-us/ [Accessed: 02-Feb-2015].

[10] Attitudes Toward Music, 2009. [Online] Available: http://www.nxtbook.com/nxtbooks/namm/2010globalreport/index.php?startid=154#/136 [Accessed: 04-Feb-2015].

[11] Share of adults taking a music lesson or class during their lifetime in the U.S. from 1982 to 2012, 2013 [Online] Available: http://www.statista.com/statistics/195460/participation-in-music-classes-and-lessons-in-the-us-since-1982/ [Accessed: 02-Feb-2015].

[12] Music Council of Australia 2001, 2002 [Online] Available: http://artfacts.australiacouncil.gov.au/music/participation-7/many-people-wish-they-had-learned-to-play-a-musical-instrument/ [Accessed: 03-Feb-2015].

[13] M. Brandao, G. Wiggins, and H. Pain, "Computers in music education", In Proceedings of the AISB '99 Symposium on Musical Creativity, pp. 82-88, 1999.

[14] M. M. Farbood, E. Passztor, and K. Jennings, "Hyperscore: a graphical sketchpad for novice composers", Computer Graphics and Applications, IEEE, vol. 24, no. 1, 2004, pp. 50-54.

[15] L. Ferrari, A. R. Addessi, and F. Patchet, "New technologies for new music education: The Continuator in a classroom setting", Proceedings of ICMPC9 and 6th ESCOM Conference, pp. 1392-1398, 2006.

[16] Music Ace 2, 2013, [Online] Available: https://www.harmonicvision.com/ma2fact.htm [Accessed: 04-Feb-2015].

[17] MiBAC Music Lessons 2, 1996, [Online] Available: http://pianoeducation.org/pnomuslr.html [Accessed: 04-Feb 2015].

[18] Practica-Musica personal music tutor, 1999, [Online] Available: http://www.ars-nova.com/practica6.html [Accessed: 04-Feb 2015].

[19] GNU Solfege, 2000, [Online] Available: http://www.gnu.org/software/solfege/ [Accessed: 04-Feb-2015].

[20] K. Ng, and P. Nesi, "i-Maestro: Technology-enhanced learning for music", Proceedings of the 8th International Conference on New Interfaces for Musical Expression, NIME, pp. 4-8, 2008.

[21] J. Yin, Y. Wang, D. Hsu, "Digital violin tutor: an integrated system for beginning violin learners", In Proceedings of the 13th annual ACM international conference on Multimedia, ACM, pp. 976-985, New York, NY, USA, 2005.

[22] S. W. Smoliar, J. A. Waterworth, and P. R. Kellock, "PianoFORTE: a system for piano education beyond notation literacy", In Proceedings of the 3rd ACM international conference on Multimedia, pp. 457-465, 1995.

[23] R. Dannenberg, M. Sanchez, A. Joseph, R. Saul, and P. Capell, "Results from the Piano Tutor Project", In Proceedings of the 4th Biennial Arts and Technology Symposium, pp. 143-150, 1993.

[24] S. Raptis, A. Chalamandaris, A. Baxevanis, A. Askenfeld, E. Schoonderwaldt, K. Hansen, D. Fober, S. Letz, and Y. Orlarey, "Imutus - an effective practicing environment for music tuition", Proceedings of the International Computer Music Conference, pp. 383-386, 2005.

[25] G. Tambouratzis, K. Perifanos, I. Vougari, A. Askenfelt, S. Granqvist, K. F. Hansen, Y. Orlarey, D. Fober, and S. Letz, "VEMUS: An Integrated Platform to Support Music Tuition Tasks", In Advanced Learning Technologies, ICALT'08, Eight IEEE International Conference, pp. 972-976, 2008.

[26] Learno Piano Online Review, 2015, [Online] Available: http://musical-instruments.toptenreviews.com/learn-piano-online-review/ [Accessed: 05-Feb-2015].

[27] List of music video games, 12-Aug-2014, [Online] Available: http://en.wikipedia.org/wiki/List_of_music_video_games [Accessed: 06-Feb-2015].

[28] M. Csikszentmihalyi, Finding Flow: The psychology of engagement with everyday life. New York, Basic Books, 1997.

[29] M. Konecki, "Learning to play musical instruments through dynamically generated lessons in real-time based on adaptive learning system", Proceedings of the 25th Central European Conference on Information and Intelligent Systems, pp. 124-129, 2014.

Intelligent Tutoring Systems for Language Learning

V. Slavuj*, B. Kovačić* and I. Jugo*
* Department of Informatics, University of Rijeka, Rijeka, Croatia
vslavuj@uniri.hr, bkovacic@uniri.hr, ijugo@uniri.hr

Abstract - **Intelligent tutoring systems (ITS) are able to provide a personalized approach to learning by assuming the role of a real teacher/expert who adapts and steers the learning process according to the specific needs of each learner. This is done by taking into account a number of learner features and deciding on the best action to follow at any point in the learning process. Within the field of computer assisted language learning (CALL), the "computer-as-a-tutor" modality has been widely accepted for some time now, although it has long been overshadowed by the "computer-as-a-tool" modality. This was mainly due to the observable lack of interactivity of language learning systems and the inability of technology to show "intelligence". However, the development of artificial intelligence in general, and natural language processing, user modeling and ITSs in particular, gave impetus for the development of the field referred to as intelligent CALL (ICALL). The paper at hand outlines and briefly discusses the issues surrounding the development and use of ITSs for language learning, taking also into account the broader context of (I)CALL, and gives an overview of such systems already in use.**

I. INTRODUCTION

When considering any form of technology-enhanced or technology-based education regarding any subject matter, it is necessary to address the issue of learner variability. For some time now, it has been recognized that the educational paradigm "one size fits all" is not the right approach when dealing with a heterogeneous population of learners, as learners may differ in a number of characteristics and individual traits, which makes each learner's approach to learning unique [1][2][3]. Taking care of learner differences is often a difficult task, especially in cases where these differences are rather prominent, and requires a great amount of planning time and effort from the teacher. A personalized approach to instruction is particularly important in educational contexts where the teacher is not immediately present (but rather distant), or is altogether absent. Such cases include, among others, online learning environments and web-based learning, the use of standalone tutorial software, and hybrid approaches to using technology. Leaving learners on their own in such environments, offering them no or loosely structured guidance through the learning process, might lead to poor attainment and subsequent failure of the teaching curriculum.

The field of computer assisted language learning (CALL) is not an exception. However, it needs to be pointed out that using computers (or other technologies such as mobile devices, CD players or TV) in the context of language learning is somewhat different than is the case with learning in other fields [4]: besides being one of the most complex undertakings, the process of learning natural languages is additionally complicated by the fact that it requires both knowledge-based and skill-based approach [5].

There is a general consensus and an abundance of research evidence that the use of technology, computers in particular, has positive effects in the language learning context (e.g. in [1] and [6]), and that technology could be used as the basis for an individualized approach to learning languages [7][8]. The main question thus is not whether to use computers for language learning/teaching, but how to employ them so that learner needs are met [9].

As suggested in [8], the main goal of CALL should be to simulate human teachers and their pedagogical competence whenever, wherever and to whomever necessary. It is possible to achieve individualization in both commonly accepted modalities within CALL, namely, computer-as-a-tool and computer-as-a-tutor. Within the computer-as-a-tool paradigm, computers are employed as mere teaching tools which increase the efficiency and/or efficacy of the language learning and teaching process [10]. Personalization in this context remains the responsibility of the teacher and should be addressed during the process of preparing and setting up learning tasks. This is achievable by, e.g., scaling the complexity of a task, suggesting the use of software packages which are likely to help learners in carrying out their tasks (such as online thesaurus or translation software), pairing or grouping learners appropriately or even choosing the right Web 2.0 tools that will be most beneficial for the development of a targeted skill. Conversely, the computer-as-a-tutor paradigm represents the situations in which the computer has temporarily taken over the role of a language teacher with the aim of addressing all learner needs as a human teacher would [10]. Moreover, the computer is responsible for tracking learner's interaction with the designated software, storing that information for future reference, making pedagogical decisions, identifying learner errors and hypothesizing about the underlying cause(s), forming appropriate feedback meaningful for the learner, and, finally, adjusting its assumption about the overall level of learner's knowledge of the subject matter. This however requires the usage of more sophisticated educational software that employs rather complex, intelligent and adaptive technologies - exemplified by user modeling (UM),

natural language processing (NLP), expert systems (ES), adaptive systems (AS) and intelligent tutoring systems (ITS) – and constitutes what is generally referred to as intelligent computer assisted language learning (ICALL) [11]. Our focus in this paper will primarily rest on ITSs as examples of systems taking over the role of real language teachers, but we will also touch upon other above stated technologies as they often appear as an integral part of ITSs.

The aim of this paper is to outline and discuss the issues surrounding the development and use of ITSs in the field of computer assisted language learning, taking also into consideration the broader context of (I)CALL. The paper also aims to give an overview of such systems developed for research purposes or successfully integrated into the language learning process.

The paper is structured as follows. Firstly, a theoretical framework for using ITSs in language learning is set. Secondly, the issues surrounding the development and use of ITSs for language learning are outlined and discussed. Thirdly, a variety of language-learning ITSs already in use are described as well as how they are employed to address different language learning skills. Finally, conclusions about using ITSs for language learning are reached.

II. Intelligent Tutoring Systems in Language Learning

The field of ITS is usually referred to as the field of research which deals with the modes of designing and developing adaptive and intelligent educational systems [12]. The term is further used to mark a range of computer-based instructional systems which keep up-to-date models containing information about what and how to teach, and use that knowledge for personalizing the access to learning [2].

Although ITSs are now generally accepted and seen as useful within the field of (I)CALL, we need to point out however that the situation was somewhat different a couple of decades ago, as practitioners mostly disregarded the tutor role of computers and only focused on the computer-as-a-tool modality [10]. The literature enumerates a number of possible reasons why language tutoring software was underappreciated and, consequently, marginalized, but we single out two reasons that in our opinion seem most plausible.

Firstly, the view that tutorial software is behavioral in nature and uses only "drill and kill" type of exercises deterred a number of teachers from using it, as it did not adhere to the dominant teaching methodology based on the notion of communicativeness [4][10].

Secondly, with the appearance of personal computers, too much was expected of CALL. Looking back at that period from our current standpoint we can clearly see that technology was not mature enough back then and software development lagged behind hardware development. As a consequence, there was a general disappointment with such software as it was rather rigid and lacked both interactivity and intelligence [5][10][13].

Notwithstanding, language tutoring systems are now widely acknowledged and are seen as being able to address more than just simple grammar and vocabulary teaching/learning, but they need to be designed keeping in mind learners' and instructors' needs [4][14].

General architecture of an ITS suggests that there are four functional components to every such system: domain or expert knowledge model, learner model, tutoring or instructional or adaptation model and user interface model [2][15]. In the section to follow we describe each of the ITS architecture components in more detail with a special emphasis on how they are applied in the language learning context. Our account will include an additional module frequently used in language learning oriented ITSs, namely the feedback module.

A. *The Domain Model*

The domain model contains all the expert knowledge of a particular domain that needs to be taught by the system [2][11]. It details different concepts of the learning domain, their hierarchy, mutual interconnection, and how they relate to educational resources to be delivered [16][17]. Broadly speaking, there are three main ways how domain models are conceptualized [16]: the set/vector model, the hierarchy model and the network model. The most popular and widely used type of domain model is the network model. In this model, the concepts are arranged into a network and are interconnected by different types of links. The most common type of links are, amongst others, prerequisite links, "is-a" or "part-of" links, and remedial links [16]. Links in both hierarchy and network models are very useful in for improving the accuracy of the learner model. For example, in case the learner does not possess knowledge related to a certain concept (e.g., formation of the Simple Past Tense), the links help in providing the most likely remedial concepts (e.g., irregular verb forms). Conversely, if the learner possesses knowledge related to a certain concept (e.g., formation of the first conditional), the links help in identifying other concepts that the learner is likely to know (e.g., the Simple Present Tense).

B. *The Learner Model*

In order to recognize individual learners' needs successfully and take appropriate pedagogical action, a piece of software needs to be able to track what learners do with it while learning a language, i.e., how they behave during usage [18][19]. Such monitoring entails the creation and constant updating of a student model which gathers all the necessary data from the learner, be it the learning path, correctness of answers to questions, time taken to offer a solution, or the overall success in solving a task [2][14][19][20]. Learner models may contain other relevant information about the learner as well, such as their previous level of linguistic knowledge, learning style preference, cognitive style, strategic competence, goals or interests, to name just a few [2][21]. Typically, the information in the model should be enough to differentiate between learners who are using the system on the basis of their knowledge, task performance, overall proficiency, goals or background, and take appropriate action to personalize their learning experience [16][22]. Hence, the main goal of a learner model is to support the process of making assumptions about the learner and to support

pedagogical decision-making concerning, e.g., the feedback to be delivered back to the learner or the choice of appropriate follow-up activities, as defined in the instructional model [23]. The two broadly recognized and used approaches to student modeling are stereotype modeling and feature-based modeling [16].

However, we should be aware that monitoring is not an end in itself [20][23][24]: in addition to carefully deciding what and when to monitor, the tracking process is further complicated by the issue of appropriate interpretation of the gathered data and correctness of inferences made from that data. Erroneous assumptions about the learner's input, his language knowledge and skills, may cause wrong action by the system, consequently leading to an inefficient learning process and poor language attainment [14].

C. The Tutoring Model

The tutoring model is made up of rules that define how a system should act during runtime, taken into consideration the data from the learner model [17]. It specifies which pedagogical action should be taken following each user activity. Therefore, this model defines what, when and how to adapt within the system [25]. In order for the adaptive action of a language learning-oriented ITS to be effective, the knowledge from language learning theory (usually First language acquisition (FLA) or Second language acquisition (SLA)) and pedagogical practice should be included as part of this model.

D. The User Interface Model

The interface model was only recently added to the above three, and is concerned with the graphical environment through which educational content is delivered to the learner [2]. In particular, it deals with the representation of instructional material, mostly focusing on multimedia and its advantages for the learner.

E. The Feedback Module

Another important and frequently included element of an ITS for language learning is the feedback module. Language teachers are often engaged in giving feedback to learners concerning their progress, success on a particular task, language areas they need to work on, etc. Similarly, tutoring systems should also be able to adequately respond to a learner activity.

Giving feedback may consist in a straightforward declaration of errors (e.g., *The answer you provided is not correct.*), request for additional clarification (e.g., *What do you mean by…*) or restating something in a correct manner [18]. In many ways this entails the usage of NLP techniques so that the computer is able to make sense of the learner's written or spoken input.

From the point of view of technology, there are three possible approaches to appropriate feedback formation: pattern matching-based approach, statistical-based approach and rule-based approach [26].

III. Issues Surrounding the Development of ITSs

Designing each of the components of an ITS for language learning requires a careful consideration of the needs of both learners and teachers. In the light of that and based on available reports from literature on specific language-learning ITSs, we point out some of the issues that may arise during the development process, briefly explain them, and, where appropriate, offer some suggestions as to possible solutions.

One of the frequently noticed issues in language-learning ITSs is over-restricting the learning domain, both horizontally and vertically. Horizontal domain restriction refers to a system's tendency to focus on a single linguistic skill (listening, reading, speaking and writing) or form (grammar, vocabulary) rather than a combination of them. Including a whole range of skills/forms or even all of them into the ITS would be more in line with the nature of language learning, which emphasizes the integration of both productive and receptive skills, visual and auditory input, written and spoken output, form and communicative competence, etc [27]. Conversely, vertical domain restriction refers to the inability of an ITS to cater for learners with different levels of language proficiency (e.g., CEFR proficiency levels A1 – C2), which decreases the number of potential system users and, consequently, system's (potential) financial value. We maintain that a possible solution and a more sensible approach would be to offer learning content designed for a wide range of skills, forms and ability levels, and use the knowledge about the learner, kept by the ITS, to detect which skills/forms are needed by each learner. Moreover, learner data should be used to identify the learner's current level of proficiency, offer learning materials that are appropriate for the identified level, and decide when the new level of proficiency has been reached.

The issue of user input evaluation is also a significant one. Whereas evaluation of certain linguistic skills and forms can be automated and independently done by the system (e.g., reading, listening and grammar), some skills are notoriously difficult for the computer to evaluate on its own. This particularly holds for productive skills, namely speaking and writing, which involve unrestricted user output. Even though natural language processing and speech recognition techniques have come a long way in terms of accurately evaluating user output, they maintain to exhibit certain flaws [13]. At present, a computer is still unable to fully replace a human being when it comes to evaluating unrestricted oral or written output. However, restricting user output to a particular domain with specific vocabulary (e.g., describe a room or a person, explain a shopping list) might make a significant difference in rising the accuracy of evaluating output.

Another issue, related to vertical domain restriction, is standardization and comparability between systems. Most of the ITSs reported in literature do not use some generalized framework of ability reference (such as CEFR or ILR scale) to define the level of learner proficiency for which the learning content of the ITS is intended, but specify such a level in terms of school grade or educational level only (e.g., college freshmen course). This significantly affects learner mobility between different ITSs, mostly because learners are unable to

objectively assess how appropriate the language learning system is for their proficiency level. Furthermore, the results of using an ITS for language learning should be observable as a shift in the proficiency level, quantified in familiar terms, and available for analysis for third parties. Aligning the content with any of the internationally set and verified standards for language learning proficiency levels can be a way to successfully address such issues.

Designing learning materials that will be presented to a learner during the educational process involve several issues that need to be acknowledged. The most notable among those is the issue of interactivity of learning materials. When using languages for communication, learners are engaged in an interactive process of sending and receiving messages through a number of communication channels. In certain cases, the meaning of messages cannot be fully grasped simply on the basis of the sum of words and their meanings, but other elements, such as tone of voice, gestures and body language, and contextual information, need to be taken into consideration. Hence, learning materials should reflect real world communicative situations, which is possible to achieve through the use of multimedia. Another issue related to learning materials is how to keep learners motivated during the learning process, and how to actively involve them to work with materials at hand. One possible way is to give small-scale task to the learner, such as "click on the right word" or "translate a particular word", which can also serve as indicators of learner knowledge and progression through the learning domain.

Feedback module design should deal with the issue of selecting or generating appropriate feedback once the learners have finished interacting with the system and the computer has evaluated their response. The feedback message in language-learning environments needs to be beyond a simple correct-incorrect statement: it needs to be able to point out the source of the linguistic problem and offer appropriate scaffolding towards obtaining correct knowledge. Additionally, the degree of feedback explicitness should depend on the already discussed level of language proficiency: more explicit feedback for lower-ability learners, and less explicit for higher-ability ones.

The final issue we point out here is how to model language teacher competences and their teaching strategies. For some time now, there seems to exist a strong tendency towards socio-cultural and socio-cognitive approaches to language learning, where the communicative function of language is strongly emphasized. As explained in [27], such an approach underlines two key concepts, namely tasks and interaction: language users and learners are seen as "social agents" who use their "communicative language competences" in numerous "language activities" to exchange meaningful messages as they try to accomplish a particular "task" at hand. The role of the language teacher in such a model seems to be the one of support, providing the learners with tasks appropriate to their level of knowledge and skill, observing their performance during the communicative act, and offering appropriate feedback. Additionally, based on the existing evidence (usually in the form of previous performance results), the teacher decides upon the best action to follow, be it a remedial task or a task involving the acquisition of new knowledge. ITSs can be considered a technology capable of successfully taking over the responsibilities of the language teacher. In order to do so, an ITS needs to be designed specifically for language learning purposes: sound pedagogy and language teaching methodology, supported by what is known from FLA or SLA theory, represents the basis on which to design an intelligent computer assisted language learning system.

The issues reflected upon above do not make up a definitive list of all the possible issues that may arise during language-learning ITS design, but they represent situations often encountered in practice. Moreover, these issues contribute to the discussion on the complexity of the process of conceptualizing an intelligent system for language learning which is one of the topics addresses in this paper.

IV. Examples of ICALL Systems in Use

There have been numerous reports on the development of ICALL systems over the years; however, it seems that most of the systems have remained in the prototype phase or have been developed purely for research purposes. Literature review has revealed three systems in particular that are confirmedly used in the language classroom, integrated into the foreign language learning curriculum, and continuously updated and improved [14]: E-Tutor for learning German as a second language, TAGARELA system for learning Portuguese at the university level, and Robo-Sensei for Japanese.

E-Tutor system is a comprehensive Web-based learning environment for university level learners of German. It employs AI and NLP techniques in order to achieve individualized learning experience. The system is mainly focused on vocabulary and grammar improvement for which it offers several types of exercises, namely translation, dictation, sentence formation and providing the missing word. Evaluation of these activities is supported by NLP and feedback is generated based on learner performance. In addition, each section has activities for listening and reading comprehension, culture and writing. Writing assessment, however, is not automated, and is done by the teacher [19][23].

TAGARELA is an ICALL system for learning university level Portuguese, intended to complement the existing pedagogical materials. It serves the purpose of an electronic workbook, and offers feedback on spelling, morphology, syntax and semantics to each individual learner. There are six activity types in total: reading and listening comprehension, picture description, rephrasing, fill-in-the-blanks and vocabulary exercises. Creators of the system point out four main tasks the system can perform: detect errors in the student input, diagnose knowledge level and skills of the learner, adapt the system accordingly and generate feedback [14][21].

Robo-Sensei is used for teaching and learning Japanese in 24 lessons, and focuses mainly on translation tasks. It receives input from the learner, performs itemization and morphological analysis, and parses the sentence syntactically. However, the sequence of

activities is the same for all learners, as is the feedback. Thus, the system does not adapt to student level of proficiency or knowledge about particular language items, but only offers NLP services [28].

Additional examples of ICALL systems address a variety of language skills. Thus, in Ref. [24] and [29] reports on WUFUN are found. WUFUN is a system for Chinese university students learning English, based on a regularly updated learner model. The system is special as it features both aural and visual input, which contributes to saliency, and, possibly, better attainment.

In [30], an ICALL system called Your Verbal Zone (YVZ) is described, which is used to support Turkish students' English vocabulary learning. YVZ contains a morphological analyzer which is able to find the root of a morphologically complex word and return its base form together with affixes. It, of course, uses NLP in order to achieve that. In order to further support students, the system contains a built-in, bilingual dictionary, a number of examples related to word use and details of function and meaning of particular affixes.

In [13], the case for WordBricks is argued, an intelligent system supporting the process of writing grammatically correct sentences. The system features a grammar checker and allows students to produce free utterances. This also permits learners to experiment with language and test their own hypotheses on language structures. Feedback generation system is also implemented.

In [31], a program called C-DA, used for improving learners' metacognitive reading strategies is described. The standalone program, also available in the Web browser, keeps track of learner answers to questions, and for each incorrect answer, it initiates the "mediation" mechanism which helps the learner solve the task. The mediation mechanism is designed to offer more implicit help at first, and if the error persists and the learner is unable to provide the correct answer, help gradually becomes more explicit.

An ICALL system called CASTLE is described in [32]. It provides an opportunity for language learners to practice their communicative skills, namely writing, through a number of predetermined role-play scenarios. Upon wrong input and depending on how the error is classified by the diagnosis module, CASTLE takes remedial action in the form of additional grammar exercises to stress certain formal language structures. In the same time, the system keeps an up-to-date student model which holds data concerning student proficiency for a particular topic, data about errors and proneness to commit them, and the estimates of the likely cause, so the system is able to adapt its behavior towards each learner.

A description of a Web Passive Voice Tutor is given in [33]. The system uses stereotype modeling to initialize the student model, applies NLP techniques for intelligent analysis of student solutions to tasks and tailors feedback and advice to each student separately. In addition, it uses the so called link annotation technique to guide the learner through the learning domain by suggesting the most appropriate activity to follow.

I-PETER, as described in [34], is an intelligent system for language learning that allows for three distinct ways of learning: identifying and mending holes in the existing knowledge of the subject matter; improving on the current level of knowledge of a selected part of the domain; selecting and practicing only selected concepts and/or subconcepts of the domain. In each case, the learner needs to give an answer to a particular question, and once it is evaluated by the system, the learner may request detailed theoretical explanation regarding the error. In addition, there is a possibility for the system to fully take charge of the learning process.

A personalized English article recommending system is presented in [35]. This system holds information about the linguistic ability of a learner and selects the most appropriate article to be read next. After reading it, learners are given a vocabulary test consisting of newly-encountered words from the article. Based on their performance on the test, the learner model is updated, linguistic ability recalculated, and a new article is chosen for delivery.

Lastly, we mention a computerized adaptive testing system for Chinese (reading and listening) proficiency [36]. This system offers each learner an individualized test in which the selection of the next test item is done "on-the-fly", and depends on the (correctness of) answers to previously generated test items. After evaluating each answer, the system updates its estimate of learner proficiency. Thus, the system needs to keep a log of the estimated level of proficiency together with all the answers to all the test items previously administered. Test items are generated until a predetermined termination rule is reached (e.g., a sufficient number of items have been given, a certain amount of time has elapsed, etc.).

The examples of software given above clearly reveal that ICALL constitutes a broad field of both research and practice. Systems differ in their focus on skill, level of complexity, degree and manner of employing AI and ITS technologies, way of generating feedback, even learner modeling. One thing they have in common however is that they are usually restricted to a single problem area instead of addressing a variety of issues. Regardless of their differences and possible shortcomings, we should evaluate a piece of software not by what it is capable of doing, but by how well it does what it is intended to do.

V. CONCLUSION

Individualizing language learning experience is certainly facilitated by the maturation of technologies from computer science, in particular AI. These technologies offer new possibilities for developing better and more efficient software that will be available even at a distance and without the presence of a teacher.

In this paper, we have given an overview of ITSs as an example of technology often used within the field of ICALL to individualize the process of language learning and teaching. Development and implementation of language-specific ITSs is not without issues which need particular attention from the development team. Addressing such issues in a systematic way should

improve the overall quality and usability of an ITS, and contribute to a better learning/teaching experience. Some of the ITSs included in our overview have successfully managed to solve certain issues raised in our paper, while other continue to persist.

Given the current state of technology, we maintain that software is not yet capable of fully replacing the language teacher, but is able to do so only to a (significant) degree. Only further developments in AI and related natural language processing technologies will see the goal of full replacement be reached.

REFERENCES

[1] M.D. Bush, "Computer-assisted language learning: from vision to reality?", CALICO Journal, vol. 25(3), pp. 443-470, 2008.

[2] M. Vandewaetere, P. Desmet, and G. Clarebout, "The contribution of learner characteristics in the development of computer-based adaptive learning environments", Computers in Human Behaviour, vol. 27, pp. 118-130, 2011.

[3] T. Heift, "Learner personas in CALL", CALICO Journal, vol. 25(1), pp. 1-10, 2007.

[4] N. Garrett, "Computer-assisted language learning trends and issues revisited: integrating innovation", The Modern Language Journal, vol. 93, pp. 719-740, 2009.

[5] M.D. Bush, "Computer-assisted language learning: from vision to reality?", CALICO Journal, vol. 25(3), pp. 443-470, 2008.

[6] U. Felix, "The unreasonable effectiveness of CALL: what have we learned in two decades of research?", ReCALL, vol. 20, pp. 141-161, 2008.

[7] M. Evans, "Digital technology and language learning: a review of policy and research evidence," in Foreign language learning with digital technology, M. Evans, Ed., London: Continuum International Publishing Group, 2009, pp. 7-31.

[8] N. Tokuda, "New developments in intelligent CALL systems in a rapidly internationalized information age", Computer Assisted Language Learning, vol. 15(4), pp. 319-327, 2002.

[9] P. Hubbard, "General introduction," in Computer assisted language learning: Vol 1 (Critical concepts in linguistics), P. Hubbard, Ed., London: Routledge, pp. 1-20, 2009.

[10] P. Hubbard and C. Bradin Siskin, "Another look at tutorial CALL", vol. 16, pp. 448-461, 2004.

[11] M. Schulze, "AI in CALL - artificially inflated or almost imminent?", CALICO Journal, vol. 25(3), pp. 510-527, 2008.

[12] P. Brusilovsky, "Adaptive and intelligent technologies for web-based education", International Journal of Artificial Intelligence in Education, vol. 13(2-4), pp. 159-172, 2003.

[13] M. Mozgovoy, "Towards WordBricks - a virtual language lab for computer-assisted language learning", Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 229-232, 2012.

[14] L. Amaral and D. Meurers, "On using intelligent computer-assisted language learning in real-life foreign language teaching and learning", ReCALL, vol. 23, pp. 4-24, 2011.

[15] M. Virvou, D. Maras, and V. Tsiriga, "Student modeling in an intelligent tutoring system for the passive voice of English language", Educational Technology & Society, vol. 3(4), pp. 139-150, 2000.

[16] P. Brusilovsky and E. Millan, "User models for adaptive hypermedia and adaptive educational systems", in The adaptive web, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds., Berlin: Springer-Verlag, 2007, pp. 3-53.

[17] P. Karampiperis and D. Sampson, "Adaptive learning resources sequencing in educational hypermedia systems", Educational Technology & Society, vol. 8(4),pp. 128-174, 2005.

[18] R. Fischer, "How do we know what students are actually doing? Monitoring Students' behaviour in CALL", Computer Assisted Language Learning, vol. 20(2), pp. 409-442, 2007.

[19] T. Heift, "Developing an intelligent language tutor", CALICO Journal, vol. 27(3), pp. 443-459, 2010.

[20] J.Bertin and J. Narcy-Combes, "Monitoring the learner - who, why and what for?", Computer Assisted Language Learning, vol. 20(5), pp. 443-457, 2007.

[21] L. Amaral and D. Meurers, "Conceptualizing student models for ICALL", Proceeding of the 11th International Conference on User Modelling, pp. 340-344, 2007.

[22] N.V. Dung, S. Sulaiman, and M.F. Hassan, "Towards the development of an online adaptive system framework for English learning using perturbation learner model", Proceedings of the 2012 International Conference on Computer & Information Science - ICCIS, vol. 1, pp. 512-516, 2012.

[23] T. Heift, "Modeling learner variability in CALL", Computer Assisted Language Learning, vol. 21(4), pp. 305-321, 2008.

[24] Q. Ma, "From monitoring users to controlling user actions: a new perspective on the user-centred approach to CALL", Computer Assisted Language Learning, vol. 20(4), pp. 297-321, 2007.

[25] S. Lee, W. Lee, H. Kim, S. Jung, and H. Lim, "A personalized CALL system considering users congnitive abilities", ICCSA 2010, Part IV, LNCS 6019, pp. 376-384, 2010.

[26] K.F. Shaalan, "An intelligent computer assisted language learning system for Arabic learners", Computer Assisted Language Learning, vol. 18(1&2), pp. 81-108, 2005.

[27] Council of Europe, "Common European Framework of Reference for Languages: Learning, teaching, assessment", Cambridge: Cambridge University Press, 2001.

[28] N. Nagata, "Robo-Sensei's NLP-based error detection and feedback generation", CALICO Journal, vol. 26(3), pp. 562-579, 2009.

[29] Q. Ma, "Empirical CALL evaluation: the relationship between learning process and learning outcome", CALICO Journal, vol. 26(1), pp. 108-122, 2008.

[30] Ö. Esit, "Your Verbal Zone: an intelligent computer-assisted language learning program in support of Turkish learners' vocabulary learning", Computer Assisted Language Learning, vol. 24(3), pp. 211-232, 2011.

[31] A. Teo, "Promoting EFL students' inferential reading skills through computerised dynamic assessment", Language Learning & Technology, vol. 16(3), pp. 10-20, 2012.

[32] M. Murphy and M. McTear, "Learner modeling for intelligent CALL", Proceedings of the 6th International Conference UM97, pp. 301-312, 1997.

[33] M. Virvou and V. Tsiriga, "Web Passive Voice Tutor: an intelligent computer assisted language learning system over the WWW", Proceedings of the IEEE International Conference on Advanced Learning Technologies, pp. 131-134, 2001.

[34] T. Read, E. Barcena, B. Barros, and F. Verdejo, "I-PETER: Modelling personalised diagnosis and material selection for an on-line English course", Proceedings of the 8th Ibero-American Conference on AI, pp. 734-744, 2002.

[35] T.C. Hsieh, T.I. Wang, C.Y. Su, and M.C. Lee, "A fuzzy logic-based personalized learning system for supporting adaptive English learning", Educational Technology & Society, vo. 15(1), pp. 273-288, 2012.

[36] H.P. Wang, B.C. Kuo, Y.H. Tsai, and C.H. Liao, "A CEFR-based computerized adaptive testing system for Chinese proficiency", The Turkish Online Journal of Educational Technology, vol. 11(4), pp. 1-12, 2012.

Exam questions consistency checking

F. Škopljanac-Mačina*, I. Zakarija ** and B. Blašković *
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** University of Dubrovnik, Dubrovnik, Croatia
frano.skopljanac-macina@fer.hr

Abstract - This paper deals with the problem of checking consistency of automatically generated exam questions in e-learning systems. We focus on the exam questions with the numeric solutions and the automatic evaluation of their answers. Each answer is evaluated as correct if it is within a specified range from the actual correct value. In our paper we check the consistency of exam questions evaluated with this simple method, i.e. are the answers calculated using wrong methods and based on false assumptions always evaluated as incorrect. We made a mutant testing algorithm for identifying inconsistent questions using model checking. At the end we present use cases based on the evaluation of students' answers on our courses Fundamentals of Electrical Engineering and Formal Methods in System Design.

I. INTRODUCTION

Computer aided learning is becoming increasingly popular, and it is already widespread in various areas of formal and informal education. It is used formally from preschools to universities, in continuing education and lifelong learning in corporate environments, as well as informally, e.g. there are many tutorials and self-learning courses available on the Internet. Recently, we have also witnessed the emergence of MOOCs (Massive online open course), a new form of semi-formal education. Most of those computer aided learning systems require objective and fair assessment of acquired knowledge. This is usually achieved using online testing, a cheap and robust way of student evaluation.

But, as e-learning and distance learning systems become more evolved, online testing becomes more complex. Different types of questions are introduced, along with familiar multiple choice and alternative response questions, such as matching type and completion questions or calculation questions with numeric solutions. This increased complexity raises questions about the consistency of the testing systems. Researchers in psychology have studied these problems very extensively with main focus on psychological testing in education, psychiatry, medicine, military and corporate field of human resources. Test reliability is the overall consistency of a test. A test is said to have a high reliability (*precision*) if it produces similar results under consistent conditions [1,2].

In our paper we will focus on a specific kind of questions – calculation questions with numeric solutions. Uploaded solutions are evaluated and accepted as correct if they are in a predefined range from the correct solutions. Using a combination of techniques and methods of mutation testing and model checking we will present an approach for checking consistency of the calculation questions. We will present an analysis of results based on example calculation questions from the course Fundamentals of Electrical Engineering obtained from our In-house developed e-learning system *WebOE* [3]. We will show how to use these results to provide students more meaningful feedback information. And finally, we will give some practical advices to increase consistency of calculation questions.

II. RELATED WORK

As we noted earlier researchers in the field of psychology are responsible for developing the theoretical foundations for describing and solving some of the general problems discussed in this paper. In the field of teaching engineering and computing a number of researchers try to detect various students' misconceptions by analyzing and studying online test and homework results. They try to overcome those misconceptions by introducing new, specific examples and tutorials in advanced e-learning systems or ITSs (Intelligent tutoring systems). Consequently, the e-learning systems can become highly personalized and better adapted to abilities and needs of each individual student [4,5].

Also, software testing and model checking are highly developed fields of computing with many useful concepts and developed methods which can be transferred and used in other areas, such as education.

Methods from system security testing can also be useful in our area. For example penetration testing which tries to detect and avoid security breaches in IT systems can be a good analogy to our problem [6]. Here we try to detect faults in a test system when a false positive answer is awarded points.

III. MUTATION TESTING AND MODEL CHECKING

A. Mutation Testing

Mutation testing is a method for software testing proposed in the 1970s (R. Lipton, DeMillo, Hamlet), and recently it has become very popular in the various areas of software testing [7].

This method is based on the little, simple faults which are deliberately inserted into program code. Those faults are very similar to the errors that the programmers make accidentally.

Technically (Fig. 1), method generates program *mutants* (*P'*), set of copies of the original program *P*, but each with some simple syntactic changes. Also, a set of carefully selected test input data set *T* is created to check original program and its mutants [7].

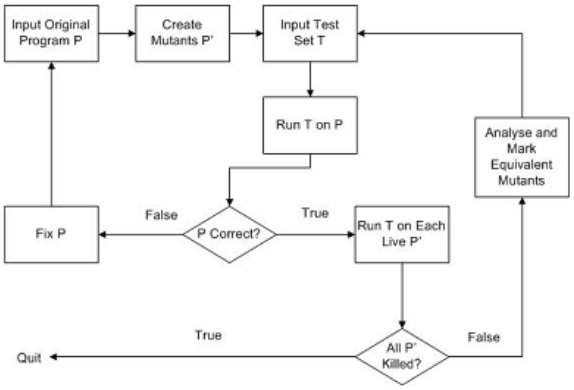


Figure 1. Basic process of mutation testing analysis [6]

If the original program passes input test then all the mutants are executed with the same test input data. If any mutant's result is different from the result of the original program then the mutant is *killed* (removed). All the mutants which give the same results as the original program have *survived*.

In the next iteration of the mutant testing a new test input data is created and original program and surviving mutants are executed with it. If mutants give different results to original program they are also *killed* and the process continues. It is important to know that some of the mutants can never be killed. Those mutants are called *equivalent mutants*, and it is also impossible to find all of them.

The mutation testing method generates also the *Mutation Score* (MS). This is the ratio of the killed mutants over the total number of non-equivalent mutants.

It must be noted that the thorough mutation testing is very computationally expensive for complex systems but it is very useful in determining new test data.

B. *Model checking*

Model checking is a method of automated checking if a model of a system meets given specifications. The most important task of model checking is to test that that the system will not stall or crash [8].

Model checking is widely used in the verification of the distributed software systems, such as network protocols. One of the most important tools in this field is SPIN. It is a tool for verifying the correctness of distributed software. It was developed by G.J. Holzmann and the Bell Labs in the 1980s.

Software models are described in special C-like language, *Promela* (Process Meta Language). *Promela* can model non-determinism in distributed software or automata. Specific temporal properties can be described in *Promela* using *Linear temporal logic* (LTL) which is transformed into Büchi automata [8].

SPIN can also serve as a simulator which gives one of the possible execution paths of the system. Also, it is very useful that it allows mixing *Promela* code with the ANSI-C code which will be extensively used in our questions consistency checking method [9].

Using SPIN's auxiliary *Modex* tool it is also possible to convert ANSI-C program code to *Promela* and to generate program verifier with SPIN [9].

IV. PROPOSED METHOD OF CHECKING CALCULATION QUESTION CONSISTENCY

We will deal with the calculation questions which require one, two or three numeric solutions. The answer to this question is evaluated as correct, and awarded points if and only if all the requested solutions are acceptable. Solutions are acceptable if they are in a predefined range (usually ± 5 %) from the correct solutions.

We based our method on the actual data from the *WebOE* e-learning system on the freshman year course Fundamentals of Electrical Engineering. In this case these questions are homework assignments but the same analysis can be applied to exam or online test questions. Each student gets one of the questions from the set of all questions for a certain homework assignment. Input parameters (usually 4 to 6), and also sometimes the figures (usually electrical circuit) are randomized so that each student gets unique input data (Fig. 2). During a semester there are 10 homework assignments, about 200 different basic questions.

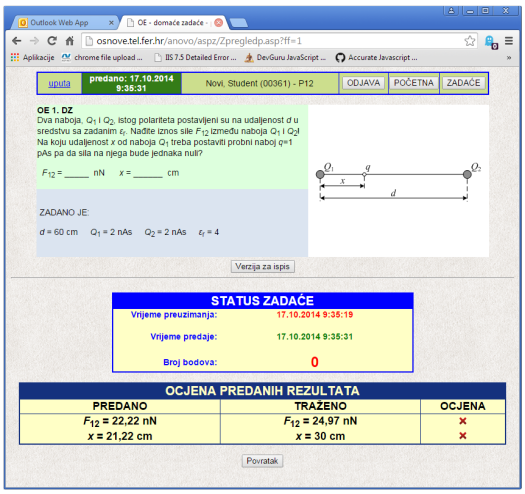


Figure 2. Student interface for homework assignments in *WebOE*

The question is whether a student can solve his/hers assignment in a wrong way (by accident – replacing input parameters or by using flawed algorithms and/or misconceptions), but still get points because the system evaluated their solutions as acceptable (inside 5% from the correct solutions).

Our method of combining concepts from mutation testing and framework of model checking will try to provide answer to that question.

Proposed method is based on the verification of the *mutated solution model*. Each numeric question in the *WebOE* e-learning system has its own unique solution program written in *VBscript*, classic dynamic web page scripting language. Our idea is to translate *VBscript* solution program to ANSI-C code. Then we use *Modex* tool to generate *Promela* process model of the solution program. Afterwards, we mutate the *Promela* process model by copying and slightly changing the solution code. Finally we insert one *assert* command at the end of the mutated *Promela* solution model. In the assertion we state that the all correct and mutated solutions are not overlapping, i.e. all the mutated solutions are outside the range ± 5 % from the correct solutions.

In the last step we use SPIN model checker to generate the verifier of the mutated solution model. The verifier is compiled using standard *gcc* compiler and then executed.

If any of the assertions did not hold then the whole verification process will fail. In that case we get the information where did the error occur and what is the complete path (*trail*) to that error. Also it is possible to find all errors and generate all their corresponding error paths.

It is also possible to automate the process of mutating solution code. Using the concept of expected errors from the mutation testing analysis we have prepared a set of ANSI-C functions which calculate correctly and wrongly (*mutants*) some basic notions from the DC circuit analysis. During the mutation process it is only necessary to copy correct solution and then change one or few function names.

V. RESULTS OVERVIEW AND DISCUSSION

We have run our analysis on a question from the DC circuit analysis as shown in the Fig. 3. The voltage U_{ab} is known as well as all resistor values (R_1, R_2, R_3 and R_4). Students must calculate equivalent resistance R_{ab} and the current I_{ab} through that resistor.

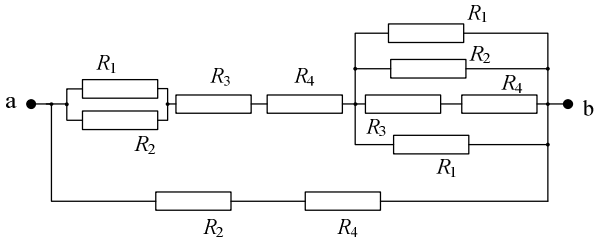


Figure 3. Example – branch of a DC circuit

In the first part of the analysis we have set the values or resistors to: R_1 = 25 Ω, R_2 = 27 Ω, R_3 = 29 Ω and R_4 = 31 Ω, and the voltage U_{ab} is 85 V. Our mutated solution predicts the accidental replacement of the resistor values and the usage of wrong methods for calculating series and parallel resistance, as well as wrong application of the Ohm's law. We have also varied the acceptable range for the correct solutions from ± 1 % to ± 10 %.

Results show us that there were exactly 3888 different solution paths (3887 wrong and of course only 1 correct). We have seen that there are indeed a small number of false positive solution paths – wrong solutions which overlap with the correct solution range and will be therefore awarded points. If the acceptable range is only ± 1 % from the correct solution then there are only 9 such paths (0.23 %). If the acceptable range is the usual ± 5 % from the correct solution then there are 42 false positive paths (1.08 %). We have determined that the number of false positives rises slowly logarithmically with the best-fit function 20.707 ln(x) + 8.2229 (R^2 = 0.9957) where x is the absolute value of acceptable range (1 – 10), as shown in Fig. 4.

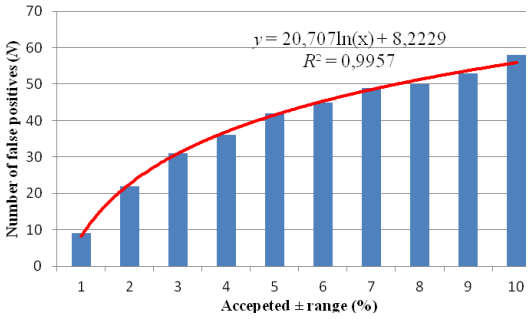


Figure 4. Number of false positives as a function of acceptable solution range

The first part of the analysis was quick (0.01 s and all the false positive solution paths were generated inside 0.4 s) and not very memory demanding (70 MB – most of it for the standard hash table).

In the second part of the analysis we have tried to generate greater number of solution paths. In this case the value of R_1 is chosen non-deterministically between 20 and 29 Ω, also the value of U_{ab} is chosen non-deterministically between 80 and 89 V. Other resistor were fixed at R_2 = 1.08 R_1, R_2 = 1.16 R_1 and R_3 = 1.24 R_1. Also we fixed the acceptable solution range to usual ± 5 %.

Results of the second part of the analysis were in concordance with the results of the first analysis. In total there were 388800 different solution paths (100 correct and 388700 wrong). There were 4170 false positive paths which is 1.07 %. This percentage is almost the same as the corresponding one from the first part of the analysis (1.08 %).

Second part of the analysis was very memory demanding 636 MB, and also time consuming (40.2 s) because of the vast search space (388800 possible paths).

It is evident that the second part of the analysis only has sense when the educator initially defines question's input parameters to better control the percentage of false positive solutions. Unfortunately it is not easy to pick input parameters which will virtually eliminate the possibility of false positives. We can only minimize this possibility by covering most of the known systemic errors. Also we could require four or five solutions, instead of two or three, which would further decrease the possibility that all of them are false positive.

But the first part of the analysis is not very resource demanding and could be a useful addition to any more complex e-learning system. This type of *a posteriori* analysis could be very beneficial feedback information to the students because it could suggest where the error in student's solution occurred. For example, from the numbers the student uploaded the e-learning system could instantly suggest the student that his/hers error was in the wrong application of the Ohm's law.

This method can also be used to generate a list of plausible wrong answers. We could use this information while composing written exams with multiple choice questions, or for devising hybrid type of questions. For example we could offer the correct answer, two or three plausible wrong answers, and one or two random wrong answer together with an input form if a student considers that a correct answer was not offered.

VI. Conclusion and Future Work

In our paper we have presented our method for checking the consistency of calculation exam questions. We have shown that it is hard to achieve total consistency because there will always be some potentially false positive answers. Also we have shown that this method can give valuable instant feedback to students or serve as a method of (semi-)automatically generating plausible wrong answers for multiple choice questions.

References

[1] V. Ruhe, B.D. Zumbo, Evaluation in distance education and e-learning: the unfolding model. New York: Guilford Press, 2009.

[2] http://en.wikipedia.org/wiki/Reliability_%28psychometrics%29

[3] A. Pavić, F. Škopljanac-Mačina, "Computer in testing the Fundamentals of electrical engineering at the University of Zagreb Faculty of electrical engineering and computing," MIPRO 2012 - 35th International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, 2012.

[4] S. Gulwani, "Example-Based Learning in Computer-Aided STEM Education," Commun. ACM, vol. 57, pp. 70–80, August 2014.

[5] T. Sheu, T. Chen et al., "Analysis of Students' Misconception Based on Rough Set Theory," Journal of Intelligent Learning Systems and Applications, vol. 5, pp. 67–83, 2013.

[6] I. Zakarija, V, Batoš, T. Domić, "A Method for the Penetration Testing in IT Environment," Proceedings of The 17th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2013.

[7] Y. Jia, M. Harman, "An Analysis and Survey of the Development of Mutation Testing," IEEE Transactions on Software Engineering, Vol. 37, pp. 649-678, September 2011.

[8] G.J. Holzmann, The Spin Model Checker, Primer and reference manual. Boston: Adisson-Wesley, 2003.

[9] http://spinroot.com/spin/whatispin.html

An Experience in Integrating Learning Management System with user's Google Drive

L. Abazi-Bexheti[*], E. Jajaga[**], M. Apostolova-Trpkovska[*], B. Ismaili[**]
[*]SEE University, Faculty of CST, Macedonia
[**]SEE University, eLearning Center, Macedonia

l.abazi@seeu.edu.mk, e.jajaga@seeu.edu.mk, m.apostolova@seeu.edu.mk, b.ismaili@seeu.edu.mk

Abstract - E-learning is usually accepted as a powerful and important addition to traditional educational programs and Learning Management Systems (LMS) are the key tools that sustain these new educational approaches. Therefore a considerable amount of study and development on both technological and educational issues in e-learning has been going on with outstanding results. In many cases the traditional LMSs are short of flexibility needed for implementing innovative educational models or simply for implementing e-learning strategies of a certain institution. South East Europe University (SEEU) is using in house LMS - Libri for almost a decade. Throughout this period the system has evolved following the trends and the user needs. One of the latest issues that we were dealing was lack of storage for the learning contents. A solution for this problem and several other enhancements were implemented in the new version of the system by integrating the LMS with Google Drive. The aim of this paper is to present the technologies, the implementation challenges and the benefits of this approach.

I. INTRODUCTION

The great interest for original pedagogical-educational practices is a reaction to the community needs for educational change. Such needs come out from the huge request and admission to higher education, the need to raise competitiveness mainly through the raise of the human potential and the necessity to take into account new approaches to learning.

These original innovative practices in education presently are mainly available through e-learning and the most pervasive e-learning tools in higher education currently are the Learning Management Systems. LMS have been present for about two decades in higher education, but yet they are still changing and evolving due to the technological advancements as well as the change in approaches towards teaching and learning in higher education institutions.

The latest reports show that LMSs have their common role, were a few standard features for supporting teaching and learning are yet the most used features [2],[4]. In parallel with their common usage there is a line of new trendy features that are present in current LMSs.

The ability of modern LMSs to integrate multimedia and Web 2.0 tools made a significant impact on instructional design and teaching practice. In particular, Web 2.0 applications may be used to enhance student interaction and collaboration, simulations and serious games are also finding their way into e-learning environment. Finally, a LMS must be adapted by using technologies in order to merge teaching, research, IT, a changing environment as well as financial accountability and managerial models.

Another interesting impact of the Web 2.0 is in finding alternative applications for the traditional LMSs. Actually, some teachers argue that because of evolving Web 2.0 applications, students can be better served by an LMS alternative, a toolbox of web resources that might include social bookmarking tools, document sharing applications, social networking sites, timeline tools, and media options available in the cloud [4].

However as LMS developers we are always looking for the best strategy/way to reach out to the students and engage them. Since Google Drive allows users to store documents in the cloud, share files, version control and edit documents with collaborators, we saw it as a possibility to create meaningful, engaging and accessible environment for all learners including teachers. It also encompasses Google Docs, Sheets and Slides, an office suite that permits collaborative editing of documents, spreadsheets, presentations, drawings, forms as well as version control. In particular, version control is an improvement provided in Google Drive that allows the users to have an overview of all saved versions, download and compare versions, restore any versions anytime without losing previous versions, see who has uploaded different versions and when, etc. On the other hand, Google unveiled Drive for Education with free, unlimited storage, which would at the same time solve our LMS storage problems without any additional cost.

II. USING ONLINE STORAGES FOR E-LEARNING PLATFORMS

When it comes to storing data, cloud storage is quickly becoming the method of choice. Storing files remotely rather than locally boasts an array of advantages for all users. In this part are presented the requirements specification and the system architecture of the new LMS which is build on the concept of using Google Drive as file repository system. In the text below, this new version of the LMS is referred with Libri_v3.

A. *Requirements Specification*

Our LMS requirements specification mainly consisted of the following constraints:

- Use online storages and online calendars
- Rich LMS tools
- Ease of use UI
- Rich statistical/reporting features

The emergence of cloud computing has enabled powerful and low-cost online storages. Accessibility, collaboration, syncing, cost and backup and recovery features are the main advantages when considering online storages as application file repository. Our application will take advantage of using Google Drive as file system repository. Google Drive API [1] offers the tools and client libraries for developing apps that integrate with Drive. Moreover, one of the most famous online calendars is Google Calendar. Google Calendar API [2] offers robust features for managing events in terms of a flexible LMS.

Standard LMS tools include: dropbox, quiz and forum module. The dropbox module enables instructors to arrange assignments delivery within specified time-frame. Since Google Drive is used as file repository system students works will be consumed in Google Drive dedicated folder. The quiz module allows instructors to design course quizzes, mid-term exams or exams. The forum module allows instructors to open class discussions about particular topic. The new version of the LMS could integrate current version's quiz and forum approaches or use third-party solutions.

In order to support ease of use, our tool will leverage the latest technologies for web-based applications: AJAX-driven interface would provide flexible and timely user interaction. Our application's course folders and files interface should be similar to Google Drive UI taking the advantage of user's Google Drive experience. On the other side, the tool should enable course evaluation features and reports e.g. how many times a student has visited a course, most downloaded file, the instructor's time consumed in Libri etc.

B. *System Architecture*

Libri's old version system architecture is based on traditional layered architecture [3]. Because of more concise control of couplings between architecture layers we decided to adopt the onion architecture proposed by Jeffrey Palermo [1] depicted in Fig. 1 and Fig. 2.

[1] https://developers.google.com/drive/

[2] https://developers.google.com/google-apps/calendar/

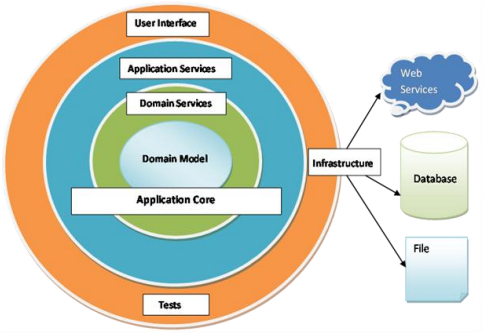


Figure 1. The onion architecture

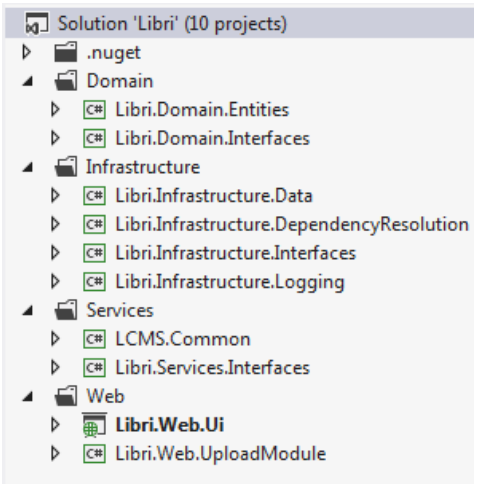


Figure 2. ASP.NET MVC implementation in Libri_v3

The main principle of this architecture is the Inversion of Control (IoC) which inverts the dependencies from lower level layers to higher layers [6].

We adopted the onion architecture implementation presented in [6] and extended it with additional modules as required e.g. logging module. Our *Domain Model* includes classes for membership management, Google interaction classes and LMS entities classes (Libri.Domain.Entities project).

Additionally, in this layer we put the interfaces to repositories (Libri.Domain.Interfaces project). In the Services layer were created two general services: authentication service and LMS service (Libri.Services.Interfaces project) and another service (LCMS.Common project) for integrating other university services. In the Infrastructure layer are placed the repository implementations (Libri.Infrastructure.Data project), dependency resolution module (Libri.Infrastructure.DependencyResolution project) and logging module (Libri.Infrastructure.Logging project). And finally, the UI layer is placed in the outer layers to model application controllers and UI views (Libri.Web.Ui project).

III. Application Design and Implementation

Following the requirements specifications presented above in section II.A, Libri_v3 was developed with the

following technology platform: ASP.NET MVC[3], Entity Framework[4] (EF), SQL Server, Google Drive API[5] and Calendar API, JavaScript, AJAX and jQuery[6]. Model View Controller (MVC) software architectural pattern was chosen because it clearly distinguishes between domain classes, business logic and user interface. Two MVC areas were defined to separate the application's administrator actions versus other user ones. EF code-first approach was adopted for flexible management of database model. Google Drive SDK offers client libraries for different platforms including the .NET[7] client library implemented in Libri_v3. JavaScript, AJAX and jQuery has been leveraged for client-side programming.

The first time a user enters Libri_v3 he is presented with the Google OAuth 2.0[8] pop-up security dialog to permit the user to do Google Drive actions. Google offers different Drive scopes [9] based on the desired app permission level. We have decided to use the drive scope which allows access to all user files. It is the only offered scope which allows course resources management from both Libri app and Drive UI. If the user authenticates our app successfully an OAuth refresh token is received and then registered in our database for performing Drive actions from Libri_v3. In order to support reporting feature requirements it was decided to use application-owned account[10] to store files. A dedicated Drive account is used to deploy all Libri_v3 courses files.

A typical Libri_v3 scenario for sharing course resources consists of the following actions:

- an administrator registers a course, assigns the course instructor and delegates the course assistants and students
- the course appears on course instructor, assistant or student side
- the course instructor (or assistant) uploads a file
- course delegated students preview or download the file

During the first step a folder is created in the Libri_v3's dedicated Drive account and shared with course assistants (edit permission) and course students (view permission). Users (be it instructor, assistant or student) can also get access to course contents by entering course PIN provided by the administrator. Course files and folders are stored in a dedicated folder on Libri dedicated Drive account.

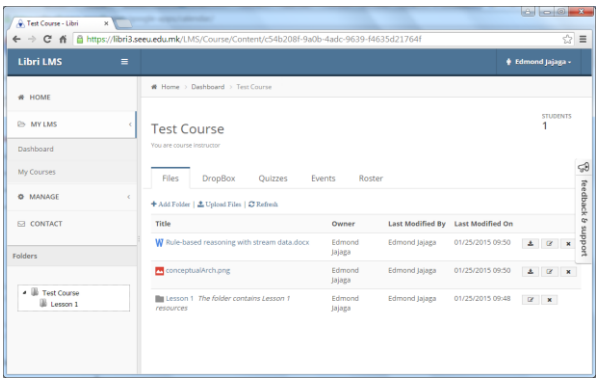


Figure 3. Libri_v3 course view: instructor side

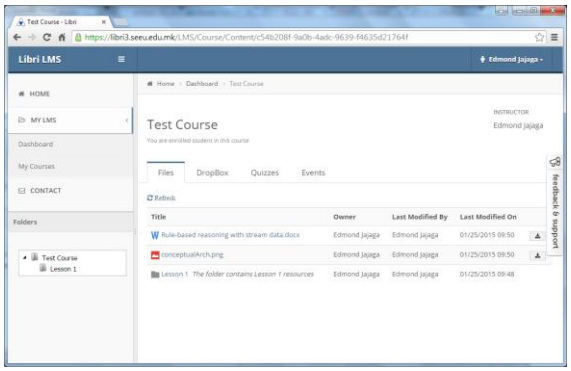


Figure 4. Libri_v3 course view: student side

Thus, each course will have its dedicated folder on Drive which will be eventually shared among users. This allows users to manage course resources through either Libri_v3 or Drive UI. In Fig. 5 is illustrated course contents Drive UI view of the same course presented in Libri_v3 (Fig. 3 and Fig. 4).

IV. FILES MODULE

Fig. 3 represents course instructor Files tab view while the Fig. 4 course students view. Drive API's method retrieveAllFiles[11] is used to get the list of course folder's resources. Before getting the list of resources the request query is modified to include only resources with parent folder ID as the course Drive resource ID:

request.Q = "'" + parentId + "' in parents and trashed=false";

An instructor can create a folder or upload a file in the current folder and can also do CRUD operations over folders/files. On the other side, a student has the option to open a folder or preview/download files which is supported by Google Drive. Enabling file preview is one of the key improvements over Libri previous version. A student would rather preview the file then download it to his/her local machine. Another improvement is the integration of a jQuery-based [12] multiple file upload feature which was absent in the previous version. An instructor or course assistant can upload multiple files

[3] http://www.asp.net/mvc
[4] http://www.asp.net/entity-framework
[5] https://developers.google.com/drive/
[6] http://jquery.com/
[7] https://developers.google.com/api-client-library/dotnet/apis/drive/v2
[8] https://developers.google.com/accounts/docs/OAuth2
[9] https://developers.google.com/drive/web/scopes
[10] https://developers.google.com/drive/web/service-accounts
[11] https://developers.google.com/drive/v2/reference/files/list
[12] https://blueimp.github.io/jQuery-File-Upload/

immediately. Each file upload is done through Drive API's method insertFile[13].

A. *Dropbox module.*

After creating a course in Libri_v3 a Drive folder containing two subfolders are automatically created. Namely, a folder for deploying lessons' files (Files) and a folder containing dropbox files (DropBox). A newly published dropbox will result in new folder in the DropBox folder. It will eventually contain student solved assignments. On the student side, active dropboxes will be listed on the dropbox tab and the student is available to upload his/her works.

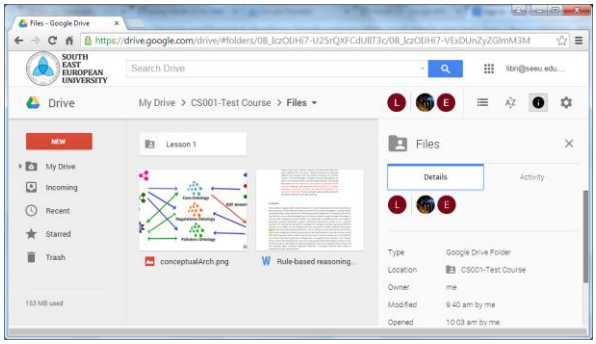


Figure 5. Libri_v3 course view: Google Drive view

Finally, the instructor or course assistant can view the students work on a particular dropbox, grade it and/or add a comment.

Dropbox time to open and close is saved in the database coupled with the dropbox Google resource ID. Moreover, for each student assignment delivery the folder's Google ID is stored in the database for grading purposes.

B. *Google Calendar Events module.*

Another Libri_v3 improvement is using Google Calendars. An instructor or course assistant can create new event e.g. first mid-term exam. By default, Google send reminders about events hence students will be additionally reminded to not miss the exam. The Calendar API method CreateEvent[14] is used to publish new event on Google. Again, as in dropbox module, time information (start and end event date) together with Google event ID is stored in our database. The list of attendees is consisted from the instructor, course assistants and students. For statistical and reporting features purposes a dedicated calendar ID is specified for all Libri_v3 events.

C. *Quiz module.*

The quiz module of Libri previous version was integrated into MVC platform. It offers definition of four kinds of questions: multiple choice, multiple response, true/false and free text answer [5]. The module was improved with support of HTML encoding of question title and the possibility to place images on it.

[13] https://developers.google.com/drive/v2/reference/files/insert

[14] https://developers.google.com/google-apps/calendar/v3/reference/events/insert

V. CONCLUSION AND FUTURE WORKS

One of the main difficulties in Learning Management Systems (LMS) is how to maintain the balance between learning and management. The primary aim of LMSs was to assist educators to manage learning. As a result of this way of thinking, mostly were promoted features such as managing content, roster students, and similar. There are numerous research studies and reports from the early period of LMS use that explore and raise the importance of the management functions of an LMS. As time goes by the accent of research is more in student experiences, assessment and efficiency of the tools. On the other hand, the storage problem is becoming a key concern especially after an LMS is in use for a longer period. Using online storages is a proven way for solving storage problem in general. In this paper we presented the experience of using online storage for an LMS.

The forum module and reporting features are subject to ongoing developments. The system will be further on enhanced based on continuous users' evaluation and feedback.

REFERENCES

[1] J. Palermo,"The Onion Architecture-Part 1", July 2008, http://jeffreypalermo.com/blog/the-onion-architecture-part-1/

[2] L. Abazi Bexheti , M. Apostolova Trpkovska, A. Kadriu, "The evaluation of the quality of online courses. " EDULEARN12, International Conference on Education and New Learning Technologies; Barcelona; Spain.

[3] L. Abazi-Bexheti, LCMS: design, development and implementing experience, Proceedings of the WSEAS International Conference on Applied Computing Conference, Istanbul, Turkey, pp.367-372, 27-30 May 2008.

[4] L. Abazi-Bexheti, M. Apostolova Trpkovska, A. Kadriu, "Learning Management Systems: Trends and Alternatives", MIPRO 2014 - 37th International Convention, CE - Computers in Education, Opatija, Croatia, 2014.

[5] L. Abazi Bexheti, E. Jajaga, M. Apostolova Trpkovska, "Online Testing Module in LMS ITI Conference 2013, 35th International Conference, Cavtat, Croatia, 2013.

[6] Modern Web Stuck "Domain driven Application design using Onion Architecture", August 2012, http://vizagtechie.blogspot.com/2012/08/domain-driven-application-design-using.html

Career and Skills Development for new jobs

Maja Lamza – Maronić, Ph.D.*, Verica Jovanovski, Ph.D.**, Jerko Glavaš, Ph.D.***
* Full Professor, Faculty of Economics in Osijek, Osijek, Croatia
** Primary school Dragutin Tadijanović, Vukovar, Croatia
*** Assistant Professor, Faculty of Economics in Osijek, Osijek, Croatia
maja@efos.hr, jovanovski.verica@gmail.com, jglavas@efos.hr

Abstract - **The purpose and aim of this paper on the perceptions of career development is to confirm or refute the main hypothesis that a career relates to personal goals rather than the goals of an employer or an organization based on the conducted research (using survey, N = 396). A major limiting factor in the research was the fact that respondents had no work experience, which reinforces the conclusion that it is necessary to strengthen the institutional and organizational support in the personal career development during education. Research methods used for the purpose of this paper include survey, statistical method, the Mann - Whitney test, comparison and compilation method, and descriptive method. The research results show the distribution of opinions on the role of an organisation in an individual's career development. Respondents believe that a career relates to personal goals, which confirms the hypothesis. It is therefore necessary to draw pupils' and students' attention to the importance of managing their careers, but also bringing it in line with the organization's goals.**

I. Introduction

We live in a chaotic transition period in a new era defined by global competition, widespread change, rapid flow of information and communication and an increasing complexity of jobs. Global competition has changed the "rules of the game" and the key to competition is no longer in the economy of scale, low prices and cheap labour, but in new technologies and innovations that create added-value products and services. Such products and services have a relatively small share of material components (raw materials, energy, raw labour, etc.), and a large proportion of inbuilt knowledge that adds new value to raw materials and capital base of the product. The code of conduct in the new economy has changed. Access to capital is no longer important, but rather access to information. Creating value refers to the creation of new knowledge and exploiting its value. The most important asset is intellectual property that has no physical characteristics. [3]

II. Career Development

Knowledge has become a product. The key to a successful business is in its ability to increase its market value and to grow continuously. In order to achieve their objectives in the knowledge economy, companies must be able to:

1. Produce - products and/or services applying knowledge adequately within the appropriate structures and processes that will facilitate the cheapest and most efficient way to deliver their products and services to the customer.
2. Respond to changes - The key to survival in the market for all companies is a quick and timely adaptation to market changes occurring in the environment.
3. Predict -Predict possible changes in the market and respond to market trends.
4. Find new ways - to pursue goals and gain profits successfully by acquiring and imparting knowledge to all employees about existing and new products, services and customers by researching and analyzing the market and customer loyalty.
5. Learn - Consider more efficient ways to change in accordance witheconomic, technological and other challenges in the environment.
6. Sustain the business –by creating a new and positive perspective through the ability to develop ways to revitalize and update knowledge and skills of the existing labour force and attract new and innovative employees. [7]

The pace of change in the next ten years will depend on changes in technology, globalization, demographic and other factors that will affect the organizational structure and culture. Modern labour markets need more, not less security; however, the security we are talking about is no longer job security. It includes the security of finding a new job and income security, in other words, the ability to retain roughly similar income after losing a job through unemployment benefits. Job security is developed through integrated policies that increase the employability of workers. [8] The employability of an individual should be increasedthrough three systems: the system of preparation for employment (professional orientation), system of employment agencies and lifelong learning as thefundamental model, and system of active policy measures through subsidized employment, retraining, additional training, and activation of the unemployed. The world of work is subject to constant change. Completely new structures and work areas are created. Entirely new occupations are starting to emerge; traditional ones disappear or change, and the world of

work is changing from day to day. As a result, the choice of occupation and education has become increasingly difficult. To know the world of work means to get to know different occupations, their main features, requirements, conditions and employment prospects. [9] Therefore, our current choice of occupation or education does not necessarily have to be set for life, but rather changes during different life stages. Many authors started to explore the issue of major changes in the nature of careers extensively in the 1990s; however, the issue is far from being fully explored. The predominant opinion is that a traditional career in an organization, which in itself used to be a hierarchical progression (climbing the career ladder) and was geared towards the planned organization's goals, no longer exists. Career development concept, rather than the job concept, provides an organization not only with better employees but individuals with much broader knowledge and skills. [1]

In this way, the competitive ability of the organization is developed. Career development ensures successful achievement of goals of both the organization and the individual. The simplest way to explain the difference between a job and a career is by defining job as something a person does in the workplace in order to get paid, whereas career means a fruitful research activity that brings pleasure. Career is a process that takes place almost the whole life – it does not start with employment or end with retirement. It practically starts in the first grade of primary school, and lasts even after the retirement, for as long as a person is active. Responsibility for career management is divided between an individual and the organization. From the standpoint of an individual, career development is a personal experienceand does not have to be connected to a single organization. However, the provision of help and support from the organization is of crucial importance in this process. [5]

Donald Super emphasizes the importance of the development of self-concept. According to Super, self-concept changes over time and develops as a result of experience. As such, career development is lifelong. Career is interpreted as a process of development and implementation of self-concept. [4] When self-concept becomes real and stable, professional choices and behaviours stabilize. People choose occupations that allow them to express their self-concept. Work represents pleasure and is linked to the stage in which people are able to express and implement self-concept. Super argues that professional preferences and competencies change with time and experience. He developed the concept of vocational maturity, which may or may not correspond to chronological age. The principles of Super's theory are shown in Figure 1. [2]

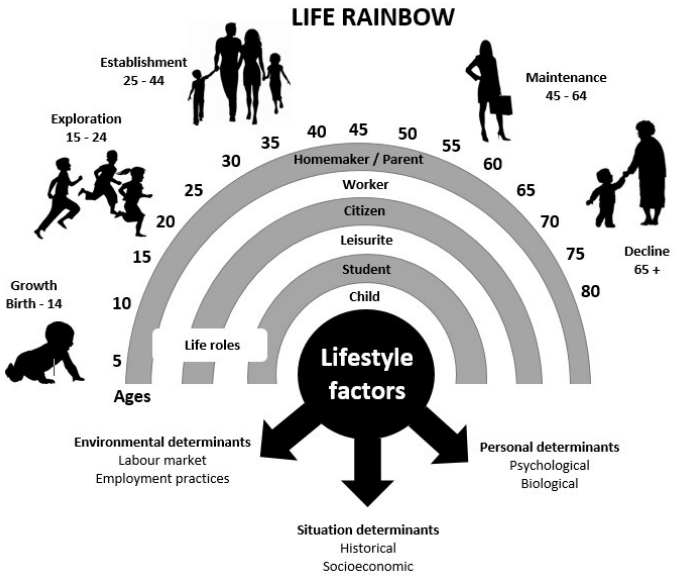


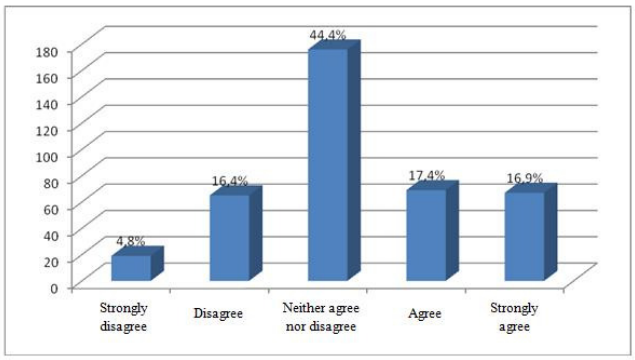
Figure 1 Donald Super's career model

Career development includes career planning, development of potential career paths, training and professional development activities, job promotion, formal human resource policies, job promotion criteria, staff evaluation, counselling and mentoring. Whereas career planning is a process dominated by an individual, the term career management is primarily associated with an organization. Career should be the subject of attention of both an individual and the organization he/she works for in equal measure. With a view to facilitate personal development and achievement of goals by means of career development planning, an organization facilitates the realization of personal goals aimed at achieving organisational goals. In this way, the career has both a subjective and an objective dimension. Subjective dimension refers to the attitudes, interests and motives of individuals, whereas the objective dimension refers to the organization's conditions, the structure of work posts, job promotion criteria, etc. It is essential to develop strategies for personal career development in order to eliminate the identified shortcomings and take advantages in achievingone's goals. [6]

III. SURVEY ABOUT PRECEPTION OF CAREER DEVELOPMENT

According to the survey into perceptions of career development conducted on a sample of 396 respondents (final-year students in secondary schools and universities) from Croatia, less than half of respondents (44. 4%) said they partially agreed with the statement that a career relates to personal goals rather than to employer's or organization's goals. Very few strongly agreed (4.8%) with this statement, while 16.9% of the respondents believe that the career is a personal matter. [2] Graph 1 shows the distribution of respondents' opinions.

Graph 1: Distribution of respondents' opinions on the role of an organisation in the development of an individual's career



Source: Authors

If we look at the difference in opinions between genders, the Mann-Whitney test showed no statistically significant difference between male (n = 209) and female (n = 191) respondents in relation to personal career goals: N = 396, U = 16819.5, Z = -1.64, p = 0.1. Such responses may be the result of lack of experience in young people, which reaffirms the necessity to strengthen institutional and organizational support in career development of individuals during their education.

Development of technology and the integration of IT and digital communications into business will continue for some time and the most sought after professions will be systems analysts, designers and computer programmers, web developers, consultants, IT managers, and hardware engineers. In addition to these, as the technology develops, new professions will appear such as: a digital architect (who designs virtual buildings for advertisers to market their products and services), avatar design-security consultant (who designs, creates and protects the virtual you);in the field of advanced manufacturing: a mechatronical engineer (who combines mechanical engineering and electronics in computer-controlled manufacturing systems), metal skin consultant(who manufactures self-healing composite materials for use on ships, aircrafts and spacecraft), a personal bot mechanic (who will repair and tune-up domestic assistants who work 24/7), powered exoskeleton engineer (whodesigns robots thatprotect people working indangerousenvironments, simplicity consultant (who simplifies and streamlines business processes, technologies and branding in an organization), LocaPreneur(who takes over and improves local products that are losing market in competition with large corporations). As the awareness of environmental protection increases, new occupations will develop such as: traceability manager (who examines global supply chain and checks for suppliers of organic products), hydrogen fuel station manager(whohas both scientific expertise and retail skills), uranium recycler(who converts bomb-grade uranium from warheads into low-enriched uranium for use in nuclear power plants), geomicrobiologist (who pieces together bits of geology, environmental science and microbiology with a view to clean up pollution). Food production will be based on ecological principles and the following occupations will appear: organic farmer (agricultural entrepreneur as skilled in genetics as in marketing), and a personal food shopper(whorecommends daily allowance targets for nutritional balance). The next area most likely to grow is the health sector. An increasing number of jobs in healthcare is directly related to the aging of population - people live longer and the proportion of ageing population is increasingly growing and as a result there is a growing need for medical and social care. In addition to traditional occupations in the health sector such as doctors, medical technicians, physical therapists, etc., new professions will develop such as experimental therapist (who connects patients with new treatments and navigates them through the maze of patient services), home companion-caretaker (who enables elderly people to stay in their homes and live with dignity). The development of nanotechnology and biotechnology will create the following professions: bioinformationist(scientist who marries genetic information with drug development and clinical techniques). Changes will occur in the education sector as well. As education becomes a necessity for any job and the traditional approach to learning is increasingly being replaced by online learning, new education institutions will emerge on the market offering their educational content which will create a new profession: online education broker (tailors a bespoke learning package for the client, dovetailing relevant modules from courses and syllabuses around the world). [10] Finally, there is a mismatch between new jobs that will be created over the next ten years and skills that current and prospective employees have, because in most countries, education systems are rigid and do not follow the global economy trends in sufficient measure. In addition to basic and highly technical knowledge, new skills for new jobs that employees will be required to have are primarily as follows:

Instrumental skills and competencies: capacity for analysis and synthesis, organization and planning; basic general knowledge, oral and written communication in the mother tongue, knowledge of foreign language, elementary computing skills, management skills, problem solving, decision-making;

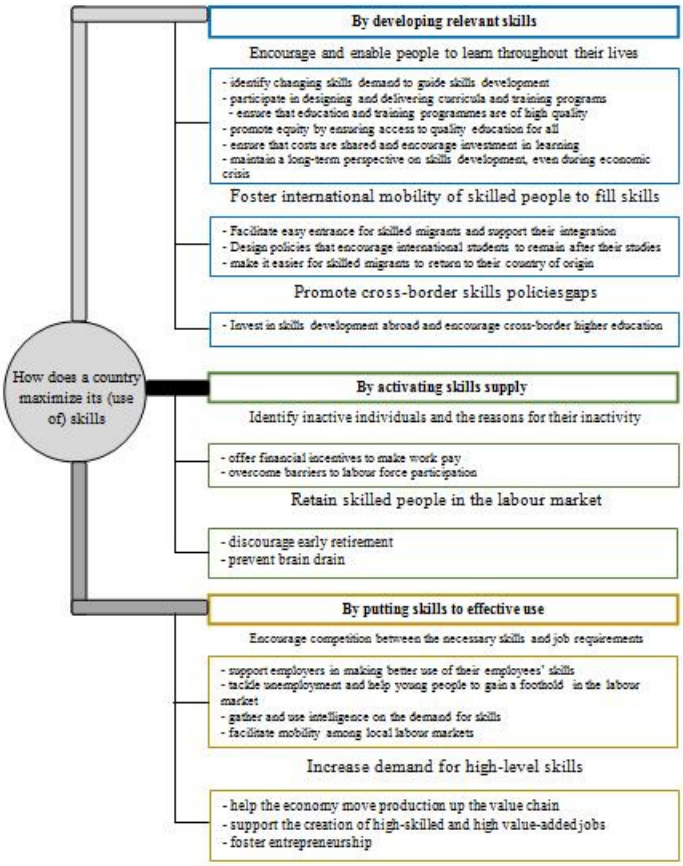
Interpersonal skills and competencies: critical and self-critical abilities, teamwork, ability to work in an interdisciplinary team, ability to communicate effectively with experts in other fields, appreciation for diversity and multiculturality, commitment to work-related ethics;

Systematic skills and competencies: capacity for applying knowledge in practice, research skills,capacity to learn, capacity to adapt to new situations, capacity for generating new ideas, leadership, ability to work on their own; project management, initiative, entrepreneurial spirit, concern for quality and the environment, success-oriented.

Recently, the European Union has launched a number of initiatives aimed at supporting the development of skills of citizens as well as improving education and professional training systems in order for them to be able to respond to the needs of the economy and society better. These initiatives include: the European framework for key competences, the European Qualifications Framework, Strategic framework for European cooperation in education and training and the New Skills for new Jobs initiative. [1]

The latest initiative aims to promote better forecasting of future skills needs of the labour market through better cooperation between the world of work and education. The existing difference between formally acquired skills and competencies and the needs of the labour market will not disappear; on the contrary, it will become even larger. When it comes to the future and strategies for professional development and career advancement, all citizens, regardless whether they are at the beginning or end of their careers, whether they are highly qualified or have only basic skills, will have to upgrade their skills continuously. Acquiring formal qualifications still remains vital for employment; however, informally acquired skills will be recognized as a mechanism that will allow individuals to have their qualifications officially recognised. Upon accession to the European Union as its 28th Member State on 1st July 2013, Croatia will adopt the programmeunder the "Europe 2020 Strategy"[1] which provides guidelines for the development of the European Union by 2020 aimed at overcoming the current economic crisis, the development of a competitive economy based on knowledge with a high level of employment. The strategy also contains recommendations highlighted by working team experts as preconditions for achieving the set priorities (Figure 2). [11]

Figure 2 Preconditions for achieving the priorities inmatching the labour market needs and developing HR skills



Source: OECD Skills Strategy, adapted by authors

One of the aims is the modernization of the education system and investing in the development of competencies and skills, and the use of creative potential of its citizens. What will happen on the European labour market? An individual looking for a job will not have a lot of information about the jobs that are offered on the labour market, and employers who want to fill a vacancy will not have a lot of information about an individual, so both sides will mobilise their resources to find a job, i.e., select the desired candidate. This will create an imbalance between unemployment and vacancies. Europe needs not only skills but the right skills matched to the nature of work and the ability to forecast future changes in the labour market because "one cannot do today's job with yesterday's methods and expect progress tomorrow" (unknown author).

IV. Final Considerations and Suggestions for Improvement

The research has shown that young people increasingly find that their career is closely linked to the employer. The respondents highlighted the attitude of support in career development that is expected from the employer. Career development through the formal education system should be harmonized with the labor market and the requirements of a particular economy, taking into consideration personality traits, knowledge, abilities, skills and aspirations of individuals, as well as the needs of future employers.

It is highly likely that the above-mentioned activities and constant mentoring of the attendees will result in a promising professional future for the business economics graduates who are entering the labor market.

That also implies the importance of the increasing awareness of lifelong career development for every individual that will arrive in the labour market after education, but also the employer's awareness of the importance of employee's career and the realization of quality human resources in his business system.

On the basis of presented research results analysis project activities can be defined even further with an aim to derive direct and indirect benefits from the project.

Direct benefits relate to competitiveness in labour market, additional practical knowledge as well as new approach to knowledge, technologies and information.

Indirect benefits imply decline of unemployment rate, increase of self-employment rate and recognition of the situation in the environment through the interaction with the attendees.

Throughout mentioned activities and constant mentoring of the attendees the more quality promising professional future for the economists that are arriving in the labour market is foreseen.

References

[1] Expert Group on New Skills for New Jobs (2010), p. 4.,Luxembourg: Publications Office of the European Union, 2012.

[2] Jovanovski, V. (2012.): "Razvoj karijere pojedinca u okružju menadžmenta znanja" (Development of individual's career in the framework of knowledge management), Doctoral thesis, Faculty of Economics in Osijek

[3] Lamza-Maronić, M., Glavaš, J., Vukašinović, A. (2011.):„Primary school education system in Croatia – improvements through a strategic management model (case study)", International Journal of Knowledge and Learning (IJKL). Vol. 7; 39-50 , DOI: 10.1504/IJKL.2011.043890.

[4] Super, D. (2004): "A Life-span, life-space approach to career development, Journal of Vocational Behaviour", Vol.16, No. 3, pp.282-298.

[5] Torringtonn, D., Hall, L., Taylor, S.(2004),"Menadžment ljudskih resursa"(Human Resource Management), Prentice Hall Europe, Data status, Beograd.

[6] http://ec.europa.eu/education/lifelong-learning-policy/doc/eqf/lithuania_en.pdf (accessed on 17 March 2013)

[7] http://ec.europa.eu/europe2020/index_en.htm (accessed on 21 March 2013)

[8] http://ec.europa.eu/social/main.jsp?langId=en&catId=958 (accessed on 9 April 2013)

[9] https://community.oecd.org/docs/DOC-1524 (accessed on 12 April 2013)

[10] http://www.guardian.co.uk/money/2010/jan/09/jobs-of-the-future (accessed on 7 February 2013)

[11] http://skills.oecd.org/documents/oecdskillsstrategy.html (accessed on 5 February 2013)

The Implementation of QR Codes in the Educational Process

M. Filipović Tretinjak*
*School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
martinafilipovic.tretinjak@gmail.com

Abstract - Quick Response (QR) codes are two-dimensional (2-D) barcodes that can contain information such as URL links (e.g. a link to YouTube video, website link) and text (e.g. contact details, product details). These square pattern codes consist of black modules on a white background. QR code generator is software that stores data (e.g. URL link, text, Google maps location) into a QR code. This encoded data can be decoded by scanning the QR code symbol with a mobile device that is equipped with a camera and a QR code reader software. QR codes have a number of purposes; they are mostly used in manufacturing (e.g. product traceability, process control, inventory and equipment management), warehousing and logistics (e.g. item tracking), retailing (e.g. sales management), healthcare (e.g. medical records management, patient identification, equipment and device tracking), transportation (e.g. ticketing and boarding passes), office automation (e.g. document management), marketing and advertising (e.g. mobile marketing, electronic tickets, coupons, payments). This paper will describe various methods for the implementation of QR codes in the educational process. Experience from the School of Electrical Engineering in Zagreb shows that QR codes supports both independent and collaborative learning and can create an interactive learning environment.

I. INTRODUCTION

The QR code was developed in 1994 at DENSO Wave Incorporated (then a division of DENSO Corporation). In 2000, the QR code was established as an international standard (ISO Standard 18004) by the International Organization for Standardization (ISO).

Since 1994, QR codes have gained wide acceptance in various industries such as manufacturing (e.g. product traceability, process control, inventory and equipment management), warehousing and logistics (e.g. item tracking), retailing (e.g. sales management), healthcare (e.g. medical records management, patient identification, equipment and device tracking), transportation (e.g. ticketing and boarding passes), office automation (e.g. document management), marketing and advertising (e.g. mobile marketing, electronic tickets, coupons, payments). With the rapid growth of smartphones and tablets, QR codes made their way into the educational process.

A. Definition and main fetaures of QR codes

Quick Response (QR) codes are two-dimensional (2-D) barcodes that can contain all types of data, such as numeric and alphabetic characters, Kanji, Kana, Hiragana, symbols, binary and control codes. Up to 7,089 characters can be encoded in one code. These square pattern codes consist of black modules on a white background.

The main features of QR codes are: high capacity data storage, small printout size, Kanji and Kana character set capability, dirt and damage resistant (QR codes have an error correction capability), readable from any direction in 360 degrees and with a structured appending feature. One QR code can be divided into up to 16 smaller QR symbols. Information stored in multiple QR code symbols can be reconstructed as a single data symbol.

B. Types of QR codes

There are different types of QR codes: QR code Model 1 and Model 2, Micro QR codes, iQR codes, SQRC and LogoQ.

Model 1 is the original QR code. The largest version of Model 1 is 14 (73 x73 modules); this version can store up to 1,167 numerals. Model 2 is an improvement on Model 1. The largest version of Model 2 is 40 (177 x 177 modules); this version can encode up to 7,089 numerals.

Due to its major feature (one position detection pattern, the regular QR code has three detection patterns located at the three corners of a symbol), the configuration of the Micro QR code allows printing in areas even smaller than the QR code. This type of code has four variations, M1 to M4. Data capacity varies for each version. Maximum data capacity for the Micro QR version 4 code is 35 numerals.

Although SQRC is a type of the QR code that has reading restriction function (e.g. can be used to store private information.), this functionality does not guarantee that the coded data is secure. It looks the same as the regular QR code.

LogoQ is a new type of the QR code that can incorporate different design features such as illustrations, letters and logos.

In this research the QR code Model 2 was used.

II. GENERATING AND READING QR CODES

Because the DENSO Wave Incorporated (the inventor of the QR Code and owner of the QR Code trademark) has allowed the patents for the QR code to be freely available to the public there are many websites that feature online QR code generators or downloadable code-generating software. Because such software is not certified by the ISO, the created QR codes may not be readable by all devices or the reading quality may be reduced. In order to avoid such problems, only code-generating software compliant with ISO Standard 18004 should be used.

A. Generating QR codes

QRdraw Ad and QRmaker Ad are DENSO Wave QR codes generation programs. The QRdraw Ad software can be used for generating QR codes on the computer. QRmaker Ad can be used to create a QR code while developing application programs with programming languages such as Microsoft VisualBasic.

QR codes that were used in this research were generated with online QR code generators such as keremerkan.net (http://keremerkan.net/qr-code-and-2d-code-generator/), Snap.vu (http://snap.vu/) and with QR Code Generator YKART, a mobile application downloaded from Google Play (https://play.google.com/store).

After creating an account or downloading an application, the creation of QR codes can start. First choose the content type (Text, URL, Email, Phone number, Contact, Geolocation and SMS), then input the content and press generate button to create the QR code image. Download the QR code and the symbol is ready for printing.

When using Snap.vu, the content document embedded in a QR code can be changed and managed at any time.

B. Reading QR codes

In order to use a QR code scanner software, your mobile device must have a built-in camera. Students download the appropriate QR code scanner on their smartphones and tablets. They mainly used the QR Code Reader Scan Inc. mobile application downloaded from Google Play.

The application is very easy to use. To scan a QR code, start the application and point the camera at the code. There is no need to take a photo or press a button, QR Code Reader will automatically recognize any QR code.

If the QR code contains a website URL, you will automatically be taken to the site. For codes that contain text, the text is immediately seen. For formats such as phone numbers, email addresses, or contact info, you will be prompted to take appropriate action.

III. IMPLEMENTATION OF QR CODES

Various methods of implementing QR codes in the classroom:

A. Adding QR codes to books and printed material

Place printed QR codes in a text book or append them to additional printed lecture material. Students scanning QR codes are taken to the lesson-related materials (e.g. videos on YouTube, SchoolTube, TedEd and KhanAcademy, web pages that contain more information on the topic presented in class, such as Pinterest). To avoid students losing the given QR codes, print them on stickers.

B. Adding QR codes to assignments or worksheets

Add QR codes to assignments or worksheets to allow students to access needed guided instructions. After completing the assignment or worksheet, students can use the QR code to check their answers.

Teachers can visually confirm when students are checking their work. QR codes can also help teachers to keep track of students' individual progress.

C. Adding QR codes to course websites

Add QR codes to your course website. QR codes can contain the teacher's contact information, links to other webpages relevant for the course, information on the course learning outcomes and grading system.

D. Placing QR codes on a classroom wall

To save space on your classroom wall, instead of placing lots of printouts, put QR codes to give students announcements, assignments or instructions.

Different subjects can use different QR code colors. Given that different teachers use the same classrooms, colored QR codes are an easy way to keep the course materials organized.

E. Using QR codes as digital portfolios for students

At the beginning of the school year, assign a QR code to each student. This QR code can contain links to the student's work done over the school year. In this way, parents have an opportunity to monitor the child's progress.

F. Creating a QR code treasure hunt

Create a QR code treasure hunt in which students have to retrieve clues from QR codes placed around the classroom in order to solve an assignment problem.

IV. SEARCH RESULTS

Research was conducted for two months on twenty-three 16 and 17 year old students. Participants were students from the vocational education program: Computer system electronic technician (experimental curriculum). Research was conducted during the course "Introduction to databases", which already has in-class computer-based activities.

Figure 1 shows the research results of the students' feedback on learning with QR codes. Most of the interviewed students (92%) agreed that activities with QR codes made learning interesting and helped them to achieve the intended learning outcomes. Only some of the interviewed students (8%) claimed that they prefer class activities that don't include QR codes.

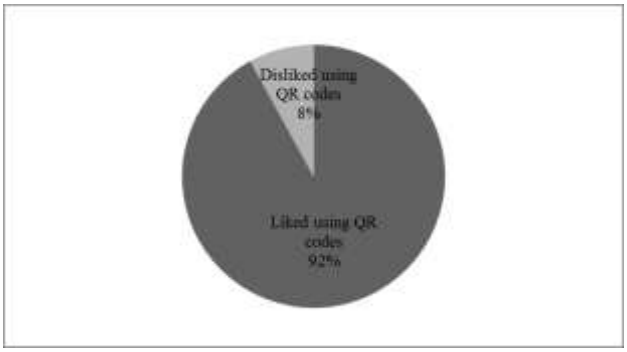


Figure 1. Students' feedback on learning with QR codes

Figure 2 shows the research results of the students' feedback on the use of QR codes. Students (98%) claimed that the use of QR codes was easy once the appropriate software was installed on their mobile device and that they learned new things about phone use. For some students (2%) the use of QR codes was difficult. The reason they gave was that they had to learn other things (for example how to choose and install appropriate software) before they can use QR codes and that reading from mobile device's display was difficult.

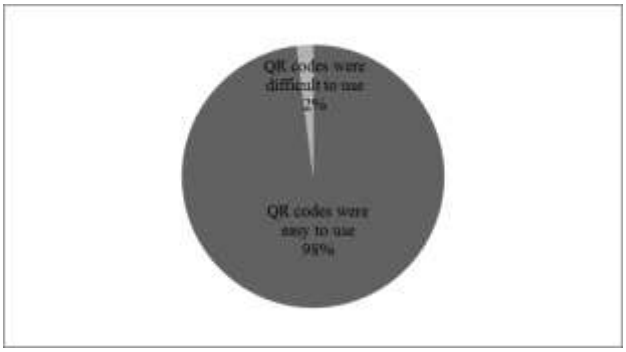


Figure 2. Students' feedback on the use of QR codes

V. CONCLUSION

Teachers are not always willing to accept new technologies. Because of their simplicity and efficiency, QR codes are becoming a popular tool in the educational process. Activities that include QR codes can help create engagement in a lesson, but also need to be implemented in a proper way.

Experience from the School of Electrical Engineering in Zagreb shows that QR codes support both independent and collaborative learning and can create an interactive learning environment.

REFERENCES

[1] http://www.bbcactive.com/BBCActiveIdeasandResources/QRcodesineducation.aspx

[2] http://www.zdnet.com/article/50-qr-code-resources-for-the-classroom/

[3] http://connectlearningtoday.com/exploring-the-educational-potential-of-qr-codes-3/

[4] http://www.scholastic.com/teachers/top-teaching/2012/09/ways-use-qr-codes-education

[5] http://www.educatorstechnology.com/2012/06/qr-codes-are-gaining-momentum-in-todays.html

[6] http://www.edutopia.org/blog/qr-codes-education-mary-beth-hertz

[7] http://www.qrcode.com/en/

[8] http://www.denso-wave.com/en/adcd/product/software/building/qrdraw-ad.html

Computers and Internet in Active Learning

A. Bednjanec *, M. Filipović Tretinjak ** and M. Tretinjak***
* School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
** School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
*** School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
andrea.bednjanec@gmail.com; martinafilipovic.tretinjak@gmail.com; mario.tretinjak@gmail.com

Summary - Active learning has the greatest effect on learners' achievements, pleasure, full understanding and correcting misunderstandings. In the classroom and at home computers and the internet support creativity and provide to teacher many design options of active learning. A teacher can respond to individual errors, errors common to the group or class to help students improve their competences. Students learn actively with greater efficiency in acquiring skills. Active learning improves results and promotes professional development of both, student and teacher.

I. INTRODUCTION

Learner - centered education also known as student - centered learning encompasses methods of teaching that shift the focus of instruction from the teacher to the student. It requires students to be active, responsible participants in their own learning. Student - centered learning emphasizes each student's abilities, interests, learning styles. The teacher is placed as a facilitator of learning for individuals rather than for the class as a whole. In the last decade there has been a growing interest in stimulating the motivation to achieve greater competence of students. Active learning teaching methods has attracted much attention from research teams. The term "active learning" has never been precisely defined in educational literature [1]. Authors [1], [2], [3], [4], [5], [6] and many others describe general characteristics commonly associated with the use of strategies promoting active learning in the classroom. Students are involved in more than listening. Less attention is placed on transmitting information and more on developing students' skills. Students are engaged in activities and greater emphasis is placed on exploration of their own attitudes and values. In this way students are involved in higher-order thinking (analysis, synthesis, evaluation).

This research is built on existing knowledge in the fields of active learning. Its purpose is to develop a framework of effective teaching in vocational education. We explored which active learning teaching methods can help students to stay motivated when they feel difficulties in acquiring the topics to learn, achieve better exam and skill results and make learning and practicing easier. Most mentioned methods require computers with access to internet in the classroom. Our research, focused on learner - centered activities, is made in the School of Electrical Engineering and it is based on the subjects that already have in-class computer based activities such as: computer networking and internet technologies, computer science, system software and technical documentation.

In the first part of our research we checked the depth of the motivation of students to acquire vocational competences. In the second part of our research we compared effectiveness of active-learning methods involving completely individual work of students,work in pairs and work in groups. After that we compared our experinces with guided exercises and e-learning.

II. MOTIVATION

The first study was conducted by surveying one hundred and twenty successful vocational students of the first, second, third and fourth grade in order to find out how they feel about the subject matter and expectations. They recounted all the competencies that are expected, and as a part of research had to choose between three options what motivates them in learning or add their, fourth.

Our findings suggest two parallel motivations that mostly drive student achievement: to improve their knowledge and competencies and to prove that competencies to others. Results are given on Figure 1.

We concluded that our test group is motivated enough to acquire competences and appropriate to compare the effect of active learning.

Authors [1], [2], [3], [4], [5], [6] indicate that different modern teaching techniques can help students achieve better exam results requiring students to 'make their own meaning' and developing their own concept of what they are learning.

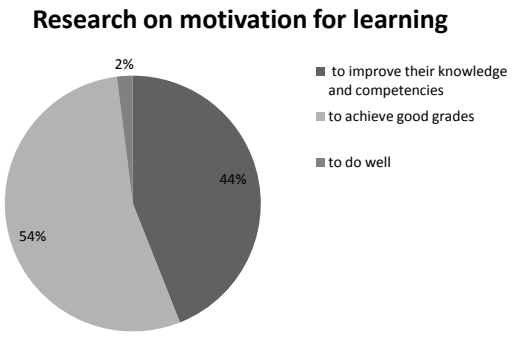


Figure 1. Research on motivation for learning

III. THE CHOICE OF TEACHING METHODS

Sometimes, students expire problems with understanding of new lectures or new topics. Sometimes it is difficult to find appropriate active learning methods that match to teenage vocational students. At the vocational education level, appropriate instructional activities are necessary to facilitate learning. Our experiments were carried out during the school-year to find out which level of cooperation in teaching is the most appropriate to vocational student and teacher. The author´s attention was focused only on the level of completed task, not on their quality details.

A. Active-learning methods involving completely individual work of students

Using active methods of learning involving completely independent work the best results are achieved in learning simple educational content, training of different skills or repetition and systematization of known content. Students made graphic organizers, questions and answers. Only part of the group is able to independently finish exploring brand new or well-known but complicated course content. The main advantages and disadvantages of the active-learning methods involving individual work of students are given in Table 1. Only between 70%-90% of students usually is able to complete tasks, depending on the educational group.

TABLE 1. THE ADVANTAGES AND DISADVANTAGES ACTIVE-LEARNING METHODS INVOLVING COMPLETELY INDIVIDUAL WORK OF STUDENTS

		Advantages		Disadvantages
Individual work of students	**Teacher**	Can detect the level of achievement of each student; Can help each student to those facilities that are needed;	**Teacher**	Needs to motivate students; Needs time and Concentration for individual work with every student;
	Students	Develop critical thinking; Learn and examine material at their own speed; The possibility of individual help from teacher;	**Students**	Lack of direct communication with colleagues; Spend more time on the computer and Internet; Loss of concentration;

B. Active-learning methods involving work in pairs

Using active methods of involving work in pairs, students make graphic organizers, questions and answers, simulation exercises, etc.). Only between 80%-90% of students is usually able to complete tasks, depending on the educational group.

The best results are achieved in repetition and systematization of known content. The main advantages and disadvantages of the active-learning methods involving individual work of students are given in Table 2.

TABLE 2. THE ADVANTAGES AND DISADVANTAGES ACTIVE-LEARNING METHODS INVOLVING WORK IN PAIRS

		Advantages		Disadvantages
Work of studentsin pairs	**Teacher**	Activates less good learners to do tasks; Can help each student to those facilities that are needed; Activates less good learners to do tasks and to contribute;	**Teacher**	Needs to motivate students; Needs time and concentration for individual work with every member of the pairs; Can´t detect the level of contribution or even achievement of each student pecisly; Can help each student to those facilities that are needed only if both of students in pair ask for help;
	Students	Develop critical thinking; Learn and examine material at their own speed; The possibility of individual help from teacher and from colleague; Can´t detect the level of contribution or even achievement of each student pecisly: To activate less good learners to do tasks and to contribute: More interesting especially if pairs are changed;	**Students**	Spend more time on the computer and internet; Loss of concentration Possibility to stay passive and to not to achieve goals;

C. Active-learning methods involving work in small groups

Using active methods of learning involving work in small groups the best results are achieved in practicing skills and repetition and systematization of known content. Only between 70%-80% of students usually is able to complete tasks, depending on the educational group.

Problems are how to grade contribution to the work of each student and how to help each student facilitating the task. The main advantages and disadvantages of the active-learning methods involving individual work of students are given in Table 3.

It has been found that learners motivated to achieve good grades prefer individual work and work in pairs. They choose work in groups only if they have influence on selecting members of the group.

TABLE 3. THE ADVANTAGES AND DISADVANTAGES ACTIVE-LEARNING METHODS INVOLVING WORK IN SMALL GROUPS

		Advantages		Disadvantages
Work of studentsinsmall groups	Teacher	Has not need to activate students;	Teacher	Can´t detect the level of achievement of each student Has to moderate a way of work of some groups Doesn´t know what kind of help each student needs May have problem to finish the exercises in time
	Students	Develop critical thinking; The possibility of individual help from teacher and from colleagues; More interesting especially if groups a changed;	Students	Spend more time on the computer and internet; Should be adapted to work in a group rate in learning and examining materials; Possibility to stay passive and leave the work unfinished;

D. Guided exercises and e-learning

Previous studies indicate that the best results in the mastering of teaching content that can be very complicated to part of students achieved by combining the workout and various forms of e-learning. This gave good results, when the students first encounter with some skills they need to adopt and practice, especially when that skills are also based on the knowledge that is needed to be adopted on the hours of theoretical classes.

Some students often for various reasons do not study the theoretical content on time, so there are problems with the application of knowledge to acquire skills. These students usually do not dare to ask teachers or colleagues for help, and the problems build on each other.

Some students try to always overcome all the theoretical contents conscientiously and on time, but in the application of this knowledge feel difficulties with part of the material. In this way they are thus spared the bad feeling that something could not be successfully designed on the basis of their work. Part of this group of students is subject to de-motivation and withdrawal from work when they run into difficulties that at the moment seem insurmountable.

Students that all the theoretical knowledge acquire in time and successfully apply them in practical skills guided exercise will also be useful. Curricula content will be practiced in controlled conditions and will have the opportunity to correct errors which may come with your own research. This mode brings them security. Such exercises are performed so that the time is adjusted to the slowest student. Each student has the opportunity to ask the teacher for any problem what complicates progress and there is no a single problem undetected. This mode raises solidarity of students. Those students who are quick to work and turn to themselves by helping others slower colleagues achieve a higher level of knowledge and skills and are rewarded grades.

The exercises take place in the time span of two hours. Guided exercise typically lasts for one hour, followed by students in the various e-learning materials that teachers are preparing for the opportunity there, or were referred for independent research on the internet.

E-learning materials are available to them at home, too, as well as numerous quizzes and educational games that are based on knowledge and skills.

Performing such exercises in combination with e-learning ensures that 100% of students complete the exercise, successfully overcome solving problems for the first time, gain self-confidence and gain the will to continue further independent learning. During the entire exercise students are actively taught, first with the help of teacher later independently. Such a mode is in the beginning for the teacher rather hard, because he must be very well prepared for practice in all groups and must prepare enough materials for e-learning but it is certainly worth of student achievements.

.

IV. CONCLUSION

On the part of vocational students overcome the more difficult teaching content is discouraging. By selecting appropriate teaching methods the teacher should act preventively to loss of motivation. Experience from the School of Electrical Engineering in Zagreb indicates that good results can be achieved by a combination of guided exercises in which the teacher helps overcome the teaching contents and different forms of e-learning.

REFERENCES

[1] Jossey-Bass Bonwell, C. C., & Eison, J. A. (1991). Active learning: Creating excitement in the classroom (ASHE–ERIC Higher Education Rep. No. 1). Washington, DC: The George Washington University, School of Education and Human Development.

[2] Davis, B.G. (2009). Tools for teaching (2nd ed). San Francisco: Jossey-Bass Publishers.

[3] Wentzel, K.R.&Wigfield, A. Handbook of Motivation at School, Taylor and Francis -library, 2009.

[4] Felder, R.M. & Brent, R. (1996). Navigating the bumpy road to student centered instruction. Retrieved, September 8, 2011 from http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Papers/Resist.html

[5] http://serc.carleton.edu/introgeo/gallerywalk/active.html

[6] http://www.crlt.umich.edu/tstrategies/tsal

Project teaching – is it achievable without the school library

Denis Vincek
Osnovna škola Ante Kovačića (Primary school)
Vladimira Nazora 1, HR-49250 Zlatar, Croatia
Phone: +385-49-466 832 +385-98-223 746 Fax: +385-49-466 370
E-mail: denis.vincek7@gmail.com

Abstract - Project teaching is one of the forms for achieving the learning outcomes in a student-directed teaching. The study conducted among the students in a Croatian elementary school from grades six to eight provided the data on information literacy and its role in problem solving in project teaching and the extent to which information technology (computers and printers the children have at home) represents an inhibiting factor for solving the information problem. A competent school librarian teaches the students how to achieve information literacy, which is necessary to efficiently participate in every school project – from raising awareness of information need to presenting the project itself – and information technology offered by the school library, which is crucial for materializing every school project as it is present in all six steps of the so-called Big 6 model, which turn the didactic triangle into a square – are a proof that the school library is an inevitable factor in a student-directed teaching, that is, in realization of every project.

I. INTRODUCTION

When discussing student-oriented teaching strategies aimed, the teaching experts of today put a special emphasis on project teaching and project learning [1].

As a part of almost every project, the students have to find and/or print certain information that can be found online. This includes visual or textual content as elements of the posters representing the end-product of the students working on the project. In addition, the teachers assign certain projects to be done at home so information technology can represent an aggravating factor for solving the project task.

The first hypothesis, that is, the first aim of this work is to prove that the school library and the information technology at its disposal represents a crucial prerequisite for the execution of project teaching – not only in terms of achieving the project task by phases but for its presentation, which includes activities up to the creation of the poster.

The second hypothesis that this work will attempt to prove is that the process of spreading information literacy being conducted in school libraries among sixth- and seventh-grade students increases the capability of identifying and locating the source of information, finding the information and evaluating it, which is a prerequisite for a successful execution of every episode of project teaching.

II. THEORETICAL OVERVIEW

A. Why project teaching and project learning

According to Matijević and Radovanović [1], project teaching is a very favourable form of teaching as the acquired knowledge is not just formal, as is the case with didactic materialism, but it is also factual, conceptual, procedural and metacognitive, that is, we can grade it according to Bloom's taxonomy of objectives. In addition, the students also acquire competencies from the affective domain, such as kindness, empathy, tolerance, proneness to team-work, communicability, punctuality and honesty.

Working in groups also breaks the monotony of learning, develops the sense of togetherness and solidarity among them, strengthens social cohesion and emphasises educational moment – the students create something together, they are more motivated, and in a more positive emotional climate in the class (that is, in the library), every student in the group can choose the task that fits him or her better.

Unlike the "old school" in which the teachers demonstrate and the students passively watch, the students in the student-oriented teaching "solve problems on their own, explore, analyze, discuss, evaluate, construct" – in other words, they take the role of the subject.

B. School library in HNOS and NOK[1]

Taking NOK into consideration, one has to ask the question what the teachers can and have to do and where the school librarians fit into in this picture. NOK takes the children as the starting point but its aim is not to teach them but to make them capable of learning and developing skills necessary to master information important for their lives [2].

The library and the librarian are here given a recognizable task and role. They represent a place that

[1] Croatian National Educational Standard and National Curriculum

has to contribute to the creation of an environment favourable for successful education.

One can clearly see from the educational objectives stated in HNOS that the school library and its features represent a place essential for every school project and is ideal for its execution [3].

There are numerous reasons why this is the case [4]: the students are more easily motivated to work in the library; the students can be approached individually through adaptation of the tasks to their aptitudes and capabilities; it allows and reinforces the creativity of students; it has the technology for finding the necessary information, their processing and presentation; it has a system of evaluation different from classic grading and the results are visible through effective presentations or displays; the students experience a certain form of public affirmation (public performances, media).

Of course, the prerequisite for this to be achieved – organizationally speaking, through support for the sources of knowledge and through professional research guidance and realizing set objectives – is a competent librarian.

C. Previous research

If we look at the first out of three competences for teachers and librarians [5] stated by Domović (working with information, technology and knowledge), one can clearly see that the school librarians have a lot to offer in term of information through the scope of the Library Education programme by making the students information literate [6].

This is particularly important as almost all objectives in project teaching are related to problem-solving, and so project teaching can often be identified with problem teaching. Here, the library represents the central hub [7] of information of the school, not only in terms of reference collections, that is, the classical sources of knowledge, but in terms of computers with an Internet connection, that is, the sources we reach by means of modern technology.

The librarians are there to teach the students how to find, evaluate, use and present the information by using the so-called Big6 model, for instance [8]. But the importance of information technology located in the school library is also highlighted by Domović.

The role of the library in the educational process was investigated by Lovrinčević in her analysis of 10-year activity of the high school library of the High School for Construction and Engineering in Osijek [9]. She emphasizes that the school library is an absolute necessity in the process of finding the way for students to become the "subjects" and making education not something happening around them but something they take part in, which are the exact traits highlighted by the student-oriented teaching and project teaching.

The results obtained by Lovrinčević are particularly prominent in indirect use of the school library, that is, in project teaching, or as the author puts it, in "classes with problem-related approach" in which the students work individually, in pairs, or in groups and use the library for research and finding the answer in form of a photograph, text etc.

In this vein, Shaw stated that, among other things, the programme objectives of the school library include [10] acquiring a whole array of educational technologies to enable searching, acquiring materials, creating original works, organizing data, solving problems, helping and organizing interdisciplinary learning and helping students in the process of learning and making their own works. Among those aims, we can see the elements of project learning.

Diana Oberg, whose career took a path from being a school librarian to being a university teacher, emphasizes [11] the importance of acquiring information literacy as a means for initiating the creation of the school project. While information literacy in Croatia is introduced in higher elementary school grades, Oberg emphasizes that it is necessary to start this process as soon as possible and that it is necessary to clearly state that the person in charge of teaching information literacy is the school librarian.

Lovrinčević, an expert in information sciences, also emphasizes that project teaching is actually guided exploratory learning, characterized by carefully planned and well-monitored work of school librarians and teachers and their students who explore the acquired knowledge according to Bloom's taxonomy.

This is actually the learning style of Carol Kuhlthau, American professor of Information and Communication Sciences [12], who states that project learning is the form of learning in which the librarian guides the students through the resources available at the school library all the while teaching them about information literacy. Ruth Small emphasizes [13] that once the students master information literacy via school library, they will not have only acquired the skills to solve the problem but will have become intellectually curious persons in the sense that exploratory behaviour will be something that they are constantly attracted to, which is a quality necessary for successful project learning.

It is also worth mentioning that in her PhD thesis from 2006, Lovrinčević, a professor at the University of Osijek, mentioned [12] new programmes, new ways of teaching, new teaching resources and aids and "trained teaching staff", but it is only five years later that Matijević and Radovanović revisit this topic and start systematically using the term student-oriented teaching, which includes project teaching among other things [1]. While the latter talk about project teaching without even mentioning the school library, Lovrinčević offered the solution in 2006 in which the school librarians are supposed to be the leaders in this new concept and the school libraries the places which carry the crucial changes in school.

Keith Curry Lance from the Colorado State Library also wrote about the influence of the libraries in terms of information literacy and technology that a library has at its disposal [14], emphasizing the cooperation between the librarian and all elements of the educational process.

In his example of class correlation between American History and American Literature in the Eaglecrest High School in Aurora, Colorado, Lance concludes: "Students are taught the research process, including accessing and using both primary and secondary sources. My students' ability to access library books and other materials and to use information appropriately in their papers is very evident in the quality of their work" [14]. In terms of technology, the studies have shown better results and student accomplishments in those schools where various databases are available, and this becomes particularly prominent when those databases can be accessed online from home.

Curry uses the example of Becky Hickox from the Silverton High School to describe a successful project on the topic of health, which started with the librarian teaching the students how to use search engines and then databases.

When it comes to the role of school libraries in increasing student accomplishments in project teaching, Michele Lonsdale from the Australian Council for Educational Research emphasizes [15] the proliferation of information, mostly online, and uses this to prove the role of school libraries and librarians in the students' accomplishments. According to her, the library is not just a source of information but a centre in which the librarian teaches the students to locate, use and evaluate the information, which gives the library a special role in the acquisition of information literacy as it creates a community of those ready for lifelong learning.

Lonsdale describes the library as a place in which, according to her, "content-based education" is transformed into "outcomes-based education". In other words, the main focus is no longer on the teaching content that the students have to acquire, and the students become the subject in the didactic triangle, as discussed by Poljak [16], in which they show what they have learned, which is precisely the student-based teaching. And this is all achievable in the library, says Lonsdale, as the librarians are more focussed on their users and their needs, and less on the library itself.

These results were confirmed by a study [17] of the effect of school libraries conducted in 22 American states and one Canadian province from 1993 to 2011. Kachel states that school libraries undoubtedly have a positive effect on students' accomplishments as they teach their students the skills of the 21st century and promote successfulness.

III. METHODOLOGY

For the purpose of this research, the "distribute and collect" questionnaire [18] was used. It consisted of 14 questions and was filled out by the students of grades sixth to eighth from the Ante Kovačić Elementary school in Zlatar in June 2014.

The research included only the mentioned grades as orientation research in the form of teacher interviews showed that project teaching is applied only from sixth grade onward, while it is seldom employed in lower grades.

The questionnaire was divided into two parts. The questions in one part were used to gain insight into information literacy of the students and its role in problem-solving in project teaching, while the questions from the second part were used to find out to which extent information technology (computers and printers the students have at home) represents an aggravating factor in solving the information problems.

IV. RESULTS AND DISCUSSION

A convenience sample was selected for the purpose of this research. The questionnaire was filled out by 151 students: 59 sixth-graders, 50 seventh-graders, and 42 eighth-graders (everyone present at school at the time the research was being conducted).

The research showed that all students have a computer with an Internet connection at home, but not everyone has a printer (Fig. 1).

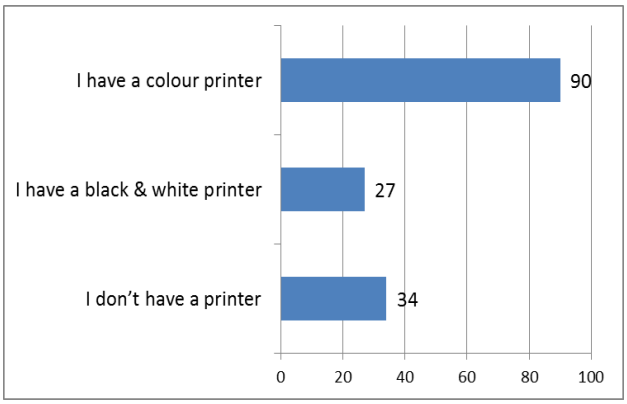


Figure 1. Distribution of students according to the fact whether they have a printer at home

As much as 93 per cent of participants stated that they had been asked to find certain data (text and/or photographs) in the scope of a project in their education up to that point. Two thirds of participants had done that at home (Fig. 2).

However, the question Where did you print the results you got? yielded some unexpected answers. Only one half of participants did this at home (Fig. 3).

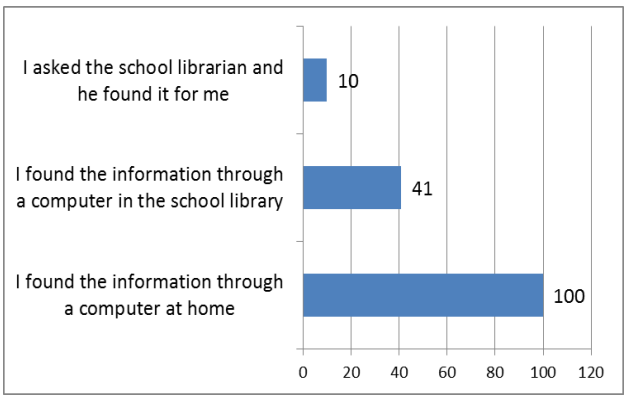


Figure 2. How the students found information for school project

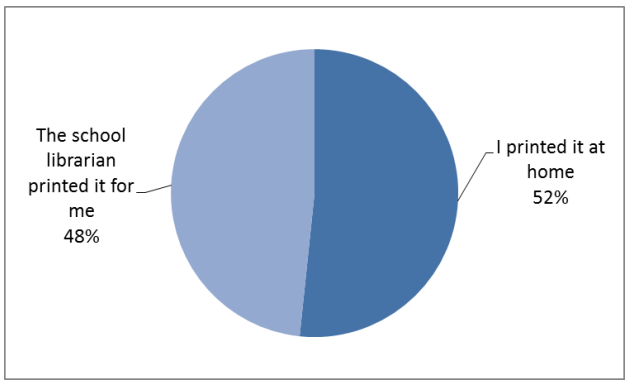


Figure 3. Where the students printed the information found for the school project they took part in

Even though all students have a computer with an Internet connection at home and could find the information for their school project from home, as much as 51 students, that is, one third of participants did so at school, in the school library to be precise. If we compare that to the question whether they have a printer at home, one can clearly see that all students that do not have a printer of any kind at home search for information (and print them) at school (individually or with the help of the school librarian) along with more than half of students that said they had a black and white printer at home.

However, out of 100 students that stated they find the information from home – and 90 of them have a colour printer at home – only 78 (out of 151) printed the information (texts and/or photographs) at home, while the remaining students asked the school librarian to print the necessary information at the library, along with those students that found the information they looked for in the school library. The following can be deduced from this: the students who use the school library to search for and print the information they need for presenting their projects are the students that do not have a printer at home or just have a black and white printer at home, while some of the students that have a colour printer at home (12) do not do the printing at home, possibly for the purpose of saving toner and/or ink.

Considering the fact that a majority of students have black and white printers, they were asked whether their teachers require the presentation materials to be printed in colour or not (Fig. 4).

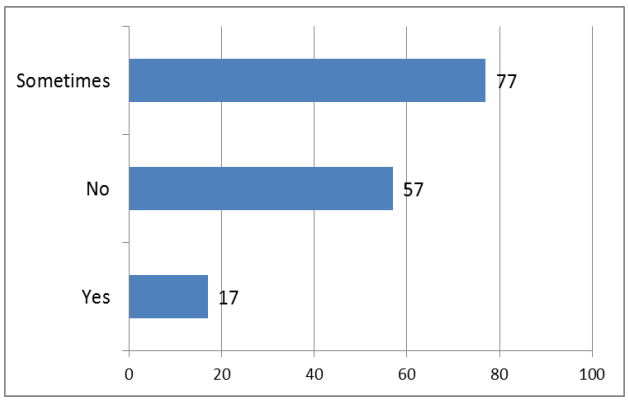


Figure 4. Do the teachers ask that the content used for project posters be printed in colour

TABLE 1. PERCENTAGE OF STUDENTS THAT ALWAYS FIND THE INFORMATION THEY ARE SEARCHING FOR DISTRIBUTED ACCORDING TO GRADES

Grade / Percentage of students that always find the information	**6th grade**	**7th grade**	**8th grade**
	44%	50%	60%

While 38 per cent participants gave a clear answer that the printed content does not necessarily have to be in colour, as much as 51 per cent said that the teachers require this from time to time. Such a high figure supports the view that the students hope their work will be given a higher grade if it is printed in colour, which is why they perform their entire task from finding information to printing at the school library.

When it comes for the skills of finding the information, 75 students said that in most cases they find the text and/or photograph they are looking for, while 76 said they always find it. However, if we distribute them according to grades, we can note that the percentage of students that always find the information they search for increases from sixth to eighth grade (Table 1).

The participants had to state whether they read the information they get before printing. As much as 92 per cent of them said that they do, but it is still unthinkable that there are 12 students who do not do that. The author of this paper notices this as soon as the students bring a Word document with parts underlined with blue (hyperlinks to other websites that were not removed) or with footnote markings left next to words.

However, the number of students that read and correct those mistakes before printing increases from the sixth to eighth grade (Table 2).

TABLE 2. NUMBER OF STUDENTS THAT DO NOT READ THE INFORMATION THEY FIND BEFORE PRINTING DISTRIBUTED ACCORDING TO GRADES

Grade / Number of students that do not read the text before printing	**6th grade**	**7th grade**	**8th grade**
	6	5	1

TABLE 3. NUMBER OF STUDENTS AND THE INTERNET SEARCH ENGINE USED TO FIND THE SOLUTION TO THEIR INFORMATION PROBLEM

Grade / Search engine	**6th grade**	**7th grade**	**8th grade**
Google	29	30	33
Google Chrome	19	6	5
Wikipedia	6	6	2
Mozilla Firefox	3	1	2
Internet Explorer	1	3	0
The student did not provide an answer	1	4	0

The only open-choice question the students were given in the questionnaire was naming the Internet search engine they used for finding the information they needed. Table 3 shows the answers the students provided.

The table clearly show that the only search engine the students use is Google, but among the answers provided, they also stated three Internet browsers and even Wikipedia, the free encyclopaedia. However, it is worth noting that the number of students that provided a correct answer to this question, i.e. who named the internet search engine, increases from sixth to eighth grade (Table 4).

TABLE 4. PERCENTAGE OF STUDENTS WHO STATED THE (CORRECT) NAME OF THE SEARCH ENGINE USED TO FIND THE INFORMATION DISTRIBUTED ACCORDING TO GRADES

Grade / Percentage of students that gave a correct answer to this question (by naming the Internet search engine)	**6th grade**	**7th grade**	**8th grade**
	49%	60%	79%

As far as the sources of information go, Wikipedia still holds the top spot (Fig. 5).

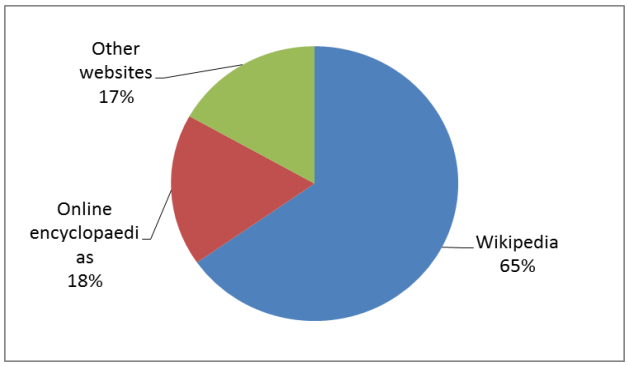


Figure 5. The online sources the students from sixth to eighth grade used to find information for their projects

However, looking from the sixth to eighth grade, the number of students who prefer Wikipedia as a source of information is in a slight decline (Table 5).

TABLE 5. PERCENTAGE OF STUDENTS THAT LOOK FOR THE INFORMATION MOSTLY ON WIKIPEDIA DISTRIBUTED ACCORDING TO GRADES

Grade / Percentage of students that look for the information on Wikipedia	**6th grade**	**7th grade**	**8th grade**
	76%	64%	59%

The percentage of students using the so-called "copy-paste" method for executing their projects is admittedly not so high as for using Wikipedia to gather information. As much as 67 students (44 per cent) still use this procedure, although a slight decline can also be noted for this from the sixth to eighth grade (Table 6).

TABLE 6. PERCENTAGE OF STUDENTS USING THE COPY/PASTE METHOD DISTRIBUTED ACCORDING TO GRADES

Grade / Percentage of students using the copy/paste method	**6th grade**	**7th grade**	**8th grade**
	46%	46%	40%

V. CONCLUSION

This research has provided the following insights for students from grades six to eight:

1. Not all students have a printer at home that they need for printing photographs or textual content necessary for creating the posters for finalizing their project tasks.

2. The students that do not have a printer of any kind at home and a portion of students that have a printer at home (black and white or colour printer) use the computers at the school library to find the information necessary for project learning. This can be affected by three factors:

a) the teachers ask that the results be printed in colour or the inclination to conserve the (black) ink cartridges from their personal printers makes the students perform their entire project tasks at the school library

b) the inclination to conserve the (colour) ink cartridges from their personal printers makes the students perform their entire project tasks at the school library

c) the students need the help of the school librarian for solving their project task

This proves the first hypothesis that the execution of the project task is impossible without information technology in the library, both in the stages of project learning and in the finalization stages, which implies the creation of the posters. In other words, the possibility of conducting project teaching without the cooperation with the school library is dubious to say the least.

As far as the second hypothesis is concerned, the following has been established:

3. The number of students who always find the information they need for solving the project (information) problem increases from the sixth to eighth grade

4. The percentage of students that do not read the information they find before printing them for the project poster decreases from the sixth to eighth grade

5. The number of students who use Wikipedia as their primary source of information decreases from the sixth to eighth grade

6. The number of students who use the copy/paste method in creation of their project poster decreases from the sixth to eighth grade

This confirms the second hypothesis. During the observed three-year period, the students' capability to identify the sources of information and choosing the best one has increased (the share of Wikipedia as the top source has decreased) as has the percentage of locating

and finding information. Reducing the employment of copy/paste method and reading the information before publishing them is nothing else but information evaluation, all of which is a proof of acquiring the information literacy skills. It is a result of intense process of acquiring information literacy conducted in the scope of the school librarian's curriculum, especially through the programme of Library Education, which is given more classes in the school library from the sixth to eighth grade, but also through correlation of other school subjects with the school library. Just like the first hypothesis, this proves that there can be no (satisfactory) project teaching without the school library.

Surely, the concept of the National Curriculum will bring many novelties in the scope of all participants in the educational processes, that is, it will make sure that this role, which has largely been outlined before, finally comes to fruition.

In this paper, this can be emphasized on the example of planning and realizing topics which involve several classes as a means of bringing together educational areas and classes into a cohesive unit. Even though the correlation between classes should have been achieved by now as it is, the author's experience shows this has so far been avoided in both the second and even more prominently in the third educational cycle of acquiring basic student competences. The curriculum of the school librarian has so far included correlation and project teaching is ideal for it, but the librarians seldom "entered the classroom" as the subject teachers (unlike class teachers) are unwilling to "give up their classes", as they tend to say.

The school library represents a crucial bond for achieving fundamental competences of lifelong learning and cross-curricular content as a part of the National Curriculum. The conditions for that are: a competent school librarian, cooperation, information technology, and various library resources [19].

All of this was shown by the study presented here. The cooperation with the students; a competent school librarian who provides the students of the upper elementary school grades with information literacy skills, necessary for them to efficiently participate in every school project and realize its problem tasks – from realizing their information needs to presenting the project itself – and the information technology offered by the school library, without which the school project is impossible as it is present in each of the six steps of the Big 6 model, thus turning the didactic triangle into a didactic square. All of this proves that the school library is an integral factor in student-oriented teaching, that is, in realization of every school project.

REFERENCES

[1] M. Matijević, D. Radovanović, Nastava usmjerena na učenika : prinosi razvoju metodike nastavnih predmeta u srednjim školama. Zagreb: Školske novine, 2011.

[2] "Nacionalni okvirni kurikulum : za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje," Internet: public.mzos.hr/fgs.axd?id=18247, Oct. 2011, [Aug. 8, 2014].

[3] J. Lovrinčević, D. Kovačević, J. Lasić-Lazić, M. Banek-Zorica, Znanjem do znanja : prilog metodici rada školskog knjižničara. Zagreb: Zavod za informacijske studije Odsjeka za informacijske znanosti Filozofskog fakulteta sveučilišta u Zagrebu, 2005.

[4] D. Kovačević, J. Lasić-Lazić, J. Lovrinčević, Školska knjižnica – korak dalje. Zagreb: Filozofski fakultet, Zavod za informacijske studije Odsjeka za informacijske znanosti : Altagama, 2004.

[5] V. Domović, "Obrazovanje nastavnika u Hrvatskoj i drugim europskim zemljama," Internet: http://www.google.hr/url?sa=t&source=web&cd=4&sqi=2&ved=0CDMQFjAD&url=http%3A%2F%2Fsiteresour-ces.worldbank.org%2FINTCROATIA%2FResources%2F3011441178203929271%2F2VlatkaDomovicObrazovanjenastavnika.ppt&ei=Hk4lTpnrCYiEwbmt_XLCw&usg=AFQjCNGPcika7u99ZcAL_5VYWSFLWG0GDQ&sig2=g5FZ-fPilVRaKzyiI4lvbg, [Aug. 14, 2014].

[6] A. Farid, Towards an effective school library : based on the writings of Mary Virgina Gaver. New Jersey: New Yersey School Media Association, 1971.

[7] C. Klingel, R. B. Noyed, School media specialists. Vero Beach, Florida : The Rourke Press, 2001.

[8] I. Stričević, "Informacijska pismenost i uloga dječjih knjižničara," presented at "Stručni skup djeca na internetu u knjižnici", Zagreb, 2011.

[9] J. Lovrinčević, Školska knjižnica kao jedan od unapređivanja odgojno-obrazovnog sustava : magistarski rad. Zagreb, J. Lovrinečvić, 1994.

[10] M. K. Shaw, "Block Scheduling and its Impact on the School Library Media Center." Internet: https://books.google.hr/books?id=t0OI10yDD0wC&pg=PR4&lpg=PR4&dq=marie+keen+shaw+block+scheduling+and+its+impact&source=bl&ots=S5RCr7J7he&sig=rlCQmDa_uqxA_efA5yaHhfFmxUY&hl=hr&sa=X&ei=mA-9VOGWK4T0OuzmgcAN&ved=0CCUQ6AEwAQ#v=onepage&q=marie%20keen%20shaw%20block%20scheduling%20and%20its%20impact&f=false, 1999, [Jan. 19. 2015].

[11] D. Oberg, "Principal Support: What Does It Mean to Teacher-Librarians?." Internet: https://www.ualberta.ca/~doberg/prcsup.htm, 1995, [Jan. 19, 2015].

[12] J. Lovrinčević, Školska knjižnica u diskursu informacijskog obrazovanja : doktorska disertacija. Zagreb: J. Lovrinčević, 2006.

[13] R. V. Small, "An Exploration of Motivational Strategies Used by Library Media Specialists During Library and Information Skills Instruction." Internet http://www.ala.org/aasl/sites/ala.org.aasl/files/content/aaslpubsandjournals/slr/vol2/SLMR_MotivationalStrategies_V2.pdf, Jan. 1999, [Jan. 19, 2015].

[14] K. C. Lance, "The Importance of School Libraries," Internet: http://www.laurabushfoundation.com/Lance.pdf, [Jan. 23, 2015.].

[15] M. Lonsdale, Impact of School Libraries on Student Achievement: A Review of the Research. Victoria : Australian Council for Educational Research, 2003.

[16] V. Poljak, Didaktika. Zagreb : Školska knjiga, 1985.

[17] D. Kachel, "School Library Research Summarized: A Graduate Class Class Project," Internet: http://sl-it.mansfield.edu/upload/MU-LibAdvoBklt2013.pdf, 2013, [Jan. 23, 2015.].

[18] A. Tkalac Verčić, D. Sinčić Ćorić, N. Pološki Vokić, Priručnik za metodologiju istraživačkog rada u društvenim istraživanjima : kako osmisliti, provesti i opisati znanstveno i stručno istraživanje. Zagreb: M. E. P., 2011.

[19] M. Zovko, "Školski knjižničar u suvremenom kurikulumu," in 22. Proljetna škola školskih knjižničara Republike Hrvatske : Zagreb, 7. - 10. travnja 2010. : zbornik radova. Zagreb: Agencija za odgoj i obrazovanje, 2010.

Benefits of using interactive whiteboards at Kajzerica elementary school

Igor Rukljač, Martina Draženović
Kajzerica elementary school, Zagreb, Croatia
Kajzerica elementary school, Zagreb, Croatia
irukljac@gmail.com, martina.drazenovic@gmail.com

Abstract - Interactive whiteboards are powerful teaching tools. In elementary school Kajzerica we are proud to benefit from interactive whiteboards and endless possibilities to create and manage our classes. Potential applications of interactive whiteboard are: using web-based resources in whole-class teaching, showing video clips to help explain concepts, manipulating text and practising handwriting, saving notes written on the board for future use, quick and seamless revision. After implementing interactive whiteboards into our classrooms, students feel more involved and therefore their motivation for the class is higher. Since the beginning of the school year, we have advanced our classes, recreating them and exploring numerous possibilities to benefit our students and encouraging them to do more. This research will show how students and teachers benefit from interactive whiteboards in our school and how we recreated particular classes using an interactive whiteboard.

I. INTRODUCTION

Introduction of interactive whiteboards into classes has enriched students' and teachers' lives. Both teachers and students do more research, constructing and interacting. Using interactive whiteboard helps teachers to explain complex and abstract concepts more efficiently and it makes them more flexible for building various class materials. Many of our teachers saw interactive whiteboards as a teaching method or technique rather than an aid in teaching or a part of an equipment. The ability to dynamically manipulate and annotate objects and the enhanced presentation of a certain matter is what differentiates classical teaching from teaching with an interactive whiteboard. We embraced digital culture and the fact that is very important in our students' lives so it is to be expected that we adjust our classes to the digital age. Our students do not expect that we continue to use chalk and green board because it becomes too abstract for them. They live in a digital age where their attention is not focused on only one thing, but it is dispersed into several things. We had to adapt.

II. TECHNICAL CHARACTERISTICS OF INTERACTIVE WHITEBOARD

There are two very distinct kinds of interactive whiteboards: The first is a "virtual" electronic version of a dry-wipe board on a computer that facilitates learners in a virtual classroom to view what an instructor, presenter, teacher or fellow learner writes or draws. It is also called an electronic whiteboard and can be found in conferencing and data-allotment systems. The second type is a huge physical display panel that can operate as an ordinary whiteboard, a projector screen, an electronic copy board or as a computer projector screen on which the computer image can be controlled by touching or writing on the surface of the panel instead of using a mouse or keyboard [1].

We use the second kind of interactive whiteboard in our school. Every classroom in our school has one laptop, projector, interactive whiteboard and Internet connection so technical prerequisites are satisfied.

III. INTERACTIVE TEACHING

Interactive teaching methods are an effective way to connect with students used to constant stimulation.

Teachers use higher order questioning and students' active contributions are valued as they test their developing understanding against collective meaning [2]. Interactive teaching encourages and expects learners to participate, use questions to stimulate discussion, emphasizing the value of answers, giving participants hands-on experience and using teaching aids to gain and retain attention.Students are revived from their passivity of merely listening to a lecture and instead become attentive and engaged [3]. These techniques are often perceived as "fun", yet they are frequently more effective than lectures at enabling student learning. Telling is not teaching, nor is listening learning. We must engage participants in learning activities that lead to a higher level of understanding and result in the participants ability to apply what he learned. Interactive teaching is a two-way process of active participant engagement with each other, the facilitator, and the content [4].

IV. METHODOLOGY

A. Method

The qualitative descriptive research method was employed to investigate the perceptions of teachers regarding teaching with an interactive whiteboard. Data was collected from teachers via interview where we divided our teachers into two groups; the first group was the teachers who are teaching from first to fourth grade, and the second group was the teachers who teach from fifth to eighth grade.

B. Participants

Participants are the teachers of our school. A total of 29 teachers were interviewed for this research. The first group of teachers had 14 members, while the second one had 15 members. Only several teachers had previous experience with interactive whiteboards. The majority of

teaching staff started using an interactive whiteboard at the beginning of the school year (September, 2014). The average age of our teaching staff is 40 years old. 86% of our teaching staff hold Master (mag. prim. Educ. or mag. Educ.).

C. Sample question from the research

1. Would you rather work in classic classroom or modern one?
2. Do you use ICT in your classroom?
3. Do you think you can benefit from ICT in order to improve your class?
4. Do you think that workshop helped you with using interactive whiteboard in your class?
5. Do you have to reorganize the way you teach your classes now when you are using interactive whiteboard?
6. Do you think you have to do the research on how to improve using interactive whiteboards in order to benefit the students you teach?
7. Do you think we should organize more workshops for using interactive whiteboard in class?

C. Analysis

The first group of teachers were the ones that teach from the first through fourth grade, while the second group were the teachers that teach from fifth through eight grade. The second group of teachers has one or two subject that they teach in higher classes, while the first group has to teach around 10 different subjects.

Since the first group of teachers had to incorporate interactive whiteboards into 10 different subjects, which means they had to do more research and practice than the second group [5]. Both groups had previos experience with implementing ICT in their classes (using computer or laptop for presentations, videos or educational games).

At the beginning of the school year, the school had organized an educational workshop for our teaching staff where our teachers learned how to use interactive whiteboards. We found out during interviews that teachers had to reorganize their classes in order to incorporate interactive whiteboard into a class curriculum. This action compelled them to do research and to extend their knowledge about using information and communication technology in the class [6].

D. Results

Results of this research compared with the results of Türel and Johnson's research are very similar. The interactive whiteboard has allowed our teachers several instructional strategies; highlighting, coloring, flipping back and forth to review previous content providing reviewing techniques better understanding, using pictures for discussion and brainstorming, collaborative writing, shared reading, peer-teaching, and collaborative problem solving, hiding and revealing, drag and drop, and matching items activities, observing different media which is essential for visual learners, accommodating lower ability and special needs, finding hidden part of a picture with spotlight or screen-shade, capturing screenshots from web pages synchronously and manipulating them, correcting mistakes in the materials and playing games [7]. Using interactive whiteboard had also given us a chance to save paper since we didn't need to print exercises and distribute them to our students so it gave us environmental benefit as well.

Students have increased their independence from the teacher as well as learning outcomes. They can vividly learn from their mistakes and they are enabled to construct and evaluate by themselves which results in higher motivation and engagement into class curriculum. We try to see students as subjects of our educational process rather than objects. With interactive whiteboards the classic frontal approach is fully replaced by interaction between teachers and students. This way we were able to change class dynamics and encourage students to actively participate in class.

Both groups of teachers have succesfully implemented interactive whiteboard in their classes.

V. EXAMPLES OF GOOD PRACTICE

A. Computer science

Teaching Computer science was elevated in several ways. Students are able to use interactive whiteboards to their full extent. We are able to combine matter from several subjects very easily (for example Mathematics, the English language and Arts are correlationing when learning and creating spreadsheets) so that the learning outcomes would be visible in every subject and not just one (in this case Computer science), which results in coherence of the syllabus.

We managed to raise the dynamics by encouraging students to be more independent and constructive in their learning outcomes with the constant support of the teacher. All of Computer science classes in Kajzerica elementary school are visible in the private student access network where they can upload and submit their notes and findings and later present it to the whole class. Some of students' works can already be seen on the school's website

B. 1st grade Croatian language

In the first grade elementary school, students learn to recognize and write letters. They become more familiar with the surroundings they live in and they are taught to express themselves.

Using interactive whiteboard has enabled teachers to combine several resources to make classes fun and creative. One teacher has created an exercise with letters of the alphabet on the far left of the screen, in the middle is the picture starting with that specific letter and on the right is the soud of that letter. The teacher would mix the pictures and the students had to match the specific letter with the picture and the recorded sound of that letter. This exercise enabled them to use information and communication technology methodically from the early age which resulted in increment of self confidence and creative thinking. Later on, they would be able to draw an object into the given box whose first letter starts with the letter in question and then to write the whole word.

C. Geography

Our Geography teacher has completely replaced textbook maps with Google Earth. While explaining the position and basic characteristic of individual countries or continent, she would point her hand to the two dimensional paper world map and the students would remember her action and do the same. Now, she launches this application and uses the interactive whiteboard so students can interact with the globe and approximate the country in question. Furthermore, they can change views and see satellite images, maps and real life footage. She can copy the screen of the footage using interactive whiteboard tools and then modify the picture, as well as flip several pictures in question for repetition and better understanding. This allows students to develop critical thinking and to compare the characteristics of several countries very vividly.

D. Foreign language (English and German language)

Teaching foreign language becomes more interesting and creative using the interactive whiteboard. Similar to the first grade example, students need to match the sounds of letters in a foreign language with the picture of the letter in question. Later they match the sound of the word with the word in question.

Older students can learn to assemble sentences in foreign languages very easily using the interactive whiteboard. The following exercise has several components: recorded sound of words in a foreign language, written words in Croatian language, written words in a foreign language, constructed sentence in Croatian language and the gaps for constructing the sentence in a foreign language. In the early stage, the teacher reproduces recorded sound of the sentence in a foreign language (the answer) so that the student needs to recognize and match the words with the sound and fill in the gaps with written words in a foreign language to construct the sentence. Later on, the recorded sound of the sentence in a foreign language becomes redundant and the students need to construct the sentence on their own.

E. Chemistry

Teaching Chemistry is difficult without visual component. Although presentations can be considered good and interesting, students cannot interact with it, nor adjust it to their satisfaction. Interactive whiteboards enable students to create molecules on their own, to connect the atoms using right connection and to see the differences in object motion. This gives us better insight to the student's understanding of the topic.

VI. Conclusion

Introducing interactive whiteboard gave our school the following benefits: increment of digital competencies of our teaching staff, our students are more motivated and proactive, our teachers have many more resources to create an interesting and fun classes, interaction between teachers and students is more dynamic and we have begun to collect materials to create our e- learning database.

References

[1] Airhunmwunde, Matthew, "Review of the Impact of ICT in School Curriculum",CCLP Worldwide Publication Volume IV Issue III, Oct 2013

[2] Jones, S., & Tanner, H., Teachers' interpretations of effective whole class interactive teaching in secondary mathematics classrooms. Educational Studies, 28(3), 265-274., 2012

[3] Diallo Sessoms, Interactive Instruction: Creating Interactive Learning Environments Through Tomorrow's Teachers, Sessoms, D. (2008). Interactive instruction: Creating interactive learning environments through tomorrow's teachers. International Journal of Technology in Teaching and Learning, 4(2), 86-96.

[4] RamzkieLumauag, Interactive Teaching Strategies, Dec 03, 2012, http://www.pinoyguro.net

[5] Francine Hart, An Elementary Teacher's Feelings of Self-Efficacy, Needs, and Pedagogical Strategies When Learning to Use the Interactive Whiteboard to Teach Science, Thesis submitted in Partial Fulfillment of the Requirements for the Degree of Master of Arts in Education with a concentration in Teaching, Learning and Evaluation, Faculty of Education, University of Ottawa, June 2014

[6] Edith Manny-Ikan Henrietta Szold Institute, Israel, Osnat Dagan KadimaMada Educating for Life, Israel, Tal Berger Tikochinski and Rachel Zorman Henrietta Szold Institute, Israel, Using the Interactive White Board in Teaching and Learning – An Evaluation of the SMART CLASSROOM Pilot Project, Interdisciplinary Journal of E-Learning and Learning Objects Volume 7, 2011

[7] Türel, Y. K., & Johnson, T. E. Teachers' Belief and Use of Interactive Whiteboards for Teaching and Learning. Educational Technology & Society, 15 (1), 381–394., 2012

Interactive Teaching with Socrative

M. Filipović Tretinjak*, A. Bednjanec* and M. Tretinjak*
*School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
martinafilipovic.tretinjak@gmail.com; andrea.bednjanec@gmail.com; mario.tretinjak@gmail.com

Abstract – Interactive teaching methods promote a learning environment that leads to achieving the expected learning outcomes through encouraging student participation. Socrative student and Socrative teacher are applications that enable teachers to implement an engaging class environment through educational exercises and games that can be used on any web enabled devices, such as smartphones, tablets, laptops and desktops. Experience from the School of Electrical Engineering in Zagreb shows that Socrative supports both independent and collaborative learning and can create an interactive learning environment.

I. INTRODUCTION

Interactive teaching methods encourage students to participate in class activities, use questions to stimulate discussions, measure student understanding. Each class activity enables students to deepen their learning and achieve the planned learning outcomes. Interactive teaching supports a student-centered approach in the learning process. The advantages and disadvantages of the interactive teaching with Socrative are given in Table 2 below.

TABLE I. THE ADVANTAGES AND DISADVANTAGES OF THE INTERACTIVE TEACHING WITH SOCRATIVE

		Advantages	Disadvantages	
Interactive Teaching with Socrative	**Teacher**	can easily check the students' understanding (e.g. Exit ticket), can create games to track the students' performance while playing (e.g. Space race), have automatically-graded multiple choice quizzes, can monitor the students' progress via Excel or Google spreadsheet, can have up to 50 students loged in at one time per room, have access to the Socrative account from different mobile devices	**Teacher**	needs to have access to the required technology, needs to have access to a computer and internet, needs time to collect and create the necessary learning materials
	Students	are involved in various activities through the use of technology, comprehend better, become more interested in course topics, develop critical thinking, have access to the Socrative from different mobile devices	**Students**	need to have access to a computer and internet, spend more time on the computer

Socrative is a free smart student response system that enables teachers to engage students in the learning process with a series of educational activities via any web enabled devices such as a smartphone, tablet, laptop or desktop. By using Socrative, teachers can immediately collect feedback from students as well as monitor their progress and help them to become self-evaluators of their work.

To be able to use Socrative, a teacher needs to create an account, students then access the Socrative virtual room by using a unique room code that is provided to them by the teacher. After logging in, students can be involved in various activities such as quizzes and answering quick questions.

II. SETTING UP A SOCRATIVE TEACHER ACCOUNT

A Socrative teacher account can be setup either through a web browser on laptops and desktops or on the Socrative teacher application on any web enabled mobile device such as a smartphone or tablet. Currently, Socrative allows one classroom per account. If additional rooms are required, the teacher needs to create and manage additional Socrative teacher accounts.

Students don't need to create a separate account or login credentials. Socrative teacher and Socrative student accounts are free for use.

A. Creating an account - using a web browser

A Socrative teacher account can be setup through a web browser, including Firefox, Chrome, Safari and Internet Explorer (http://b.socrative.com/login/teacher/). After entering email and password you will enter the Socrative site (Dashboard) where you will receive a unique Room Code (e.g. ROOM: f306d0a7), which students will use to access the created activities.

B. Creating an account - using the Socrative teacher application

A Socrative teacher account can also be registered through the Socrative teacher application (http://www.socrative.com/apps.php to download the application) via a smartphone or tablet.

After downloading and installing the Socrative teacher app, start the application, enter email and password and enter the Socrative site (Dashboard) where you will receive a unique room code (e.g. ROOM: f306d0a7), which students will use to enter a virtual room and access the created activities.

III. LOGGING IN TO SOCRATIVE

Teachers log in to Socrative by using their created teacher accounts (http://b.socrative.com/login/teacher/ and enter email and password) via a web browser or the installed Socrative teacher application. Teachers that have a Google email account can log in by using the "Sign in with Google" button.

Students can log in to Socrative by using a web browser or the installed Socrative student application. To access the teacher's virtual room, students need to enter the room code (given by the teacher) and then select the "Join Room" button. After logging in, students can access various activities.

IV. ACTIVITIES

Students learn through their participation during the lesson. Educational quizzes can help stimulate interest and promote further learning or can be used as an assessment tools.

Each activity below provides students with opportunities to deepen their learning. To teachers these activities can provide feedback about the students' learning.

After logging in to Socrative, teachers can start different activities such as: start a quiz, run a space race, run an exit ticket, run quick questions (e.g. multiple choice and true/false quick questions, short answer quick questions).

Within Socrative, only one activity can be active at the same time and this activity will stay open until the teacher selects finish or chooses to start another activity.

Teachers can see the activity results through an opened browser or application. Even if the teacher chooses to shut down his/her computer, the activity stays opened. To end the currently running activity, the teacher needs to press the button Finish.

A. Start a quiz

At the beginning of the lesson students individually or as a team can take a quiz about the class topic they have already learned. The teacher estimates the students' comprehensions of the topic and gaps in understanding.

Quizzes like evaluation exercises can contain multiple choice questions, short answers or can be a combination of the two. Teachers can create a new quiz, import a quiz from another teacher, import a quiz from Excel and access quiz reports through the Dashboard (main screen in Socrative) using Manage Quizzes button.

When importing a quiz from an Excel file, teachers need to download the Excel template first, populate it with quiz question and answers according to the template headings and save it to a computer with a unique name (e.g. SOC-XXXXXX). A quiz can then be imported into Socrative as an Excel file.

When starting a quiz, teachers have three options to choose from: Student Paced – Immediate Feedback, Student Paced – Navigation or Teacher Paced.

When using the Student Paced – Immediate Feedback quiz type, students receive immediate right/wrong feedback and explanations after answering each question. Students answer questions in a particular order and cannot skip them or change their answers. Teachers are able to monitor the students' progress via the Live Results Table.

When using the Student Paced – Student Navigation quiz type, students have the ability to edit questions, skip questions and navigate the quiz in their own way. Once they've completed the activity they can submit the entire assessment. Teachers are able to monitor the students' progress via the Live Results Table.

When using the Teacher Paced quiz type, teachers are able to control the flow of questions. Teachers can send one question at a time and visualize the responses as they happen. Teachers can also skip and revisit certain questions.

Additional quiz settings allow teachers to disable student names and randomize the order of questions or answers.

B. Run a Space Race

Quizzes can be deployed as a fast-paced rocket race game. Students are engaged in fun exercises that track their performance while they're playing. This Socrative feature allows students to work in pairs or small groups in order to answer the quiz questions.

Teachers can choose the number of teams (there can be up to 20 teams), team colors can be automatically assigned or selected by students. Each correct answer moves a rocket across the screen. The team of students that have answered the most questions correctly will win the space race.

Before activating the Space race quiz, students need to be divided into teams, while only one person from each team needs to be logged into Socrative. The teams need to choose a team name before the game is started.

There can be teams that believe they have no chance of winning so they quit the game. If that happens, each student needs to submit his/her answers on a piece of paper after the game to help them get back into a good mindset.

If there is a small number of students (up to 14) in a class, students can participate in the space race separately.

C. Run an Exit Ticket

Exit tickets are quizzes that are mainly used to check the students' understanding at the end of a class. Exit tickets provide students with an opportunity to reflect on what they have learned, to analyze their learning efforts.

For teachers, Exit tickets provide feedback on an instructional strategy used during a lesson. It is an easy way to gather the students' responses, access level of mastered learning outcomes and their interests. Based on the data collected, teachers can at the beginning of the next lesson clarify small misunderstandings and adjust the pace for further lessons.

Here are some examples of Exit ticket questions:

- Name one important thing you learned in the class today.
- Write/ask one question about the content of today's lesson - something that has left you puzzled.
- What questions do you still have about today's lecture?
- I used the blackboard extensively today. Were its organization and content helpful in your learning? Why or why not?
- Which of the readings you did for the today's class was most helpful in preparing you for the lesson? Why?
- We did a concept map activity in the class today. Was this a useful learning activity for you? Why or why not?

D. Quick question activities - "Right now" polling

Quick question activities enable teachers to do a quick pool during the lesson. Pool questions can be asked aloud, but it is much better to give students a paper copy of the questions. If a classroom is equipped with a projector, questions can be projected on the screen.

Teacher can ask a question with the option Short answer and students then give their answer. After students have answered a question, the teacher uses the option Start Vote. The class can then vote on their favorite responses.

V. REPORTS

Reports are used to keep students activity results. After students have completed the activity, the teacher needs to press the Finish button and the Report pop-up window will open. From the Report pop-up window, various report types can be selected, such as the Whole Class Excel, Individual Student(s) PDF or Question Specific PDF.

To download a selected report to the computer, select the Download option. Because the download option is not available on tablets or smartphones, the report can be received via email and Google Drive or can be viewed in the live results Chart (option View Chart). To get reports sent to the Google Drive, the email associated with your account must be a Google Domain email address and you must log in via Google Single-Sign-On.

Then select the report you want (Whole Class Excel, Individual Student(s) PDF or Question Specific PDF) by clicking the grayed-out OFF and toggling it to ON. To collect all the data in one spreadsheet, it is best to use the Whole Class Excel report. The Individual Student(s) PDF refers to the report with a completely graded quiz for each student. The Question Specific PDF report shows response data for each question in a full quiz format. To receive the selected report, click the Submit button.

To access all previously created reports, go to the Dashboard and click on the Manage Quizzes button.

VI. SEARCH RESULTS

Research was conducted for two months on seventy-four students from 17 to 18 years of age. Participants were students who attended subjects from the vocational education program: Computer system electronic technician. Research was conducted during courses that already have in-class computer-based activities.

Students already use different platforms such as Edmodo (https://www.edmodo.com/) and Zondle (https://www.zondle.com/publicPagesv2/). They always like to participate in various activities that can be carried out through the use of smartphones, tablets, laptops and desktops.

Figure 1 shows the research results of the students' feedback on using Socrative. Most of the interviewed students (82%) liked using Socrative and claimed that Socrative gave them a better insight into what they have learned.

A small number of interviewed students (18%) didn't like participating in Socrative activities. They explained that this was because there was too much emphasis on technology. They would rather participate in activities that include working with printed materials and textbooks.

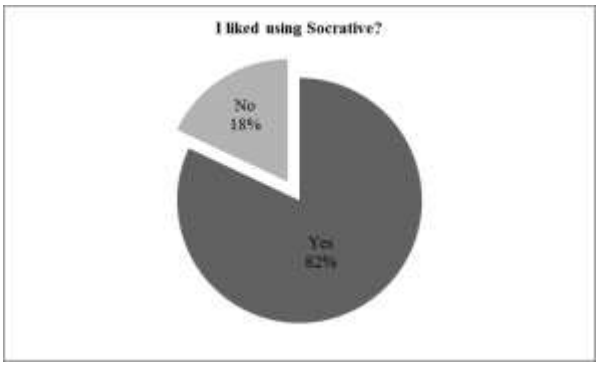


Figure 2. Students' feedback on using Socrative

Figure 2 shows that the 83% of the high school students agreed that Socrative activities were an interesting new way to learn. A small number of interviewed students (17%) didn't like using Socrative. They said they would rather use Edmodo learning platform.

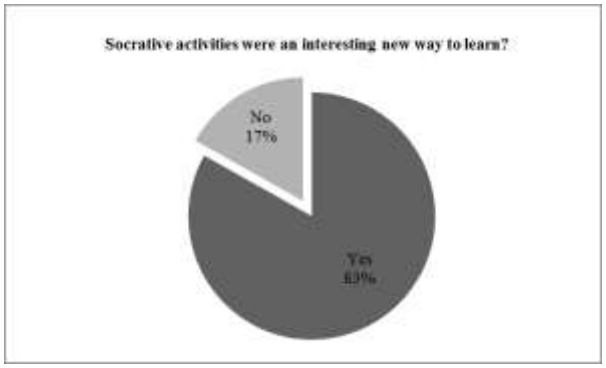


Figure 1. Students' feedback on using Socrative

Here are some examples on what student wrote about Socrative:

- *"The Socrative activities were fun!*"
- *"It was nice to learn in a different way."*
- *"I think that it was a totally useless activity."*
- *"Space race got me motivated to learn more and it was fun.*"
- *"I would rather use my textbook and printed materials."*
- *"I have learned a lot."*
- *"Space race is my favorite class activity."*

VII. CONCLUSION

Students perceive and assimilate information in different ways. Socrative allows the implementation of the interactive learning environment in which students repeat core information and practice basic skills through various games.

Quizzes throughout the lesson can help teachers assess the effectiveness of their instruction, as well as student understanding of the concepts taught.

Socrative allows teachers to engage and assess their students through the educational activities on tablets, laptops and smartphones. Through the use of real time questioning and instant result aggregation, teachers can easily evaluate the effectiveness of their instruction and the students' current level of understanding.

Experience from the School of Electrical Engineering in Zagreb shows that the use of Socrative increases student participation during class and provides instant feedback (to both the students and the teacher) on the achieved learning outcomes, thus creating an interactive learning environment.

REFERENCES

[1] http://www.bedu.com/interactive.html

[2] http://www.socrative.com/

[3] http://www.socrative.com/apps.php

[4] http://www.socrative.com/resources.php

[5] http://www.makingtechsimple.com/socrative

[6] http://www.makingtechsimple.com/socrative

[7] http://garden.socrative.com/

[8] http://www.thelandscapeoflearning.com/2012/02/11-ways-to-use-socrative-as-formative.html

[9] http://catlintucker.com/2014/01/my-favorite-mobile-apps-2-socrative-lesson-ideas

[10] http://www.nicolechubb.com/2014/07/31/getting-set-up-on-socrative/

[11] http://www.brown.edu/about/administration/sheridan-center/teaching-learning/effective-classroom-practices/interactive-classroom-activities

Using cloud computing in Stjepan Benceković elementary school

Branka Smolković Cerovski, Martina Draženović
Stjepan Benceković elementary school, Zagreb, Croatia
Stjepan Benceković elementary school, Zagreb, Croatia
brankasc@gmail.com, martina.drazenovic@gmail.com

Abstract - Cloud computing means storing and accessing data and programs over the Internet instead of your computer's hard drive. By introducing cloud computing in 2012, the school has managed to ease access and sharing, enhance collaboration between teachers and school management and was able to adopt and consummate another digital competence in order to benefit our school curriculum. Another crucial benefit from implementing cloud technology to our school is saving money. Stjepan Benceković elementary school is a small school with very limited resources and it is very important to us to explore and use technology in order to save money and time. This paper will show how did we implemented cloud computing, how much money has been saved and how did cloud computing evolve school curriculum.

I. INTRODUCTION

Stjepan Benceković elementary school introduced cloud storage in 2012. It changed the way pedagogical documentation is stored, it enhanced collaboration between teaching and administrative staff and it opened the door for many more possibilities in order to increase digital competencies of teaching and administrative staff and to enhance students' learning outcomes.

II. SCHOOL'S CHARACTERISTICS

Stjepan Benceković elementary school is a small school with limited resources. It is not self sustainable and is almost fully dependent of Zagreb city educational office. The building is very old and it is not very well equipped. It has a total of 6 classrooms and a gym. Class is organized in two shifts.

Technical characteristics of the school are not at all satisfying; our computers (30) are 10 years old on average and the Internet connection has not been set in every classroom. Every projector that school possesses doesn't function properly or it doesn't function at all.

Despite the disadvantages, the school is very active in various projects (over 20 in the past 2 years), it constantly develops their curriculum and innovates, annually publishes school newspaper and cares about the community.

III. CLOUD STORAGE

Cloud storage is a way to save information to the web. Cloud computing services are online services which store this information in a safe and secure location online; they offer a quicker, easier and safer backup than conventional methods of file storage.

Using cloud computing, you can access your files and programs from any computer anywhere in the world; all you need is an internet connection.

IV. IMPLEMENTATION PROCESS

A. *The idea*

At first, the school management started using Skydrive (today's OneDrive) to store various school documentation which later evolved into the idea to store school documentation and resources and making it transparent for the whole teaching and administrative staff without using paper in any way.

B. *Pedagogical documentation in Croatian elementary schools*

Pedagogical documentation in Croatian elementary school consists of: school curriculum, school's annual work program, school's annual subject program, school's annual civil program plan, school's monthly subject program, various certificates of development of teaching and administrative staff, calender of written tests, various laws and regulations. The school had to print out everything listed above which would consist of thousands of sheets of paper every year.

C. *Tools*

We are using Microsoft technology since the Croatian Mistry of science, education and science closed the deal that all Croatian schools have benefited from Microsoft technology which includes operating systems and Microsoft Office tools.

The school is now using Microsoft OneDrive and Office 365. We have 15 GB of storage at our disposal.

D. *Workshop – learning how to use cloud storage*

The school management organized workshop for all teaching and administrative staff in order to learn how to use cloud storage. In the beginning of the school year

2013/2014 the whole teaching and administrative staff was using OneDrive to store school's documents.

The workshop consisted of several parts: introduction, managing workspace, uploading and downloading files, creating folders and sharing folders and/or files.

V. FINANCIAL BACKGROUND

School's financial resources are very limited. The educational office of the city of Zagreb endorsed around 1700,00 kn monthly for office material which contained of paper, printing toners, pens and other.

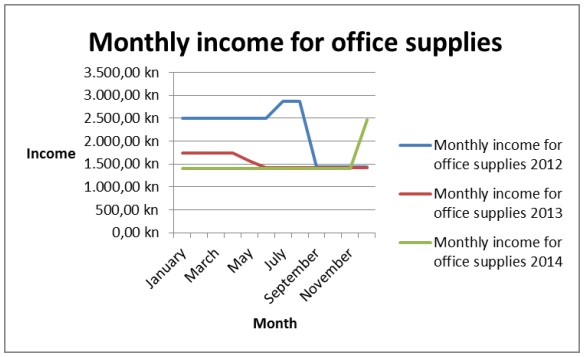


Figure 1. School's monthly income for office supplies

From Figure 1. we can see the decrement of income throughout out the years. The average monthly income in 2012 for office material was 2208,67 kn, in 2013 it was 1540,17 kn and in 2014 it was 1495,33 kn.

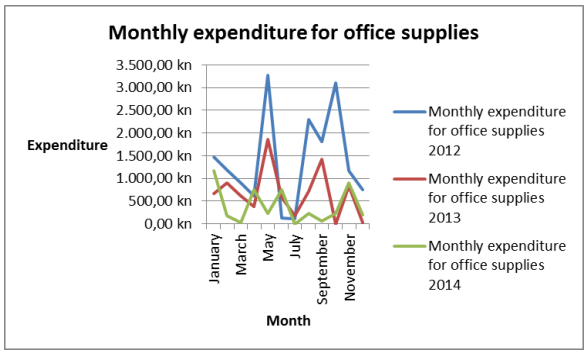


Figure 2. School's monthly expenditure for office supplies

Analyzing Figure 2. we can see that monthly expenditure in 2012 is significantly bigger than 2013 and 2014. Average monthly expenditure for office supplies in 2012 was 1405,64 kn, in 2013 it was 682,36 kn and in 2014 it was 391,82 kn.

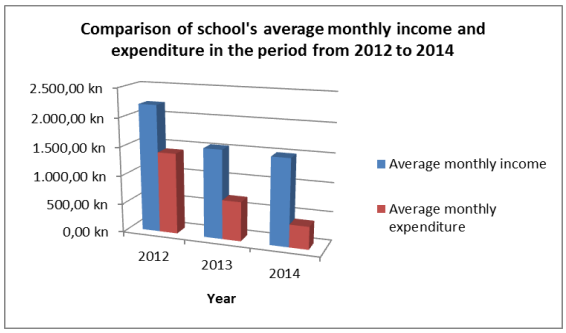


Figure 3. Comparison of school's monthly income and expenditure in the peroid from 2012 to 2014

Figure 3. shows significant decrement of income and expenditure for office supplies in 2014. It also shows that the difference between income and expenditure, despite decrement of the income, becomes bigger every year. In 2012 it was 803,03 kn, in 2013 it was 857,81 kn and in 2014 it was 1103,51 kn. This is the direct consequence of implementing cloud storage in school because we downsized the need for using paper, pens, printing toners and other office supplies.

VI. RESULTS OF IMPLEMENTATION PROCESS

A. *Increasing digital competencies of teaching and administrative staff and improving collaboration*

By implementing OneDrive and introducing Microsoft Office 365 tools in general into our school, we developed and increased digital competences of our teaching and administrative staff. We improved collaboration between our staff by sharing resources and now we can see more clearly the collaboration between subjects (through individual subject plans) and how it has been done.

B. *Saving money*

Since the school has very limited resources, we are very eager to save money, however we can, but the imperative was that our students don't suffer in this process. Now we can direct the money we saved for student's projects and needs which we had to decline in the past because we had no resources to endorse them.

C. *Transparency*

The whole school documentation became transparent, which enormously helped our teaching staff. We can see how our school curriculum develops during the school year, what other teachers do with their students and how they present their matter so that we can correlate easier and efficiently.

D. *Care for the environment*

Saving paper shows our care for the environment which is very important for our school since it is a part of our curriculum to awake our students to care for the nature in general and to educate others to care following their example.

E. *Future plans*

In the future we have plans to save enough money so that every classroom could have Internet access which would propel us to develop our e - learning model. We believe that it would influence student's motivation in class and that it would make them more proactive.

VII. CONCLUSION

Implementing cloud storage had many benefits for our school; it increased digital competencies and care for the environment, saved money and made us more determined to try to implement more components of information and communication technology in order to benefit our students and grant them the conditions in which they can all progress.

REFERENCES

[1] Logica Banica, Cristian Stefan, Doina Rosca, Florentina Enescu, Moving from Learning Management Systems to the e-Learning Cloud, 3rd World Conference on Innovation and Computer Sciences 2013, AWER Procedia Information Technology & Computer Science, Vol 04 (2013) 865-874

[2] Chen-Feng Wu, Learning Attitude and Its Effect on Applying Cloud Computing Service to IT Education, International Journal of u- and e- Service, Science and Technology, Vol. 6, No. 1, February, 2013

[3] Geir Ottestad, School Leadership for ICT and Teachers' Use of Digital Tools, Nordic Journal of Digital Literacy 01-02 / 2013

Project-based learning : Students (and teachers) as authors / co-authors of interactive multimedia educational content (Space Teacher)

Krunoslav Bedi
School of Civil Engineering, Art and Craft Čakovec
Športska 1, HR-40000 Čakovec, Croatia
bedi.krunoslav@gmail.com

Abstract - SPACE TEACHER is a project of an interactive and educational multimedia CD / DVD. The aim of this paper is to present the extracurricular activities of an IT group of secondary school students who have developed a multimedia CD / DVD over the period of three years. The project was based on the knowledge acquired in geography classes (teaching unit: the Earth and the solar system).

The purpose of this paper is to examine the students' attitudes towards acquiring knowledge in extracurricular activities as well as development and implementation of educational and interactive multimedia content in school and their future job. This particular educational and interactive multimedia CD / DVD promotes multimedia and e-learning.

Keywords: multimedia, constructivism, e-learning, extracurricular activities, interaction, interactive multimedia educational content, CD/DVD

* In 2011, the application "Space Teacher" was presented at the national competition Info kup – Rabac – 2012. The trailer of the game can be accessed via – http://www.youtube.com/watch?v=qVCNUEoBw4k

I. INTRODUCTION

Widespread usage of computers and the media has created conditions for higher quality innovation of educational and teaching processes. Everyday life media presence is not only important from the aspect of communication but also from a social and educational one.

The digital era encourages new ways of mass communication. The digitalization process started on a global level has wiped out the borders. In such circumstances new media use affects identity. Aspects of authorship and co-authorship interweave in a public communication space. New media begin to merge with the existing cultural convention of the teacher as the unambiguous author of educational content. Time has shown that everyone, even the students themselves, can be authors / co-authors.

Learners must develop their thinking processes such as reasoning, creative and analytical thinking, systematical problem solving as well as computer skills[6].

Students are able to apply their technical knowledge, acquire practical skills in programming, get involved into team processes and understand in some cases even so called soft factors in project management[4].

However, nothing can replace the teacher and their role in the educational process. It is every teacher's wish that their students do not only have good studying results, but also enjoy studying[8].

Project-based learning provides students with opportunities to pursue their own interests and questions, make decisions about finding answers and solve problems. In the classroom, project-based learning provides many unique opportunities for teachers to build relationships with students. Shared student work includes documentation of the learning process as well as the student's final projects with other teachers, parents, mentors, and the business community who all have a stake in the students' education[3].

II. MULTIMEDIA TEACHING AND CONTRUCTIVIST APPROACH TO LEARNING

Multimedia material is generally defined as the use of a combination of media forms, such as pictures and words, or animation and sound[5].

Multimedia contents combine text, image, sound, animation and videos previously used in various ways although recently a multimedia computer is commonly used to play multimedia files whereas CD-ROM or the Internet provide means of data storage[2].

The term multimedia classroom has been coined just recently. The entire system of multimedia teaching is being build up with the sole purpose of functionally and technically integrating different and numerous media into the educational process. Of course, the need for modern teaching should be accompanied by trained teachers.

Maximizing multimedia applications is the best practice of the contemporary educational system with various measures used to obtain the available knowledge and reference materials. Knowledge is therefore not limited to just one source. Indirectly, it provides a new way for educators and students to find additional teaching and learning materials. Paperbacks are not convenient for browsing and do not have the advantage of various information presented in audio-visual and high impact graphics[7].

The researchers, on the one hand, constantly raise the question of the impact of new technologies on educating, learning and teaching, while on the other hand the question of participants' willingness to accept new IT technologies comes up. Both issues are immensely significant and topical for the teaching process.

However, the most significant change in the framework of developing new creativity is no longer produced by the established media publishers. For example, the most watched clips and videos are not and need not be made by professionals. They are individual works of anonymous people, new media users (the Internet, primarily).

The focus of this research is on the constructivist approach, as it is considered that the solutions necessary for change in contemporary schools can be found within it.

The constructivist teaching-learning approach which has been worded widely in educational settings in recent years focuses on students taking responsibility of their own learning while constructing their own versions of knowledge [9].

The assumption derived from the constructivist approach is that students can construct knowledge from their own experience and in such an educational process the question *what* do children learn transforms into *how*. Constructivism as a theoretical concept perceives building knowledge as a personal-based experiential process making acquiring knowledge unique for each individual. Emphasis is put on students selecting and processing information individually as well as generating hypotheses and decision-making based on their own mental models (cognitive structures, schemata) organizing personal experience. According to constructivism, the instructor should encourage students to discover the principles and laws of the contents learned on their own while previously transforming the information they need into a form tailored to the level of their acquired knowledge[1].

The goal of a contemporary school is to prepare students for life and use of modern technologies. Various forms of educational software are increasingly being used as sources of knowledge in everyday educational process. With an indispensable source of knowledge at hand students may have a different role in class.

III. EXTRACURRICULAR ACTIVITIES - DEVELOPMENT ENVIRONMENT

Apart from mandatory pedagogical work, extracurricular activities (groups) have an important role in the educational process. The aim of such activities, that is, each group is (or should be) students' personal development and spending quality leisure time.

The results of extracurricular student activities (IT group) are presented in this paper. The students have independently used the knowledge acquired during the three years of schooling and created a multimedia interactive CD / DVD, which can be used in secondary school geography classes.

In such an environment, teaching problem-based learning requires big changes in teachers' usual roles[10].

Along with their own experience, students used the subject curriculum of geography. The geography teacher was also at their disposal as an advisor.Implementation schedule was designed by the students themselves, keeping in mind the objectives of the curriculum as well as previous experience and knowledge.

It is worth mentioning that problem-content structuring is significant for the study of natural phenomena and the development of an interactive multimedia CD / DVD which entices students' curiosity and intellectual activities.

Research activities were crucial in the development of the CD / DVD. While planning the curriculum topics an integrated thematic approach demanded subject correlation and incorporating other curricula. In addition to learning about basic concepts, the aim of the interactive multimedia CD / DVD was developing curiosity, interest and ability to actively explore the environment as well as perception of people, phenomena and environmental processes while at the same time using all previously acquired IT knowledge in the making of the CD / DVD.

One of the background objectives of educational, interactive and multimedia DVD is primarily game-based

education. A multimedia DVD provides plenty of information about the universe and our solar system. Knowledge can be tested in a quiz. In addition to games Space Teacher includes a music player, a puzzle, as well as a photo and video gallery.

IV. STUDENT COMMENTS

... "Our personal experience from elementary school, primarily geography classes, brought this idea to mind. The subject teacher is in a difficult situation when they need to clarify the size and properties of the universe to the students.

It has been proven that the majority of environmental information is received through the eyes (visual) - a picture speaks a thousand words. A major problem in describing the universe is that it is unthinkably large and very difficult to show visually.

The universe is ultimately so large that its size exceeds the limits of human imagination. These limits make portraying the universe and its fundamental components in a consistent and non-abstract manner very difficult. This non-abstract project aims for multimedia effects on its users, expanding the boundaries of their imagination as well as showing and explaining the universe and its fundamental components in detail.

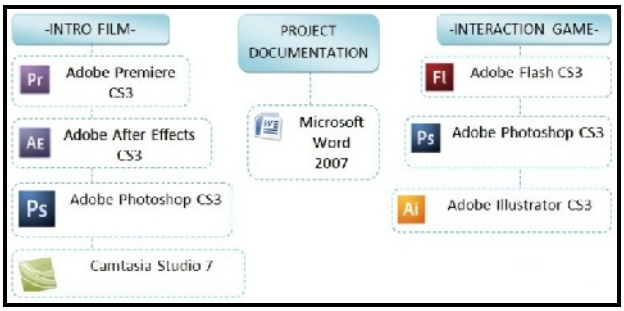
Educational, multimedia DVD was created with the goal of closely presenting geography to students, allowing them to learn both theoretical and practical subject material more easily and accurately. The focus of our educational and interactive DVD is on multimedia and e-learning.

The process of creating the application was not easy at all. Time spent on preparation and completion of the project confirms this fact. A whole year has passed from the initial to the final product stage, which indicates just how much time was devoted to research and development of the product so that it would be better and more convenient ... "

V. PACE TEACHER

The multimedia interactive CD / DVD starts with an introductory film (trailer) which is actually an application instruction manual. The trailer offers an application overview making further use much easier – the user learns how to use the application and solve problem tasks. The start menu was made with the aim of greater convenience and simplicity of use. Planets are the main elements of the menu. When the user moves the mouse cursor over any of the planets, the application automatically opens an additional window offering relevant educational content on the respective planet. At the bottom of the menu there is a practical musical player and the fun part of the multimedia CD / DVD consists of the following submenus: Picture Gallery, Puzzle, Quiz, Video gallery, Solar system, The universe, Black holes and Shoot The Moon (game).

A number of applications were used in the making of Space Teacher and supporting documents: Adobe Flash CS3 Professional, Adobe Premiere Pro CS3, Adobe Photoshop CS3, Adobe After Effects CS3, Adobe Illustrator CS3, Microsoft Word 2007, Camtasia Studio 7 (Picture 1).



Picture 1. Development environment

Windows
Intel® Pentium® 4 **(0.8 Ghz minimum)**, Intel Centrino®, Intel Xeon®, **or** Intel Core™ Duo
Microsoft® Windows® XP Service Pack 2 or Windows Vista™ Home Premium, Business, Ultimate**, Windows 7**
256 MB RAM
1,024x768 screen resolution with 16-bit video card
DVD-ROM drive

Picture 2. Minimal configuration required for DVD use

Picture 3.Development team

Picture 4. Application screenshots

VI. RESEARCH METHODS

The purpose of this survey was to determine the amount of multimedia content and effects that secondary school children may include in their extracurricular work and measure their satisfaction in such work. The focus of such acquired knowledge is on identifying the activities that can contribute to the improvement of teaching process in future work assignments / projects.

A questionnaire with eight questions was conducted on 109 students undergoing education for occupations web designer and media technician in the field of multimedia. First- to fourth-grade students participated. Students expressed their opinions on Likert scale where mark 1 stands for strongly disagreeing with the above statement, 2 disagreeing with the statement, 3 undecided, 4 agreeing with the statement and 5 strongly agreeing with the statement.

VII. RESEARCH RESULTS

The opinions of the students after they had taken part in the activities of the project of making a multimedia interactive DVD entitled Space Teacher:

Table 1. The opinions of students

I think that….	
…I have mastered a satisfactory level of knowledge and skills of creating multimedia applications.	
Mean	Std. Dev
4.333	1.019
…I have mastered a satisfactory level of practical skills required for getting a job in the production sector of economy.	
4.521	0.988
…I have mastered the use of (Adobe® CS4 Photoshop, Adobe® Flash CS4, Adobe® Premiere Pro® CS4 v4.0, Adobe® After Effects® CS4 v9.0. Cinema 4D®, 3D Studio Max etc.), the applications similar to those used in the production sector of economy.	
3.903	1.112
…I can successfully take part in all the stages of planning and developing of a multimedia application.	
4.001	1.433
…I can successfully cooperate in a team work.	
4.153	1.105
…my knowledge and skills level in the area of planning and organizing resources (time, space, material, equipment) needed for fulfilling activities are satisfactory.	
3.539	1.303
…I do have the skills of finding and using the relevant information for challenging multimedia problems.	
4.166	1.403
…I do have the skills for a creative solving of multimedia problems.	
4.278	1.129

Results show a high percentage of students in the field of multimedia (Mean = 4.333, Std.Dev = 1.019). Their perception of possibility of employment in the field is equally high (Mean = 4.521, Std.Dev = 0.988). As for multimedia tools: Adobe® CS4 Photoshop, Adobe® Flash CS4, Adobe® Premiere Pro® CS4 v4.0, Adobe® After Effects® CS4 v9.0. Cinema 4D®, 3D Studio Max, students' attitudes are even higher (PB3 Mean = 3.903,Std.Dev = 1.112). Other factors also show that students engaging in this type of work appreciate activities more when concrete assignement are involved.

VIII. CONCLUSION

As far as school is concerned, experts agree that the student must be in the center of the teaching process. In such circumstances, students using new IT technologies should be given the opportunity to develop their skills (and responsibility). In addition to developing their learning methods through specific educational forms, students should be given the choice of learning materials.

Regardless of instructions arising from a particular subject's curriculum it will (only?) be possible in specific extracurricular activities.

Knowledge acquired in this way is personal, with success in practice. This knowledge is a form of preparation for a dynamic world in which school is no longer an isolated, closed and rigid but rather a flexible and multidisciplinary institution.

At the same time, maintenance and development of this type of education will result in greater appreciation of the teacher and their work. It is the teachers themselves who represent the society for which the school prepares its students.

REFERENCES

[1] http://www.carnet.hr/referalni/obrazovni/mkod/pedagogijkonstr, uploaded November 8, 2014

[2] http://www.carnet.hr/referalni/obrazovni/mkod/pedagogija/multimedija.html, uploaded December 14, 2014

[3] Fezile Özdamli, The experiences of teacher candidates in developing instructional multimedia materials in project based learning, Procedia - Social and Behavioral Sciences, Volume 15, 2011, Pages 3810-3820

[4] Robert Pucher, Martin Lehner; Project Based Learning in Computer Science - A Review of More than 500 Projects, Procedia - Social and Behavioral Sciences, Volume 29, 2011, Pages 1561-1566

[5] Warren Davies, Kathryn Cormican; An Analysis of the Use of Multimedia Technology in Computer Aided Design Training: Towards Effective Design Goals, Procedia Technology, Volume 9, 2013, Pages 200-208

[6] Charuni Samat, Sumalee Chaijaroen; Design and development of constructivist multimedia learning environment to enhance computer skills for computer education learners, Procedia - Social and Behavioral Sciences, Volume 46, 2012, Pages 3000-3005

[7] Mohamad Zaid Mohd Zin, Ahamad Asmadi Sakat, Nurfahiratul Azlina Ahmad and Azri Bhari; Relationship Between the Multimedia Technology and Educationin Improving Learning Quality, Procedia - Social and Behavioral Sciences, Volume 90, 10 October 2013, Pages 351-355

[8] Beáta Krajčovičová, Martin Cápay; Project based education of computer science using cross-curricular relations, Procedia - Social and Behavioral Sciences, Volume 47, 2012, Pages 854-861

[9] Tolga Erdogan, Nuray Senemoglu; Problem-Based Learning in Teacher Education: Its Promises and Challenges, Procedia - Social and Behavioral Sciences, Volume 116, 21 February 2014, Pages 459-463

[10] Sofia Morgado, Laurinda Leite; Science and Geography Teachers' Conceptions Regarding Problem-based Learning Related Concepts**,** Procedia - Social and Behavioral Sciences, Volume 106, 10 December 2013, Pages 2343-2347

We like digital – using mobile and web applications in high-school project teaching

Krunoslav Bedi
School of Civil Engineering, Art and Craft Čakovec
Športska 1, HR-40000 Čakovec, Croatia
bedi.krunoslav@gmail.com

Abstract - The aim of this paper is to present the possibilities and ways of secondary school teaching provided by new information and communications technology with the growing trends of mobile device development taken into account.

The process of making a mobile device application is illustrated with student projects "MY CAR" and „MOBILE WEB BANKING". The purpose of this paper is to examine the extent and range of mobile applications use among student population.

Keywords: mobile applications, smartphone, learning, project-based classroom

* In 2012, application "Moj Auto" (My car) was presented and won the 3rd Prize at Info kup – Primošten – 2012. The trailer of the game can be accessed via – http://www.youtube.com/watch?v=ArNzj0CgTXY

* In 2012, application "MWB - Mobile Web Banking" was presented and won the 6th Prize at Info kup – Primošten 2012. The trailer of the game can be accessed via – https://www.youtube.com/watch?v=PpGJX_Vx1KU

I. INTRODUCTION

Mobile applications are types of application software designed for use on smartphones, tablet computers and other mobile devices.

As mobile devices and wireless technologies continue to rapidly grow and significantly advance, Web services technology recognizes mobile computing as an area to which it should expand. When combing smart mobile devices and Web services; i.e. running a Web service on a mobile device, one can greatly increase the functionality of mobile devices to interact with its environment. This makes a wide range of new functionalities and features possible especially in the field of ubiquitous computing[5].

Competition and conditions on mobile device market have resulted in a great improvement in computer power and possibilities they offer. In this paper two types of mobile devices are examined more closely; smartphones and tablet computers.

These devices can have a multicore processor, large RAM, full HD screens and built-in peripherals such as high-quality cameras. Nowadays, separate graphic chips in mobile devices are not uncommon. Simultaneously, operating systems for such devices have been improved, making programming easier and faster.

However, measuring student's attitudes towards new modes of learning interaction and Information Seeking in technology-pervasive learning environments is of particular interest to educators and learning technologists who seek to leverage the power of ICT for effective teaching and learning[3].These types of devices are most commonly used by younger population, more precisely, people engaging in a new form of information processing.

Classroom use of web applications has several advantages, such as involving students in learning activities and interactions along with facilitating learning reflection and presentation. Furthermore, a significant advantage is that the web applications have been widely used by most students in daily life. Therefore, the proposed approach could delight the majority of participants in using the web applications in a new and meaningful way [1].

Mobile applications are surely the drivers of communicational development. The market embraces everything that is innovative, useful and reasonably priced, regardless of whether it is made by a large manufacturer or an innovative individual. The most important things are a good idea and knowledge needed for its technical implementation so that global success and earnings can be achieved with a minimal investment. ☺

Students bring these technology anywhere, at anytime for their daily affairs. Educators should look upon this phenomenon as an opportunity though indeed it is a challenge as well. The concept of 'anytime' and 'anyplace' of mobile learning should be utilized in enhancing the pedagogical activities in delivering lessons[2].

There are two phases in the process of developing mobile device applications. The first and most essential one is selecting an operating system or platform on which the application will run, followed by defining the appropriate technology for its development. There is a large number of incompatible operating systems for mobile devices today (Android, Apple (iOS), BlackBerry, Nokia (Symbian), Windows Phone / Windows Mobile ...) therefore each developed application is not guaranteed to work on each of these devices.

Learning through mobile devices has enhanced a lot with the advent of pervasive computing technologies. In this age of progress and great change, we tend to easily adapt to the technologies and pedagogies that emerge [6].

II. MOBILE APPLICATION VS. MOBILE WEBSITES

Mobile apps have many advantages over mobile websites and are much easier to use. For example, an average well-made mobile application provides a much better user experience than the best mobile website. However, the costs associated with the development of mobile applications still remain an issue.

The development of mobile applications is usually more demanding (and more expensive) than the process of developing an equivalent multi-platform mobile website.

Mobile applications are better in terms of functionality and user experience while mobile websites development costs are lower. The main feature of mobile application development is difficult switching of the application to another platform which demands a new programme code.

With the number of mobile applications on the market increasing, it is evident that the field is very promising and offers many opportunities.

All mobile application developers who design and upload applications on mobile stores (Ovi Store, App Store, Android Market, Windows Marketplace ...) share an equal chance of success. If you have good programming skills, innovative spirit and are willing to invest some effort, you can create high-quality applications for a large number of users.

Therefore, the well-known web applications could be a potential new way to engage the instructors and students in meaningful teaching and learning activities [1].

III. PROJECT APPLICATIONS (STUDENT WORK UNDER MENTOR SUPERVISON) „MOBILE WEB BANKING" AND „MOJ AUTO" („MY CAR")

A. „Mobile Web Banking" - „MWB"

Fast pace and different lifestyle of the contemporary man accompanied by a rapid technical and technological development result in uncertainty and necessity to change the way of doing business in all financial institutions, including banks. MWB application provides the opportunity for establishing stronger relationships with customers and increasing bank revenue while reducing operating costs at the same time. This type of doing business has great potential and banks are facing a great challenge. If the possibilities of mobile technology are used completely, mobile banking could soon become a major channel for banking services distribution, even more popular than Internet banking. Mobile banking has great potential and it already offers the same possibilities as electronic banking. Additionally, the number of mobile phone users worldwide is increasing, with accompanying services becoming more extensive and easier to operate.

MWB is based on the majority of modern desktop, smartphone and tablet platforms.

Options offered by MWB applications:

1. System login, application is blocked after three failed attempts
2. Accounts overview
3. Daily transactions overview
4. Search all accounts
5. Option of fixed-term depositing
6. Cards overview
7. Loans overview
8. Transfers between accounts
9. Creating payment orders
10. Collective orders
11. Bank statements
12. Creating special reports
13. Automatic exchange rates update
14. Locating the nearest bank branch and directions to the branch or an ATM
15. Application design personalisation
16. Portrait and landscape view mode

In addition, older browsers are supported with a cross-platform solution.

Technologies used:

Server page

- ASP.NET 4.0 / C#
- MS SQL
- XML

Tools

Project developing is oriented towards Microsoft development tools:

- Microsoft Visual Studio 2010
- Microsoft Expression Studio

Client page

- HTML5
- CSS3
- jQuery Mobile
- Ajax
- JavaScript

Operating systems

B. „MY CAR“

Mobile learning aplication is made for educational purposes only. It is also convenient for the automobile industry, the public sector and driving schools. The application offers a marketing opportunity with advertising space given to a particular sector which uses it. It provides an overview of driving schools by county, offers choice and overview of over 60 car companies, notes, laws and road signs overview, knowledge quiz and, of course, fun. Users might also be interested in the opportunity to save their own notes and send them via email or social networks.

The application can be categorized as part of the latest and most effective market communication channels. When compared to traditional communication channels, the user is offered a greater degree of interactivity, wider scope, choice and range of information, as well as fun.

On the other hand, advertisers are offered a new, direct and more attractive communication channel. This is why smartphones are the foundation of the future, with mobile applications i.e. apps being the most contributing factor to a sharp rise of the smartphone market and the rise of the mobile revolution.

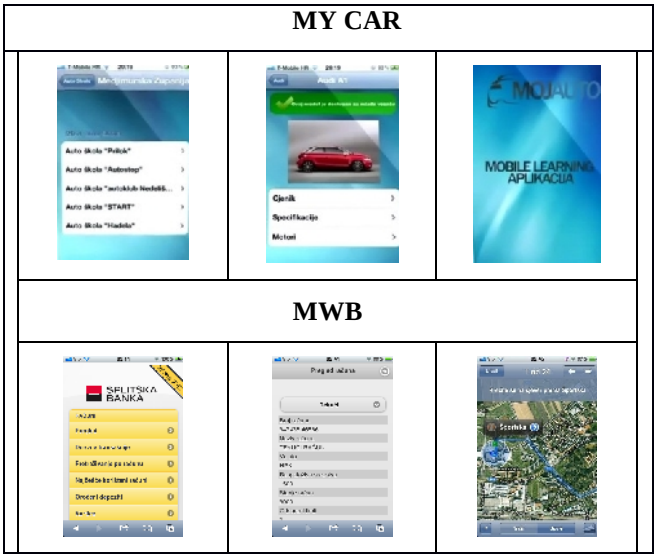
Following software was used in the making of iPhone application - xCode and Objective-C as well as Microsoft Visual Studio 2010, Cienma 4D, 3D Studio Max, Adobe CS4 Premiere, Adobe CS4 Photoshop, Adobe CS4 Illustrator and Adobe CS4 After Effects.

Two students (third and fourth grade respectively) made the applications **„Mobile Web Banking“ (MWB) and „My car“** under mentor supervision. A wide range of software normally used throughout the four-year course was implemented. For the implementation of this web-designer/media technician course all of the necessary software has been provided by the Ministry of Education and Sports of the Republic of Croatia. The work itself was team-organized as it had been proven as far more efficient.

One of the students did the programming and the other visual designing (MWB), while for the purpose of the application "My car" one student did both the programming and the designing. The students worked on the projects in their optional courses with the key role of the coordinator assigned to the mentor, i.e. the optional subject teacher, who is also the author of this paper.

Teaching mode and adapt to the constructivist learning theory in the multimedia network environment, can be summarized as follows: to students as the center, in the whole teaching process, the teacher as organisers, teacher, help and service providers, use of multimedia network environment resources through cooperative learning, students give full play to one’s own initiative, enthusiasm and entrepreneurial spirit, and make students the effective ways to achieve the goal of construction of realization, which means they have learned. In this model, the students are of positive significance constructer knowledge[7].

C. *Examples of application wallpapers*



IV. RESEARCH METHODS

In order to obtain relevant data, a research involving a sample of 303 students (first, second, third and fourth grade respectively) was conducted.

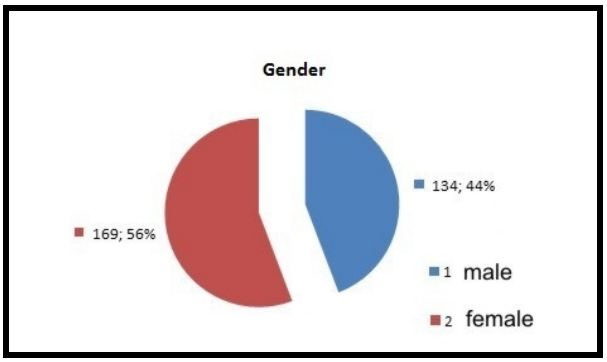
All of the students involved attend the following vocational programmes in School of Civil Engineering, Art and Craft Čakovec: media technician, web-designer, interior architecture designer and ceramics designer. The research was conducted by survey.

The survey sample included both sexes, 134 male and 169 female students. The students were given the possibility to withdraw at all times, to participate or decline participation, therefore one can assume that the obtained data is relevant.

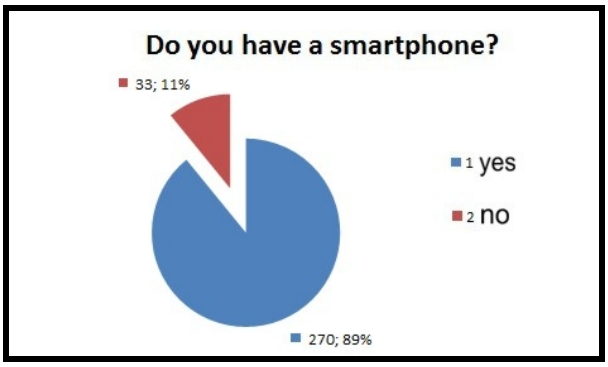
The aim of the survey was to get feedback on the following:

- do students own a mobile device
- using mobile devices for Internet access
- monthly Internet traffic and expenses paid for accessing the Internet via mobile device
- operating systems used on mobile devices
- mobile applications use
- number of mobile applications installed
- expenses paid for using a mobile application
- most commonly used mobile applications

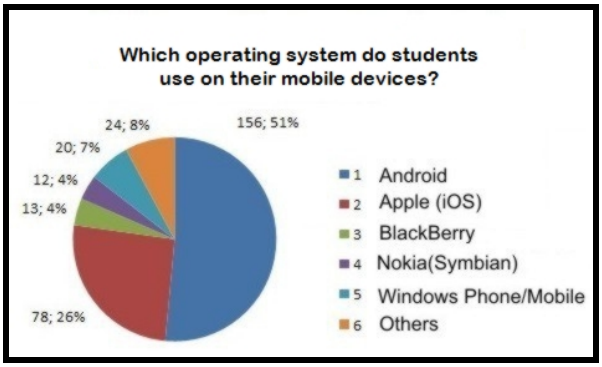
V. RESEARCH RESULTS



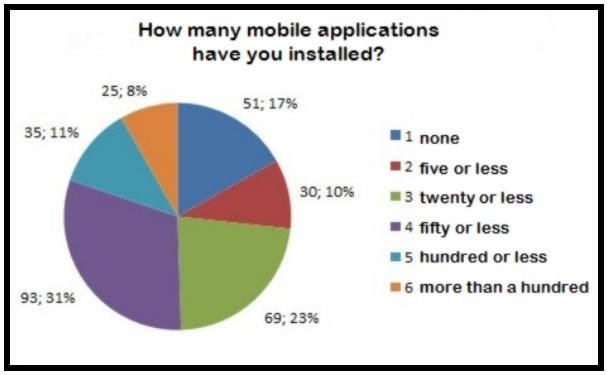
Graph 1. The number and percentage of students participating in the survey



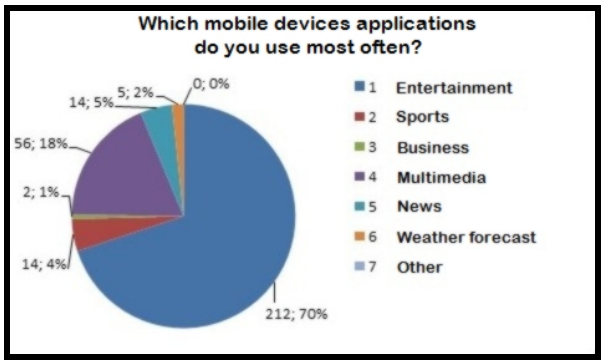
Graph 2. Do you own a smartphone?



Graph 3. Operating systems used on students' mobile devices



Graph 4 Number of mobile applications installed



Graph 5 Types of mobile device applications used

Statistical data show that the majority of students own a mobile device - 89%, and more than a half of them – 51% use Android operating system.

In addition, the data show that only 17% of students have not installed a mobile application while most of them, 83%, have already installed it. This confirms that most students already possess the knowledge needed for installing applications on mobile devices. The data confirm the assumption about the type of content installed - 70% entertainment, sports 4% and 18% multimedia features.

VI. CONCLUSION

Contemporary users are supposedly more devoted to web than mobile applications. Is investing in mobile applications a sensible move? By downloading mobile applications on their mobile phone, the user chooses to follow a particular service, which is probably the most important reason why companies create mobile applications of their own.

However, the user can have full access to a particular content just by entering a short website address, which is undeniably easier. They can enter just about any address and not opt for brand loyalty (as is the case with mobile applications).

The number of mobile applications available, as well as the number of users, are likely to increase in the future. Mobile applications have potential in several areas such as games, recreation applications, various readers and editors, etc.

The number of mobile applications which can be used offline will presumably increase while web applications will continue to dominate online.

The rapid developments of apps have shifted the way information delivered especially in educational field to adapt with the current technology and to be used in teaching and learning[4].

As for mobile applications and mobile websites, if we take all of the previously mentioned advantages, disadvantages and differences in mobile devices into account, it is necessary to adjust web applications to the mobile platform which would enable easier use for both programmers and other users.

Of course, a number of procedures need to be made. Obviously, capacity is the biggest drawback of mobile devices. Although they are closing in on desktop computers every day, mobile devices are still somewhat weaker and therefore not capable of running and processing data as fast as standard personal computers.

REFERENCES

[1] Yen-Ting Lin, Min Jou; A Web Application Supported Learning Environment for Enhancing Classroom Teaching and Learning Experiences, Procedia - Social and Behavioral Sciences, Volume 64, 9 November 2012, Pages 1-11

[2] Samsiah Bidin, Azidah Abu Ziden; Adoption and Application of MobileLearning in the Education Industry, Procedia - Social and Behavioral Sciences, Volume 90, 10 October 2013, Pages 720-729

[3] Leila A. Mills, Gerald Knezek, Ferial Khaddage; Information Seeking, Information Sharing, and going mobile: Three bridges to informal learning, Computers in Human Behavior, Volume 32, March 2014, Pages 324-334

[4] Nurul Farhana Jumaat, Zaidatun Tasir, Integrating Project Based Learning Environment into the Design and Development of Mobile Apps for Learning 2D-Animation, Procedia - Social and Behavioral Sciences, Volume 103, 26 November 2013, Pages 526-533

[5]KamalEldin Mohamed, Duminda Wijesekera; Performance Analysis of Web Services on Mobile Devices, Procedia Computer Science,Volume 10,2012,Pages 744-751

[6] P.V. Vinu, P.C. Sherimon, Reshmy Krishnan; Towards pervasive mobile learning – the vision of 21st century, Procedia - Social and Behavioral Sciences, Volume 15, 2011, Pages 3067-3073
[7] Lin Qingsong; The Application of Multimedia Technology in Web Education, Physics Procedia, Volume 33, 2012, Pages 1553-1557

Anonymity ensurance in creation of satisfaction surveys - experience of Polytechnic of Šibenik

Frane Urem *, Krešimir Fertalj **, Biljana Šupe *, Ivan Livaja *
* Polytechnic of Šibenik/Department of managemet, Šibenik, Croatia
** Faculty of Electrical Engineering and Computing/Department of Applied Computing, Zagreb, Croatia
e-mail: frane.urem@vus.hr, kresimir.fertalj@fer.hr, bsupe@vus.hr

Abstract - Quality assurance in higher education implies different student surveys that examine teaching, learning environments and educational outcomes. It is crucial to convince students that survey is fully anonymous because it is an essential prerequisite for getting the right answers from students about teaching process.The application of information technology enables further progress in the anonymous survey but the problem of student perceptions about the actual anonymity of participants still remains. This paper presents the recommendations of other authors to achieve that goal and conducting experience in student surveys at Polytechnic of Šibenik.

I. INTRODUCTION

The quality strategy at Polytechnic of Šibenik [1] (as in most higher education institutions in the Republic of Croatia) is defined with formal structure which consists of three independent bodies: Committee for quality assurance, Committee for internal evaluation of the quality assurance system and Committee for the implementation and processing of student surveys.

This form of organization arises from regulatory authorities in the field of higher education in Croatia, such as Ministry of science, education and sports of the Republic of Croatia (MZOS) [2] and Agency for science and higher education (AZVO) [3], but also from the quality systems which are applied in higher education institutions like standards and guidelines for quality assurance in the European higher education area (ESG) [4] or ISO 9001 [5].

Particularly important part for the quality strategy is a Committee for the implementation and processing of student surveys (in the rest of the text Committee) which (among others) implements teacher evaluation survey which evaluates the work of teachers from the perspective of students. The objective of this survey is assessment of teachers' work in a particular course. The results of this survey enable the separation of teachers who have been assessed as the best and worst from a student perspective. In the survey, students can share their observations about what's good and / or is not well in the work of teachers, and the questionnaire also contains standard parts on the quality of the work and observations of teachers and students services. At Polytechnic of Šibenik this surveys for a long period have been conducted on paper forms whose processing and subsequent analysis was prolonged and subject to errors due to rewriting the results because there was no possibility of automatic processing by optical readers. All that time Committee is particularly emphasized the problem of insufficient response of students to such surveys. In the meantime, Committee had some discussions with student representatives about that issue and the problem of student perceptions about the actual anonymity in existing survey procedure appeared as the main cause. The Committee acknowledged previously mentioned problem and proposed survey solution that would be as close as possible to the actual vote elections because it is assumed that improvement of student perceptions of the actual anonymity in survey will also increase survey response rate. It is also accepted the need for simpler and faster processing of the survey results and online survey method is suggested.

This paper presents the results achieved with online survey solution that is implemented according to the aforementioned recommendations by the Committee for quality assurance.

II. DEFINITIONS

According to [6] survey is defined as study where researcher collects data by asking people questions. Anonymous surveys do not record either the name of the respondent or the name of any institution that the respondent represents [6].

For anonymous online survey following rules must be applied [7], [8], [9]:

- respondent should not provide explicit contact information (like email, telephone ...)
- survey website settings must exclude the collection of online respondent's email address
- survey website settings must exclude the capture of online respondent's IP address

The above recommendations for the case of totally anonymous online survey mean that no link (technically: foreign key relationship) is available between survey invitations (for an example https link sent with an e-mail to respondent) and the responses. That can be in conflict with one major problem: *How to be sure that one respondent takes the online survey only once?*

III. METHODOLOGY

If the Committee for the implementation and processing of student surveys wants to avoid fraud and make sure that one person can take the online survey only once, then some kind of tracking system must be

implemented that will recognize that respondent has already took the survey. There are many technical solutions for this issue where online survey invitation is realized with an e-mail message witch contains authorization link or token number for anonymous survey but the problem is how to convince respondent that received authorization will not be linked with his survey answers and e-mail address later.

It may sound irrational, but respondent in this situation can only believe the Committee on their word because he doesn't have a full access to the technical system or proper technical knowledge for review. The main challenge is how to establish trust between two entities (Committee – student) in a way that does not compromise respondent's interest for anonymity. Therefore it is good to design a methodology closest to real voting experienced by respondents on the actual elections (parliamentary, presidential, local ...). That methodology is obliged to provide the following elements:

1. Personally voting and the principle of immediacy (there is no representative of the respondents, Committee determines that the respondents has the right to participate in survey because he is on the list of potential respondents - student is enrolled in the course for which the survey is carried out)
2. The online survey content
3. The method how respondents express their opinion in online survey (for an example one answer selection for a particular question)
4. Time (duration) for the online survey
5. Securing peace and order during online survey
6. Location with computers where online survey would be filled

The first and most important element from this methodology at the Polytechnic in Šibenik is applied with *LimeSurvey* system [8] unique functionality called "dummy token". With this functionality Committee can generate the total number of tokens according to the number of students registered on a specific course before the survey. Every token is printed on separate paper and placed in a sealed box that opens in front of students just before survey starts. Students get printed token from the box and completing survey unlocked with printed code. Since the token drawing is a random process there is no way to connect students and their survey responses in any way.

The entire process of surveying formally is described as follows in two major steps:

1. Survey preparation
   - Student office is delivered number of students enrolled for the course to be used for survey
   - One token for every enrolled student is generated and printed
   - Survey content is made
   - Survey date and time for specific course is defined by the Committee and posted on web and e-learning system

2. Survey conduction
   - Students and Committee members enter the computer room at the previously determined date and time
   - System administrator opens the survey
   - Students get printed token from the box, open survey with token and completing survey
   - System administrator closes the survey and printing report for Committee
   - The Committee determines the number of used tokens and closes unused tokens in the envelope
   - Students and Committee members leave the computer room

IV. CASE STUDY

The teacher evaluation surveys at Polytechnic of Šibenik were conducted by paper-pencil method until academic year 2013/2014. According to available data for the academic years 2010/2011, 2011/2012 and 2012/2013, the stable response for organized surveys (which did not exceed 25% of the total number of potential respondents) is shown in Table 1.

The student response was acceptable but Committee was not satisfied with the existing situation. Displayed turnout on student surveys for the mentioned period is not specifically analyzed, but it is certainly considered that probable reasons were that presence in the survey is not mandatory and the survey is generally conducted within the last week of classes when the students' attendance at lectures is significantly less. The same conditions were present during online survey and according to the comparison shown in Table 2 response in the online survey was significantly higher.

Therefore, subsequently research on possible additional reasons is conducted. A random sample of 78 third year students of 247 who certainly had the opportunity to meet the survey in both ways is questioned and results are presented in Table 3.

The results from the Table 3 indicate the possible different reasons for the worse turnout in the previous surveys. Among other things, it is evident that students easier express a worse opinion in new survey methodology about teachers' work in a particular course because they are obviously feel anonymous during survey and they are considered that manipulation of the survey results at new system is more severe.

TABLE I.. RECORDED STUDENT RESPONSE FOR PAPER-PENCIL SURVEYS

Academic year	**2010/2011**	**2011/2012**	**2012/2013**
Survey response	22,52%	18,20%	21,55%

TABLE II. COMPARISON OF THE SURVEY RESPONSE FOR BOTH METHODS

N	Study program	Method: paper-pencil	Method: online	Comp.
		Academic year 2012/2013	Academic year 2013/2014	
		Number of processed survey questionnaires	Number of processed survey questionnaires	INDEX III / IV (%)
I	II	III	IV	V
1.	Management	1048	1550	147,90%
2.	Administration	340	588	172,94%
3.	Transport and traffic	405	651	160,74%
TOTAL		1793	2964	165,31%

TABLE III. SUBSEQUENT RESEARCH RESULTS ON STUDENT CONFIDENCE FOR BOTH SURVEY METHODS

Question	Answers	
	Yes	No
The new online survey is complicated to use than previous paper-pencil methodology	8,97%	91,03%
My impression is that manipulation with results in the new methodology is possible	3,85%	96,15%
My impression is that manipulation with results in the old methodology was possible	94,87%	5,13%
I freely expressed a poor opinion about teachers in the old survey methodology	6,41%	93,59%
I want reintroduction of the old survey methodology	8,97%	91,03%

If more students believed the new survey methodology then more respondents who give lower grades to teachers are included and certain number of teachers will have lower grades. This hypothesis is further validated by subsequent analysis of the survey results for all undergraduate degree programs and presented in Table 4. Before the analysis, the results are considered only for the courses that:

- surveyed with both methodologies, with more than ten students
- held the same teachers for both methodologies
- recorded higher response in new methodology

According to the results from the Table 4 it is clear that significant number of considered courses have achieved worse results for the case of increased response in online survey. Although such results cannot be statistically processed, Committee also found that students are given quite sharp and direct final comments (enabled as text box with free entry) about teacher's work, in the online survey.

TABLE IV. INDICATORS OF INCREASED RESPONSE OF STUDENTS WHO GIVE LOWER MARKS AFTER NEW SURVEY METHODOLOGY IMPLEMENTATION IN ACADEMIC YEAR 2013/2014

	I	II	III	IV
Study program	Total number of cosidered courses	Number of courses with a worse grade	**Index II/I [%]**	Number of courses with a better or the same grade
Management	23	17	**73,91%**	6
Administration	9	7	**77,78%**	2
Transport and traffic	13	9	**69,23%**	4
TOTAL	45	33	**73,33%**	12

V. CONCLUSION

According to the available and processed results presented in a case study, implemented online methodology has proved to be a better solution for teacher evaluation surveys at Polytechnic of Šibenik. Both previously established objectives have been met because survey response is increased and post processing of the survey results is simplified. The Committee for the implementation and processing of student surveys is obviously well assumed that student perception of the actual anonymity in survey was the problem in a past. The result was more sincere and freely answering to survey questions. Also online survey is not sufficient in itself and it was justified to pay attention on the implementation methodology details like real voting experience (authorized place where online survey can be filled on computer and avoiding of authorization token delivery to the students by e-mail message).

REFERENCES

[1] Polytechnic of Šibenik, "The quality assurance rules at Polytechnic of Šibenik", http://www.vus.hr/wp-content/uploads/2014/11/Pravilnik-o-sustavu-za-osiguravanje-kvalitete-VUS.pdf, 2013.

[2] Ministry of science, education and sports of the Republic of Croatia, ""Evaluation of Higher Education Institutions and Study Programmes", http://public.mzos.hr/Default.aspx?art=5867&sec=2510 , 2005.

[3] Agency for science and higher education ,"Quality Assurance", https://www.azvo.hr/en/quality/quality-assurance, 2009.

[4] The European Association for Quality Assurance in Higher Education (ENQA), "The Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG)", http://www.enqa.eu/index.php/home/esg, 2009.

[5] Marija J. Rosa, Claudia S. Sarrico, Alberto Amaral, "Implementing Quality Management Systems in Higher Education", Quality Assurance and Management, ISBN 978-953-51-0378-3, Publisher: InTech, 2012.

[6] Alan Blackwell, "Survey Methods in Technology and Physical Sciences", University of Cambridge, https://camtools.cam.ac.uk/wiki/site/e30faf26-bc0c-4533-acbc-cff4f9234e1b/survey%20methods.html, 2009.

[7] SurveyMonkey, "Making Surveys Anonymous", http://help.surveymonkey.com/articles/en_US/kb/How-do-I-make-surveys-anonymous, 2014.

[8] LimeSurvey, "LimeSurvey Manual", https://manual.limesurvey.org/Tokens, 2014.

[9] Google, "Google Forms: Get speedier survey results", http://static.googleusercontent.com/media/www.google.com/hr//support/enterprise/static/gapps/learningcenter/Google-Forms-Get-speedier-survey-results.pdf , 2013.

Implementation of Computer Programs in Visual Arts Education in Croatian Primary Schools

Zvjezdana Prohaska, BFA*, Silvija Benković, BFA**, Ivan Uroda, PhD***
* Education and Teacher Training Agency, Office Branch Rijeka, Rijeka, Croatia
** Primary School Zamet, Rijeka, Croatia
*** University of Rijeka, Faculty of Economics, Department of Computer Sciences, Rijeka, Croatia
zvjezdana.prohaska@azoo.hr, silvija.benkovic@skole.hr, ivan.uroda@gmail.com

Abstract - Based on two previous studies how teachers and pupils generally evaluate the introduction of ICT in Visual Arts education in primary schools in Croatia, this follow-up research investigates what methods and computer programs were actually introduced in classrooms.

A survey was carried out into how pupils evaluate pros and cons of using such programs and whether that improves their learning process in Visual Arts.

I. INTRODUCTION

Program for primary school subject Visual arts was defined in Croatian Curriculum for Elementary Schools by Ministry of Science, Education and Sports in the year 2006 and it emphasizes the importance of research-based learning and creation process. Nowadays, such learning process is impossible without the utilization of information and communication technologies as well as informatics.

Curriculum contents encourage visual curiosity, openness towards new experiences, sensitivity to problems as well as originality in problem solving by artistic and technical means. Also, they encourage the development of all degrees of divergent thinking in the areas of visual and artistic upbringing and education. [1]

By active inclusion of pupils in teaching and learning process that is also at the same time a communication processes, their creativity is encouraged which is in harmony with goals of all teachings as well as the teaching in the Visual Arts.

Contemporary, effective as well as efficient teaching and learning implies the utilization of various methods, by technology adopted means and aids, but it also demands flexible and creative teacher.

All of this needs some "new teachers" in a school because without them none of these innovations or changes can be realized successfully, anyway. That person, besides well-known professional and pedagogical knowledge, skills and abilities, should certainly possess mediator's dexterity between cultures and prosocial link between people (pupils). Concisely, such teachers should be the persons for intercultural dialogue. [2]

With teaching as well as learning strategy directed towards pupil and with cultural change of teaching practice teachers are becoming intermediaries in the beginning of the dialogue and in the communication process between pupils and contemporary learning and teaching means as well as aids.

Computers and didactically developed materials are directly reflecting on teaching and learning through rationalization and optimization of time spent in pupils' learning process. Such learning can be easier, faster and more productive what is in harmony with fast changes in contemporary society.

Life and work in contemporary society of fast changes and fierce competition requires new knowledge, skills abilities, values and attitudes i.e. new individual's competences that put the emphasis on the development of innovativeness, creativity, problem solving, development of critical thinking, enterprise, computer literacy as well as social and other competences. [3]

Accordingly, for active and creative relationship with surroundings as well as for development of visual perception and visual thinking competence, teaching and learning practice needs the foundation that will also lead to the formation of pupils' visual and artistic speech.

Teaching and learning in Visual Arts is based on literal observation, exploration of immediate surroundings and artwork, creative processing of information, research, encouragement of cognitive processes as well as on valorization. Visual and artistic speech is developed by the reorganization of internal relationships between visual signs and meanings in personal experience of creative processes. It is formed by observation, research, artistic creation and by application of knowledge on visual as well as on artistic elements and their relationships. [4]

Visual experience is transformed into relevant information that is appropriate for further processing, while utilization of computers in teaching and learning favors visual representation of data. Solving visual problems and encouraging visual curiosity is presumption for the activation of creative process in pupils.

Expression through creative process contributes to wholesome child's development: multiple intelligences, creative and symbolic thinking, expressive and practical forming abilities as well as personality. In addition, it encourages self-confidence, spontaneity and will for free research, it develops attention, concentration, sensitivity to measure and cyclic patterns plus it enhances the ability for self-expression and critical thinking. [1]

Former experiences of many Visual Arts teachers related to preparation and formation of didactic materials show that all teaching and learning themes and contents from the afore mentioned Curriculum are much more acceptable for pupils if they are presented in a way that is close to their everyday activities and to work on the computer.

By utilizing presentation software and computer program Paint, every teacher can very efficiently and quickly switch from one teaching stage to another, while not losing the essentials such as time articulation and class dynamics.

Equally, all available contents in Visual Arts that can be found on website created by Miroslav Huzjak's (who is both author and performer) (http://likovna-kultura.ufzg.hr, Likovna kultura – metodički centar, Faculty of Teacher Education in Zagreb) are suitable for utilization and formation with Paint.

There is almost no teaching unit where Paint could not be utilized with a purpose of easier explanation and understanding of keywords, processes and in pupil's practical work. Paint is fully utilizable in Visual Arts teaching and learning.

II. Implementation and Utilization of Computer Programs in Teaching and Learning Visual Arts

There are many ways in which computer program Paint can be implemented and utilized in teaching and learning Visual Arts. However, besides everything previously mentioned, there are 2 instances in which Paint is especially suitable:

1) it is useful for teacher during the preparation of presentations in presentation software and afterwards during their demonstrations

and

2) it is useful for pupil who can draw in this program during research phase or during practical work phase.

Examples in Paint and presentation software that teacher utilizes in various teaching units are presented in 4 subheadings below.

A. Example from 5th (fifth) grade

Plane – Static and dynamic planes

Key terms: static/dynamic, graphic, cardboard print, high print, matrix, impress, graphical sheet

Teacher can draw work phases in a process of creating cardboard images in order to demonstrate the procedure.

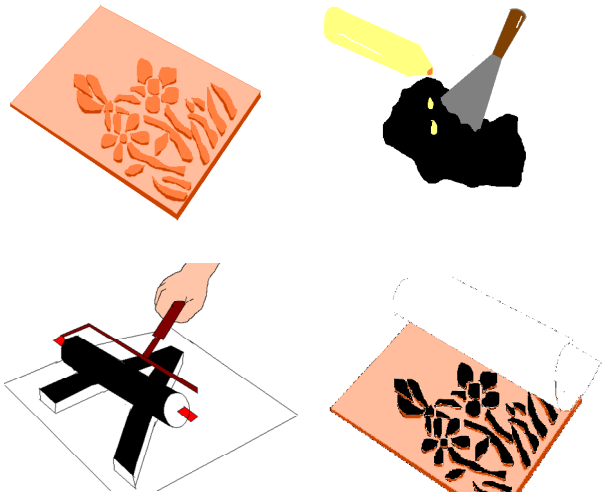


Figure 1. Graphical representation of example from 5th (fifth) grade

B. Example from 6th (sixth) grade

Color – Local color, color grading

Key terms: local color, harmony (harmony of colors), tonal gradation

Using prepared photography that is copied inside Paint in order to cut out the details in which tonal gradation is visible.

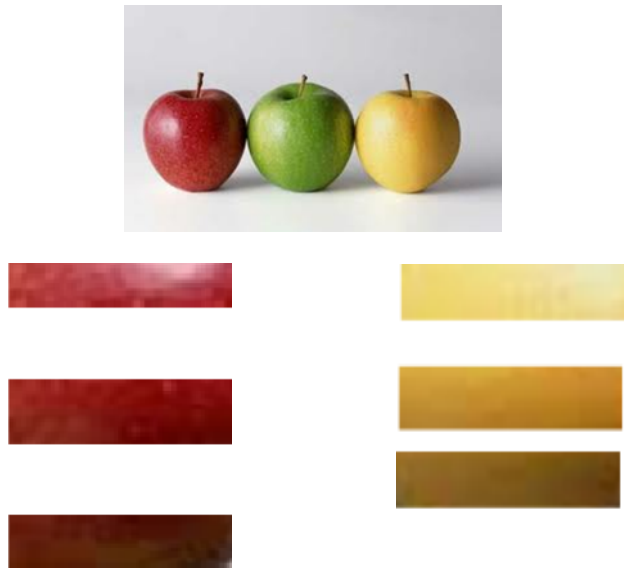


Figure 2. Graphical representation of example from 6th (sixth) grade

C. *Example from 7th (seventh) grade*

Linear perspective

Key terms: linear perspective, standpoint, eye relief, horizon, observation angle, field of vision

Using prepared photography that is copied inside Paint and drawing on the photography. This is pupil's work in the exploration of directions.

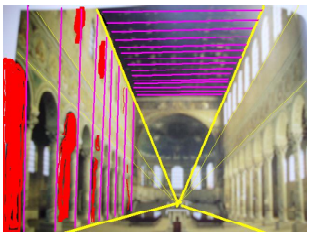


Figure 3. Graphical representation of example from 7th (seventh) grade

D. *Example from 8th (eighth) grade*

Harmony of shapes

Key terms: horizontal, vertical, diagonal, circular/oval, pyramidal composition

Using predefined geometrical shapes in Paint with a purpose to create layout so that various kinds of compositions can be formed.

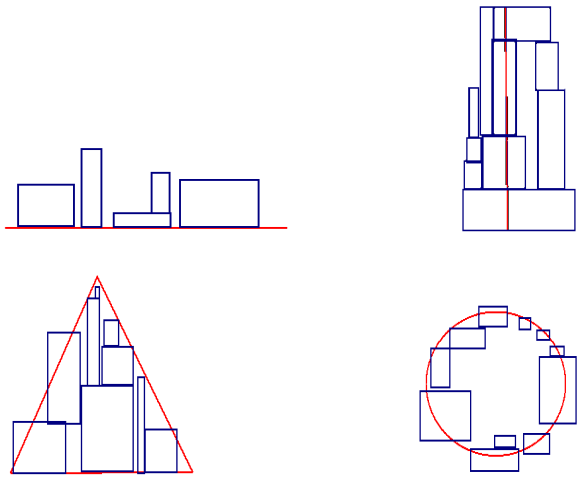


Figure 4. Graphical representation of example from 8th (eighth) grade

III. SURVEY INTO COMPUTER PROGRAMS IN TEACHING AND LEARNING VISUAL ARTS

Especially for research conducted in this paper the completely new and original survey was carried out. Briefly, in this comprehensive survey pupils were able to evaluate pros and cons of using computer programs in teaching and learning Visual Arts.

This survey was carried out in primary school Zamet in Rijeka, Croatia, Europe on a representative sample of pupils from 5th to 8th grade. Additionally, both male and female pupils were answering the same questions because this survey was carefully, specially and properly prepared as well as created in order to completely and adequately cover both genders.

Survey consisted from a total of 24 questions, whereas 1 question was considering pupils' gender, 1 question was considering pupils' grade, 20 questions were classical Likert scale questions considering pupils' evaluation (5 being the highest grade and 1 being the lowest grade), while 1 question was open type question considering pupils' evaluation.

Furthermore, questions considering pupils' gender and grades were analysed and represented as numbers and percentages, Likert scale questions considering pupils' evaluation were analysed and represented as 2 statistical measures i.e. mean and standard deviation, along with open type question considering pupils' evaluation that was represented in a form of the most numerous and the most representative groups of answers.

Notably, all tabular representations of data in answers 1-21 and 23-24 are not tables by Mipro standard, because data represented in them cannot be reduced to font size 8.

1. Pupils' gender

GENDER	M	F	Σ
n	39	53	92
%	42,39%	57,61%	100,00%

The majority of total number and percentage of pupils who participated in this survey were female.

2. Pupils' grade

GENDER	M		F		Σ	
GRADE	n	%	n	%	n	%
5th GRADE	12	44,44%	15	55,56%	27	100,00%
6th GRADE	14	46,67%	16	53,33%	30	100,00%
7th GRADE	7	38,89%	11	61,11%	18	100,00%
8th GRADE	6	35,29%	11	64,71%	17	100,00%

Even though number of pupils varies from grade to grade, it is evident that females represent majority in each of the classes that participated in this survey.

3. Evaluate your motivation to utilise Paint in class.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,62	1,32	3,67	1,17	3,64	1,25

Such motivation values are actually quite high for primarily school pupils. It is observable that males have slightly lower motivation, but that is probably because they consider other types of software (games, advanced programs, etc.) more challenging and therefore somewhat more motivating.

4. Evaluate your knowledge of Paint.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	4,35	1,00	3,96	1,04	4,16	1,02

These quite high values for both males and females are actually encouraging for teachers who plan to implement and utilize Paint in teaching and learning Visual Arts. Also, the lowest standard deviation implies that differences among the pupils are minimal and that is just another plus in the same regard.

5. Evaluate the value of Paint in class.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,66	1,49	3,56	1,11	3,61	1,30

The explanation for those values can be the same like in 3rd question (3.), but somewhat higher standard deviation among the males implies that differences in opinion are even more emphasized. Overall, this is also a very encouraging set of values in the context of this paper.

6. To which extent would you like Paint to be implemented in class?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,67	1,31	3,35	1,25	3,51	1,28

Similar to 3rd question (3.) and 5th question (5.) these values are definitely in favor of ideas promoted in this paper, despite the subtle differences in opinions as it is represent by standard deviation values.

7. Implementation of Paint in class would make the creation of Visual Arts exercises easier for me.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,55	1,34	3,41	1,48	3,48	1,41

With values similar to those in 6th question (6.), despite a tad higher standard deviation values for females this question essentially implies the same conclusion.

8. Implementation of Paint would make class more interesting.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,69	1,52	3,36	1,38	3,52	1,45

Instantly observable difference in mean values between males and females in this question only suggests what is traditionally well-known and accepted i.e. females are (in general) more inclined towards Visual Arts and therefore they need less incentive such as computer programs in this regard.

9. Evaluate the former level of Paint utilisation in class.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	2,81	1,69	2,63	1,41	2,72	1,55

Even though the former level of Paint utilisation in class was fairly low within drawn sample, this is the exact issue that this paper aims to address and, consequently, promote so that Visual Arts classes could become even more interesting and rewarding for the future generations.

10. How often does your teacher utilise Paint while teaching new subject matter?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	2,13	1,54	2,15	1,26	2,14	1,40

Almost the lowest mean values and rather high standard deviation values implies why this paper is aiming to be(come) one of the essential incentives for Visual Arts teachers in this respect.

11. To which extent would you like your teacher to utilise Paint for while teaching new subject matter?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,42	1,46	3,57	1,36	3,49	1,41

Larger than average mean values and standard deviation values in this question definitely show that pupils are interested in the proposed idea, but there are obviously some differences in opinions that should be sorted out.

12. How often does your teacher demonstrate way of applying computer programs for drawing and painting in class?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	2,55	1,60	3,06	1,59	2,81	1,60

Similar to 9th question (9.) and 10th (10.) question, values for this question shows that there is a lot of space for additions and improvements in this regard.

13. How often did you utilise Paint for freehand drawing in class?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	2,50	1,55	2,25	1,28	2,37	1,41

Even lower values than the ones in the previous question just increasingly emphasize everything previously written.

14. How often did you utilise predefined shapes from Paint (rectangles, ovals, stars, ….) in class?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	2,46	1,50	2,07	1,27	2,26	1,39

Predefined shapes are often very handy and convenient for pupils. Therefore, their usages should be properly taught as well as learned in Visual Arts to a larger extent. Rather high standard deviation values in males show that differences within the sample are rather high in this specific regard.

15. Painting on a computer screen is something that I consider interesting.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,50	1,36	2,92	1,51	3,21	1,43

Conclusion in this question is actually quite similar to the one on the 8th question (8.), but even greater difference in mean values and remarkably high standard deviation values imply that differences within this sample should be taken into account very carefully.

16. Paint enables me to enjoy drawing and painting.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,32	1,60	3,13	1,45	3,22	1,52

All values in this question are somewhere in the region between 8th question (8.) and 15th question (15.) what leads to conclusions in the specific region which is actually a combination of those two.

17. Utilisation of tools from Paint speeds up my process of producing artistic idea.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,39	1,07	3,06	1,27	3,22	1,17

With fairly larger than average mean values and low to average standard deviation values, this question actually leaves some valuable space for improvements in both teaching and learning.

18. Color palette in Paint provides me with wider selection of hues than the ones that we physically mix on a palette.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,65	1,34	3,45	1,36	3,55	1,35

Higher than average mean values and standard deviation values show that pupils essentially appreciate what Paint has to offer them. However, once again, differences in opinions should be taken seriously and, therefore, approached carefully.

19. By adding text to a picture we would enrich the production of ideas in graphical design.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,53	1,50	3,31	1,30	3,42	1,40

With only slightly lower values than in the previous question i.e. 18th question (18.), explanation is essentially the same, but it can be concluded that pupils value Paint's text capabilities as, more or less, mediocre ones.

20. I consider font selection as acceptable in production of graphical design exercises.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,41	1,24	3,15	1,26	3,28	1,25

In accordance with previous 2 questions i.e. 18th question (18.) and 19th question (19.), Paint's font selection is approximately in the same rank according to these pupils' opinions.

21. How do you evaluate teacher's support in utilisation of Paint for doing your homeworks?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,29	1,61	2,46	1,60	2,87	1,60

Surprisingly low mean values and almost the highest standard deviation values imply that there is a lot of space for improvements, equalizations and innovation. Moreover, because homeworks are the important part of learning process in primary schools, these values are even more significant and motivating.

22. What are the most common problems that happened to you during the utilisation of Paint in class?

The most numerous and the most representative groups of answers (formed with the help of keywords) that were given by pupils are presented in the list below.

NO ANSWER (EMPTY BOX)	*29 out of 92*	*= 31,53%*
no problems	*10 out of 92*	*= 10,87%*
when whole screen is painted	*8 out of 92*	*= 8,70%*
when something is deleted	*3 out of 92*	*= 3,26%*
color spillover	*2 out of 92*	*= 2,17%*
eraser is too small	*2 out of 92*	*= 2,17%*
hard to draw straight lines	*2 out of 92*	*= 2,17%*
when mouse stops	*2 out of 92*	*= 2,17%*
OTHERS (1 ANSWER)	*34 out of 92*	*= 36,96%*

23. By utilising Paint Visual Arts problems have become clearer to me.

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	2,71	1,66	2,79	1,52	2,75	1,59

Significantly lower mean values and pretty close standard deviation values in comparison with the 21st question (21.) signify, once again, room for desired, considerable and genuine improvements.

24. Would you participate in further researches related to the introduction of Paint in class?

GRADE	5th to 8th		5th to 8th		5th to 8th	
MEAN & ST. DEV	MEAN	ST. DEV.	MEAN	ST. DEV.	MEAN	ST. DEV.
VALUES	3,40	1,57	2,98	1,44	3,19	1,51

Such values were pretty much expected, especially because it is not sensible to expect from a sample of elementary school pupils to be seriously and greatly interested in participation in further researches. It is mainly because at least a certain number of them cannot be fully aware of the importance and the implications that research like this one can have on future generations, particularly if it is properly disseminated and continued. However, significantly lower interest by females is something that was somewhat unexpected.

General remarks regarding the conducted survey:

- the lowest mean value was 2,07
- the highest mean value was 4,35
- the average mean value was 3,21
- the lowest standard deviation value was 1,00
- the highest deviation value was 1,69
- the average deviation value was 1,35
- this survey and its encouraging results will definitely be the incentives for a number of new surveys by the same authors in the coming years

IV. CONCLUSION

Computer program Paint is fully utilisable for teaching and learning in Visual Arts. There are literarily no teaching units where Paint could be utilised with a purpose of easier explanation and understating of key terms and processes of practical work, but also in the practical work itself.

Pupils occasionally utilise Paint i.e. when they express a wish for it or if they estimate that they could better realise their idea in Paint. Paint has also proven to be a good program for pupils with difficulties, especially the ones with graphomotoric disorders because they feel safer (when working) in Paint, whereas they usually successfully realise the (given) theme (in the class). Paint is also good for hyperactive pupils because this program can calm them and it requires concentration as well as persistency from them.

In addition to all of these (and numerous other) cognitions, observations and experiences throughout the years, authors have recently carried out completely new and original survey 24-question survey in primary school Zamet in Rijeka, Croatia, Europe on 92 pupils from 5th to 8th grade. Results of this survey, which were elaborated in detail under the separate heading of this paper, were absolutely and distinctly positive, encouraging and motivating for authors. It is so, mainly because they have definitely proven that computer program Paint alone and in the combination with other computer programs is something that pupils value, appreciate and (slowly, but surely) require as the welcome novelty and addition in their teaching and learning process.

Finally, this very survey and its afore mentioned results are definitely the incentive for a number of new surveys by the same authors in the coming years. That way authors will able to gather more data on issues presented in this paper as well as to disseminate, direct and expand this (and other related) research in directions that will be of most interest and benefit as well as for prosperity of future generations.

REFERENCES

[1] Croatian Curriculum for Elementary Schools, Republic of Croatia, Ministry of Science, Education and Sport, Zagreb, 2006, pp. 5

[2] Previšić, V., Interkulturna obzorja suvremene škole, Zbornik radova 15. Križevačkih pedagoških dana, Croatian Literary-Educational Association, Bjelovar-Križevci, 2009, pp. 26

[3] Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje, Republic of Croatia, Ministry of Science, Education and Sport, Zagreb, 2010, pp. 11

[4] Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje, Republic of Croatia, Ministry of Science, Education and Sport, Zagreb, 2010, pp. 152

E-Learning of Students in Course „Fundamentals of Engineering Design“ Using the System of Evaluation and Maintaining the Quality of Education and Scientific Research

L. I. Baltova
Technical University of Sofia/ Department of Fundamentals and Technical Means of Design, Sofia, Bulgaria
lbaltova@tu-sofia.bg

Abstract - This paper presents the organization of education in the course “Fundamentals of Engineering Design” in the Department of Fundamentals and Technical Means of Design, the Faculty of Mechanical Engineering at the Technical University of Sofia.

In the paper are shown helpful applications developed in the Department which presents good practices for training students using System of Evaluation and Maintaining the Quality of Education and Scientific Research as measures to be applied to increase the knowledge and motivation and reducing the impact of risks identified during the training of the students.

I. INTRODUCTION

The Department of Fundamentals and Technical Means of Design, the Faculty of Mechanical Engineering at the Technical University of Sofia has developed varieties of assignments and test-papers that are used to control the level of acquired of knowledge in the training of First-year students at the course “Fundamentals of Engineering Design”.

According to Baltova (2013) [7] the potential risks are analyzed that were identified during training the students and measures are shown to reduce their influence: a risk of loss of interest in the subject, a risk the information presented in the tasks to be out of the date, a risk of incorrect terms of the tasks, a risk of inaccuracy in the difficulty of the tasks, a risk of incorrectness in the methodical arrangement of tasks, a risk of incorrectness of deadlines for implementation of tasks, a risk of lack of motivation of teachers and students, a risk of unsuccessful realization of the students in their specialty after graduation. In order the students to be able to increase their knowledge and motivation the annual assessment of each student is composed by summing the points obtained from all course tasks and tests made during the two semesters of the academic year. Each semester they make four course tasks and four control work and each task is described in [8]:

At First semester – the first is Electrical Circuits; the second is Heuristics I (develop Morphological Analysis and Synthesis, the Focus Objects, Denial of Signs, Empathy – the goal is to find original ideas to change the appearance, to add an additional function or to propose another realization of a function of the object); the third is Design of Parts and Mechanical Assemblies (design of bracket and fundament for internal and external assembly of signal lamp in the technical environment); the fourth is Design of Electrical Compounds for Volumetric Assembly (design of wire bundle connecting the devices in a car, according to the scheme of the compounds and assignment instructions).

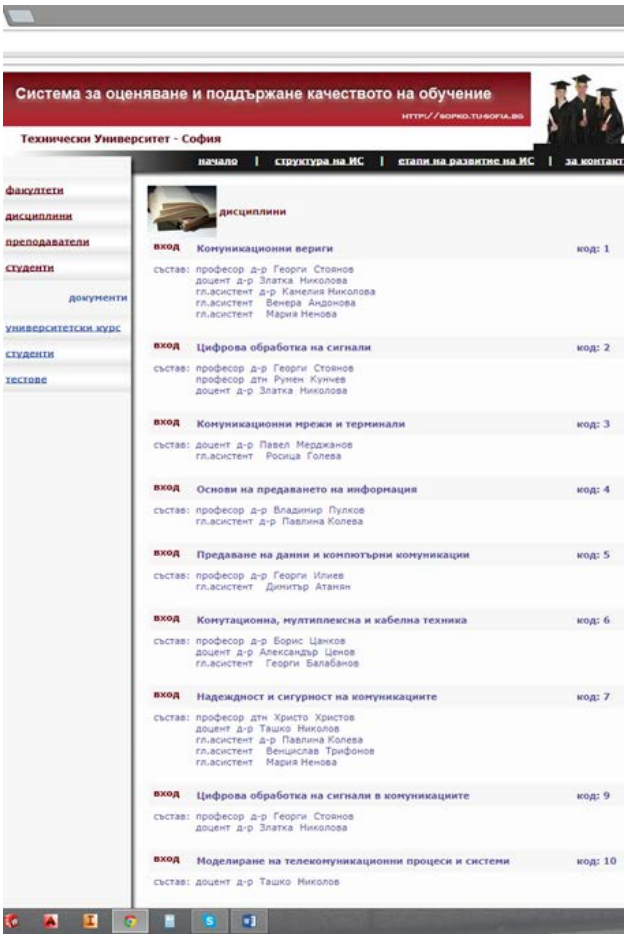


Figure 1. Login the System SEMQESR

At Second semester the students make next course tasks: the first of them is Engineering Analysis of Device (Structural analysis – the outlines of assign part and interaction with neighbouring parts of the device are determined; Constructive analysis – the graphic symbol and lettering in electrical scheme are determined, engineering effects, order of assembling, possible reasons for the failure of the device; Functional-Value analysis – the major and basic functions are determined, the value of parts that it participate in respective function as it looking for expensive function with greater costs than the significance and make a constructive modification in order to improve the quality and reducing the price of this function and of the electrical device; Dimensional analysis – the type of assemblies specified in the drawings of the General type are analyzed and the limit deviations of the dimensions of the parts involved in the assemblies presented graphically are determined and making a choice of smoothly cylindrical compound); the second is Heuristics II (Total analogy, Synectica, Modification of an Idea); the third is Design of Electro-Technical Devices (design of assembly units of relay – group of contacts, an electromagnet, base and lid); the fourth is Design of electronic device (at a specified principle electrical scheme design Printed Circuit Board - PCB and the box is designed for fixing and protecting the PCB).

In the Department have been developed variants of assignments for control work used at the end of each semester to review the acquire knowledge of students gained from the completion of course tasks.

The students redraw drawing of a part that is given by the assistant using AutoCAD. They draw assembling units consist of two parts by nature that are given by the assistant and they have to show the drawing of the assembled unit using SolidWorks. They draw PCB-design from electrical scheme that is given by the assistant using CADStar. They make graphical handmade control work. They answer in written to theoretical questions on the topics of lectures and in one of the tests we used for System of Evaluation and Maintaining the Quality of Education and Scientific Research (SEMQESR). This system is on the web site of the Technical University of Sofia.

In [1] are shown the tests developed in the Department are: 1. Fundamental of design; 2. Engineering analysis of the product; 3. Design of electrical compounds for printed circuit board (PCB Design); 4. Electrical schemes; 5. Mechanical assembly in the design on a technical object. Draw of designed technical object; 6. Dimension, roughness and threading in the design of technical object; 7. Design of parts, assemblies and electrical compounds for volumetric installation.

Assistants have developed over 48 questions for each test. When a student starts test the system generate 12 questions randomly and change the locations of the correct answers, which prevents mechanical learning of correct answers and stimulate thinking and rapid decision-making based on accumulated knowledge.

Each test can be used as: a trainer without timing (in the initial solving students seek answers to particular questions in lectures, guide and textbook that takes time); a trainer with timing (the control work is examination test which is with time registration necessary the students to have preliminary preparation to be able to respond quickly); examiner test (done in the computer room and the assistant generates a special code).

Figure 2. Login for student registration

According to E. Gerganov [4] the tests should cover the topics that are taught during the semester in tasks and lecture. The questions contain graphic and/or descriptive sections and four answers only one of them is correct. From a methodological point of view it is very important in the process of training the student to be able quickly to choose the correct answers and for that the response time is in conformity with the difficulty and scope of the explanatory part of the question. Students receive points for each correct answer, and in the end the system shows the final assessment depending on the ratio between a number of questions with correct answers compared to the total number of questions: for Excellent – 90 ÷ 120 points, for Very good – 72 ÷89 points, for Good – 54 ÷71 points, for Satisfied – 36 ÷ 53 points and less - Fail.

The beginning of work in SEMQESR is a click to the button “entry” with the specific number 65 for the course shows to fig. 1.

Students must be registered for the course by clicking the "registration" button to access the tests shown in fig. 2. In fig. 3 students fill in their personal data. After that the registered students must be confirmed by the assistant leading laboratory in order to activate their registration and set the period of activity of the registration as shown in fig.4.

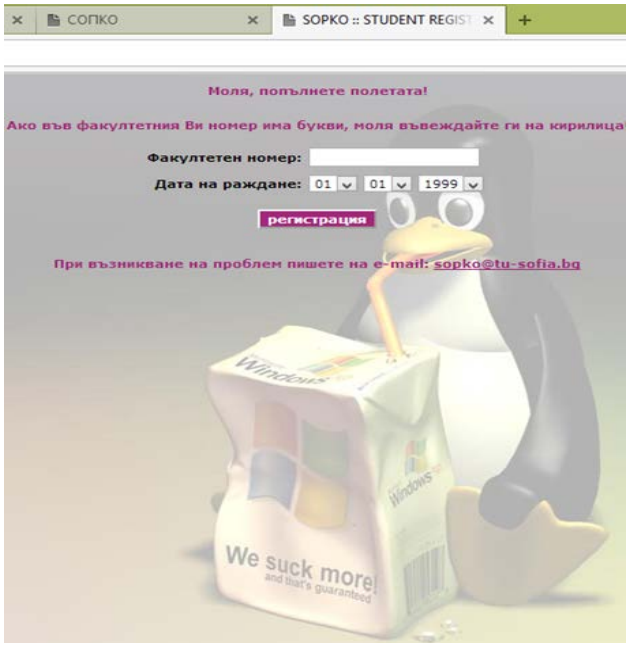


Figure 3. Process of student registration

After the successful start each registered student has access to the tests and other useful additions according to R. Pranchov, H. Hristov (2006) [3] and [2] by which we aim to help the students and to increase their knowledge and motivation: “Most Common Mistakes” in course tasks are shown in fig. 5; “Frequently Asked Questions” in the submission of each course task are shown in fig. 6; the questionnaires for each semester including the topics taught in the lectures are shown in fig. 7; consultation hours and teachers e-mail; dates of examinations; etc..

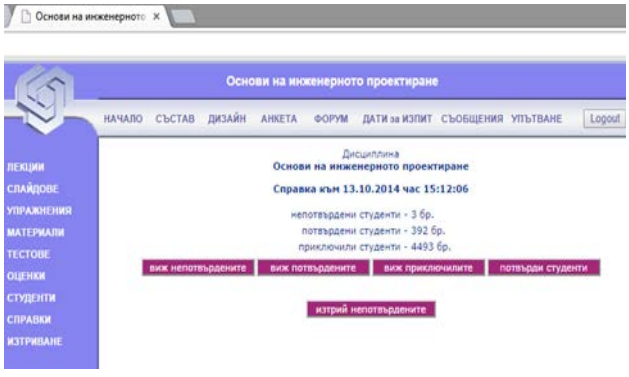


Figure 4. Confirm of students

The annex "Most Common Mistakes" describes errors that students make in the completion of each of the course task most frequently are shown in fig. 5. They are published in the paragraph "Materials" in SEMQESR.

The utility is expressed in that the questions of the students related to the performance of the course tasks became more specific and shown drafts became more correct and we achieved an increase in knowledge of students with.

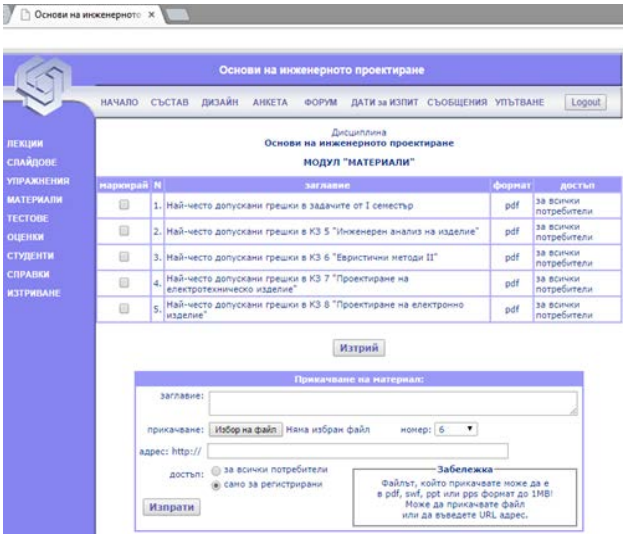


Figure 5. “Most Common Mistakes”

Students could check draft of course task before showing it during assistants’ consultation hours. Students must correctly eliminate the errors that have been discussed with the assistant and submitted the course task within the due time with maximum points if they fail to make the proper correction or does not respect the time of submission obtained points are reduced. Submission of course task occurs during consultation hours of assistants and then the manner of the implementation of the course discussed with each student.

In the annex "Frequently Asked Questions" are described the questions that assistant asks students in the submission of each course task and they are shown in fig. 6. They are published in the "Exercises" in SEMQESR.

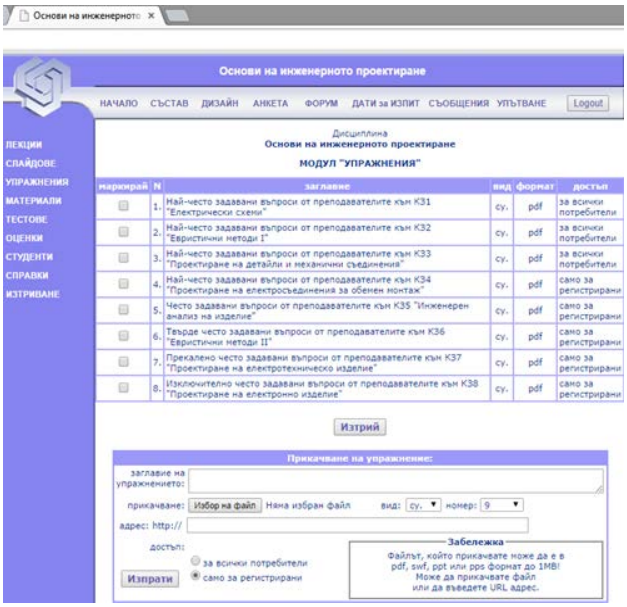


Figure 6. “Frequently Asked Questions”

This application focuses on the most important questions of the course tasks that help students to prepare for defense of each of them for the submit of course task during the consultation hours of the assistants if they fail to respond to the questions obtained points are reduced.

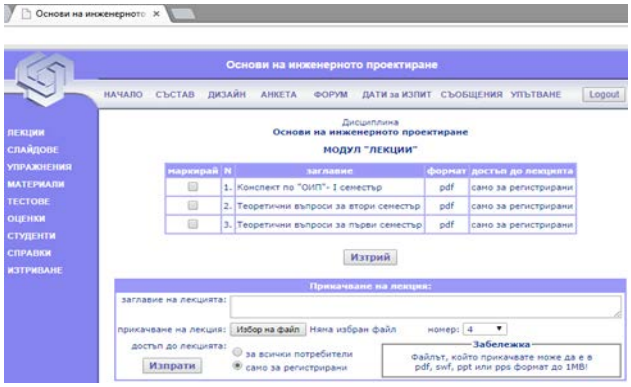


Figure 7. Questionnaires with the topics on the lectures

Fig. 8 shows the means how the introduction of a question from a respective test by teachers is organized and fig. 9 shows how the same question of the same test looks when it is resolved by the student.

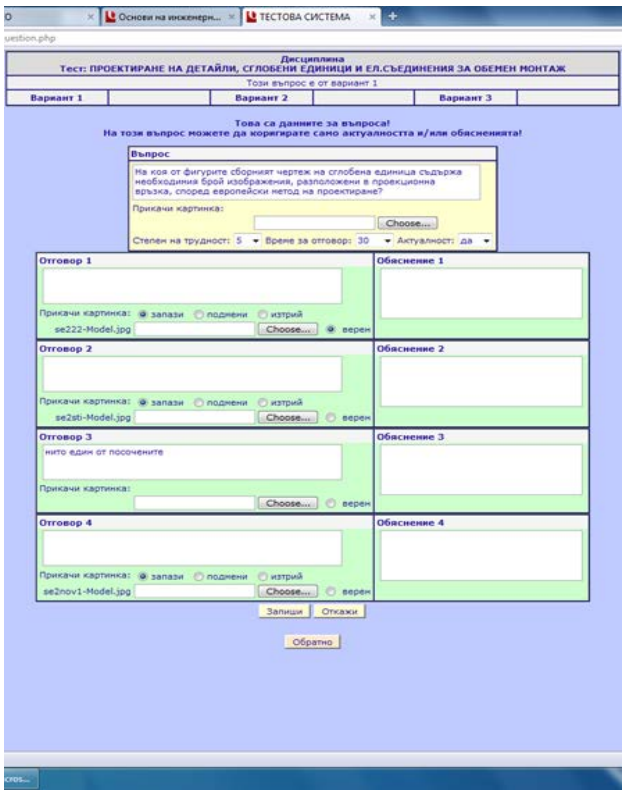


Figure 8. Developing a question by the teacher

In fig. 10 is shown the possibility to change the content of question of the test in order to clarify and update the explanatory part or change drawing in the graphic part to change the difficulty. If necessary corrections they have to be made at the beginning of the semester and any corrections are inadmissible after activation of test.

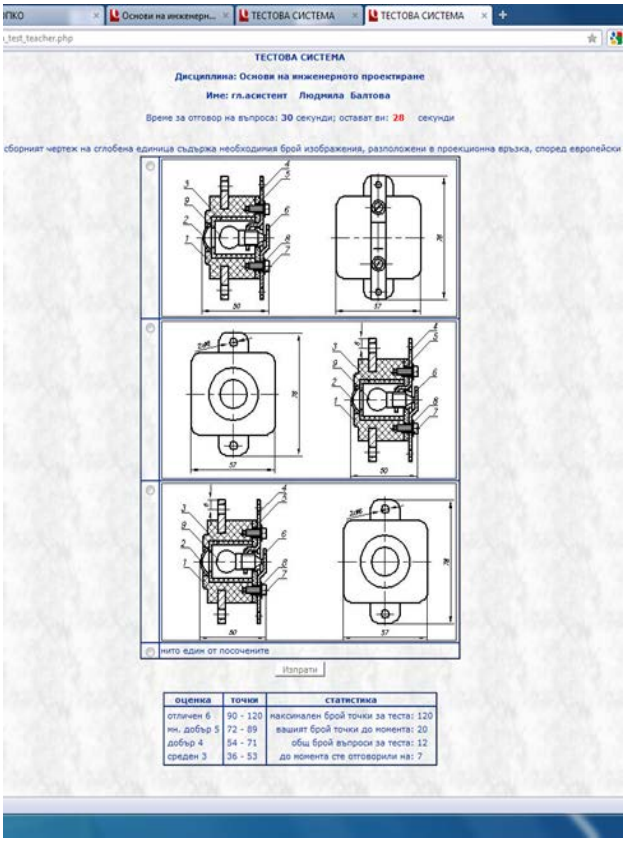


Figure 9. Solve a question of student

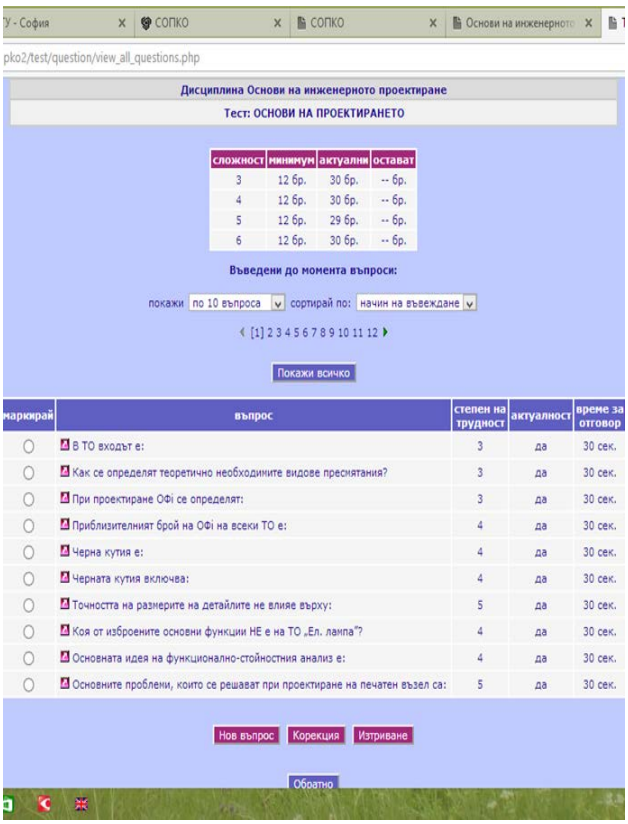


Figure 10. Ability to change the content of question

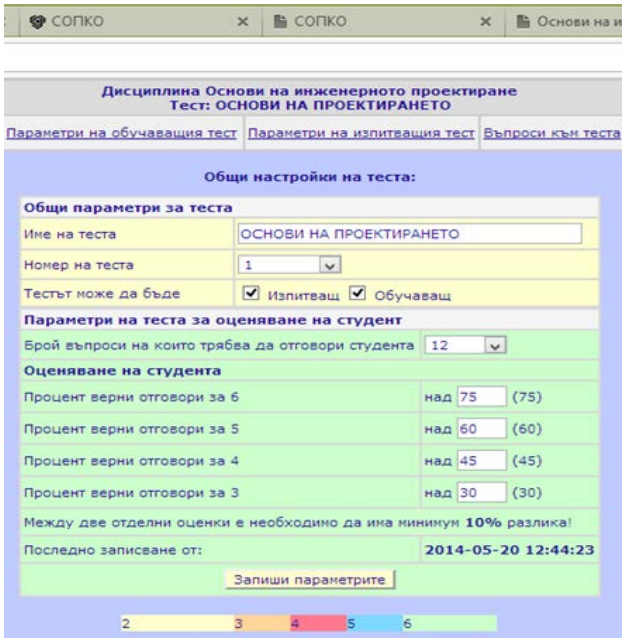


Figure 11. General settings on the test

In [5] and [6] are presented the conditions for the formation of graphic and descriptive parts of the questions in different courses respectively on “Fundamentals of Design and CAD” for mechanical Engineer and the “Basics of Design” for the specialty 'Industrial Engineering'. It is shown how to present erroneous graphical part so that students can quickly identify errors and quickly click on the correct answers. The results of the tests are shown and analyzed there.

In the general settings of the test shown in fig. 11 it is possible the type of test to be set: teaching test without timing, teaching test with time registration or test examiner.

Another possibility is to set which tests to be active during the relevant semester as within the first semester are active the four of them and within the second semester are active other three. The activity of each of the tests is related to the course tasks that students made during the semester.

Other possibility is in assessing the student in the system to set different percentages to the number of correct answers to various assessments and it is necessary to have minimum 10% difference between two separate assessments: for Excellent – over 75%, for Very good – over 60%, for Good – over 45%, for Satisfied – over 30% and less than 30% - Fail.

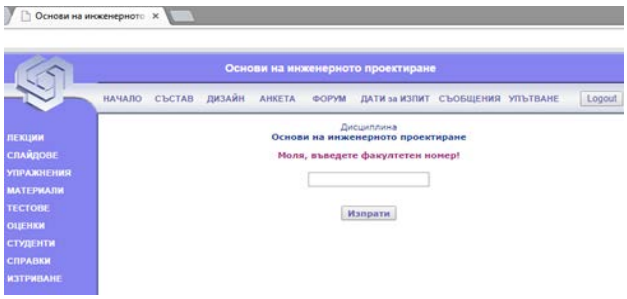


Figure 12. Statistical information of the results

At any time each registered student and assistant could make reference about the results in solving tests by introducing student’s faculty number shown in fig. 12.

II. Conclusion

This paper presents the organization of education in the subject “Fundamentals of Engineering Design” in the Department of Fundamentals and Technical Means of Design at the Faculty of Mechanical Engineering at the Technical University of Sofia.

The paper shows the tests and supporting annexes “Most Common Mistakes” in course tasks and “Most Frequently Asked Questions” in the submission of each course task developed at the Department representing the best practices for teaching students using the System of Evaluation and Maintaining the Quality of Education and Scientific Research as measures to be applied to reduce the impact of the risks identified during the training of students and to increase the knowledge and motivation.

References

[1] www.tu-sofia.bg System of Evaluation and Maintaining the Quality of Education and Scientific Research (SEMQESR) at the Technical University of Sofia, 2009.

[2] www.tu-sofia.bg H. Hristov, “A view on common requirements to Systems of Evaluation and Maintaining the Quality of Education and Scientific Research at the Technical University of Sofia”, Quality of higher education, 2006.

[3] www.tu-sofia.bg R. Pranchov, H. Hristov, “Scope of Systems of Evaluation and Maintaining the Quality of Education and Scientific Research at the Technical University of Sofia”, Quality of higher education, 2006.

[4] www.tu-sofia.bg E. Gerganov, “Test and expert assessment”, The assessment center, New Bulgarian University.

[5] G. Dinev, R. Angelova, V. Stancheva, L. Baltova, “Test control in design documentation on the subject Fundamentals of Design and CAD”, 8th International Conference on Advanced Manufacturing Operations, 18-20 June 2008, Kranevo, pp. 391-394, Faculty of Machine Technology, Technical University of Sofia, Bulgaria; DMT Product Sofia, ISSN 1313-4264.

[6] G. Dinev, E. Chalakova, “Test control in the Basics of Design for the specialty ‘Industrial Engineering’”, Bulgarian Journal for Engineering Design, Publisher Mechanical Engineering Faculty, Technical University-Sofia, issue №7, February 2011, pp. 39-42, ISSN 1313-7530.

[7] L. Baltova, “A practical application of the system of assessment maintenance of the quality of education in subject “Fundamentals of Engineering Design””, 8th Research/Expert Conference with International Participations “QUALITY 2013”, Neum, B&H, June 06-08, 2013, Year VIII, №1, pp. 367-372, Publisher: University of Zenica, Faculty of Mechanical Engineering in Zenica, Printed by Stamparija Fojnica, ISSN 1512-9268.

[8] N. Ganeva, M. Leparov, G. Stanchev, “Fundamentals of Engineering Design”, Student’s Handbook, Technical University of Sofia, 2013, ISBN 978-954-334-059-0.

The developments in mobile learning and its application in the higher education including libraries

Radovan Vrana
Department for information and communication sciences,
Faculty of humanities and social sciences, University of Zagreb, Zagreb, Croatia
rvrana@ffzg.hr

Abstract – As technological devices become more present in education, more researchers and educators are looking into different aspects of technology enhanced learning. Most recently, focus in technology enhanced learning is put on use of mobile devices i.e. mobile learning or m-learning. Mobile learning promotes concept of anytime anywhere learning which is important for students who frequently change places of learning and need access to the learning material as they move from one place to another. Such concept of learning is possible because more and more students have some type of a mobile device and mobile devices are mostly cheap, portable and flexible, have no start-up time, and require virtually no maintenance. While the mobile technology seems to be very attractive to students and usable in the learning process, there are also problems which burden further development of mobile learning or at least they are not helping its development: technological problems and obsolesce, lack of digital content made for use on mobile devices, lack of training of proper use of mobile devices in the learning process etc. Most of these and similar problems could be overcome if universities put more effort in adoption of mobile technologies among students and teaching staff.

I. INTRODUCTION

"Mobile and communication technologies are increasingly impacting the lives of individuals and are a major factor in helping reshape our culture and societies" [1]. As technological devices become more present in education [2],[3] more researchers are looking into different aspects of technology enhanced learning. Most recently, focus in technology enhanced learning is put on use of mobile devices i.e. mobile learning or m-learning. Mobile learning is learning delivered or supported solely or mainly by mobile technologies [4]. Mobile technologies enable people to be mobile and yet accessible [5].

The mobile technology i.e. mobile devices "(…) are generally defined as units small enough to carry around in a pocket, falling into the categories of PDAs (personal digital assistants), mobile phones, and personal media players" [2]. Mobile learning and mobile devices are directly connected as mobile learning is hardly possible without use of some kind of a mobile device. Therefore, mobile learning if often described as "the intersection of mobile computing and e-learning which offers: accessible resources wherever you are, strong search capabilities, rich interaction, powerful support for effective learning, and performance-based assessment, e-learning independent of location in time or space" [6] i.e. anytime and anywhere learning. Anytime anywhere learning is central to the concept of mobile learning. For Martin and Ertzberger [7] mobile devices open "(…) the door for a new kind of learning and performance support in the field, providing anytime and anywhere access to information, processes, and communication". For Tatar, Roschelle, Vahey and Penuel [8] "Mobile learning holds the promise of offering frequent, integral access to applications that support learning anywhere, anytime". Generally speaking, mobile learning is important in education for achieving academic goals while its final goal is helping the development of the knowledge society [3].

In the education sector, mobile technologies are becoming crucial because they "(…) can provide frequent and comprehensive access to systems and applications that support formal and informal learning" [9]. Mobile learning is present in higher education institutions as mobile devices have become ubiquitous at universities [10]. This widespread use of mobile devices has become possible because they are cheap, portable and flexible, have no start-up time, and require virtually no maintenance [8]. As a result, use of mobile phones and other mobile devices that allow data collection anywhere and anytime has become a routine [1].

Based on these facts and the current educational practice in higher education institutions, mobile learning seems to be generally attractive, especially to younger generations of students during their daily education related activities, it remains to be seen how strong it will influence programs of study and individual courses at universities. This paper will discuss potentials of use of mobile learning in higher education based on characteristics of mobile learning.

II. MOBILE LEARNING

Mobile learning is defined as "any sort of learning that happens when the learner is not in a fixed, predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies" [11]. According to Brown and Metcalf [12] "mobile learning (commonly referred to as m-learning) as all "knowledge in the hand." It includes the use of mobile/handheld devices to perform any of the following: deliver education/learning; foster communications / collaboration; conduct assessments / evaluations; provide access to performance support/knowledge. Mobile learning uses the benefits of smaller and faster computing which enables learning anywhere and anytime [13].

Mobile learning comes in several "flavors". In their chapter on learning design with mobile and wireless technologies, Kukulska-Hulme and Traxler [14] enumerated categories of mobile learning with description for each category:

- "Technology-driven mobile learning – some specific technological innovation is deployed in an academic setting to demonstrate technical feasibility and pedagogic possibility
- Miniature but portable e-learning – mobile, wireless and handheld technologies are used to re-enact approaches and solutions already used in 'conventional' e-learning, perhaps porting some e-learning technology such as a VLE to these technologies or perhaps merely using mobile technologies as flexible replacements for static desktop technologies
- Connected classroom learning – the same technologies are used in classroom settings to support collaborative learning, perhaps connected to other classroom technologies such as interactive whiteboards
- Informal, personalized, situated mobile learning - the same technologies are enhanced with additional functionality, for example location-awareness or video-capture, and deployed to deliver educational experiences that would otherwise be difficult or impossible
- Mobile training/performance support - the technologies are used to improve the productivity and efficiency of mobile workers by delivering information and support just-in-time and in context for their immediate priorities
- Remote/rural/development mobile learning - the technologies are used to address environmental and infrastructural challenges to delivering and supporting education where "conventional" e-learning technologies would fail, often troubling accepted developmental or evolutionary paradigms".

Cheon, Lee, Crooks and Song [10] offered another view on mobile learning in which mobile learning facilitates four type of learning: individualized learning, situated learning, collaborative learning, and informal learning, which is excellent for a rather new technology. Each type of learning can benefit from use of mobile technologies; however, mobile technologies are not available to all students in all type of learning, so the actual use of mobile technologies depends of individual student's financial abilities to acquire a mobile device and knowledge to use it properly for learning. Finally, Lippincot [15] investigated the influence of mobile technologies on higher education and suggested that mobile devices with compelling new features and wireless connectivity are almost ubiquitously available and that we may be on the verge of a revolutionary phase of mobile device impact on higher education and libraries.

III. BENEFITS OF MOBILE LEARNING

Popularity of mobile devices and increased use and reliance on broadband networks have marked the beginning of a new era of education. Everyone wants to benefit from fast access to the internet and digital content available online for learning, both formal and informal. Some of the most obvious and direct benefits of mobile learning are equal-opportunity access, ubiquitous connectivity, multigenerational users and uses, expanded services for mobile workers and enhanced access to services for mobile learners" [16]. As mobile learning is becoming more popular, its benefits become more visible: cost savings, ubiquitous communications, study aids, and location-based services [10]. Mobile learning offers expansion in terms of the learners' experiences with an emphasis on device ownership, informality, movement and context [14]. Additionally, Klopfer, Squire and Jenkins [17] singled out features of mobile devices that can further enhance quality of mobile learning:

- Portability – students can take the mobile device to different sites and move around within a site
- Social interactivity - students can exchange data and collaborate with other students face to face
- Context sensitivity - students can gather data unique to their current location, environment, and time, including both real and simulated data
- Connectivity - students can use mobile devices to connect their handhelds to data collection devices, other handhelds, and to a common network that creates a true shared environment
- Individuality - mobile devices can provide students unique opportunity of scaffolding that is customized to the individual's path of investigation.

"Mobile learning can enhance and support more traditional learning modes, making it more portable and accessible. Mobile devices can also serve as powerful data collection tools and cab also facilitate the capture of user created content [12]. Mobile learning depends of access to

digital content. This access is made possible by wireless mobile and wifi networks at universities. However, it should be pointed out that the quality of digital content used in education can vary and in such cases university libraries could play an important role of facilitator to quality content for learning.

By using their mobile devices, students can learn at any time and at any place; they can learn as if they are in a real classroom environment; the learning environment is equipped with sensors and other tiny devices embedded in learning objects; mobile devices are capable to identify the current context and other learning objects of the students so the students are provided with services according to his/her context and situation; the communication between students and instructors has become more ubiquitous [1].

Mobile learning supports individual learning as well as collaboration learning [18] which makes it applicable in different learning environments and situations. Mobile devices can be used in laboratories to gather and control data; in field trips mobile devices can be used to collect pictorial and textual data; in distance learning mobile devices support the delivery of synchronous and asynchronous learning while in informal settings the devices support incidental and accidental learning [18]. The current versatility of use of mobile devices suggests that possibilities of their use will only increase in near future. These are the most observable benefits at the moment and this list could be expanded with new features of new generations of mobile devices.

IV. DOWNSIDES OF MOBILE LEARNING

In addition to benefits, mobile learning has met some obstacles. From the technological point of view, mobile devices are different according to their specifications and usability (limited memory, screen size, battery life, storage space, and backlighting making use over long periods uncomfortable, formats such as Adobe Flash are not always supported and PDF or graphics may not display clearly on the screens, input methods are often not user-friendly), the lifecycle of the devices (with each new mobile device, users must continuously learn how to use it and may not have time to exploit the full functionality of the particular device), the diversity and lack of standards in this area, cost could be an issue for users, especially because of the rapidly changing technology and short lifecycle of the devices [2],[19].

Some authors like Keough [20] think that there are serious reasons why mobile learning will not work: it is technology driven; we know too little about what mobile devices are used for; we know too little about learning relationships/networks or the transactional analysis of mobile relationships; it will not change entrenched institutionalized education models; it relies on nascent consumer technology: mobile devices are inherently dissatisfying by never quite meeting every promised need for the consumer; there are no standards to overcome cultural differences; there is lack a bobigogy i.e. teaching and learning models.

As it was previously mentioned, mobile learning depends on quality information resources. Publishers and libraries should provide access to information and support all learners in use of educational and other types of information resources regardless of differences of mobile devices they own and differences in data formats or how much technical support students get in their use of such devices [2]. Part of the problems with content can be solved by libraries and their digital collections. Many libraries already have Web sites adapted for access by mobile devices so the digital content in library collections could be used in the learning process.

V. LIBRARIES AND MOBILE LEARNING

In the system of higher education, libraries have an important role of facilitator to printed and digital knowledge necessary for research and teaching. More frequently than ever, libraries are adopting mobile technologies to provide new and innovative services, as well as introducing ways for users to incorporate library services into their daily lives [21]. Today, libraries offer a variety of services. Some of them include: including mobile library websites and MOPACs (Mobile OPACs), mobile collections, mobile library instruction, mobile databases, mobile library tours, mobile learning, library SMS notifications, mobile library circulation, QR codes, access to services etc. [21]. Furthermore, mobile devices can be used in libraries for the following activities: voice and video calling, sending and receiving e-mail, SMS text messaging, searching the internet, searching databases of scholarly information, organizing citations, accessing a course management system, reading or listening to books and articles, taking photos, playing videos, making videos, setting an alarm clock, using a GPS navigation system and playing games, reading, watching, listening to, and producing digital content that will have the most impact on libraries [15]. Some of these activities are more related to the learning process than other. In libraries which are part of higher education institutions students can use rich digital resources in digital libraries for instruction, assessment, and professional development [22]. In combination with mobile technologies, students can study when they want to or need to and transport their learning materials [22] with them. Mobile technologies help students not only in using the already created digital content but as an instrument for creation of digital content [15].

With proliferation of use of mobile devices "(…) libraries must now decide how much time and funding they should devote to the development and support of academic library resources in formats that can be accessed on such devices" [2]. Once again, libraries can prove the values of their services because they have already traditionally served as a public good, providing resources and services to all, including those who could not afford to purchase some types of content or services on their own [15]. In case of mobile learning and other services related to use of mobile technologies, libraries are unable to equip every library user with and adequate device. It is therefore

highly unlikely that libraries would buy smartphones to its users, but many of them do offer tablets for use in library premises accompanied by the digital content. This is a major step forward towards popularization of use of mobile devices in libraries as well as use of digital content. Libraries would be hardly useful without librarians. The role of librarians can be very interesting and innovative. "Librarians might encourage, through contests or other means, student development of apps that make library content and services more useful for specific groups of users" [15]. They are not only knowledge managers. They are also promoters of new technologies that can be used in libraries and in academic institutions for research and teaching. As a result, many libraries in the world have embraced mobile technologies and support mobile learning.

VI. DEVEOPLMENTS AND CHALLENGES OF MOBILE LEARNING

While there are many aspects of mobile learning that are already known to scientist and professional involved in research and implementation of mobile learning, there are still some issues that remain unresolved or unknown due to the novelty of technology used in this type of e-learning. Conole [23] wrote about organizational aspects of e-learning which can be applicable to mobile learning as well. Theses aspects should be taken into account for mobile learning to be successful: 1.) understanding of how stakeholders (academics, support staff, administrators, senior managers and students) currently work 2.) mechanism and procedures for developing shared knowledge banks of expertise and information 3.) outline of roles and responsibilities for mobile learning activities? management, technical, research, dissemination, evaluation, training 4.) What are the different views of e-learning and its role? academics vs support staff 5.) how are institutions dividing roles and responsibilities for mobile learning and how much training and support are staff getting.

Other problems related to mobile learning and libraries are: the growing quantity but not all content of library holdings is prepared and adapted for use on mobile devices. In addition to lack of digital content in libraries made for use on mobile devices, there is also a lack of digital repositories with educational materials. Furthermore, teaching staff is not equipped and trained for use of mobile devices in the teaching process, there is a lack of standards for evaluation of use of mobile devices in teaching etc.

Implementation of mobile learning is a complex endeavor and many factors must be taken into account if educators and technology developers want it to be successful [24]: context: gathering and utilizing contextual information may clash with the learner's wish for anonymity and privacy; mobility: the ability to link to activities in the outside world also provides students with the capability to 'escape' the classroom and engage in activities that do not correspond with either the teacher's agenda or the curriculum.; learning over time: effective tools are needed for the recording, organization and retrieval of (mobile) learning experiences; informality: students may abandon their use of certain technologies if they perceive their social networks to be under attack; ownership: students want to own and control their personal technology, but this presents a challenge when they bring it in to the classroom.

When speaking about mobile learning, it is almost impossible not to mention social networks and learning management systems [25]. Students at the Faculty of humanities and social sciences (in Zagreb, Croatia) are using social networks in search of the learning material and for communication with other students and teachers while learning management system is used for access to the learning material created by teachers and students at this Faculty [25]. Since social networks are well adapted to access on mobile devices, everyday educational practice demonstrates that every communication between students and teachers on social networks is well accepted by students who use their mobile phones for communication related to the learning process as well for informal communication with family and friends [25]. The latest version of the learning management system Moodle which is called Omega at the Faculty of humanities and social sciences is also accessible on mobile phones and tablets and it is used in the learning process on daily basis. This helps students to stay in touch with the latest announcements or published content related to courses they enrolled in the academic year. The problem is the course related content which is not made to fit the small screens on smartphones or smaller tablets but is made for desktop PCs and laptops. This is one of the frequent problems related to mobile learning in an environment which is not fully adapted for implementation of mobile learning. Furthermore, the lesson plans must include students who own mobile devices as well as students who don't own mobile devices or their mobile device is too old to access and display course content properly if at all. To enable use the course content more extensively on mobile devices in future, content creators will have to produce several versions of content which will be displayable on various devices.

VII. CONCLUSION

In the educational sector, mobile technologies are being introduced into the learning process more and more frequently and are becoming important for its development. More specifically, mobile phones can expand educational access and they can support instruction, administration and professional development [27]. The results of research conducted in 2012 [25] suggest that mobile devices are already present in the learning process in higher education in Croatia. Generally speaking, mobile technologies open the door for a new

kind of learning in which students can learn at any time and at any place. As technological devices become more present in education, more researchers are looking into different aspects of technology enhanced learning. Most recently, focus in technology enhanced learning is put on use of mobile devices i.e. mobile learning or m-learning. Current versatility of use of mobile devices in learning suggests that the possibilities of their future use will only increase as mobile devices become more feature rich and as new types of digital content becomes available and usable on mobile devices. Further implementation of mobile technologies in the learning process would probably result in better learning materials and systems [26] and this will in the focus of future research. Use of mobile devices in education represent a challenge because it is difficult to predict what the next generation of mobile devices will bring from the technological and pedagogical point of view. In addition to students who are keen on using mobile devices, libraries are also adopting mobile technologies more frequently than ever to provide new and innovative services, as well as to introduce new ways for users to incorporate library services into their daily lives With proliferation of use of mobile devices libraries must now decide how much time and funding they should devote to the development and support of academic library resources in formats that can be accessed on mobile devices. While there are many aspects of mobile learning that are already known to scientist and professional involved in research and implementation of mobile learning, there are still some issues that remain unresolved or unknown due to the novelty of technology used in this type of learning.

REFERENCES

[1] P.V. Vinu, P.C. Sherimonb and R. Krishnanc, "Towards pervasive mobile learning – the vision of 21st century", Procedia - Social and Behavioral Sciences, vol. 15, pp. 3067–3073, 2011.

[2] G. Parsons, "Information provision for HE distance learners using mobile devices", The Electronic Library, vol. 28 no. 2, pp. 231-244, 2010.

[3] A.A. Economides, "Adaptive context-aware pervasive and ubiquitous learning", International Journal of Technology Enhanced Learning, vol. 1 no 3, pp. 169-192, 2009.

[4] A. Kukulska-Hulme and J. Traxler, "Learning design with mobile and wireless technologies". Retrieved January 4, 2015. from http://kn.open.ac.uk/public/getfile.cfm?documentfileid=9882

[5] C. BenMoussa, "Workers on the move: new opportunities through mobile commerce". Retrieved January 4, 2015. From http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=22A8C32247093BAA30F82A8083714119?doi=10.1.1.202.3560&rep=rep1&type=pdf

[6] C. Quinn, "mLearning: mobile, wireless, in-your-pocket learning", Retrieved January 4, 2015. from http://www.linezine.com/2.1/features/cqmmwiyp.htm

[7] F. Martin and J. Ertzberger, "Here and now mobile learning: An experimental study on the use of mobile technology", Computers & Education, vol. 68, pp. 76–85, 2013.

[8] D. Tatar, J. Roschelle, P. Vahey, and W.R. Penuel, "Handhelds go to school: Lessons learned", IEEE Computer, vol. 36 no. 9, pp. 30-37, 2003.

[9] P. Rodríguez-Arancón, C. Arús and C. Calle, The Use of Current Mobile Learning Applications in EFL, Procedia - Social and Behavioral Sciences, vol. 103, pp. 1189–1196, 2013.

[10] J. Cheon, S. Lee, S. Crooks and J. Song, "An investigation of mobile learning readiness in higher education based on the theory of planned behavior", Computers & Education, vol. 59, no. 3, pp. 1054–1064, 2012.

[11] C. O'Malley et al. Guidelines for learning/teaching/tutoring in a mobile environment,2005. Retrieved January 4, 2015. from http://kn.open.ac.uk/public/getfile.cfm?documentfileid=7488

[12] J. Brown and D. Metcalf, "Mobile learning update, Eliott Masie's Learning Consortium Perspectives". Retrieved January 4, 2015. from http://209.151.89.205/usdla.org/public_html/cms/mini_websites/old/MobileLearningUpdate.pdf

[13] J. Hahn, "Mobile learning for the twenty-first century librarian", Reference Services Review, vol. 36 no. 3, pp. 272-288, 2008.

[14] A. Kukulska-Hulme and J. Traxler, "Learning Design with Mobile and Wireless Technologies", Retrieved January 4, 2015. from http://kn.open.ac.uk/public/getfile.cfm?documentfileid=9882

[15] J.K. Lippincott, "A mobile future for academic libraries", Reference Services Review, vol. 38 no. 2, pp. 205-213, 2010.

[16] E. Wagner, "Mobile Learning: The Next Step in Technology-Mediated Learning", Retrieved January 4, 2015. from http://www.clomedia.com/articles/mobile_learning_the_next_step_in_technology_mediated_learning

[17] E. Klopfer, K. Squire and H. Jenkins, "Environmental Detectives: PDAs as a Window into a Virtual Simulated World", Proceedings of the IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'02), Los Alamitos, Ca.: IEEE, 2002, pp. 95-98.

[18] P.B. Muyinda, "MLearning: pedagogical, technical and organisational hypes and realities", Campus-Wide Information Systems, vol. 24 no. 2, pp. 97–104, 2007.

[19] F.A. Kutluk amd M. Gülmez, "A research about mobile learning perspectives of university students who have accounting lessons", Procedia - Social and Behavioral Sciences, vol. 116, pp. 291–297, 2014.

[20] M. Keough, "7 reasons why mlearning doesn't work", Retrieved January 4, 2015. from http://www.mlearn.org.za/CD/papers/McMillan-Keough.pdf

[21] C.Y. Wang, H.R. Ke and W.C. Lu, "Design and performance evaluation of mobile web services in libraries", The Electronic Library, vol. 30 no. 1, pp. 33–50, 2012.

[22] S. Ju-Ling, H. Gwo-Jen, C. Yu-Chung and C. Chien-Wen, "An investigation- based learning model for using digital libraries to support mobile learning activities", The Electronic Library, vol. 29 no. 4, pp.488–505, 2011.

[23] C. Conole, "E-Learning: the hype and the reality", Journal of Interactive Media in Education, vol. 2004 no. 12, 2004. Retrieved January 4, 2015. from http://jime.open.ac.uk/article/download/2004-12/212

[24] L. Naismith, P. Lonsdale, G. Vavoula and M. Sharples, "Literature Review in Mobile Technologies and Learning", Retrieved January 4, 2015. drom http://www2.futurelab.org.uk/resources/documents/lit_reviews/Mobile_Review.pdf

[25] R. Vrana, "Modes of use and preservation of digital information resources as a support to the learning process at the Faculty of humanities and social sciences in Zagreb", in Proceedings of the 36th International Convention MIPRO 2013, P. Biljanović Ed.. Rijeka : MIPRO, 2013, pp. 685-690

[26] J. Attewell, "From Research and Development to Mobile Learning: Tools for Education and Training Providers and their Learners" . Retrieved March 3, 2015. From http://www.mlearn.org.za/CD/papers/Attewell.pdf

[27] Mobile learning for teachers: global themes. Paris: UNESCO, 2012.

Open science, open access and open educational resources: challenges and opportunities

Radovan Vrana
Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
rvrana@ffzg.hr

Abstract – This paper discusses the concept of openness in general and its implementation in access to research and teaching material used in academic institutions. Though idea of openness i.e. open science and open access to information sources is not new, it is still not universally accepted and it is not understood equally in all parts of the world. The greatest benefits of open access can be observed in academic activities i.e. in research and teaching at academic institutions. However, open access is not understood and presented equally everywhere. There are differences in openness and rights of users in access and use of scientific and educational materials in open access digital repositories. While the number of open access digital repositories increases there are many questions waiting to be answered from quality of content, roles in management of open access information resources, responsibility for maintenance of digital information resources, financing of digital information resources, quality of scientific and educational digital resources etc.

I. INTRODUCTION

The access to internet has become a human right [1] and an essential tool in science and education. Inability to access textbooks, presentations, journal articles, conference papers, books, reports and other digital publications on the internet for any reason, creates a phenomenon called digital divide. The term digital divide was coined in 1990s, and it distinguishes people who have access to information and communication technology from those who don't [2]. Inability to use information and communication technology results in inability to access information resources too. While one part of the world has access to the great variety of paid information resources while the other part depends mostly on free of charge information resources available on the internet.

In both cases, members of educational sector (in this paper the focus is on higher education community) are interested in materials that already incorporate content with a specific didactical or pedagogical approach. These materials are often referred to as digital learning materials [3] and are becoming increasingly significant in education globally. Digital learning materials are available for access on various personal, corporate and institutional Web pages on the internet as well as in digital repositories [4] including open sources digital repositories. Open access means that information resources are digital, online, free of charge, and free of most copyright and licensing restrictions" [5]. In the last 20 years, open access initiative has played an important role in dissemination of scientific and educational material as it offers easy access to the peer reviewed journal articles, book, conference papers and other types of information resources usually found in academic institutions' libraries [6]. This initiative supports the idea of open science which is gaining on popularity as open access information resources grow in number and become more available.

Open access has gained good reputation because it has potential to reach "(…) a larger potential readership and thus establish the premises for potentially greater impact" [6]. Open access information resources are becoming more extensively used in research and teaching in higher education institutions in Croatia as a replacement for fee based information resources [7]. As a result, more members of academic community in Croatia and people outside of it are getting acquainted with open access initiative and benefits it could bring to their professional and personal lives. This paper will have focus on open access, open science and open educational resources and their role in science and teaching.

II. OPEN SCIENCE

Open science is "(…) the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process" [8]. It is also "The conduction of science in a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, with terms that allow reuse, redistribution and reproduction of the research" [9]. Open science can be described by its four fundamental goals: 1. transparency in experimental methodology, observation, and collection of data, 2. public availability and reusability of scientific data, 3. public accessibility and transparency of scientific communication and 4.using web-based tools to facilitate scientific collaboration [10]. Similar view on open science was offered by Smith [11]: open science should be related to transparency of the scientific methods, observations, data collection, data access, communication, collaboration and research tools. "Instead of limiting the sharing of the practice of science to publication of selected results, the entire scientific process should be exposed to potential users, collaborators and extenders of the work" [11]. The benefits of open science include sharing of knowledge, especially the knowledge that is publicly funded and the ability to use and re-use the results of scientific research including research data for the purpose of new research and teaching not only in higher education

institutions but elsewhere where quality information resources are needed. In addition to benefits, open science is also meeting some challenges, related frequently to the funding of scientific research in private sector where research results are not always published or publicly announced until such time when these research results can produce some profit to the company which financed it. Open science is greatly depending on open science information resources which create opportunities for facilitating publishing and access to human knowledge. As a result the idea of open science started to spread globally including proliferation of open access information resources as its most prominent manifestation.

III. OPEN ACCESS

Fee based (paid) information resources have become one of the major barriers in work of the higher education institutions in the world, primarily because of the high annual scientific journal subscription rates which are far too expensive for many libraries at universities [12]. Students, teaching staff and researchers need ubiquitous daily access to information resources which must satisfy the following characteristics: they must be free of charge, they must have peer reviewed content and be easily accessible, they must use common file formats readable on many electronic devices etc. As financial conditions at universities and in academic libraries around the world deteriorated, the open access initiative became more and more attractive to facilitate the access to scientific information resources used for research and teaching. Open access to scientific information resources such as scientific journal articles means they are freely available on the public internet, with permission to any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself [13]. In its document on open access and Horizon 2020 projects, European Commission points out that definitions of open access "(…) describe "access" in the context of open access as including not only basic elements such as the right to read, download and print, but also the right to copy, distribute, search, link, crawl, and mine" [14]. For Sawant [15] "The open access literature is digital, online, free of charge, and free of most copyright and licensing restrictions". For Björk [16] open access "means that a reader of a scientific publication can read it over the Internet, print it out and even further distribute it for non-commercial purposes without any payments or restrictions". All these definitions aim to describe open access as extensively as possible to make this initiative more acceptable to the scientific community and general public.

Open access is sometimes understood and defined very idealistically not taking into account many conceptual, financial, technical and legal factors which determine the degree of access and use in open access information resources. In real life, the understanding of open access depends on the level of openness governed by different access policies and licenses. Wiley [17] too pointed out that openness in open access is determined by licenses which state the degree of access or rights to use content that is considered open. In such situations users have little or no doubt what they can do with this information resource. Openness of digital information resources could be measured by the following rights a user of the content is granted [17]: 1. reuse – the right to reuse the content in its unaltered / verbatim form, 2. revise – the right to adapt, adjust, modify, or alter the content itself, 3. remix – the right to combine the original or revised content with other content to create something new, 4. redistribute – the right to make and share copies of the original content, your revisions, or your remixes with others.

In spite of sometimes different approaches to understandings of the open access initiative, the global academic community has accepted open access information resources as relevant for research and teaching. They participate in enlarging the quantity of open access information resources by self-archiving of already published scientific journal articles, conference papers, books, reports etc. in digital repositories. Today, more than 60% of journals already allow authors to self-archive content that has been peer-reviewed and accepted for publication [18]. Another way of participation is publishing of scientific works in open access journals or digital repositories thus starting new generations of journals committed to free and unrestricted access to scientific and professional information [19].

Definitions of open access indicate that the mainstream idea of open access focuses mostly on access to scientific information i.e. peer-reviewed scientific journal articles and scientific research data [20]. In addition to scientific content, a growing quantity of open access information resources include educational materials accessible with little or no restrictions which is very important for more stronger acceptance of open access in higher education.

Open access offers many opportunities and benefits to the general public, such as greater access to research outputs without the often prohibitive costs associated with the traditional publishing industry; no requirement for passwords or other forms of authentication; greater access to academic research funding for those researchers, academic and professional, who work outside academia; ensuring that those who actually provide the money for publicly-funded research (i.e. tax-payers) have access to the outputs they have funded" [21]. Researchers can benefit from open access in the following ways: their articles are freely accessible online; the authors hold copyright for their work and grant anyone the right to reproduce and disseminate the article, provided that it is correctly cited; a copy of the full text of each open access article is permanently archived in an digital repository separate from the journal; authors are assured that their work is disseminated to the widest possible audience, the widespread availability of articles will enhance literature searching; the results of publicly funded research will be accessible to all interested readers and not just those with access to a library with a subscription; a country's economy will not influence its scientists' ability to access articles [22].

While there are many benefits of open access, this initiative is not without challenges inherited from the still existing fee based system of scientific communication [23]. Quite frequently, open access publications are mimicking printed fee based scientific publications. Consequently, open access publications are suffering from similar problems as printed publications: costs related problem during the process of publication of journal articles (this problem is known as the article processing charge which is usually paid by authors themselves or their academic institutions); online articles can be as short or long, however, many articles are still limited in length (similar to printed journal articles); online journal issues are still limited in number of articles while they can contain as many or few articles as the editors wish etc. In spite of the problems, open access introduced many advantages over the existing system of scientific publishing and communication thus enabling easier access to scientific and educational material used for research and teaching [24].

Although there are still many challenges ahead of it, open access initiative has already gained significant popularity resulting in over 3900 open access digital repositories available to the internet users in academic institutions and outside of them [25].

IV. OPEN EDUCATIONAL RESOURCES

Realization of idea of open educational resources started to develop more rapidly a decade after open access initiative was born, somewhere after 2001 when MIT's started the OpenCourseWare initiative, which was followed by hundreds of universities worldwide collaborating in the advancement of open educational resources. Organizations like UNESCO, World Bank, OECD, the Commonwealth of Learning and the European Union also supported the development of open educational resources [26].

Most generally, educational materials that are available on the internet with little or no restrictions are usually referred to as open educational resources or OER. Open educational resources can be further defined as "(…) materials used to support education that may be freely accessed, reused, modified and shared by anyone" [27]. According to UNESCO, open educational resources are "(…) technology-enabled, open provision of educational resources, for consultation, use and adaptation by a community of users for non-commercial purposes. They are typically made freely available over the Web or the Internet. Their principal use is by teachers and educational institutions to support course development, but they can also be used directly by students in their daily academic activities. Open educational resources include learning objects such as lecture material, references and readings, simulations, experiments and demonstrations, as well as syllabi, curricula and teachers' guides" [28]. According to Van Acker, Frederik, Vermeulen, Kreijns, Lutgerink and van Buuren [3] open educational resources include "(…) full courses, course materials, modules, textbooks, videos, tests, software, and any other tools, materials or techniques used to support access to different kinds of knowledge. The complexity of the open educational resources may vary from very simple (e.g., a single video clip or a single illustration) to rather complex (e.g., comprising a series of lessons)". The main motive for sharing educational materials can be different for everyone included in this process. Van Acker, Frederik, Vermeulen, Kreijns, Lutgerink and van Buuren [3] suggested that altruism is a very reasonable factor for sharing open educational resources.

Open educational resources offer benefits and opportunities for teachers such as use of openly licensed materials, license educator's own open educational resources so others can use it, promotion of teachers' work to a global audience and benefits for students such as provision of supplemental learning materials for courses, making decision what classes to enroll in and be better prepared for classes [29]. Open educational resources provide a great opportunity for "(…) reducing the cost of expanding quality education" [30]. Most important benefit of open educational resources is access to educational materials with little or no obstacles very similarly to open access.

Like open science and open access, open educational resources are also facing a number of challenges. In their report "A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities", Atkins, Brown and Hammond [30] presented a number of challenges that should be resolved in order for Open educational resources to remain available to users: sustainability (connected mostly to long term funding), curation and preservation of access (relying on proper infrastructure for curating and preserving access), object granularity and format diversity (the size of the digital objects that can be individually tagged, referenced, found, and re-used under appropriated attached terms and conditions – the possibility that even the smallest part of an digital object is identified and used), intellectual property issues (material that is intended for publishing on the internet must be checked for any content that could potentially break copyright laws), content quality assessment and enhancement (by providing open community review). Furthermore, open educational resources suffer from: missing components that require additional work, different degree of completeness of content, non existing appropriate academic peer review, missing learning objectives, problems with licensing, if open educational resources are not created or edited with developed with reuse in mind, there may be internal references that are meaningful within the original course context but make little sense in the new context, re-purposing of online material previously used or updated but without version numbering, learning material can contain old URLs or references etc. [26]. All these challenges can be overcome in time but only if there is general agreement about standards which would describe better what open educational resources should look like and how they could be used.

V. DIGITAL REPOSITORIES AND OPEN ACCESS

A digital repository in an academic institution is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its

community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution [32]. Digital repositories offer a wide range of access opportunities such as remote access, 24-hour access, and multiple users for single sources, of course, all in open access.

Digital repositories have an important role in preservation and access to scientific and educational material, making other internet information resources less relevant for research and education [33], [34], [35], [36]. They serve universities to systematically organize the Web-based scientific and educational material produced by the faculty [16]. Digital repositories can contain e-prints of scientific papers, research data, but also e-learning materials and other forms of institutional intellectual outputs, which are generally not published or preserved elsewhere [37]. Digital repositories in higher education institutions could store: course information, teaching and learning materials, student admission data, student academic record, resource information that can be shared across departments, research output, research projects content, academic staff and expertise information, accreditation data etc. [38]. As a result, digital repositories are helping to manage and capture all or almost all intellectual assets in an academic institution making them more visible and available to users as part of information strategies of academic institutions [38]. In addition to content stored locally, digital repositories also provide link lists to freely available educational resources on the internet which are not stored in the digital repository of an academic institution [16] thus expanding access to scientific and educational materials.

Digital repositories represent most concrete opportunity for researchers and teachers to store and use different types of material but they are also meeting challenges similar to those of open science and open access initiative. Their development is facing an opposition in academic institution which can lead to situations where the quality of digital content stored in digital repositories may vary creating extraordinary problems for the digital repository if it wants to preserve these materials [39].

VI. CONCLUSION

In the era of economic crisis, access to scientific and educational resources without restrictions is particularly important. It is essential for almost all activities in academic institutions. Since the advent of the open access idea, it has gone through many changes including the transformation from somewhat idealistic notion of unlimited and fee free access to all information resources to more realistic approach which includes systematic care for creation, access and use of quality information resources for research and teaching. During the last two decades, many new digital forms of knowledge have appeared and publishers and academic institutions have started finding mutual language in agreeing upon most important factors that would enable academic institutions to publish and/or access articles, books and educational material without so many financial and legal restrictions and enable researchers and students to access and use wide range of content with as little limitations as possible. However, the ideas of open science and open access are still not universally accepted and there are many challenges in their development such as: international legal restrictions including digital rights management, diverse and sometimes murky business models of open access publishing and access to published materials, new formats of digital content that are still not accepted or there are problems in their use, problems of peer review for authors trying to publish their papers in open access, evaluation of quality of open access information resources is still undeveloped and questionable, article processing charge has become a major problem in open access publishing, inaccessibility of computer and other electronic devices for access to open access resources, inability to access the internet etc.

Open education resources are a special case. A sufficient amount of open educational resources should become available supported by a sufficiently large community to share digital learning materials for open educational resources to become successful [3]. As more students gain access to the internet, more open educational resources will be needed. Higher education institutions can contribute to the amount of digital educational resources by providing incentives for teachers to deposit their educational resources in digital repositories. They should also have written strategies and policies regarding preservation and use of education materials in digital format. Educational institutions could have and important influence on governments for wider introduction of digital educational resources in general and open educational resources in particular as well as introduction and implementation of digital technologies in education.

REFERENCES

[1] S.B. Wicker and S.M.Santoso, „Access to the Internet Is a Human Right“, Communications of the ACM, vol. 56, no. 6, pp. 43-46, 2013.

[2] A, Kassam, M. Iding, P. Hogenbirk, “Unraveling the digital divide: Time well spent or 'wasted'?”, Education and Information Technologies, vol.18, no.2, pp. 215-221, 2013.

[3] F. Van Acker, M. Vermeulen, K. Kreijns, J. Lutgerink, H. van Buuren, “Why do not teachers share OER? The role of evaluation apprehension and trust”, Computers in human behavior, vol. 39, pp. 136-144, 2014.

[4] D.A. Koutsomitropoulos, A.D. Alexopoulos, G.D. Solomou, T.S. Papatheodorou, “The use of metadata for educational resources in digital repositories: Practices and perspectives”, D-Lib Magazine, vol. 16, no. 1-2, 2010. Retrieved December 12, 2014, from http://dlib.org/dlib/january10/kout/01kout.html

[5] P. Suber, “Open access”. Cambridge, Massachusetts; London: The MIT Press, 2012.

[6] F. Mann, B. Von Walter, T. Hess, R.T. Wigand, “Open access publishing in science”, Commmunications of the ACM, vol. 52, no. 3, 135-139, 2009.

[7] R. Vrana, “Access to digital information resources as a support to academic achievement” in central European Conference on Information and Intelligent Systems - CECIIS 2014., Varaždin : Faculty of Organization and Informatics, University of Zagreb, 2014, pp. 144-151.

[8] “Open science”, Retrieved December 12, 2014 from http://openknowledge.stanford.edu/science

[9] “Open Science Definition”. Retrieved December 12, 2014 from https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition

[10] D. Gezelter, “What, exactly, is Open Science?”, Retrieved December 12, 2014 from http://www.openscience.org/blog/?p=269

[11] K. Smith, “What is Open Science?”, Retrieved December 12, 2014 from http://blogs.library.duke.edu/scholcomm/2010/09/13/what-is-open-science/

[12] M. Krishnamurthy, “Open access, open source and digital libraries: A current trend in university libraries around the world”, Program: electronic library and information systems, vol. 42, no. 1, pp. 48-55, 2008.

[13] “Budapest Open Access Initiative (BOAI). Retrieved December 12, 2014 from http://www.budapestopenaccessinitiative.org/read downloaded December 27th 2013.

[14] “Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020”, Retrieved December 12, 2014 from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[15] S. Sawant, “Open access resources useful in LIS education”, Library Hi Tech News, vol. 30, no. 7, pp. 16-20, 2013.

[16] B.C. Björk, “Open access to scientific publications - an analysis of the barriers to change?”, Information research, vol. 9, no. 2, 2004, Retrieved December 12, 2014 from http://www.informationr.net/ir/9-2/paper170.html

[17] D. Wiley, “Iterating toward openness: Defining ‘open.’”, Retrieved December 12, 2014 from http://opencontent.org/blog/archives/1123

[18] R. Van Noorden, “The true cost of science publishing”, Nature, vol. 495, pp. 426-429, 2013.

[19] S.L. Zhang, “The flavors of open access”, OCLC Systems & Services: International digital library perspectives, vol. 23, no. 3, pp. 229-234, 2007.

[20] N. Koutras, M. Bottis, “Institutional repositories of open access: a paradigm of innovation and changing in educational policies”, Procedia - Social and behavioral sciences, vol. 106, pp. 1499-1504, 2013.

[21] P. Suber, “Open Access in 2007”, The Journal of Electronic Publishing, vol. 11, no. 1. Retrieved December 12, 2014 from http://dx.doi.org/10.3998/3336451.0011.110

[22] P.J. Chiao, C. Schmidt, “Open Access gains attention in scholar communication”, Molecular Cancer, vol. 23, no. 3, pp. 1-2, 2004.

[23] J. Cirasella, S. Bowdoin, “Just Roll with It? Rolling Volumes vs. Discrete Issues in Open Access Library and Information Science Journals”, Journal of Librarianship and Scholarly Communication, vol. 1, no. 4, p. 1086, 2013. Retrieved December 12, 2014 from http://dx.doi.org/10.7710/2162-3309.1086

[24] R. Rufai, S. Gul, T.A. Shah, “Open Access journals in library and information science: the story so far”, Trends in Information Management (TRIM), vol. 7, no. 2, pp. 218-228, 2011.

[25] Registry of open access digital repositories, Retrieved December 12, 2014 from http://roar.eprints.org/

[26] I. DeVries, “Evaluating open educational resources: Lessons learned”, Procedia - Social and Behavioral Sciences, vol. 83, pp. 56–60, 2013.

[27] S. Downes, “Open Educational Resources: A Definition”, Retrieved December 12, 2014 from http://www.downes.ca/archive/11/07_18_news_OLDaily.htm

[28] UNESCO. (2002). UNESCO promotes new initiative for free educational resources on the Internet. Paris: Author. Retrieved from Retrieved December 12, 2014 from http://www.unesco.org/education/news_en/080702_free_edu_ress.shtml

[29] “Open Educational Resources: Benefits for Faculty and Students”, Retrieved December 12, 2014 from http://open.umich.edu/sites/default/files/3659/PDFs/OER-benefits-handout.pdf

[30] A. Marcus-Quinn, Y. Diggins, “Open Educational Resources, Procedia - Social and Behavioral Sciences”, Procedia - Social and Behavioral Sciences, vol. 93, pp. 243–246, 2013.

[31] A. Daniel, J.S. Brown, A. Hammonds, “A Review of the Open educational Resources (OER) Movement: Achievements, Challenges and New opportunities”, Retrieved December 12, 2014 from http://www.hewlett.org/uploads/files/ReviewoftheOERMovement.pdf

[32] C. Lynch, "Institutional repositories: essential infrastructure for scholarship in the digital age", ARL Bimonthly Report, no. 226, Retrieved December 12, 2014 from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

[33] R. Vrana, “Access to digital information resources as a support to academic achievement” in Central European Conference on Information and Intelligent Systems - CECIIS 2014, Varaždin : Faculty of Organization and Informatics, University of Zagreb, 2014, pp. 144-151.

[34] R. Vrana, “Digital Repositories of Scientific Information at the Croatian Universities: Developing the Bridge towards e-Science” in Proceedings of the ITI 2013, 35th International Conference on Information Technology Interfaces, Zagreb : University Computing Centre ; University of Zagreb, 2013, pp. 145-150.

[35] R. Vrana, “Digital repositories at the Croatian universities: providing support for the enhancement of research and teaching” in MIPRO 2010 33rd International convention on information and communication technology, electronics, and microelectronics, Rijeka : Croatian society for information and communication technology, electronics and microelecronics, 2010, pp. 141-146.

[36] R. Vrana, “Digital repositories and possibilities of their integration into higher education” in MIPRO 2009 32nd international convention on information and communication technology, electronics and microelectronics: computers in education, Rijeka : Hrvatska udruga za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku - MIPRO, 2009, pp. 94-99.

[37] H. Hockx-Yu, "Digital preservation in the context of institutional repositories", Program: electronic library and information systems, vol. 40, no. 3, pp. 234-245, 2006.

[38] F. Sarker, H. Davis, T. Tiropanis, “The Role of Institutional Repositories in addressing Higher Education Challenges, EDUCAUSE Evolving Technologies Committee white paper”, Retrieved December 12, 2014 from http://eprints.soton.ac.uk/271694/1/The_Role_of_Institutional_Repositories_in_addressing_Higher_Education_Challenges.pdf

[39] P. Conway, "Modeling the digital content landscape in universities", Library Hi Tech, vol. 26, no. 3, p. 342, 2008.

The emerging strategy of social network analysis: an overview of indicators, concepts and implications

Anita Papić*
* Faculty of humanities and social sciences/Department of information sciences, Osijek, Croatia
apapic@ffos.hr

Abstract - This paper describes an emerging strategy called social network analysis (SNA) and points out some of its possible implications. SNA is an interdisciplinary technique developed under many influences and the most important ones came from mathematics and computer science. Mathematical graph theory is used for modelling social networks and it is a starting point for many of the most fundamental ideas of SNA. The paper gives an overview of main concepts of SNA such as local centrality (degree), betweenness, global centrality (closeness), density and centralization. Also, the mathematical notation of the most important indicators of SNA are presented in the paper. The main research question was to determine the prevalence of SNA in scholarly publications in last decade across different scientific areas and different geographical areas. According to obtained empirical data the usage of SNA strategy in scholarly publications has ascending pathway.

I. INTRODUCTION

This paper gives theoretical insight into a strategy called social network analysis (SNA) and points out some of its possible implications. The origins of SNA dates from Moreno's sociograms and the sociometric star – person who held a position of great popularity and leadership [11]. A sociogram helps to visualize the channels through which information could flow from one person to another and through which one individual could influence another. SNA is an interdisciplinary technique developed under many influences and the most important ones coming from mathematics and computer science. In informetrics, information scientists study citation and co-citation networks, collaboration structures and other forms of social interaction networks and underlying any network lies a graph [7]. Mathematical graph theory is used for modelling social networks and it is a starting point for many of the most fundamental ideas of SNA so terms of simple graph, directed graph, weighted graph etc. are described. Decomposing networks into their constituent sub-groups leads to terms as cliques, clusters, or blocks namely to so called groups in front and groups in back [12]. A clique describes a particular configuration of informal interpersonal relations. The concept of point centrality originated in the sociometric concept of the star [13]. Furthermore, a family of point centrality measures consists of: local centrality (degree), betweenness and global centrality (closeness). The concepts of density and centralization refer to differing aspects of the overall compactness of a graph. Density describes the general level of cohesion in a graph and centralization describes the extent to which this cohesion is organized around particular focal points. Centralization and density are important complementary measures. Density plays a powerful role in the comparative study of social networks.

The great mathematician Leonhard Euler (1707-1783) considers as the person who grounded the graph theory because he solved till then unsolved problem of Konigsberg's bridges in year 1736 [3]. In honor to Euler the tour in which we pass through every edge once till we return to the starting point is called Euler's tour. Konigsberg was in 18 century town in Russia and through the town passed two little rivers of the river Pregel surrounding the island Kneiphof. At the river were seven bridges which connected four parts of the town separated by the river. The problem was is it possible to make a walk through the town starting at one part of the town, encompassing all seven bridges just once and ending in a starting point. Euler solved that problem in a way that he firstly made a model shown at the Fig. 1 which we today call a graph.

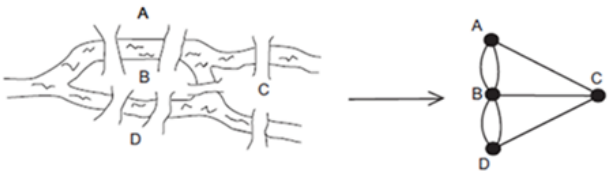


Fig. 1. Konigsberg's bridges [9]

Graph G is ordered triple G = (V(G), E(G), φ_G) consists of un empty set V=V(G), which elements are vertices of G, set E = E(G), which elements are edges of G and incidence function φ_G which every edge of G connects not ordered couple (not necessarily different) vertices of G [2]. The most used combinatorial structure is a graph. The simplicity of this structure enables that a lots of practical problems can be modeled in terms of graphs. The graphs are mathematical abstraction of different problems from real world and on them we can implement

all proven theoretical facts. Although, foundations of graph theory have their roots in 18 century this mathematical discipline as theory begins to develop in second half of 20 century. The actuality of graph theory can be seen in many of today's open problems. Namely, problems in graph theory can be very simple to formulate them but sometimes these problems are very difficult to be solved. The approach with graph theory is useful for modeling and explaining also the social phenomenon. Some applications of graphs are for example: (a) in different scientific areas and different disciplines especially in informatics and applied mathematics, (b) for representation of real network configurations for example road maps, electrical circuits etc., (c) for representation of abstract objects such as databases, project's activities etc., (d) for representation of social relationships and organization's hierarchies etc.

At the following figures it can be seen some classifications of graphs. Simple graphs don't consist loops not multiple edges. Some examples of simple graphs are shown at Fig. 2.

Fig. 2. Some examples of simple graphs [9]

Multigraphs allow multiple edges but do not allow loops. Some examples of multigraphs are shown at Fig. 3.

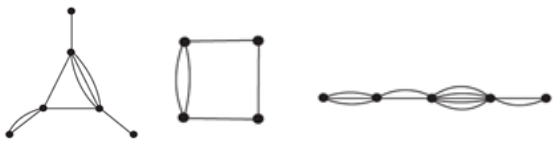


Fig. 3. Some examples of multigraphs [9]

Pseudographs are (multi)graphs which allow (multiple) loops. Some examples of pseudographs are shown at Fig. 4.

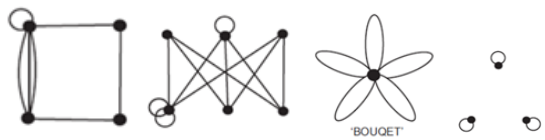


Fig. 4. Some examples of pseudographs [9]

Some examples of special graphs are shown at Fig. 5.

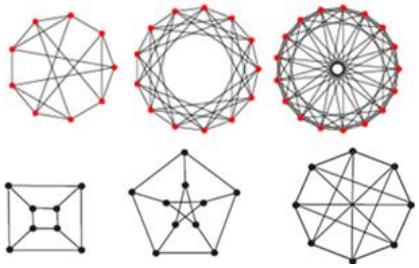


Fig. 5. Some examples of special graphs [9]

Digraph or directed graph is ordered triple (V(D), A(D), φ_D) which consists of un empty set V(D) of vertices, set A(D) of arcs (or directed edges) and incidence function φ_D which every arc *a* connects ordered couple (not necessary different) vertices *u*, *v* which *a* connects [2]. Vertex *u* is the starting vertex and vertex *v* is the ending vertex of *a*. Namely, digraph is the graph which edges have direction. Some examples of digraphs are shown at Fig. 6.

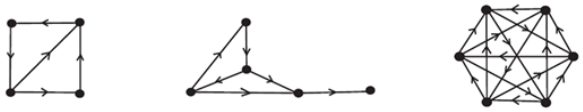


Fig. 6. Some examples of digraphs [8]

Weighted graph is graph in which every edge e is connected to not negative number *ω(e)*, which we call the weight of edge e. Directed and undirected graphs can be weighted. Un weighted graphs can be considered as weighted graph if all edges have the same weight for example 1. The main domain of weighted graphs implementation is in area of optimization. The shortest path between two vertices in weighted graph we consider the path with the lowest weight. The psychiatrist Jacob Moreno is considered the person who grounded the area of social networks because he published foundations of sociometrics which later becomes famous as social network analysis – SNA in his book *Who shall survive?* [11]. Social networks can be represented by graphs. The nodes represent social entities (actors): people, groups, organizations, countries etc. The edges represent relationships or connections between actors: communications, coauthorship, business or political relationships etc. In cases of directed connections we use directed graphs and in cases of measurable connections we use weighted graphs. Fig. 7 shows two same graphs: (a) un weighted and (b) weighted.

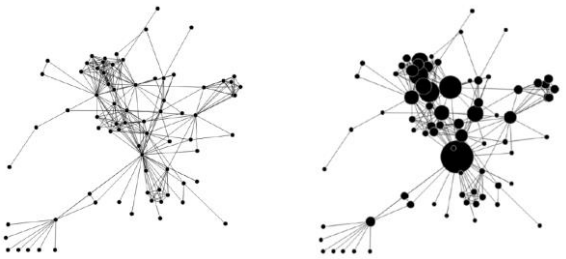


Fig. 7. Visualisation of actor's interactions in Victor Hugo's LES MISERABLES [4]

Information visualization is the area of graphical representation of information or infographics. Infographics uses computer supported tools for exploration of huge amounts of abstract data. Fig. 8 shows an example of infographics Greg Kolb's "New maps of influence – 10 visualisations of the social graph".

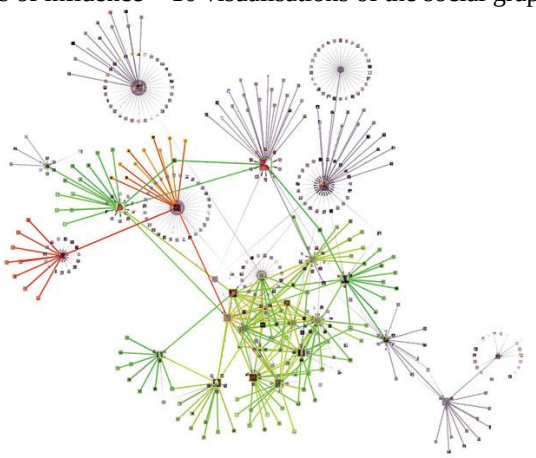


Fig. 8. Greg Kolb's "New maps of influence – 10 visualisations of the social graph" [5]

II. AN OVERVIEW OF MAIN CONCEPTS OF SNA

A. *Local centrality (degree)*

Local centrality or degree centrality is number of connections which this specific node posses namely the number of edges neighbour with that node. The formula for local centrality or degree centrality

$$d_i = \frac{\sum_j a_{ij}}{(n-1)}$$

where a_{ij} shows (not)existance of connection between node *i* and node *j* in a way that $a_{ij}=1$ if exists connection between node *i* and node *j*, while $a_{ij}=0$ if does not exist connection between node *i* and node *j*,where *n* is number of nodes [10].

B. *Betweenness*

Betweenness centrality is defined as number which shows how many times a node needs specific node to come to another node namely this is number of shortest paths which pass through certain node. Betweenness centrality measures the level till the node enables a flow in a graph. The formula of betweenness centrality is

$$b_i = \sum_{j,k \wedge i \neq j \neq k} \frac{g_{jik}}{g_{jk}} \Big/ \frac{(n-1)(n-2)}{2}$$

where g_{jk} is number of the shortest paths from node *j* to node *k*, g_{jik} is number of the shortest paths from node *j* to node *k* which pass through node *i*, where *n* is number of nodes [10].

C. *Global centrality (closeness)*

Global centrality or closeness centrality is equal to total distance in a graph, of this specific node from all other nodes. The formula of global centrality or closeness centrality

$$c_i = \frac{(n-1)}{\sum_j e_{ij}}$$

where e_{ij} is number of connections in the shortest path from node *i* to node *j*, where *n* is number of nodes [10].

D. *Density*

Density describes the general level of graph's cohesion. The formula of density is

$$\frac{l}{n(n-1)/2}$$

where *n* is number of nodes and where *l* is number of edges [10] .

E. *Other indicators*

- the clique is a subgraph in which every node is in connection to all other nodes of that subgraph;
- the claster is an area of greather cohesion in a graph;
- the diameter is the longest path in a graph;
- the radius is distance from the focus of graph till the most distant nodes of the graph's centar [10].

III. METHOD

For the purpose of conducting research in this paper scientific method of content analysis was selected. Content analysis is flexible research method and it can be qualitative or quantitative but also within it can be used combination of these two methods. Namely, in this paper was used quantitative content analysis of data obtained from Web of Science database according to searching strategy: "social network analysis" OR SNA.

Content analysis is research method which objectively and systematically describes written, verbal or visual communication [1]. Krippendorff describes content analysis as research method for derivation of repeatable and valid conclusions from text within its context [6]. Content analysis helps to determine presence of certain words, phrases or concepts within text or collections of texts. Texts can be from books, speeches, interviews, essays, historical documents, scientific articles, pictures, web pages, databases etc.

Content analysis can be implemented to any written or other form of human communication and that's why in recent decade it is often used in marketing, cultural studies, journalism, sociology, politics, psychology, cognitive science, economy, information sciences etc.

IV. INTERPRETATION OF RESULTS

The main research question was to determine the prevalence of SNA in scholarly publications in last decade across different scientific areas and different geographical areas. In this paper was used quantitative content analysis of data obtained from Web of Science database according to searching strategy: "social network analysis" OR SNA. Table 1 shows list of authors whose papers dealt with social network analysis such as Park H. W., Wang L., Latkin C. A. etc.

TABLE 1. PREVALENCE OF SNA IN SCHOLARLY PUBLICATIONS - LIST OF AUTHORS

Field: Authors	Record Count
PARK HW	32
WANG L	27
LATKIN CA	25
LEYDESDORFF L	24
ZHANG Y	22
CARLEY KM	21
HOSSAIN L	21
SNIJDERS TAB	21
THELWALL M	20
LEE S	19
MORENO Y	18
PARK J	18
CHRISTAKIS NA	17
FOWLER JH	17
TUCKER JS	17
CHEN HC	16
CHEN Y	16
EICKHOFF SB	16
LEE J	16
KIM S	15
LI X	15
VALENTE TW	15
CROFT DP	14
JAMES R	14
LITWIN H	14
LIU Y	14
WANG Y	14
BORGATTI SP	13
KIM J	13
WU B	13
ABBASI A	12
DE MOYA-ANEGON F	12
DOREIAN P	12
GREEN HD	12
JUNG JJ	12
KAWACHI I	12
KIM H	12
LEE H	12
PENTLAND A	12
STEGLICH C	12
WANG B	12

Table 2 shows list of different geographical areas or countries where authors wrote papers regarding social network analysis such as USA, England, China, Australia etc.

TABLE 2. PREVALENCE OF SNA IN SCHOLARLY PUBLICATIONS - LIST OF DIFFERENT GEOGRAPHICAL AREAS/COUNTRIES

Field: Countries/Territories	Record Count
USA	956
ENGLAND	263
PEOPLES R CHINA	193
AUSTRALIA	174
GERMANY	169
JAPAN	144
SPAIN	132
ITALY	127
NETHERLANDS	127
CANADA	126
SOUTH KOREA	126
TAIWAN	98
FRANCE	82
BELGIUM	60
BRAZIL	60
SWITZERLAND	58
SWEDEN	54
TURKEY	40
FINLAND	36
NEW ZEALAND	32
IRELAND	28
SCOTLAND	27
INDIA	25
AUSTRIA	24
GREECE	23
DENMARK	22
POLAND	22
SLOVENIA	17

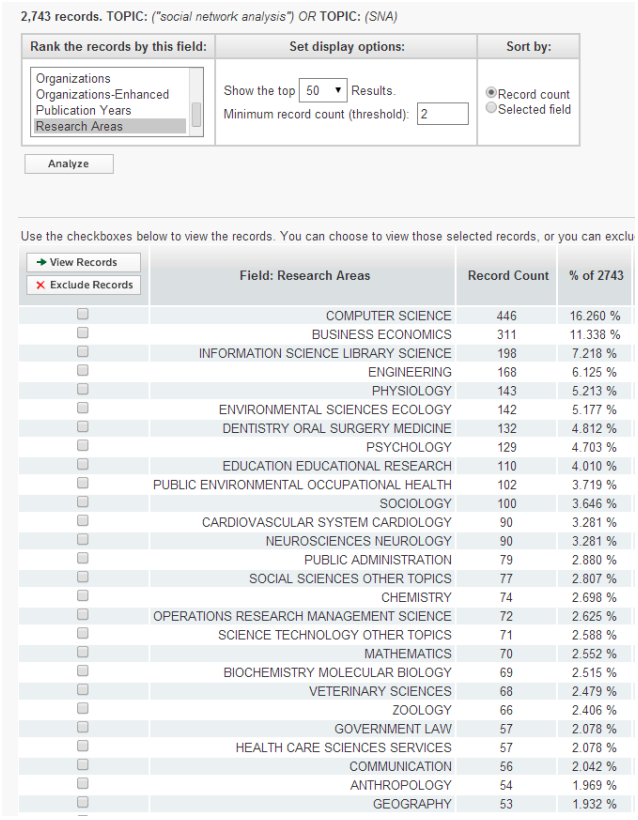
Table 3 shows number of papers regarding social network analysis per publication years where it can be seen an ascending pathway.

TABLE 3. PREVALENCE OF SNA IN SCHOLARLY PUBLICATIONS - NUMBER OF PAPERS PER PUBLICATION YEARS

Field: Publication Years	Record Count
2013	531
2012	457
2011	395
2010	301
2009	273
2008	217
2007	173
2006	136
2005	117
2004	99

Table 4 shows the prevalence of SNA in scholarly publications in last decade across different scientific areas such as computer science, business economics, information science and library science, engineering etc.

TABLE 4. THE PREVALENCE OF SNA IN SCHOLARLY PUBLICATIONS IN LAST DECADE ACROSS DIFFERENT SCIENTIFIC AREAS



2,743 records. TOPIC: *("social network analysis") OR* TOPIC: *(SNA)*

Rank the records by this field:	Set display options:	Sort by:
Organizations Organizations-Enhanced Publication Years Research Areas	Show the top 50 Results. Minimum record count (threshold): 2	Record count Selected field

Analyze

Use the checkboxes below to view the records. You can choose to view those selected records, or you can exclu

View Records
Exclude Records

Field: Research Areas	Record Count	% of 2743
COMPUTER SCIENCE	446	16.260 %
BUSINESS ECONOMICS	311	11.338 %
INFORMATION SCIENCE LIBRARY SCIENCE	198	7.218 %
ENGINEERING	168	6.125 %
PHYSIOLOGY	143	5.213 %
ENVIRONMENTAL SCIENCES ECOLOGY	142	5.177 %
DENTISTRY ORAL SURGERY MEDICINE	132	4.812 %
PSYCHOLOGY	129	4.703 %
EDUCATION EDUCATIONAL RESEARCH	110	4.010 %
PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH	102	3.719 %
SOCIOLOGY	100	3.646 %
CARDIOVASCULAR SYSTEM CARDIOLOGY	90	3.281 %
NEUROSCIENCES NEUROLOGY	90	3.281 %
PUBLIC ADMINISTRATION	79	2.880 %
SOCIAL SCIENCES OTHER TOPICS	77	2.807 %
CHEMISTRY	74	2.698 %
OPERATIONS RESEARCH MANAGEMENT SCIENCE	72	2.625 %
SCIENCE TECHNOLOGY OTHER TOPICS	71	2.588 %
MATHEMATICS	70	2.552 %
BIOCHEMISTRY MOLECULAR BIOLOGY	69	2.515 %
VETERINARY SCIENCES	68	2.479 %
ZOOLOGY	66	2.406 %
GOVERNMENT LAW	57	2.078 %
HEALTH CARE SCIENCES SERVICES	57	2.078 %
COMMUNICATION	56	2.042 %
ANTHROPOLOGY	54	1.969 %
GEOGRAPHY	53	1.932 %

According to obtained empirical data the usage of SNA strategy in scholarly publications has ascending pathway.

V. CONCLUSION

SNA is an interdisciplinary technique developed under many influences and the most important ones coming from mathematics and computer science. Mathematical graph theory is used for modelling social networks and it is a starting point for many of the most fundamental ideas of SNA. The paper gives an overview of main concepts of SNA such as local centrality (degree), betweenness, global centrality (closeness), density and centralization. Also, the mathematical notation of the most important indicators of SNA are presented in the paper. The main research question was to determine the prevalence of SNA in scholarly publications in last decade across different scientific areas and different geographical areas. According to obtained empirical data the usage of SNA strategy in scholarly publications has ascending pathway. Also, the wider use of SNA as multidisciplinary approach is encouraged to uncover the power located in the networks that structure society.

REFERENCES

[1] B. Barelson, Content analysis in communication research. New York : Free Press, 1952.

[2] Goodaire, E. G. & Parmenter, M. M. (2002). Discrete Mathematics with Graph Theory. New York: Prentice Hall.

[3] Graph Theory Resources. http://www.graphtheory.com/ (2015-01-18)

[4] Knuth, D. E. The Stanford Graph-Base: A Platform for Combinatorial Computing. Addison-Wesley, 1993.

[5] Kolb, G. New maps of influence – 10 visualisations of the social graph. http://ccc.georgkolb.com/?p=28(2015-01-18)

[6] K. Krippendorff, Content analysis: an introduction to its methodology. Thousand Oaks : London : New Delhi: Sage Publications, 2004.

[7] Liu, X., Bollen, J., Nelson M. L. & Van de Sompel, H. (2005). Co-Authorship Networks in the Digital Library Research Community. http://arxiv.org/pdf/cs/0502056.pdf. (2015-01-18)

[8] Math Archives: Discrete Mathematics. http://archives.math.utk.edu/topics/discreteMath.html. (2015-01-18)

[9] Mathworld. http://mathworld.wolfram.com/topics/DiscreteMathematics.html. (2015-01-18)

[10] Otte, E. & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science, 28, 6, 441–453.

[11] Scott, J. (1987). Social network analysis: a Handbook**.** London: Sage Publications.

[12] Serrat, O. (2010). Social network analysis. Washington, DC: Asian Development Bank.

[13] Streeter, C. L. & Gillespie, D. F. (1992). Social Network Analysis. Journal of Social Service Research, 16, 1/2, 201-222.

Children's interest in the media - empirical research

M. Nadrljanski* and V. Domitrović **
* Faculty of Maritime Studies – Split, Split, Croatia
** University College of Inspection and Personnel Management in Maritime Affairs, Split, Croatia
*mila@pfst.hr
**veronika.domitrovic@icloud.com

Abstract - The 21st century has been marked by the media which have a significant influence on adult population. This paper analyzes the effects of the media on the youngest population in Croatia. The entire research is oriented empirically on the random sample of two age categories of children. The first age category includes children at the age of 4 to 7, and the other one children at the age of 8 to 11. A special analysis was performed on positive media influence with quantitative data on the statistical sample. Negative media effects were studied on the same sample as well. As the media influence children's personality formation, the research has shown it is necessary to control their use of the media in order to get proper education.

Key words: media, children, media influence, research on media influence

I. INTRODUCTION

This paper discusses relations between the media and children, as well as its role in their everyday life. Nowadays, every new generation of children grows up surrounded by new media, different from the media that influenced the previous generation, which leads to the conclusion that impact of the media on children is inconsistent, it changes with each generation. That implies the necessity for more frequent conducting research on the media influence on children.

The media are an endless source of information and play an important role in forming people's opinions, and children's opinions as well. The media are mainly mentioned in a negative context, and have been considered the cause of alienation in society since the sudden development of technology. Today we have smart phones, android operating system, iPhone, iPad, tablets and similar devices. People spend more time with their devices than socializing with other people. It is reflected in the family, and in this context, children are a part of the social-media era. The media is the dominant feature of the 21st century. This fact leads to some sort of alienation, creates isolation, and obliterates any form of conversation.

In this regard, special emphasis of this paper is in exploring and presenting behaviour of different age groups under the influence of the media. The tasks are:

(a) to determine which types of media children are most interested in,

(b) to describe the media content ,

(c) to define time spent with the preferred medium, and

(d) to perceive the difference between younger children's perception of the media and older children's perception of the same.

The media tend to form people's opinions, therefore it is important to comprehend the way it happens with children as they are more susceptible to negative influences than adults. They accept information from the media at conscious and subconscious level and such information definitely has influence in creating of their interactions and comprehension in future society.

For this research we used a survey, which is a quantitative form of research, and also one of the most popular methods. Quantitative research focuses on figures, thus being a kind of research on which statistical analysis is based. One of the basic characteristics of quantitative research is a set of identical questions for all respondents [2].

Besides, it should be kept in mind that the value of the survey is limited. The obtained information depend on the sincerity of the respondents and their ability to respond to questions.

We defined the goal, set the hypotheses, determined the sample size and compiled the questionnaire of 26 questions. The survey was conducted in writing and online, using polydaddy.com.

II. RESEARCH PROBLEM

The main problem of this research is to examine the role of the media in children's everyday life and state what type of media they use and prefer (newspapers, radio, television or the Internet).

As it was stated, media can influence opinion formation, and children are more susceptible to media influences than adults. They consciously or subconsciously absorb information from the media, and such influence will have an impact on their future personality and their future judgement.

Three hypotheses were set:

The first hypotesis (H1)

- Children use the media (newspapers, radio, television and the Internet) on a daily basis, but under control of their families, friends or other people; they read articles, literature, watch programs and shows and use the Internet in accordance with their age.

The second hypotesis (H2)

- Children use the media (newspapers, radio, television and the Internet) on a daily basis, often without the control of their families, friends or other people; they read articles, literature, watch programs and shows, and use the Internet in accordance with their age.

The third hypotesis (H3)

- Children use the media (newspapers, radio, television and the Internet) on a daily basis, without any control of their families, friends or other people; they read articles, literature, watch programs and shows, and use the Internet in a way which is inappropriate for their age.

III. RESEARCH INSTRUMENT - QUESTIONAIRRE

The instrument used in this research was a questionnaire consisting of 26 questions, arranged in five groups:

- General questions (age, gender, parents' employment and occupation, using of media)
- Questions about the print media
- Questions about the radio
- Questions about television
- Questions about the Internet

The questionnaire was designed using: multiple choice answers, dichotomous questions (Yes or No answers), and open format questions (without predetermined set of responses). There was a paper edition, as well as an on-line questionnaire.

Data are presented in tables and graphs using absolute and relative number of incidences, the arithmetic mean with standard deviation, and the range of values. Difference analysis between the observed groups was performed by the chi-square test with the Yates's correction for small samples, and Fisher's exact test. As the level of statistical significance of differences, we used 95% of confidence interval or the value of the tests with $p<0.05$.

The analysis was performed using the statistical package for sociological research – IBM SPSS Statistics v 21.0 (Chicago, Illinois, the USA).

IV. RESEARCH PROCEDURE

Respondents were informed about the anonymous character of the survey. The questions were explained to younger respondents, and the questionairre with that group was conducted by asking each of them individual questions and sub-questions. Children used their own flow of speech. Their answers were recorded and interpreted.

The downsides of the survey were: low response rates to the online survey, answering survey questions by parents who were persistently trying to make an impression with erudite responses unlike responses of a child, and children's incomprehension of the term "print medium". Respondents did not consider that reading comics is reading the media, but in subsequent responses they answered what they read and use, and what they are interested in. This information is considered in further analysis of results.

The positive side was the fact that online respondents were not part of a group or an assembly that would give the same answers to questions as it would occur in an educational group, or class at school, where all the children know each other, socialize, have the same interests and fancy same things.

V. RESEARCH RESULTS

The research results started with answers to general questions. It was necessary to determine social characteristics: children's age, gender, parents' employment and parents' occupation. The following results are divided by type of media, starting with the print media, followed by the radio, television and the Internet. Each medium has 5 compulsory questions. Each question is shown individually as follows, in the same order as in the survey.

Question: How old are you?

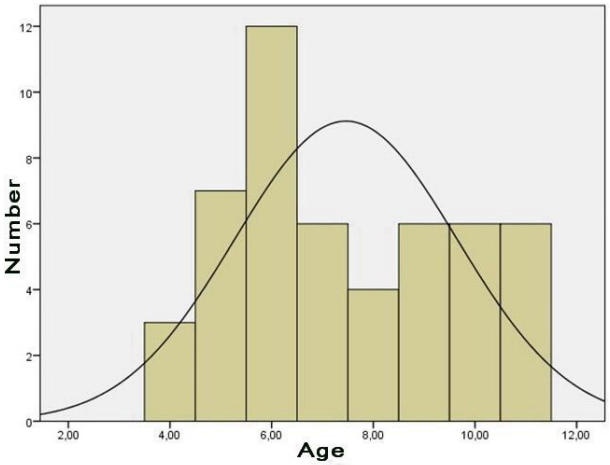


Figure 1. Age of respondents chart

The sample included 50 children aged 4 to 11, with a mean age of 7.5 ± 2.2. The respondents were divided into two groups on the basis of their age, in order to prove the hypothesis: children who can not read or who have just learned to read, aged 4 to 7 (28 or 56.0% of the total sample), and children who have already mastered reading and writing, aged 8 to 11 (22 or 44.0% of the total sample).

This research explores which media are used by children, which content they prefer in these media, how

much time they spend with the media, and what are the circumstances when they do it regarding their age, meaning if they are alone or accompanied by friends or under the supervision of parents. Media included in this research are print media, radio, television and the Internet.

Fifty respondents participated in the research "Children's Interest in the Media", including 56% male and 44% female children. Subjects were divided into two groups (aged 4-7 and 8-11) depending on their ability to read and write. For the purposes of this research, social characteristics were also determined. The results have shown that children live in families with both parents employed and mostly highly educated.

VI. TELEVISION AND THE INTERNET

Of all media included in this research, most commonly used were television and the internet (100.0%). The research explores what programmes and shows respondents prefer, who decides about the content, who they watch television with, and how much time they spend watching it. Children generally watch children's programmes, series, sports and music programme on television, while computer is mainly used for games. Older children show increased interest in YouTube and social networks. They use both media on a daily basis. The analysis results show that television is still more prevalent (around 10%), thus confirming its dominance. Regarding control of using or forbidding to use television or internet, the data show that parents influence in a large percentage on selection of television programmes in both groups. There is a significant difference between age groups, namely, older subjects watch television mostly alone (59.1%) and use the Internet alone (40.9%). They choose the contents on the Internet mostly themselves (90.0%).

Question: I usually watch TV

TABLE I. THE WAY OF WATCHING TELEVISION

Age		4-7 years	8-11 years	Total
With parents	N	23	5	28
	%	82,1	22,7	56,0
Alone	N	3	13	16
	%	10,7	59,1	32,0
With friends/peers	N	2	4	6
	%	7,1	18,2	12,0
Total	N	28	22	50
	%	56,0	44,0	100,0

$\chi 2=12,542$; $p=0,024$

Table I shows important questions: who respondents watch television with, whether they do it alone, in company of their parents, or perhaps with peers or friends. As it can be seen from Table I, there is statistically significant difference of younger respondents watching television mostly in the company of their parents (82.1%) whereas older respondents watch it alone (59.1%). Older children are expected to watch television alone.

Question: I usually use the Internet

TABLE II. THE WAY OF USING THE INTERNET

Age		4-7 years	8-11 years	Total
With parents	N	20	8	28
	%	71,4	36,4	56,0
Alone	N	7	9	16
	%	25,0	40,9	32,0
With friends/peers	N	1	5	6
	%	3,6	22,7	12,0
Total	N	28	22	50
	%	56,0	44,0	100,0

$\chi 2=25,421$; $p=0,001$

Table II shows the way respondents use the Internet, alone or under some form of supervision (parents) or with company (friends, peers). With a statistically significant difference, younger respondents use the Internet more in the company of their parents (71.4%), while older use it more independently (40.9%) or in the company of friends and peers (22.7%). It can be concluded that there is a substantial exposure to this medium and that parents believe that children should learn to use computers as early as possible. It can be assumed that a computer is an ordinary part of the household, just like a television set.

The first hypothesis, which suggests that children use the media every day, but under supervision of family, friends or other people, and that they read articles and literature, watch programmes and shows, and use the Internet in accordance with their age, has not been confirmed for both groups regarding television and the Internet, as these results indicate. The second hypothesis, which suggests that children use the media on a daily basis without supervision of family, friends or other people, and that they read articles and literature, watch programmes and shows and use the Internet in accordance with their age has not been confirmed for the younger age group, because they are still under supervision of adults who influence their choice. In the older group hypothesis was partially confirmed as subjects use television on a daily basis and mostly by themselves, although the

selection of programmes is not entirely theirs. As for the Internet, the hypothesis is also partly confirmed because the respondents choose themselves what to search / visit on the Internet on a daily basis (90%), and they mostly use it alone, unlike the younger group of respondents. The third hypothesis, which suggests that children use the media on a daily basis without guidance from family, friends or other people, and their reading articles and literature, watching programmes and shows, and using the Internet is not in accordance with their age, has not been confirmed for television and the Internet in the younger age group. Respondents are still subjected to adult supervision, although they use media on a daily basis and watch programmes that are not tailored to their needs. The hypothesis is confirmed for the older group, though.

VII. PRINT MEDIA

Concerning print media and radio the results were a bit different. In the younger group print media are read by 85.5%, while in the older age group 100.0% of respondents read them. The younger age group mainly read the media which suit their age, while older respondents read newspapers as well. Also the younger age group stated they prefer sports and informative content, while the older group showed increased interest in sports. In younger age group parents influence the selection of the content, while the older age respondents select the content mainly themselves and lend from the library as well.

Table III shows the type of print media that respondents most commonly used. As expected, there was statistically significant difference between age groups regarding the print media they read ($p < 0.01$), and we observe that most children read media in accordance with their age, although there are a few who read inappropriate contents. Certain number of respondents in older group (22.7%) read Slobodna Dalmacija (which can be assumed as their parents' influence), while the younger age group read the print media intended for their age with the exception of comics (8.3%).

Older age group read print media every day (54.5%) or more than twice a week (27.3%), while in the younger age group 25% read on a daily basis and 25% read more than twice a week. It may be observed that the first hypothesis is confirmative for the younger age group, taking into account their age. They read media designed for their age, their parents help them choose the type and content, and most of them read print media on a daily basis or more than twice a week. Therefore, the other two hypotheses are unconfirmed for this age group. As for the older age group, the first hypothesis is unconfirmed because children choose themselves what they want to read. Although parents are those who buy the print media, there is no such a parental control as with younger age group. Moreover, respondents of this group are interested in reading newspapers and magazines that are not appropriate to their age and and they use print media on a daily basis (54.5%). Although there are a few of them who are interested in inappropriate print media (assuming these are older participants in the group), most of them read books (probably for book reports, considering they are grades 1-5 of primary school), which confirms the second hypothesis. If there had been more inappropriate content for their age, then the third hypothesis would have been confirmed for the print media.

Question: What print media do you read?

TABLE III. PRINT MEDIA THAT SUBJECTS READ

Age		4-7 years	8-11 years	Total
Books	N	0	10	10
	%	0,0	45,5	21,7
Comics	N	2	3	5
	%	8,3	16,7	10,9
Garfield	N	2	0	2
	%	8,3	0,0	4,3
Tom and Jerry	N	2	0	2
	%	8,3	0,0	4,3
My Planet	N	0	3	3
	%	0,0	16,7	6,5
Picture Books	N	1	0	1
	%	4,2	0,0	2,2
Teen	N	0	2	2
	%	0,0	9,1	4,3
OK	N	0	3	3
	%	0,0	16,9	6,5
Smib	N	4	0	4
	%	16,7	0,0	8,8
Prvi izbor	N	1	0	1
	%	4,2	0,0	2,2
Slobodna Dalmacija	N	0	5	5
	%	0,0	22,7	10,9
Mikimax	N	2	2	4
	%	8,3	9,1	8,8
Novels	N	0	2	3
	%	0,0	9,1	6,5
Moj izbor	N	0	1	1
	%	0,0	4,5	2,2
Total	N	24	22	46
	%	52,2	47,8	100,0

$\chi 2=23,254$; $p=0,001$

VIII. RADIO

We wanted to learn whether the respondents listen to the radio, which radio shows they listen to, who helps

them selecting radio shows, whether they listen to the radio alone and how much time they spend listening to it.

Question: I usually listen to the radio

TABLE IV. THE WAY OF LISTENING TO THE RADIO

Age		4-7 years	8-11 years	Total
With parents	N	16	5	21
	%	94,1	27,8	60,0
Alone	N	1	9	10
	%	5,9	50,0	28,6
With friends/peers	N	0	4	4
	%	0,0	22,2	11,4
Total	N	17	18	35
	%	60,7	81,8	100,0

$\chi 2=18,2574$; $p=0,001$

Table IV shows whether respondents listen to the radio alone or accompanied by their parents, friends or someone else. Similar to using print media, young children mostly listen to the radio in the company of their parents (94.1%), whereas older children tend to listen to the radio alone (50%), with statistically significant difference between the observed age groups. One can assume that the radio is mostly listened to as a secondary activity while something else is being done, which means that music is preferable content.

Related to the younger age group, 60% of respondents listen to the radio, out of who 94% listen to the radio in the company of their parents, the content being mostly music. 58% of respondents listen to the radio on a daily basis and their parents choose what to listen (70%). In older age group, 81% of respondents listen to the radio, most of the time alone (50%), the content being music and sports, but also cultural and news programmes. 44% of them choose for themselves what to listen and 50% of respondents listen to the radio on a daily basis. The above shows that the first hypothesis is confirmed for the younger age group for everything assumed in the hypothesis: children use the media every day, but under supervision of family, friends or other people, they also consume the contents in accordance with their age. In older age group, the first hypothesis was not confirmed because there was a significant difference between their responses and those of younger age group. The facts that they mostly listen to the radio alone, choose for themselves what to listen, and listen to the radio on a daily basis, confirmed the second hypothesis that children use the media on a daily basis but without supervision of a family, friends or other people, and consume the contents in accordance with their age. The contents they referred to were mostly music and some cultural and news programmes, but not enough to confirm the third hypothesis which suggests that children use the media every day, but without control of a family, friends or other people, and read articles or literature, watch programmes and shows and use the Internet content which is not in accordance with their age.

The observed differences between these two age groups were in the control of their use of media, the time spent using the media, and their preferences concerning content. The results have revealed that as children get older, parents have less influence on their children's use of the media. In the younger group parents still have a certain influence, which changes as their children start school where environment takes over and begins its influence.

IX. CONLUSION

It can be concluded that sociological characteristics did not play an important role in this research, although family environment and family relationships should be explored more meticulously in order to reveal a significant impact of the media and media relations. This is a consequence of the lack of research on parents' education and their role as educators. The research did not confirm the incorrect assumption of highly negative impact of the media. It was concluded that much younger children start using the media, especially the Internet, which was unexpected at that early age. The scope of effects the media will have on children, depends primarily on their usage. Children should be adequately educated and guided for using the media, and this is primarily responsibility of parents, then schools which still show inadequate concern about the media, and finally the rest of society, especially those who create media products.

REFERENCES

[1] Bertrand Claude J.: Deontologija medija, Zagreb, 2007, str 20-21.

[2] Nadrljanski, M.: Komunikologija i menadžment, Istraživanje komunikoloških aspekata obrazovanja menadžera, Split, 2010, str. 34-35.

[3] Erjavec, K.; Zgrabljić, N.: Odgoj za medije u školama u svijetu, Hrvatski model medijskog odgoja, Medijska istraživanja, (god, 6. br. 1), 2000, str. 89.

[4] Nadrljanski, Mila.; Nadrljanski, Đorđe.; Bilić, Mirko.: Digitalni mediji u obrazovanju, 2007, str 527-537., 1.međunarodna znanstvena konferencija "The Future of Information Sciences (INFuture2007): Digital Information and Heritage": zbornik radova / Seljan, S.; Stančić, H.; (ur.). - Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu, Središnji državni ured za e-Hrvatsku, Zagreb , 2007., str 527-537.

[5] Miliša, Z.; Tolić, M.; Vertovšek, N.: Mediji i mladi, Prevencija ovisnosti o medijskoj manipulaciji, Sveučilišna knjižara, Zagreb, 2009.

[6] Klaić, B.: Riječnik stranih riječi, Nakladni zavod matice hrvatske, Zagreb, 1981.

[7] Sambol, D.: Kako mediji utječu na djecu, Poslovni savjetnik, Marketing, lipanj 2010, str 56-57. (20.10.2014) URL: http://www.poslovni-savjetnik.com/sites/default/files/dir_marketing/PS%2067.pdf]

[8] Kurtović, D.: Tiskani medij, knjižničarka IV. OŠ Bjelovar, 27.4.2010. URL: http://www.scribd.com/doc/241979062/Tiskani-mediji-Dejana-Kurtovi%C4%87-ppt (9.2014)

[9] Aufderheide, P.: A report of the National leadership conference an media, 1992, litetacy, URL: http.//www.interact.uoregon/MediaLit/FA/aufderheide/report.html.

[10] Povijest medija- od tiska do interneta URL:http://ahyco.uniri.hr/Seminari2009/povijest_medija/index.html#nogo (20.09.2014)

URL: http://ahyco.uniri.hr/Seminari2009/povijest_medija/Televizija/tvPovijest.html URL: http://ahyco.uniri.hr/Seminari2009/povijest_medija/Internet/internetPovijest.html

URL: http://ahyco.uniri.hr/Seminari2009/povijest_medija/Tisak/tisakPodjela.html

URL: http://ahyco.uniri.hr/Seminari2009/povijest_medija/Radio/radioPovijest.html

URL: http://ahyco.uniri.hr/Seminari2009/povijest_medija/Radio/radioDanas.html

[11] URL: http://hr.wikipedia.org/wiki/Mediji#Utjecaj_na_dru.C5.A1tvo (20.9.2014.)

[12] URL: http://hr.wikipedia.org/wiki/Mediji (20.9.2014)

[13] URL: http://hr.wikipedia.org/wiki/Masovni_mediji (20.9.2014)

[14] URL: http://hr.wikipedia.org/wiki/Anketa (29.9.2014.)

[15] URL: http://hr.wikipedia.org/wiki/Televizija (7.10.2014.)

[16] URL: http://hr.wikipedia.org/wiki/Tiskarski_stroj (18.10.2014)

[17] URL: http://hr.wikipedia.org/wiki/Radio (18.10.2014)

[18] URL: http://hr.wikipedia.org/wiki/VHS (18.10.2014)

[19] URL: http://hr.wikipedia.org/wiki/DVD (18.10.2014)

[20] URL: http://hr.wikipedia.org/wiki/Internet (18.10.2014)

[21] URL: http://hr.wikipedia.org/wiki/Smje%C5%A1ko

[22] URL: http://hr.wikipedia.org/wiki/Elektronicki_mediji (18.10.2014)

[23] Zakon o medijima: Opće odredbe, članak 2, Broj: 01-081-03-3343/2, Zagreb, 7. listopada 2003.

[24] URL: http://narodne-novine.nn.hr/clanci/sluzbeni/306926.htm (19.10.2014)

[25] URL: http://bosnian.irib.ir/teme/item/139242-utjecaj-medijana-odgoj-djece]subota 30 juni 2012 na net (19.10.2014)

[26] URL: http://hr.wikipedia.org/wiki/Knjiga (19.10.2014)

[27] URL: http://hr.wikipedia.org/wiki/Hrvatska_radioteleviziji By_Božena_Begović početak_ emitiranja_ programa. (19.10.2014)

[28] URL: http://en.wikipedia.org/wiki/Samsung (22.10.2014)

[29] URL: http://hr.wikipedia.org/wiki/Hipertekst (19.11.2014)

[30] URL: http://hr.wikipedia.org/wiki/Internet#Internetski_slang (19.11.2014)

[31] Službene stranice: Djeca Medija, Društvo za komunikacijsku i medijsku kulturu,

URL: http://www.djecamedija.org/?p=671 8.9.2012, (19.11.2014)

URL: http://www.djecamedija.org/?p=666 8.9.2012

URL: http://www.djecamedija.org/?p=657 8.9.2012

URL: http://www.djecamedija.org/?p=604 8.9.2012

[32] URL: http://en.wikipedia.org/wiki/SOS (19.11.2014)

[33] URL: http://en.wikipedia.org/wiki/Reality_television (19.11.2014)

Slavomir Vukmirović
Economics Faculty
University of Rijeka, Croatia
vukmirovics@gmail.com

Marko Čičin-Šain
Jadrolinija, Rijeka, Croatia

Ivan Host
KBC Rijeka, Croatia

Modeling and analysis of the polyhedral pyramid structures by method of modified traveling salesman problem (TSP)

Abstract

In this paper we discuss the features and benefits of usage the concepts of polyhedral pyramidal structure in design and analysis of flexible and adaptive organizational structures. Pyramidal polyhedra are considered as a network consisting of nodes and flows where nodes correspond to points of polyhedra, and edges correspond to flows. In this concept flows include all possible links between the points of polyhedra. Polyhedral pyramid structure is considered from the perspective of the possibilities for flexible and adaptive connecting between nodes with the possibilities of achieving more combinations that generate optimum solutions, related to optimal organizational structure. In considering of dynamic potential of pyramidal polyhedral structure uses a modified method of traveling salesman problem (TSP).

1. Introduction

This paper analyzes advantages and possibilities of use of the concepts of polyhedral pyramid in designing and analyzing flexible and adaptive organizational structures.

A regular polyhedron is an angular geometric shape whose all sides (surfaces) are made up of congruent regular polygons. A regular polygon is a planar figure with all sides (edges) and all angles congruent.

In this paper, we discuss the pyramidal polyhedron as a network that consists of nodes and flows, wherein the nodes correspond to the points in polyhedron and the polyhedron edges correspond to the flows. The flows indicate all the possible links between the polyhedron points. This means that the analysis of polyhedral structures covers both the edges that represent external links and internal links such as diagonals of the base and height of a polyhedron.

In this way, with the above mentioned attributes of regular polyhedra describing static features (points, nodes), we create the conditions for addition of a new attribute that is related to the flows (distances) between the points (nodes) of the polyhedra and indicates a dynamic aspect of their structures.

In considering flows as dynamic features of polyhedral pyramid structures we use a modified method of traveling salesman problem. From the viewpoint of combinatorial optimization, Travelling Salesman Problem (TSP) can be formulated as a given set of **n** cities (nodes, polyhedron points) C and pairwise distances $\mathbf{d_{ij}}$ between the nodes to be the tours from city **i** to city **j**. TSP is to find the best possible way of visiting all the cities exactly once, starting from a given city, determining a route whose total length is minimal. When applied to polyhedral pyramid structures, optimal solution is the minimum sum of all the lengths (distances) between the points.

In practice, Travelling Salesman Problem is focused on a single optimal solution and refers mainly to models with a large number of nodes and flows. In this paper, the models of polyhedral pyramid structure are restricted to a limited number of nodes, where the focus is not on a single optimal solution, but the ability to achieve a larger number of optimal solutions (combinations).

Based on the above, a research hypothesis can be formulated: "The number of optimal solutions and the share of optimal solutions in the total number of permutations (all possible combinations of solutions) in the calculation of flows between the network nodes is an important indicator of flexibility and adaptability of the organizational structure of a certain type of pyramidal polyhedron."

Flexibility and adaptability of transportation networks can be achieved by optimizing routes for the movement of vehicles between a starting point and a certain number of destinations. The most important operational decision related to transportation in a supply chain is the routing and scheduling of delivers. An analogy can be made between transportation network and polyhedral structure which implies that the number of optimal solutions may be considered a factor of flexibility and adaptability of polyhedral structures.

In this context, we have identifed two indicators of flexibility and adaptability of polyhedral structures from the viewpoint of the traveling salesman problem: 1) NO - Number of optimal solutions (NO) and 2) Flexibility Coefficient of Polyhedral Structure (FCPS), which is calculated as the ratio of the number of optimal solutions (NO) and the total number of possible values calculated (NT).

Methodological frame of use of Visual Basic as a development tool in the visual modelling of exhaustive search algorithm in VBA for Excel can serve as an incentive in creating new highly sophisticated algorithms, which will enable us to compute multiple optimal solutions [6].

2. Travelling Salesman Problem (TSP)

Travelling Salesman Problem (TSP) can be formulated in the following way: TSP, is to find the best possible route based on given criteria, where minimum distance between cities or minimum travelling cost are the most commonly used optimization criteria. TSP, is to visit all the cities and return to the starting point under the condition of visiting each city only once. The travelling salesman problem (TSP) is one of the most studied problems in management science. Optimal approaches to solving travelling salesman problems are based on mathematical programming. But in reality, most TSP problems are not solved optimally. When the problem is so large that an optimal solution is impossible to obtain, or when approximate solutions are good enough, heuristics are applied. Two commonly used heuristics for the travelling salesman problem are the nearest neighbour procedure and the Clark and Wright savings heuristic [3].

In terms of combinatorial optimization, the Travelling Salesman Problem (TSP) can be formulated in the following way: Given a list of ***n*** cities C and distance $\mathbf{d_{ij}}$ from city ***i*** to city ***j***; TSP, is to find the best possible way of visiting all the cities by visiting each city only once finding minimum total travel distance. In analogy to the above definition, the following formulations are valid: 1) Travel distance or distance between cities is symmetric: (1) $d_{ij} = d_{ji}$; 2) Final list of cities is defined as incoming variable by the formula C = (c_1 … c_n), while distance matrix containing distance between city c_i and city c_j for each pair i,j is defined by $d(c_i,c_j)$. In view of the fact that distance matrix is symmetric, the following equation is valid: $c_{ij} = c_{ji}$; 3) Permutations or in other words all permuted relations that can be achieved for a given number of cities are computed as resulting variables. Permutations p(1), …, p(n) in the list 1, …, n are calculated and compared to give the minimum sum [1], [4].

It follows from the above, that the following equation can represent the travelling salesman problem [1]:

$$\sum_{1 \le i \le n-1} d(c_{\pi(i)}, c_{\pi(i+1)}) + d(c_{\pi(n)}, c_{\pi(1)})$$

Above equation is the sum of length of routes that begin in city $c_{p(i),\ i=1}$ visiting each city in defined pattern ($c_{p(i),}$ $c_{p(i+1)}$) and returning to the starting point $c_{p(1)}$. This algorithm gives travel distance for all possible routes and finds the shortest route. In addition, the number of possible routes is the factorial of *n* cities, or number of permutations of *n* elements.

3. Square Pyramid Model in Generating Multiple Optimal Solutions

Model of a square pyramid base allow us to demonstrate and analyze possibilities of generating a large number of optimal solutions that have equal or similar values [7]. The base of a square pyramid is a square having equal sides and equal angles. The base of a square pyramid is also made of diagonals that intersect at right-angles. The intersection of diagonals is the central node of the base that is equally distant from all other nodes. Square Pyramid Model is described in the Figure 4. The nodes and arcs have been marked up according to the Figure 1. Table 1 describes an example of use of the square pyramid model in resolving and analyzing the travelling salesman problem.

Figure 1 is an example of use of the square pyramid model in resolving and analyzing the travelling salesman problem. Table 2 shows input data and the solution obtained in VBA for Excel [5]. The cell range K1:O5 contains transportation route lengths (arcs) that correspond to the dimensions indicated in the pyramid diagram in the Figure 4. The cell range J1:J24 offers all possible solutions, which have been categorized from the lowest to the highest. The Table 2shows that there are 8 optimal solutions which result in equal minimum values of the total transportation route (22). Considering the matrix from the cell range K1:O5 is symmetric, it follows that the model has generated 4 optimal solutions. Figure 2 is a graphical display of optimal solutions for the square pyramid model [7].

Figure 1. Figure 1. Model of a square pyramid base

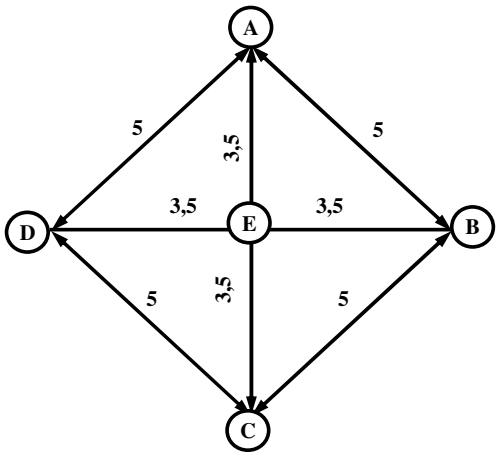
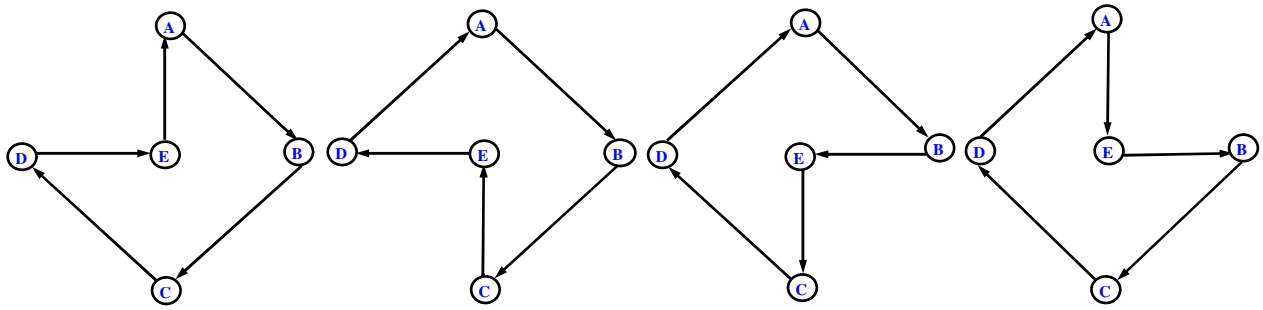


Table 1: Finding solution by use of Visual Basic for Excel

	A	B	C	D	E	F	J	K	L	M	N	O	P	Q
1	1	2	3	4	5	1	**22**	1,0	5,0	7,0	5,0	3,5	**1**	**A**
2	1	2	3	5	4	1	**22**	5,0	1,0	5,0	7,0	3,5	**2**	**B**
3	1	2	5	3	2	1	**22**	7,0	5,0	1,0	5,0	3,5	**3**	**C**
4	1	5	2	3	4	1	**22**	5,0	7,0	5,0	1,0	3,5	**4**	**D**
5	1	5	4	3	2	1	**22**	3,5	3,5	3,5	3,5	1,0	**5**	**E**
6	1	4	5	3	2	1	**22**	**1**	**2**	**3**	**4**	**5**		
7	1	4	3	5	2	1	**22**	**A**	**B**	**C**	**D**	**E**		
8	1	4	3	2	5	1	**22**							
9	1	2	5	4	3	1	**24**							
10	1	2	4	3	5	1	**24**							
23	1	3	5	2	4	1	**26**							
24	1	3	2	4	5	1	**26**							

Figure 2: Optimal Solution Graphical Display



Source: Vukmirović, S , Et. Al. (2014)

Square Pyramid Model in Generating Multiple Optimal Solutions is presented in the form of a video animation [8] that has been placed on the website **www.infosustav.com/video/squarepyramid.mp4**

4. Analysis of the polyhedral structure by usage the method Traveling Salesman Problem (TSP) in the case of pyramidal octahedron

An n-gonal bipyramid or dipyramid is a polyhedron formed by joining an n-gonal pyramid and its mirror image base-to-base. The referenced n-gon in the name of the bipyramids is not an external face but an internal one, existing on the primary symmetry plane which connects the two pyramid halves. The face-transitive bipyramids are the dual polyhedra of the uniform prisms. A bipyramid can be projected on a sphere or globe as n equally spaced lines of longitude going from pole to pole, and bisected by a line around the equator. Bipyramid faces, projected as spherical triangles, represent the fundamental domains in the dihedral symmetry. The tetragonal bipyramid with identical edges, or regular octahedron, counts among

the Platonic solids. The regular octahedron is a special case of tetragonal bipyramuid [2].

A comparison of the two constructions of pyramidal octahedron is shown in Figure 3 and Figure 4. The first construction was shown in Figure 3 and second construction of pyramidal octahedron was shown in Figure 4.

The Figure 3 depicts a regular octahedron with all edges equal. As can be seen, there are 12 equal edges: AB, AD, BC, CD, AE, BE, CE, DE, AF, BF, CF and DF. On the right side of Table 2 are distances between the points, based on which it can be concluded that all the edges are equal. On the left side of Table 2 are calculations of the total relation length (route) between all of the points of the regular octahedron made by method of traveling salesman problem.

The table 2. shows that the calculated total number of all the possible permutations of points is 720, and the total number of optimal solutions is 16. The flexibility coefficient of polyhedral structure (FCPS) of a regular octahedron is calculated as the ratio of the number of equal values of the optimal solutions (NO) and the total number of all possible solutions (NT). In our case, the generic equation FCPS = NO/NT for regular octahedron is 16/720. FCPS = 16/720 = 2.22%

Figure 3 Regular Octahedron – First Construction

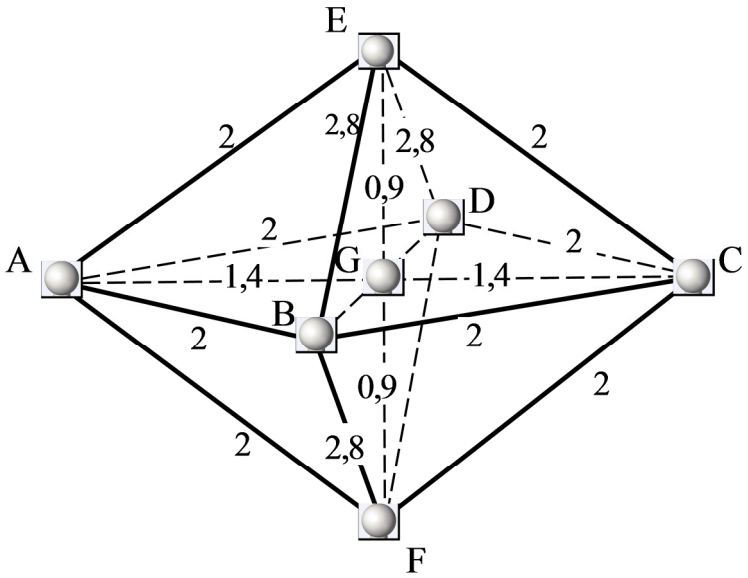


Table 2 Calculations

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	1	2	3	4	5	7	6	1		**11,7**	10	2	2,82	2	2	2	1,41	**1**	**A**
2	1	2	3	4	6	7	5	1		**11,7**	2	10	2	2,82	2	2	1,41	**2**	**B**
3	1	2	3	6	7	5	4	1		**11,7**	2,82	2	10	2	2	2	1,41	**3**	**C**
4	1	2	3	5	7	6	4	1		**11,7**	2	2,82	2	10	2	2	1,41	**4**	**D**
5	1	2	5	7	6	3	4	1		**11,7**	2	2	2	2	10	1,7	0,85	**5**	**E**
6	1	2	6	7	5	3	4	1		**11,7**	2	2	2	2	1,7	10	0,85	**6**	**F**
7	1	6	7	5	2	3	4	1		**11,7**	1,4	1,4	1,4	1,4	0,85	0,85	10	**7**	**G**
8	1	5	7	6	2	3	4	1		**11,7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**		
9	1	5	7	6	4	3	2	1		**11,7**	**A**	**B**	**C**	**D**	**E**	**F**	**G**		
15	1	4	3	2	6	7	5	1		**11,7**									
16	1	4	3	2	5	7	6	1		**11,7**									
17	1	2	3	4	5	6	7	1		**11,95**									
18	1	2	3	4	6	5	7	1		**11,95**									
719	1	3	6	5	2	4	7	1		**14,15**									
720	1	3	5	6	2	4	7	1		**14,15**									

Source: Authors

Figure 4 depicts a polyhedral structure of a modified regular octahedron. The specificity of the modified octahedron as shown in the figure is that the lengths of the base edges differ from the values (lengths) of the lateral edges, and the values of the base diagonals of the base are equal to the values of the lateral edges.

On the right side of Table 3 are distances between the points based on which it can be concluded that the values (length) of the lateral edges of AE, BE, CE, DE, AF, BF, CF and DF are equal to the values of the diagonals of the base of AD and BC. On the left side of Table 3 are calculations of the total relation length (total route) between all of the points of the octahedron made by method of traveling salesman problem (TSP).

The table indicates that the total number of all the possible permutations of points is 720, and the calculated number of optimal solutions is 96. The flexibility coefficient of polyhedral structure (FCPS) of a regular octahedron can be calculated as the ratio of the number of equal values of the optimal solutions (NO) and the total number of all possible solutions (NT). In our case, the generic equation FCPS = NO/NT for regular octahedron is 96/720. FCPS = 96/720 = 13.33%

Figure 4. Modified Octahedron – Second Construction

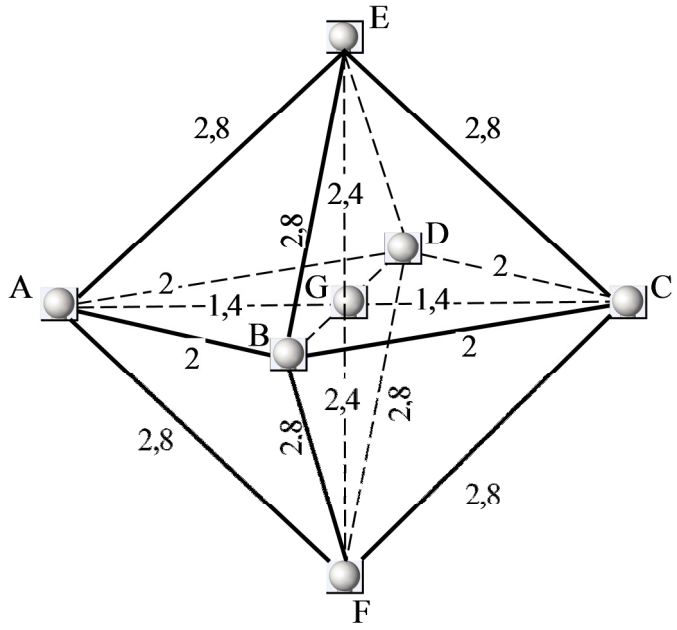


Table 3 Calculations

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	1	7	2	6	3	4	5	1		**16,1**	10	2	2,82	2	2,82	2,82	1,41	**1**	**A**
2	1	6	2	3	7	4	5	1		**16,1**	2	10	2	2,82	2,82	2,82	1,41	**2**	**B**
3	1	6	2	7	3	4	5	1		**16,1**	2,82	2	10	2	2,82	2,82	1,41	**3**	**C**
4	1	6	2	3	5	4	7	1		**16,1**	2	2,82	2	10	2,82	2,82	1,41	**4**	**D**
5	1	7	2	5	3	4	6	1		**16,1**	2,82	2,82	2,82	2,82	10	4,88	2,44	**5**	**E**
6	1	6	2	5	3	4	7	1		**16,1**	2,82	2,82	2,82	2,82	4,88	10	2,44	**6**	**F**
7	1	5	2	6	3	4	7	1		**16,1**	1,41	1,41	1,41	1,41	2,44	2,44	10	**7**	**G**
8	1	5	2	3	6	4	7	1		**16,1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**		
95	1	6	3	7	2	5	4	1		**16,1**	**A**	**B**	**C**	**D**	**E**	**F**	**G**		
96	1	7	3	2	6	4	5	1		**16,1**									
97	1	2	3	6	4	5	7	1		**16,3**									
98	1	2	3	6	4	7	5	1		**16,3**									
719	1	3	2	4	7	5	6	1		**19,2**									
720	1	3	2	4	5	6	7	1		**19,2**									

Source: Authors

5. Conclusion

Using the method of traveling salesman problem (TSP) based on the exhaustive search algorithm, and a graphical modeling in spreadsheet interface, we have analyzed flexibility and adaptability of tetragonal bipyramid (pyramidal octahedron) structures. Flexibility and adaptability were analyzed through the comparison of the flexibility coefficients of the two tetragonal bipyramid models constructed.

The first structure represents a regular octahedron with the lateral edges equal to the base edges. The second structure is an octahedron with the lateral edges equal to the diagonals of the base, but different from the base edges. A comparative analysis of the results reveals that the flexibility coefficient of the second model constructed (13.3%) is six times greater than the flexibility coefficient of the first model constructed (2.2%).

The reason for such a large difference in favor of the second model constructed may be relationship between internal and external links. In the first model, all external links (edges) are equal, but there is no equivalence between external and internal links (flows). This means that there is equivalence between external dimensions (links), but there is no equivalence between external and internal links. In the second model, we observe equivalence between external and internal links, which means that structure flexibility applies not only to the outer part (surface area), but the entire area (volume) of the pyramidal octahedron.

Literature:

1. Advanced Algorithms and Complexity, CM30073 http://people.bath.ac.uk/masnnv/Teaching/AAlg12_1.pdf
2. Bipyramid, http://math.wikia.com/wiki/Bipyramid
3. Heizer, J., Render, B.: Operations Management, seventh edition, Pearson Prentice Hall, ISBN 0-13-120974-4, New Jersey, 2004.
4. Lawler, E.L et al., The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization, Wiley, Chichester, 1985
5. Vukmirović, S., Kovač, D., Mežnarić, I., Object-oriented programming for graphical modeling on the example of logistics network design, MIPRO, 2013 Proceedings of the 36th International Convention, Opatija, 2013
6. Vukmirović,S., Pupavac, D., The Travelling Salesman Problem in the Function of Transport Network Optimalization, Interdisciplinary Managemet Research IX, Josip Juraj Strossmayer University in Osijek, Fakulty of Economics in Osijek, 2013.
7. Vukmirovic, S.; Pupavac, D.; Drašković, M., Identification and analysis of specific problems in the rationalization of macro-level logistics networks using travelling salesman problem, Montenegrin Journal of Economics, Voll. 10, No 2, 2014.
8. Vukmirović, S., Modeling and analysis of the polyhedral pyramid structures by method of modified traveling salesman problem (TSP), Video Animations, 2015., www.infosustav.com/video/squarepyramid.mp4

Slavomir Vukmirović
Zvonko Čapko
Ana Babić
Economics Faculty
University of Rijeka, Croatia
vukmirovics@gmail.com
zvonko.capko@efri.hr
ana.babic@efri.hr

Geometric programming in designing of mental models on the example of strategic thinking between synergies competition and cooperation

Abstract

In this paper we consider the mental models in context of encouraging innovative and creative development and shift in strategic thinking. Mental models encourage multiperspectival approach to the process of representing imagination in strategic thinkng. Visual designing of mental models can show how a new way of thinking may works in different contexts and situations. Geometric programming is considered in the function of mental models designing. The model of lemniscate is used in visualization and modeling of synergies between competition and cooperation in business in the context of strategic thinking.

1. Introduction

Mental models can serve as a useful tool to develop, integrate and achieve synergies among constructive and innovative concepts and ideas while fostering creativity, innovation and freedom of expression. Using mental models can be achieved a number of positive effects, such as [3]:

- Storing information systematically and comprehensively in long-term memory, not cram the information;
- Organizing client information;
- Finding inspiring and effective solutions to each business problem;
- Preparing meetings and organizing business activities;
- Delivering concise, clear and interesting presentations;
- Developing a business plan;
- Organizing conferences and business meetings;
- Improving workspace;
- Optimizing and improving their business.

The Introduction is the first part of this paper. In the second part of this paper, titled 'Mental models and programming', we define and describe mental models and discuss the features and advantages of programming within the context of abstract thinking and shaping mental models. In the third part of this paper, titled 'Competition and cooperation", the concepts of competition and cooperation are analyzed by comparative and complementary methods, and the concept of coopetitive model is considered in the context of linking and achieving synergy between competition and cooperation. In the fourth part, titled 'Designing a mental model of synergy between competition and cooperation', we define elements of a mental model and describe a geometric model for linking and acheiving synergy between competition and cooperation, we define and analyze how flows of elements of mental models are linked, and describe a complete mental model process for achievement of synergy between competition and cooperation. The Conclusion is a brief summary of the paper's major points.

2. Mental models and programming

A context for learning programming in mental model development will facilitate improvement of personal metacognitive skills and development of abstract thinking also in areas outside programming.

In the development of programming skills is very important to recognize mental models of basic programming concepts held by his/her students. Learning programming involves construction of viable mental models of basic programming concepts. To improve mental models it is suggested such an approach to teaching programming that emphasizes constructivism, not objectivism. The assumption is that many of students, before enrolling in introductory computer programming courses have already deeply rooted ideas about some computer concepts such as assigning command. Constructivism claims that traditional teaching approach which is based on teaching lessons and textbooks is too passive and is not doing enough to challenge pre-defined ideas and assist students in the development of viable mental models [1]. Constructivism advocates active construction of

students' knowledge by combining experience with existing cognitive structures.

Teaching strategy that specifically challenges existing ideas is Festinger's cognitive conflict strategy. This strategy may encourage developers of mental models to use programming in their understanding thus promoting improvement of non-viable models. The cognitive conflict strategy is one of the key teaching strategies based on constructivism, and developed on the assumption that the prior knowledge and current beliefs might affect the way students acquire new knowledge and construct new cognitive structures. Cognitive conflict is a state in which the developer perceives a gap between his/her cognitive structure and external conditions, or between parts of his/her cognitive structure. A learning model that integrates cognitive conflict strategy with visualizing algorithms proved to be effective for improving the interest and involvement in constructing viable mental models [1] .

Constructivism claims that traditional teaching approach which is based on teaching lessons and textbooks is too passive and is not doing enough to challenge pre-defined ideas and assist students in the development of viable mental models. On the other hand, constructivism argues that developer actively construct their knowledge combining experimental work with existing cognitive structures [1].

The key teaching strategy is based on constructive perspective of cognitive conflict, which clearly challenges existing ideas to encourage mental model developers to recognize errors in their understanding and push them to improve their non-viable models. Cognitive conflict by itself may not be sufficient to achieve changes in non-viable models. The developer of mental model must be supported in the process of creating viable models; the concepts should be delivered in a manner and style to allow for true constructivism of dependent models.

3. Coopetitive model

Linguistically speaking *coopetition* is neologism coined to describe cooperative competition, which are two well known and often used words in economic terminology: *cooperation* and *competiton*. The development of information systems contributed to the emergence of a new paradigm in which information, networking and time are key factors that influence running a business. Information is richer in quality and quantity which encourages collaboration between market participants. The world-wide connectivity between individual networks via Internet, Intranet or Extranet eases barriers to trade for new competitors attempting to enter the market and is fostering global hyper competition. The ability to quickly access market data reduces the asymmetry of information between the active and potential (new) market match participants in which process the ability to sense and respond quickly to market changes emerges as a necessity. As a result, companies are facing a high level of competition in the market where new rivals appear overnight in unexpected places or, until recently, stable markets. This change in business dynamics promotes collaborative value creation in the coopetition model which is more necessary than ever [2].

In a model of balancing and sinergy between competition and cooperation, a business system is discussed by considering how each one integrates into the overall relevant environment which includes potential and active business partners. It is important to analyze business environment as a network in which the company functions in which process a good starting position for business is being sought. The strategic use of information systems extends the concept of customer orientation in such a way that the company is simultaneously working with customers and suppliers on the principle of interaction and joint creativity. In this process the totality of business networks and relationships that include relationships with suppliers and third-party service providers is considered.

In contemporary business paradigms, business entities are only part of the ecosystem in which they must develop if they want to survive, because the economy also consists of unpredictable organisms in constantly changing environment. In such an environment business organizations should have a range of different relationship types in their portfolio that can change at different time points from competition, through coexistence to cooperation, or simultaneous existence of both - competition and cooperation.

Coopetitive approach, therefore, is based on the belief that the organization can be successful through synergy between cooperation and competition with the others at the same time. It is important that the company recognizes that customers, suppliers, distributors, competitors and other organizations are equal associates and partners in the process of value creation. Establishing and maintaining relationships with customers and partners is a philosophy or strategy that is vitally important for competitive progress and economic development.

4. Object oriented program Visual Basic for Application (VBA) in geometric programming and designing mental model of synergy between competition and cooperation

In geometric programming and designing mental model of synergy between competition and cooperation program was used object oriented program Visual Basic for Application (VBA) in Excel spreadsheet interface.

Object oriented programming for graphical modeling encourages integration and interaction of computer programming and graphical modeling knowledge, where psychomotor skills needed in traditional hand drawing and modeling are replaced with programming skills. Visual Basic for Excel as an object-oriented visual programming language used for integrated graphical modeling and programming [5]

enabling us to better understand and visualize the process of geometric modeling. The usage of programming language VBA in geometric programming and designing mental model of synergy betweeen competition and cooperation can be presented on the example of creating lemniscate. Program for drawing a lemniscate by using VBA in spreadsheet interface is shown in Figure 1. In the right part of figure 1 are shown the coordinates of lemniscate.

Figure 1 Program for drawing a lemniscate

```
Private Sub Lemniscate_Click()
For j = 1 To 2
For i = 2 To 31
y = 20
x = 20
a = Cells(i, 6) * 1.2
b = Cells(i, 7) * 1.2
v = i - 1
   With Cells(y, x)
        lbx = .Left
        lby = .Top
    End With
    If j = 1 Then
    a = a
    End If
    If j = 2 Then
    a = -a
    End If
Shapes.AddShape(1, lbx + a - 5, lby + b, 10, 10).Select
Selection.ShapeRange.Line.Weight = 2#
Selection.ShapeRange.Line.ForeColor.RGB = RGB(0, 0, 0)
Selection.ShapeRange.Fill.ForeColor.RGB = RGB(0, 255, 255)
Selection.ShapeRange.Fill.BackColor.RGB = RGB(0, 255, 255)
Next i
Next j
End Sub
```

	E	F	G
1	**No**	**X**	**Y**
2	1	3	3
3	2	8	8
4	3	14	13
5	4	20	19
15	14	98	11
16	15	100	0
17	16	-100	0
18	17	-98	-11
19	18	-93	-20
28	27	-20	-18
29	28	-14	-13
30	29	-8	-8
31	30	-3	-3

The entire process of the designing and development of mental models using VBA program and geometric programming is presented in the form of a video animation [6] that has been placed on the website

www.infosustav.com/video/mentalmodel1.mp4

According to video animation can be presented methodological framework of designing a mental model of synergy between competition and cooperation, which is described in the next section.

4. Designing the mental model of synergy between competition and cooperation

In designing the mental model of synergy between competition and cooperation has been proposed a methodological framework:

1. Defining the elements of a mental model
2. Designing a geometric model for linking and acheiving synergy between competition and cooperation,
3. Defining and analyzing how flows of elements of mental models are linked,
4. Analyzing and interpretation a complete mental model process for achievement of synergy between competition and cooperation

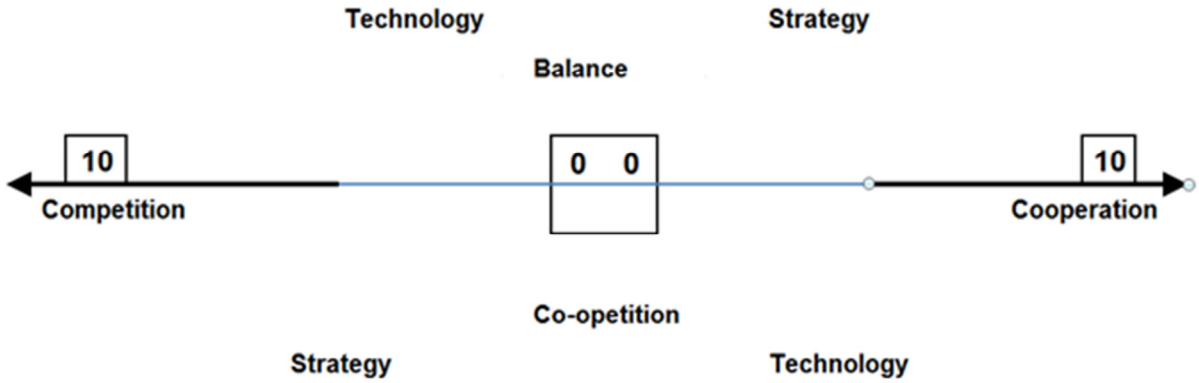
4.1. Elements of a mental model of synergy between competition and cooperation

The first step proposes basic elements of a mental model of synergy between competition and cooperation: strategy, technology, competition, cooperation, balance, coopetition and rating scale. The elements have been positioned virtually using a rating scale. At the center of the rating scale 0 represents centre of gravity that virtually maintains the balance

10 to 10. The model also shows the range of a competition rating scale of -10 to 0, and the range of a cooperation rating scale of 0 to 10. The range of the

and cooperation in a business system.

Figure 1. Elements of mental model



Source: Authors

The model reveals mutually conflicting features of contrariety and complementarity between competition and cooperation. Coopetition indicates that there exists a synergistic interaction between competition and cooperation. A strategy shapes concepts and methods of implementation, development, use and maintenance of competition and cooperation. The elements of mental models will be briefly described below.

Business **strategy** includes objectives with respect to products, services and markets for a certain strategic business unit, and points to the future activities of the company in specific sectors.Functional strategies focus on maximizing resource efficiency and integrating functional activities to facilitate the implemenation of corporate and business strategies.After selecting a strategy on how a company endeavors to accomplish set goals, the company can proceed to the next stage of preparing procedures, programs, projects and tactical plans required to implement the strategy.

Technology can be defined as the totality of knowledge, methods, techniques and tools used in the manufacturing process, a means to transform inputs into outputs and knowledge on how to do something. Technology may have a positive influence on competitive advantage, if it has a significant role in determining the relative cost position or during differentiation. Since technology is included in each value added activity and is involved in the process of

between competition and cooperation. A score of -10 on the left side of the rating scale in the model indicates the maximum orientation towards competition, while a score of 10 on the right side of the rating scale indicates the maximum orientation towards cooperation. According to the model presented in Figure 1, and the foregoing, it can be seen that the range of the rating scale of synergies between competition and cooperation has been estimated from -

rating scale also marks a value domain. The Figure 1 depicts a mental model of synergy between competition

integrating such activities, it may have a powerful effect on cost and differentiation. Technology may have a positive influence on cost or differentiation if it has an influence on cost drivers or value added activity drivers. Technology used in value added activity is often the result of other drivers, such as volume, timing, or interrelationship.

Competition can be defined as a capability for creating differential advantage that competition was not able to copy, and described as a set of unique features of a company and its products perceived as significant and superior to competitive ones in the target market.

Cooperation as a foundation of interorganizational relationships in markets, may be defined as a coordinated joint action of two or more companies or their specific functions aimed at improving business relationships, and solving identical or similar tasks and common issues, while its goal is to improve company's position in the market, or any market segment. Economically speaking, cooperation is defined as joint work of independent companies by their free will with the intention of, with only partial abandonment of discretion, achieving a higher level of accomplishment of common goals, but in the case of independent actions.

Coopetition is based on the balance and synergy between cooperation and competition

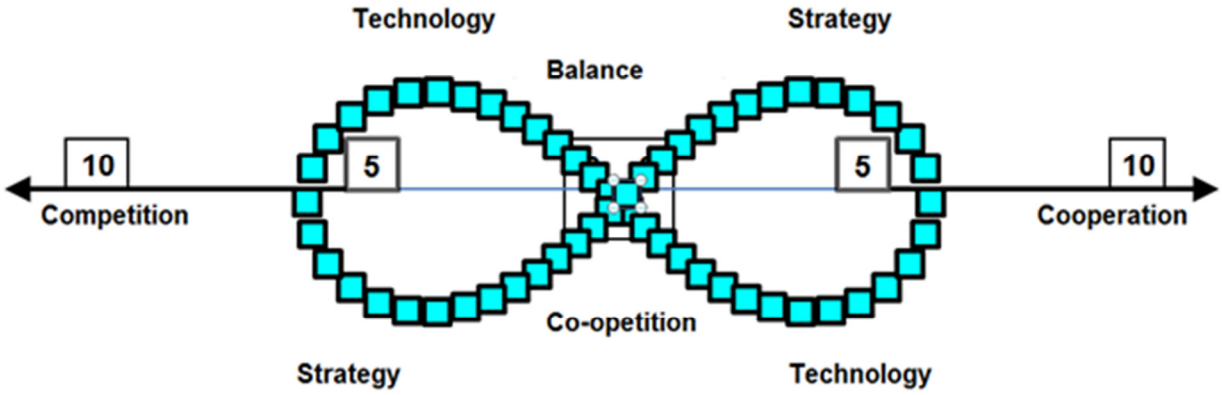
4.2. Lemniscate as a geometric form of the achievement of synergy between competition and cooperation

A lemniscate like model (having a shape of an algebraic curve) offers a useful visual representation of the state and trends in complex phenomena [4]. Scheme 1 depicts the lemniscate as a model of dynamic balance between competition and cooperation. The focus is on visualization of the possibilities and ways of connecting key components in achieving synergy and balance between competition and cooperation. In the example, the arrows mark and indicate the intertwined flows such as: conceptual, mental, process and information. Strategies and technologies have been included in the model as key determinants in achieving balance and synergistic relationship.

A lemniscate can be obtained as the inverse transform of a hyperbola, with the inversion circle centered at the center of the hyperbola. The length of an arc of a lemniscate is expressed by elliptic integrals, as was discovered in the eighteenth century. Around 1800, Carl Fridriech Gauss studied elliptic functions, which are inverse of these integrals (his work was not published at the time, but it was alluded to in the notes of Gaussian Disquisitiones Arithmeticae). A lattice network has a special shape, proportional to the Gaussian integers. For this reason, a case of elliptic functions with complex multiplication in some sources is called the lemniscate case [4].

Figure 2 depicts Lemniscate as a geometric model of linking and acheving synergies between competition and cooperation.

Figure 2. Lemniscate as a geometric model of linking and acheving synergies between competition and cooperation



Source: Authors

A score of 5 on the rating scale is an example of defining a level of cooperation at a given level of competition. For example, based on the level of competition as estimated it is possible to decide on the level of cooperation required to maintain a balance and synergy between competition and cooperation and achieve a given level of competition that involves the development and integration of strategy and technology.

As can be seen in the scheme, only the horizontal axis, which is specified by competition and cooperation, has been defined. The vertical axis has not been explicitly defined, but may be represented as another dimension of a model of synergy between competition and cooperation. For example, the diagram's vertical axis can be specified by strategy and technology.

The vertical axis may also include multiple dimensions which can be analyzed by comparative methods and related to the primary (reference) dimension of the synergy between competition and cooperation, which is shown on the horizontal axis. Vertical dimension can be made of control and executive subsystems which can be positioned in the upper part of the vertical axis.

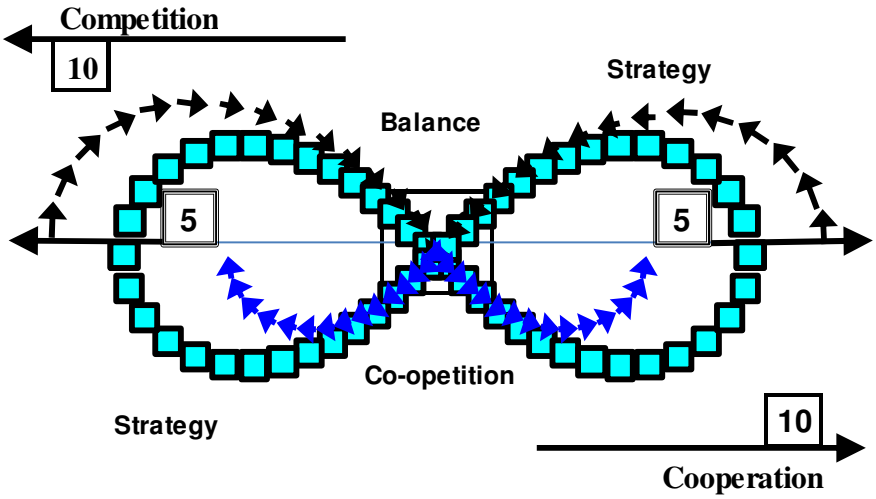
As can be seen in the present mental model in Figure 2, strategy and technology are changing their vertical position in relation to the horizontal axis position.

4.3. Flows in a model of synergy between competition and cooperation

Figure 3 depicts the flows in the model of synergy between competition and cooperation. External flows indicating processes and activities between the business system and environment as well as processes and changes in the relevant business environment, are marked in black color. On the other hand, internal flows indicating processes and activities within the business system are marked in blue color.

The model depicted by the Figure 3 illustrates the process of generating internal and external flows, and their integration. As can be seen, internal flows are generated within the business system, while external flows are generated within the business environment.

Figure 3 The flows in the model of synergy between competition and cooperation



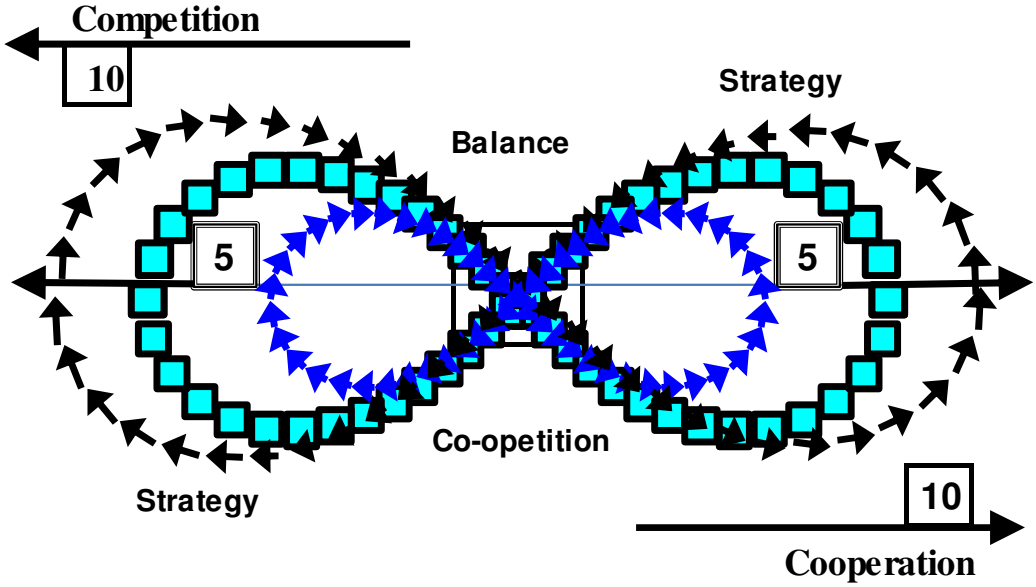
Source: Authors

4.4. The entire model of synergy between competition and cooperation

The flows show dynamic relationships among the elements of the model based on iterative and incremental development and give an image that reflects a coopetition model. The traditional system, which is characterized by vertical relationships in which competition and cooperation are mostly separate categories can be displayed by linear relationships. In a coopetition model, business organizations should have a range of different relationship types in their portfolio that can change at different time points from competition, through coexistence to cooperation, or simultaneous existence of both - competition and cooperation.

Flows are generated and developed simultaneously. The model in Figure 3 shows internal and external flows that are linked and intertwined in the centre of gravity of the model of synergy between competition and cooperation

Figure 4 The entire model of synergy between competition and cooperation



Source: Authors

In this context, the flows indicate dynamic relationships and a continuous interweaving of the model's basic elements: strategy, technology, business system, environment, competition and cooperation.

As can be seen, the lemniscate shows a geometric model of synergistic relationship between competition and cooperation, wherein arrows indicate non-linear internal and external flows that follow the path of lemniscate shape with some deviation.

5. Conclusion

Mental models can serve as a useful tool to develop, integrate and achieve synergies among constructive and innovative concepts and ideas while fostering creativity, innovation and freedom of expression. A context for learning programming in mental model development will facilitate improvement of personal metacognitive skills and development of abstract thinking also in areas outside programming.

In designing the mental model of synergy between competition and cooperation has been proposed a methodological framework:

1. Defining the elements of a mental model
2. Designing a geometric model for linking and acheiving synergy between competition and cooperation,
3. Defining and analyzing how flows of elements of mental models are linked,
4. Analyzing and interpretation a complete mental model process for achievement of synergy between competition and cooperation.

Presented mental model reveals mutually conflicting features of contrariety and complementarity between competition and cooperation.

Coopetition indicates that there exists a synergistic interaction between competition and cooperation. Proposed mental model may generates ideas, concepts and methods of implementation, development, use and maintenance of balance and sinergy between competition and cooperation.

Literature:

1. Bubica, N., Mladenović, M., Boljat, I., Programiranje kao alat za razvoj apstraktnog mišljenja https://bib.irb.hr/datoteka/702093.Programiranje_kao_alat_za_razvoj_apstraktnog_miljenja-CUC-zbornik.pdf
2. Dagnino, G.B., Padula, G., Coopetition Strategy: A New Kind of Interfirm Dynamics for Value Creation, Paper Presented at the Second EURAM Annual Conference, Stockholm, May 9-11., 2002.
3. Domitran, I., Rješavanje poslovnih problema pomoću mentalne mape, Poslovna učinkovitost d.o.o., http://www.poslovnaucinkovitost.eu/korisni-sadrzaji/kolumne/988-rjesavanje-poslovnih-problema-pomocu-mentalne-mape
4. Lemniskata, http://hr.wikipedia.org/wiki/Bernoullijeva_lemniskata
5. Vukmirović, S., Kovač, D., Mežnarić, I., Object-oriented programming for graphical modeling on the example of logistics network design, MIPRO, 2013 Proceedings of the 36th International Convention, Opatija, 2013
6. Vukmirović, S., Mental model of synergy between competition and cooperation, Video Animations, 2015. www.infosustav.com/video/animations.htm

Valuation of Options by Using Excel

Zdenko Prohaska, Ph.D.*, Bojana Olgić Draženović, Ph.D.*, Ivan Uroda, Ph.D.**

* University of Rijeka, Faculty of Economics, Department of Finance and Banking, Rijeka, Croatia

** University of Rijeka, Faculty of Economics, Department of Computer Sciences, Rijeka, Croatia

zdenko.prohaska@ri.t-com.hr, bolgic@inet.hr, ivan.uroda@gmail.com

The presented results are part of the University of Rijeka Research project No. 13.02.1.2.02

Abstract - Options represent the right to buy or to sell a specific number of underlying assets (stocks, indexes, commodities etc.) at a given price in a predetermined period of time.

To analyze the value of options the Black and Scholes model will be presented and special spreadsheets will be developed using only "plain vanilla" Excel, i.e. its basic financial functions, for call and put options in the case of stocks which are paying or not paying dividends.

The reason for using only "plain vanilla" Excel is a better portability of such spreadsheets for students at faculties of economics and in the public sector on different platforms or computers with various operating systems.

I. INTRODUCTION

In financial courses at universities educational software is important for students not only to better understand fixed income securities, like bonds or equity securities, like stocks, but also to especially recognize possible profits and immense risks being immanent to derivatives like options.

Valuation of options was a tricky issue until the development of the famous Black and Scholes model which required a lot of computation. Therefore, in this paper the authors are presenting the Black and Scholes model for valuation of call and put options for stocks without and with payments of dividends. Instead of using Visual Basic for Applications (VBA) script language, in order to obtain better portability on other computers and operating systems in the future, so called "plain vanilla" Excel spreadsheets containing only ordinary financial and statistical functions were programmed.

It has also to be pointed out that students at universities and especially at faculties of economics or business schools are usually familiar with Excel, a lot of good textbooks exist, so the time for learning such models and adopting them should be a relative short one. [1], [10], [11]

II. THE BASICS OF INVESTMENTS IN OPTIONS

In case of investing in stocks an option is the contract that provides the right to buy or to sell a specific amount of stocks at a predetermined price (called exercise price) during a given period of time. Options are available for different sorts of securities, currencies, interest rates or market indexes. The two major types of options are so called call and put options.

A call option is a contract that gives the holder the right to buy a given number of stocks of a security (in the USA normally 100 stocks per unit) at a predetermined price, i.e. the exercise price. The buyer of the call option pays the writer or seller a premium. The writer of the call option agrees and has the obligation to sell the given number of stocks of the underlying security at the exercise price within the predetermined time period.

A put option is a contract that gives its holder the right to sell a given number of stocks of a security (normally 100 stocks per unit) at a given price, the so called exercise price. The buyer of the put option pays the writer or seller a premium and the writer of the put option agrees to buy the given number of stocks of the underlying security at the exercise price within the predetermined time period.

The option's premium is influenced by many factors. It is affected positively by market value and price volatility of the underlying security, risk-free rate of interest and time to expiration. On the other hand it is affected negatively by the exercise price and security cash dividend. The possibility of gain or loss for an option is directly related to the price volatility of the underlying security, time to expiration and the interest rate. [2], [3], [9]

The most straightforward and commonly used model for valuation of options is the Black - Scholes Option Valuation Model.

III. THE BLACK AND SCHOLES OPTION VALUATION MODEL

The main parts and equations of Black and Scholes option valuation model are presented below.

A. *Call options*

The Black-Scholes model defines the value of an option by applying a specific mathematical calculation of the relationship between the market price of the underlying security, i.e. a stock and the option`s exercise price, time to expiration, interest rate and the price volatility of the underlying stock. The value of a call option can be computed using the Black and Scholes model which consists of three equations: [5]

$$C = SN(d_1) - Xe^{-rT}N(d_2) \quad (1)$$

$$d_1 = (\ln(S/X) + (r + 1/2\sigma^2)T)/\sigma\sqrt{T} \quad (2)$$

$$d_2 = d_1 - \sigma\sqrt{T} \quad (3)$$

where:
C = price of the call option
S = current price of the stock
X = exercise or strike price
r = risk-free rate of interest
T = time (in years) remaining before expiration of the option expressed in decimal form
e = base of the natural logarithm, or 2.71828
$N(d_n)$ = probability that a normally distributed random variable with a mean of 0 and a standard deviation equal to 1 will have a value less than or equal to d
$ln\ ()$ = natural logarithm
σ and σ^2 = standard deviation and variance of the continuously compounded annual rate of return on the stock.

Taking into account the model presented above it can be concluded that the value of call options has a direct relation to the value and the price volatility of the underlying stock, risk-free rate of interest and the time remaining to expiration, but has an inverse relation with the exercise price.

Other properties of call options are the following:

a) The price of a call option C has an upper bound, therefore C is always less than the price S of the stock.
b) The price of a European call option C also has a lower bound, since it is always greater than the current stock price S minus the present value of the exercise price X.
c) The option becomes worthless if the stock is worthless.
d) When the stock price S is very large, the call option value C approaches the stock price less the present value of the exercise price X.

It can be pointed out that a call option becomes more valuable, the higher the stock price, the lower the exercise price, the higher the interest rate, the longer the time to expiration and the greater the volatility or risk of the stock. Another two facts should be mentioned about options. First, the riskier the stock, the greater the standard deviation and therefore the higher the value of the option. So risk seems to be desirable, because more risk provides higher expected option value. Second, the willingness of the investor to bear risk, he could be a risk seeker or risk averse investor, has no impact on the value of the option.

Beside that, it has to be stressed that the Black and Scholes model is based on quite a lot of assumptions such as no cash dividends, no transactions costs or taxes, constant risk-free rate of interest, no penalties for short selling, the stock market operates continuously, the option can be exercised only at expiration date, and so on.

Even with these assumptions, the Black and Scholes model has been found to be the best available predictor of actual options prices. However, biases have been noted. First, the model tends to overvalue so called out-of-the-money call options or options on stock whose current market price is less than the exercise price. It also undervalues so called in-the-money call options or options on stocks whose current price of stock is greater than the exercise price. Second, the model tends to underprice call options on low-variance stocks and overprice call options on high-variance stocks. Nevertheless, it is a better and more robust model than anyone presented earlier.

B. *Put options*

In the case of put options and based on the principle of the put-call parity and assuming that the put and call options are on the same security or stock with the same exercise price and the same time to expiration, the Black and Scholes model can be used to calculate the value of a put option. The put-call parity is the relationship between the values of put and call options when no arbitrating opportunity exists. When the put and call options are in parity which each other, the difference between their prices should be equal to the difference between the present value of their common exercise price and the current market value of the underlying security. [12]

Or put in another way once the value of a call option is known, determining the value of a put option is quite easy. Based on the put-call parity mentioned above, it becomes evident that the value of a call, plus the present value of the exercise price, equals the value of the put plus the market price of the stock. Therefore the value of a put option is equal to: [5]

$$P = -SN(d_1) + Xe^{-rT}N(d_2) \quad (4)$$

where:
P = value of the put option
and the rest of the variables are defined previously.

So, the value of the put option has a direct relation to the exercise price and the volatility of the stock price, but has an inverse relation to the stock price and risk-free interest rate. To the time remaining to expiration there are two possibilities. As time to expiration T increases, normally the value of the put option P increases; but if the stock price S is significantly below the exercise price X, then as time T increases, the value of the put option P decreases.

Other properties of put options are:

a) The price of a put option P is always less than the present value of the exercise price X, which is an upper bound.
b) The price of a put option P is always greater than zero, except at expiration.
c) When the stock price S is very small, the value of the put option P approaches the present value of the exercise price X minus the stock price.

Similar to call options, put options become more valuable the more volatile the price of the underlying security or the stock is. Differently from call options, however, the lower the stock price, the greater the value of a put option.

C. *Options on dividend paying stocks*

In case of cash dividend payment the Black and Scholes models for call and put options for non-dividend paying stocks have to be adopted. Generally speaking, the current market value of the underlying stock minus the present value of cash dividends to be paid prior to expiration, in place of the current market value of the stock in the Black and Scholes model is to be used.

D. *Call options on dividend paying stocks*

For call options the presence of cash dividend tends to reduce the value of a call option, since the option must be exercised for the dividend to be claimed. Under the assumption that only one cash dividend is expected to be paid before expiration of the option, the value of a call option can be calculated as follows: [5]

$$C^* = \left(S - \frac{D}{e^{rT}} \right) N(d_1) - \frac{X}{e^{rT}} N(d_2) \tag{5}$$

where:
C^* = value of a call option on a dividend paying stock
D = dividend in cash that will be paid before the expiration of the option
T = time in years (decimal form) until the dividend is expected to be paid.

The rest of the variables are presented before. This equation differs from the equation for call options for non-dividend paying stocks only in part in which the market price is reduced by the present value of the dividend that is expected to be received before the expiration date of the option. This is so because an option holder does not have a claim on dividends. On the contrary, the holder suffers a loss in value if the stock pays dividends before the call option is exercised. [6]

E. *Put options on dividend paying stocks*

For put options the modified equation is: [5]

$$P^* = C^* + \frac{X}{e^{rT}} - P + \frac{D}{e^{rT}} \tag{6}$$

where:
P^* = value of a put option on a dividend paying stock
and the other variables are discussed previously.

The only changes from the Black and Scholes model for put options on non-dividend paying stocks are: to use the value of a call option on dividend paying stocks C^* and to add the present value of the dividends expected before the option is exercised.

It has to be pointed out that when cash dividends are paid before the put option can be exercised, the value of the put option increases. This is because payment of a cash dividend lowers the market price of a stock and therefore makes the put option more valuable. If there will be more than one dividend payout until expiration of the option, then the last expression in the equation above must be modified to reflect the present value of all dividends to be paid before expiration. [7]

Taking all that into account it can be concluded that the Black and Scholes model is demanding and that it requires the use of computers for fast and effective calculations. Nevertheless, it is crucial for the valuation of options.

IV. EXCEL SPREADSHEETS

For determining the value of call and put options for non-dividend and dividend paying stocks according to the Black and Scholes equations presented above, ordinary or "plain vanilla" Excel spreadsheets were developed applying only standard financial and statistical functions built in Excel. [4] For explanation of the models two examples are presented below. Beside that, two of the authors have already developed two spreadsheets for bond investments and presented it with another coauthor at MIPRO Conference in 2013. [8]

A. *Call and put options on non-dividend paying stocks*

Given the following information about the stock of the "DD" company and the options on the stock, calculate the current value of the call and put options using the Black and Scholes model.

Current market price of the "DD" stock = kn 75.00
Exercise price of the option = kn 60.00
Time remaining to expiration = 5.4 months (0.45 years)
Risk-free rate of interest = 8.0%
Standard deviation of stock return = 0.75
No dividend payment

After typing all the necessary data in the "input data" field of the spreadsheet model (options.xls) the results for call and put options of the company "DD" stock are calculated and displayed in the "output data" field. (Fig. 1)

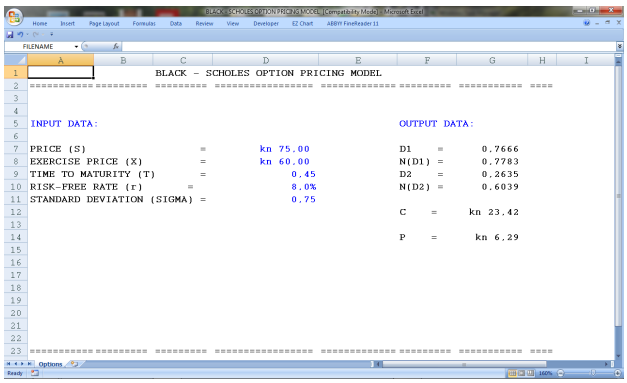


Figure 1. Spreadsheet screenshot for call and put options on non-dividend paying stock

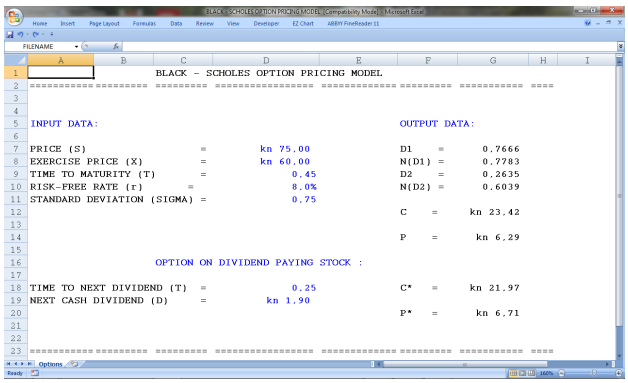


Figure 2. Spreadsheet screenshot for call and put options on dividend paying stock

The results show that the value of the call option is higher and amounts to kn 23.42 and the value of the put option is kn 6.29. So there is a higher risk premium to be paid for buying call options rather than put options on the same stock of "DD" company.

B. Call and put options on dividend paying stocks

Given the same basic information about the stock of the "DD" company and the options on the stock, but this time with payment of dividends, calculate the current value of the call and put options using the Black and Scholes model adapted for dividends.

Current market price of the "DD" stock = kn 75.00
Exercise price of the option = kn 60.00
Time remaining to expiration = 5.4 months (0.45 years)
Risk-free rate of interest = 8.0%
Standard deviation of stock return = 0.75
Next cash dividend = kn 1.90
Time to next dividend = 3 months (0.25 years)

After typing all the necessary data in the "input data" field of the spreadsheet model (options.xls) the results for call and put options of the company "DD" stock are calculated and displayed in the "output data" field for "option on dividend paying stock". (Fig. 2)

Taking into account payment of dividends the results show once again that the value of the call option is higher and amounts to kn 21.97, while the value of the put option is lower and amounts to kn 6.71. Again there is a higher risk premium to be paid for buying call options rather than put options on the same stock of "DD" company paying dividends.

Comparing both call options for the stock without and with payments of dividends it can be concluded that the call option without dividend payment is more expensive because payments of dividends lower the value of that call option. The opposite is true in the case of put options without and with payments of dividends.

V. INSTALLATION OF THE SPREADSHEET

Since all mentioned procedures and calculations are implemented in a single Excel spreadsheet (options.xls) it is only necessary to copy this file in the working directory of the PC where Microsoft Excel program can access them. Of course, installation of the Excel Analysis Toolpak Add-In could help to speed up the programming of some functions.

VI. CONCLUSION

In this paper the valuation of options, both without and with payments of dividends, according to the Black and Scholes model are presented. For this purpose special Excel spreadsheets were developed using only "plain vanilla" Excel, i.e. by implementing standard financial and statistical functions and not relying on Visual Basic for Applications (VBA) script language. The results of options calculations with the Black and Scholes model could make the students more aware of the possible profits, but also of the risks involved in options valuation and investments.

REFERENCES

[1] Benninga, S., Principles of Finance With Excel, Oxford University Press, New York, 2006

[2] Bodie, Z., Kane, A., Marcus, A.J., Investments, McGraw-Hill, New York, 2009

[3] Fabozzi, F., Investment Management, Prentice Hall, Englewood Cliffs, New Jersey, 1995

[4] Holden, C.W., Spreadsheet Modeling in Investments, Prentice Hall, New Jersey, 2002

[5] Hull, J.E., Options, Futures and Other Derivatives, Pearson Prentice Hall, New Jersey, 2006

[6] Prohaska, Z., Upravljanje vrijednosnim papirima, Infoinvest – Poslovna knjiga, Zagreb, 1994

[7] Prohaska, Z., Uvod v finančne trge – metode analize in instrumenti, Univerza v Ljubljani, Ekonomska fakulteta, Ljubljana, 1994

[8] Prohaska, Z., Olgić Draženović, B., Šarić, V., Calculating Duration and Convexity of Bonds Using Excel, in Čičin-Šain, M. et al., 36th International Convention MIPRO 2013, Proceedings Vol. IV, Computers in Education, Opatija, 2013

[9] Reilly, F.K., Investments, The Dryden Press, Orlando, 1996

[10] Underdahl, B., Excel Expert Solutions, Que Corporation, Indianapolis, 1996

[11] Walkenbach, J., Excel 2010 Formulas, Wiley Publishing, New Jersey, 2010

[12] Wilmott, P., Paul Wilmott Introduces Quantitative Finance, John Wiley & Sons, New York, 2001

Exploring the use of computers in classroom teaching

Branka Kontošić, mag.prim.educ.
Osnovna škola Šijana
43. istarske divizije 5, Pula, Hrvatska
E-mail: b_kontosic@net.hr

Mr.sc. Alen Kontošić, mag.oec.
Sveučilište Jurja dobrile u Puli
Zagrebačka 30, Pula, Hrvatska
E-mail: akontosi@net.hr

Sanja Biloslav, mag.prim.educ.
Osnovna škola – Scuola elementare Milana Šorga
Matka Laginje 25, Oprtalj, Hrvatska
E-mail: sanja.biloslav@skole.hr

Summary – The computer is becoming more used and prevalent in learning. It is increasingly being used in knowledge transfer and acquisition of knowledge. The textbooks items in the classroom teaching are enriched with CDs. How many pupils in lower grades use a computer for learning new stuff? Do they learn more easily using textbooks or CDs? How much time do pupils spend on the computer to get some information as for learning on the Internet?

The data were obtained by means of a questionnaire in which pupils were anonymously surveyed. The aim of the research is to determine whether pupils use the computer for learning and in which school subjects the most. The data analysis shows that pupils mostly use the computer for the adoption of the content of educational .subjects.

Keywords: classroom teaching, pupils of classroom teaching, computer use, subjects of classroom teaching.

I. INTRODUCTORY REMARKS

A. Introduction

Looking back to the history of education, generations were educated through oral tradition, then by means of a book, and today it is increasingly acquired to use information and communication technology (ICT). Computers and tablet, or educational content in digital form are more and more present in the process of acquiring knowledge. Digital textbooks are enriched with lots of facilities while classic textbook does not have such presentations, videos, games, quizzes and similar materials. They can be used in all stages of the lesson, in any type of lesson from presenting new materials to training and repetition, especially with independent learning. Today the children from an early age get in touch with several digital forms of teaching materials, starting with their first photographs, learning the first songs and enriching their vocabulary through stories and cartoons. That is why it is expected that children grow up using digital forms for the acquisition of new knowledge, disseminate learning perspective and evaluation of acquired knowledge. Before some information and answers to questions were looked for in encyclopedias, lexicons, dictionaries, collections, but today searching for data in digital form by means of a web search is more appropriate.

B. Purpose and objectives of the research

The survey on the use of computers in classrooms was desired to determine the extent to which pupils in lower grades of elementary school use the computer. Its aim was to determine what percentage of pupils in the primary school use computers for learning different subjects and in which the most. Pupils in lower grades have three compulsory subjects (Croatian language, Mathematics, Science), three compulsory subjects (Art, Music, Physical education) and a foreign language. The surveys also want to explore and get to the data of how many pupils of surveyed population use a computer, the Internet and digital content for each of the above mentioned subjects in class teaching. In addition to this one of the goals was also to determine where pupils of the surveyed population in classrooms love more to learn, and how often they learn using the computer.

II. RESEARCH METHODOLOGY

The study was carried out by means of a questionnaire in which pupils of classroom teaching were anonymously surveyed in several schools of Istria County. The study used an interview with a meter survey lists of a total of 14 questions (of which 12 questions with in advance offered YES-NO answers) while in two questions pupils circled in advance offered answers: where they prefer learning from (from textbooks or using the computer); how often they use the computer for learning (every day, often, rarely, never).

In the research, besides these survey methods, other scientific methods such as statistical methods, methods of description, analysis and synthesis are applied. The data were analyzed using MS Excel and statistical method for measuring the central tendency (frequency and percentage).

III. THE RESULTS AND ANALYSIS OF THE RESULTS

The study was conducted in several elementary schools in the Istria county on a sample of 109 pupils from a total of 141 pupils and in the analysis of the results worth mentioning large percentage (p = 0,773) surveyed school pupils. Taking into account that the completed questionnaires returned so many pupils, the rate of return can be considered very satisfactorily. In similar studies it moves in a much smaller percentage. Therefore, the standard error of the proportion is small considering that *n* is large (n> 30). Size arithmetic mean is approximately normally distributed, and therefore, as a rule two sigma, achieved with a probability greater than 0,95 all values within the interval ± two sigma around the middle.[1] Taking into account that nearly four out of five pupils answered the questionnaire, the pattern can be representative as for assessment and can be the basis for making appropriate conclusions concerning the population.

The data analysis has led to results which show that 92,66% of the total number of pupils have a computer at home. Internet access has 89,91% of the total number of surveyed pupils. In addition to this, the survey results indicate that 66,97% pupils use computers for learning. Internet for the purpose of learning used 67,89% of the total number of surveyed pupils. 40,37% of the surveyed pupils use computer for learning Croatian language. 49,54% of the surveyed pupils use the computer for mathematics exam. 69,72% of the surveyed pupils use the computer for learning Science. 26,61% of the surveyed pupils use computer for learning Music. The same number of pupils, or 21,1% of the surveyed pupils use the computer for learning Art and Physical education. Almost three-quarters of the surveyed pupils (73,39%) use computer for learning a foreign language. 79,82% of the surveyed pupils use Internet for the purpose of acquiring new knowledge. Survey data from the first to the twelfth questions are presented in the table format 1. and in the graph in Figure 1.

	Question	Pupils of classroom teaching:		
		YES [%]	NO [%]	∑ [%]
Q1	Do you have a computer at home?	92,66 %	7,34 %	100,00 %
Q2	Do you have an access to the Internet?	89,91 %	10,09 %	100,00 %
Q3	Do you use a computer for learning?	66,97 %	33,03 %	100,00 %
Q4	Do you use the Internet for learning?	67,89 %	32,11 %	100,00 %
Q5	Do you use a computer to learn Croatian language?	40,37 %	59,63 %	100,00 %
Q6	Do you use a computer to solve mathematics tasks?	49,54 %	50,46 %	100,00 %
Q7	Do you use a computer for learning Science?	69,72 %	30,28 %	100,00 %
Q8	Do you use a computer for learning Music?	26,61 %	73,39 %	100,00 %
Q9	Do you use a computer for learning Art?	21,10 %	78,90 %	100,00 %
Q10	Do you use a computer for learning Physical education?	21,10 %	78,90 %	100,00 %
Q11	Do you use a computer for learning a foreign language?	73,39 %	26,61 %	100,00 %
Q12	Do you use the Internet for getting new information?	79,82 %	20,18 %	100,00 %

Table 1. Tabular presentation of the results - answers to the questions 1. – 12.

[1] LJ. Baban, S. Jelinić, M. Lamza-Maronić, K. Ivić, "Primjena Metodologije znanstvenog istraživanja", Ekonomski fakultet Osijek, Osijek, 1993.

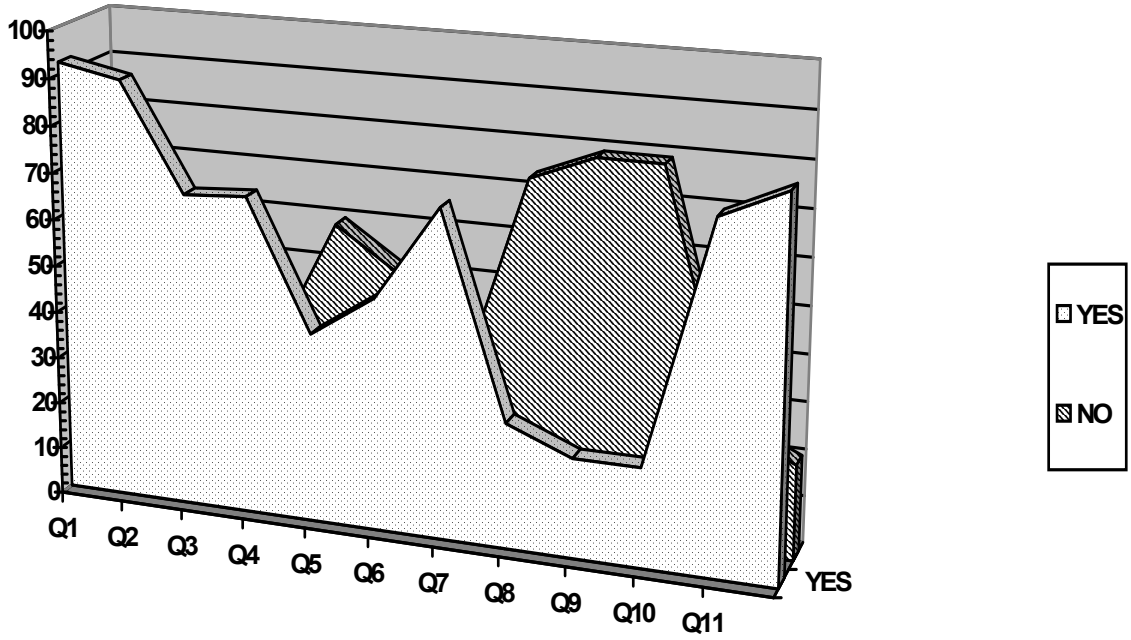


Figure 1.Total positive and negative answers of pupils on the questions (in %)

In addition to 12 questions with in advance offered YES-NO answers in the survey there are two questions: pupils are asked to circle the default answers to questions. The answers to question number 13: "Where do you like to learn from?" were textbook and computer. The order of response is determined by the frequency and intensity displayed in a table format no. 2.

	Question	Order of responses as for intensity and frequency of answers:	Pupils of classroom teaching: [% by frequency of given answers]:
Q13	Where do you like to learn from?	1. Textbook [%]	38,53%
		2. Computer [%]	61,47%

Table 2. Tabular presentation of the answers to the 13th question

As for frequency and intensity and considering the total number of respondents which is displayed in a table format no. 2., 38,53% of pupils like to learn from the textbook while 61,47% of pupils like to learn using the computer.

The answers to question number 14: "How often do you use computer for learning?" were every day, often, rarely and never. The order of response is determined by the frequency and intensity displayed in a table format no. 3.

	Question	Order of responses as for intensity and frequency of answers:	Pupils of classroom teaching: [% by frequency of given answers]:
Q14	How often do you use computer for learning?	1. Every day [%]	15,60%
		2. Often [%]	42,20%
		3. Rarely [%]	25,69%
		4. Never [%]	16,51%

Table 3. Tabular presentation of the results answers to the 14th question in class teaching

As for frequency and intensity and considering the total number of respondents which is displayed in a table format no. 3., 42,2% of surveyed pupils often use computers for learning, 15,6% of surveyed pupils use computer for learning every day, 25,69% of surveyed pupils answered that they rarely use computer for learning and 16,51% of surveyed pupils answered that they never use a computer for learning.

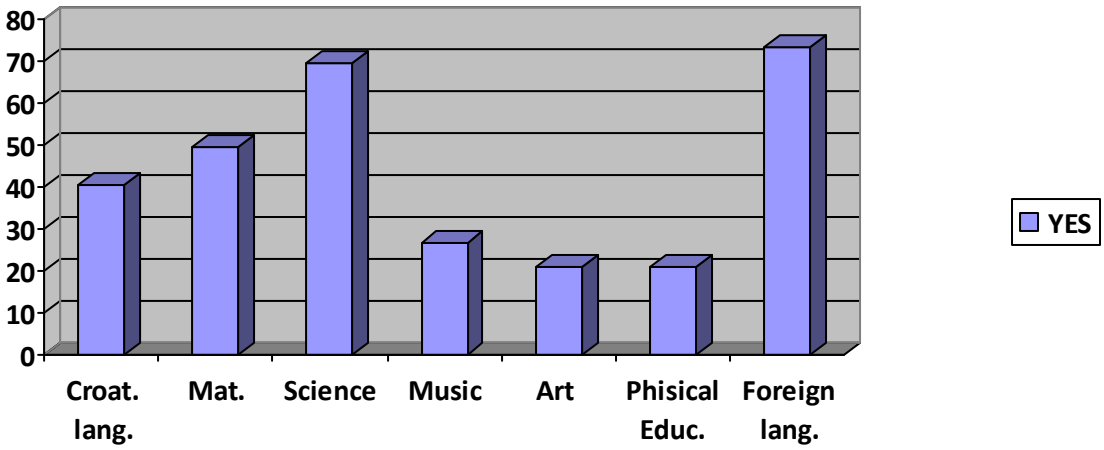


Figure 2. Comparison of positive responses to the questions according to the subjects (in %)

IV. CONCLUSION

The results obtained by processing of the data are shown in the graph in Figure 2 and give answers to the questions what percentage of elementary school pupils use computers for learning different school subjects and in which the most. Pupils mostly use computer and Internet for new knowledge acquisition and learning foreign languages. Slightly less pupils use the computer for learning Science (almost 70% of surveyed population), and then even less (almost half of those surveyed) use the computer for problem solving in mathematics. A little less (about 40% of surveyed pupils) use the computer for learning Croatian language. It is interesting to note that only one of four surveyed pupils uses the computer for learning Music, while every fifth surveyed pupil uses the computer for learning Art and Physical education. The results of the survey prove that today we encounter a computer already at birth, and it becomes an integral part of growing up of a child. The survey confirms that more than three fifths of surveyed populations prefer learning with computers instead learning from textbooks, while the question How often do they use computer for learning responded positively almost 85% of the surveyed pupils.

REFERENCES

[1] LJ. Baban, S. Jelinić, M. Lamza-Maronić, K. Ivić, "Primjena metodologije znanstvenog istraživanja", Ekonomski fakultet Osijek, Osijek, 1993.

Intelligent Assistant for Helping Students to Learn Programming

Mario Konecki *, Nikola Kadoić * and Rok Piltaver **
* Faculty of Organization and Informatics, Varaždin, Croatia
** Jožef Stefan Institute, Ljubljana, Slovenia
mario.konecki@foi.hr; nikola.kadoic@foi.hr; rok.piltaver@ijs.si

Abstract - Students experience constant problems in their introductory programming courses. Along with difficulties to cope with abstract programming concepts students tend to experience fear and lack of motivation to learn programming. In order to make programming more interesting and to make programming concepts easier to comprehend an intelligent assistant which enables students to ask question about their programming lessons in a simple and familiar way just as they would ask the lecturer is proposed. In this way the process of getting the right information and answer gets faster and also more interesting and engaging compared to classic approach. Along with getting the desired answer faster this kind of approach increases motivation of students to spend more time in learning programming because of interesting new technology that is used as an aiding tool in their programming education. The results about students' attitude towards such approach and its effect on students' motivation and amount of time spent in learning are presented and discussed.

Keywords **- programming, education, problems, intelligent assistant**

I. INTRODUCTION

Students have been struggling with their programming courses for years. Despite of all efforts that have been made programming is still considered as hard to learn and students experience fear about programming even before they enroll into these courses. Over the years one fact has become a common opinion among both students and teachers of programming courses and this fact is that programming indeed is hard to learn [1; 2; 3; 4; 5]. The fact that various efforts in the area of aiding programming students and especially introductory programming students haven't produced any long-term results asks for more focus on this problem and more research in order to find some sort of usable solution.

The problems that students of programming courses experience come from many different sources and reasons. The solution for these problems includes multiple factors and aspects that need to be considered. These factors can be observed on either teachers' of students' side as well as on the side of programming itself. Teachers are responsible for establishing a sound and effective teaching methodology and this is when talking about programming very challenging task. Beside the fact that the right course of action is not always clear or not clear at all, abstract nature of programming and its concepts requires a lot of effort as well as various technical knowledge and skills in order to produce the right materials in the design and organizational sense that will make programming concepts understandable to students. Teachers also have problems in adapting their lectures to every particular student since students differ in their preferable learning style but also in their advancement speed.

Students on the other hand have responsibility to focus on lectures and actively listen in order to adopt presented matter. Another responsibility of students is to learn and study because all efforts of teachers and all innovative approaches will prove as futile without the cooperation of students and their willingness to learn. Motivation of students to learn is one of major factors that have influence on the success rate of programming courses [6]. It is important for students be aware of importance of programming for their future careers but it is also important for them to know why they enrolled their faculties and programming courses because this general direction and being committed to it has a major effect on subgoals that include learning programming and passing programming courses with good grades.

Programming itself is another factor that contributes to various problems in programming courses. Programming is abstract and that brings a lot of different aspects which need to be incorporated in order to learn programming properly. Because of its abstract nature programming asks for more effort from teachers, more complex methods and techniques of teaching, greater focus of students and different way of learning that asks for prolonged learning rather than learning of a lot in a short time. The reason for this is that programming is not only a knowledge but a skill [7] and a skill requires some time to develop. Students are generally not used to this kind of learning since most of their education included mostly memorization of facts because of different nature of most other school subjects.

Programming is important and new computer experts are needed on a daily basis because of rapid growth of business industry. Improving state regarding programming courses passing rates and quality of students' knowledge and skills is something that needs more attention and more focus. In this paper problems that programming courses experience and reasons of these problems as well as methods of their improvement are presented and

discussed. A new approach that includes the usage of intelligent assistant as a mean of increasing motivation of students to spend time learning programming is proposed and research results about the effectiveness of this kind of approach are presented and elaborated. Some pointers and possible approaches for future research are also mentioned.

II. PROBLEMS IN PROGRAMMING EDUCATION

Failure and dropout rates in programming courses and especially in introductory programming courses have been reported as relatively high throughout the years [8; 9]. Programming courses have simply become something that students are used to struggle with and something that is frequently a reason of dropouts. Research results show that students in many cases don't know how to produce a valid programming code even at the end of introductory programming courses [10].

Programming is abstract and challenging for students. It requires a way of learning and reasoning that is new and not habitual for students. There are three very important factors that influence the results of programming courses:

- Level of developed problem-solving skills
- Students' motivation
- Teachers' efforts

One of key prerequisites for students to learn programming are properly developed problem-solving skills or in other words developed algorithmic way of thinking [11]. It is hard for students to produce a valid programming code if they don't know how to understand and analyze the problem in order to be able to construct its algorithmic solution. Since programming requires prolonged way of learning students have to be committed to the learning process. This requires students to be properly motivated with a clear goal in mind. Knowing why it is important to enroll programming courses and why it is important for students' future career is something that affects students' motivation and needs to be clear.

Many researches have been conducted regarding reasons of problems in programming courses with different results. Hawi [12] has defined 10 casual attributions which have emerged as a result of conducted research among undergraduate students: learning strategy, lack of study, lack of practice, subject difficulty, lack of effort, appropriate teaching method, exam anxiety, cheating, lack of time, and unfair treatment. Gomes and Mendes [13] have identified several possible reasons of problems in programming courses:

- Programming demands a high abstraction level.
- Programming needs a good level of both knowledge and practical problem-solving techniques.
- Programming requires a very practical and intensive study, which is quite different from what is required in many other courses which are more based on theoretical knowledge, implying extensive reading and some memorization.
- Usually teaching cannot be individualized, due to common classes' size.
- Programming is mostly dynamic, but usually taught using static materials.
- Teachers' methodologies many times don't take into consideration students' learning styles. Different students have different learning styles and can have several preferences in the way they learn.
- Programming languages have a very complex syntax with characteristics defined for professional use and not with pedagogical motivations.

Different approaches have been researched in order to try to improve results in programming courses but they have shown variating results. Jenkins [7] proposes several steps that should increase success rate in programming courses:

- Programming should never be taught before the second year of any course
- The language used should be chosen for pedagogic suitability and not because it is popular in industry
- Programming should be taught by those who can teach programming and not by those who can program
- Programming courses should be designed to be flexible to allow different students to learn in different ways
- There should be no summative (continuous) assessment to ease pressure on students
- Departments should acknowledge that programming is difficult and supply adequate support to students

Importance of a proper way of thinking that is needed in order to learn programming is also confirmed by Don Norman who claims that there is a huge gap wide as Grand Canyon between the way of thinking applied by programming novices and the way of thinking that is required by computer in order to perform valid instructions. [14]. Students use many concepts which are used in programming in their everyday life but many of these concepts are mostly unconscious because of great ability of human brain to perform complex tasks and process constraints and conditions is such a speed that one is virtually not aware of it. On the other hand computers need clear boundaries, constraints and conditions in order to produce valid programs that will be robust and usable for different cases and scenarios in some particular area of interest. Students frequently omit different important elements which in the end results in faulty programming code. There is a need to devise new approaches, methods and tools that will promote algorithmic way of thinking and train students in problem-solving skills which will in the end result in better performance during their programming courses. Several summative steps can be

proposed in order to aid students in their programming education [15]:

- Introduce additional programming course prior to introductory programming course that would promote algorithmic way of thinking
- Increase motivation of students for learning programming
- Explain to students that programming is a skill, not merely knowledge
- Introduce elements of constructivism into teaching process
- Introduce learning by example
- Introduce animation and other visualization techniques combined with interaction
- Introduce interactive visual simulations
- Include support for multiple learning styles

Solving problems in programming education is a challenging task that asks for more effort form both teachers and students. Further research is needed in order to improve situation in programming courses which have the main goal to produce new computer experts who are able to write valid computer programs.

III. USING INTELLIGENT ASSISTANT IN PROGRAMMING EDUCATION

When talking about various aiding tools and techniques that are mostly mentioned as means of trying to make programming more understandable and more interesting the most commonly mentioned are various visualization tools and techniques that are aimed at making programming less abstract [16]. These visualization tools and techniques can be either static or dynamic in nature [17]:

- Static:
  - Code lists
  - Flowcharts
  - Diagrams
  - Pictures
- Dynamic:
  - One-way presentations
    - Movies
    - PowerPoint slides
    - Flash animations with sound explanation and music
  - Two-way interactive tools:
    - Executable flowcharts
    - Algorithm animations
    - Program visualizations

Many existing visualization tools have been used in practice such as XTANGO [18], JavaVis [19], JHAVE [20], BlueJ [21], Jeliot3 [22], Raptor [23] and ALVIS LIVE! [24]. Visualization tools have shown to be beneficial to students but another aspect that has proven to be useful for capturing students' attention is interaction [25].

In order to combine visual aspect and interaction into one aiding concept that should promote greater quality of students' programming knowledge an intelligent assistant has been proposed and applied in programming education. Intelligent assistant incorporates elements of visualization, interaction and expert systems as well as segments of artificial intelligence. Intelligent assistant is designed as a system that provides students with a possibility of interaction by using the natural language. Students are able to ask intelligent assistant regular questions just as they would ask their teacher or colleague. Intelligent assistant is developed as an aiding tool and it is web based. It consumes only a small portion of overall screen real estate and it has its own avatar representation. Its main purpose is to capture students' attention with its novelty as a technology, to speed up the process of getting desired information, to increase motivation and the time that students spend in learning programming and to make learning process more interesting. Intelligent assistant supports many different areas of questions, from more general to more specific, for example:

- Who are you?
- How are you?
- What is programming?
- I don't know what is variable.
- What is the difference between for and while loop?

Question and answer from intelligent assistant are shown in Fig. 1.

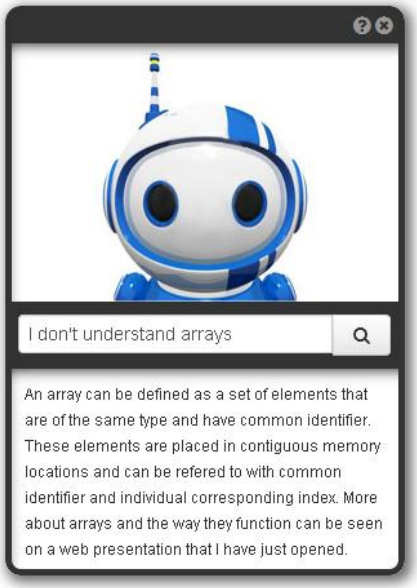


Figure 1. Asking intelligent assistant a question and getting the right response

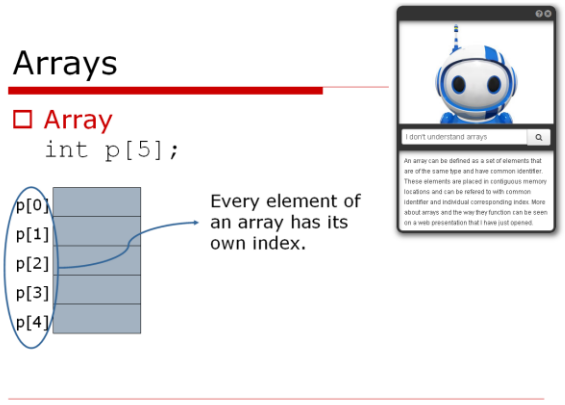


Figure 2. Opening of web presentation as an addition to textual answer to given question

Intelligent assistant supports different variations of the same question. For example, it is possible to ask "What is the difference between for and while loops?" but also "How while and for loops differ?" or "While/for loops difference". There is a large amount of tolerance to various forms of questions incorporated into intelligent assistant in order to provide students with better experience and easier usage that is more natural. Intelligent assistant also supports feature of variations in answers' form so the answers to some particular question are not always of the same form.

Beside answering questions in textual form intelligent assistant has also a feature that enables opening of web pages or web presentations which is shown in Fig. 2. These web resources are complementary to already given answer and they give a deeper insight into area that is related to the asked question. In this way students are able to ask a question regarding any area of programming and they are given instant answer and appropriate web resource is opened where students are able to get a more detailed and clearer visual insight into some particular matter. In this way the process of learning is made quicker and more engaging for students compared to classic presentations and searching for some piece of information among a number of slides.

In the current stage of development intelligent assistant is customized with various knowledge rules by using a simple settings interface where one is able to define a keyword or a set of keywords and their priority as well as one or more forms of essentially the same answer. These forms of answer are presented randomly to make intelligent assistant appear more interesting and less predictable. Every knowledge rule can be paired up with one or more web resources that are presented along with the answer and which can be used to make answer even clearer and explained in more detail.

In order to test how intelligent assistant and its usage influences students' motivation to learn programming, does it make students more willing to learn programming by spending more time since learning becomes more interesting and whether intelligent assistant makes getting the right piece of information quicker a research has been conducted among computer science students who have gone through all lessons of their introductory programming course. This population was chosen in order to get more objective feedback from students who have already gone through some form of more or less classic lectures and learning process so they are able to compare it to learning by using the intelligent assistant. 68 students have participated in this research and they have been given intelligent assistant with support for learning lessons about arrays.

Students also had classic presentations to browse and read but all practice and learning was supported by intelligent assistant. Students have been given intelligent assistant for a week since this is time that is available to students between lessons in classic lectures approach. At the end students were given a questionnaire in order to assess whether intelligent assistant had any impact on mentioned aspects of interest. Results of conducted research are shown in Table I.

TABLE I. RESULTS OF QUESTIONNAIRE RESEARCH

Questionnaire item	Mean	Std. dev.
Intelligent assistant is interesting and engaging learning tool	4,25	0,58
Intelligent assistant makes learning of programming more interesting for me	4,09	0,63
Intelligent assistant makes learning of programming easier for me	4,41	0,82
Intelligent assistant makes me more motivated to learn programming	4,12	0,49
Intelligent assistant makes me willing to spend more time in learning programming	3,82	1,04
I am able to get answers quicker by using intelligent assistant than by using classic presentations search approach	4,53	0,72
I am able to better comprehend programming concepts by using intelligent assistant	3,97	0,87
I prefer classic way of learning compared to using intelligent assistant	1,47	0,42
I think that intelligent assistant would also be beneficial for me in other subjects	4,11	0,73
I would recommend usage of intelligent assistant to my student colleagues	4,34	0,54

The research results have shown that students perceive intelligent assistant as a useful and interesting learning tool and aiding mean. Intelligent assistant has also been reported as a facilitator of more interesting way of learning, easier learning, greater motivation and willingness of students to spend more time in learning of programming. Students have also stated that they prefer this kind of learning over classic learning approach and that this kind of learning promotes better understanding of programming concepts. Students have shown that they perceive intelligent assistant as something that could be useful for them in other areas and subjects and as something that they would recommend to their colleagues.

IV. CONCLUSION

Programming education is faced with reoccurring problems of students that are not able to follow lectures and to do their programming tasks. Students experience fear about programming even before they enroll these courses which is based upon experience of past

generations of students. Abstract nature of programming makes these courses additionally challenging both for teachers and students. In order to try to make programming concepts clearer to students various approaches have been incorporated into programming courses with various success. One of the approaches that has proven to be beneficial in some extent to students was approach that includes visualization tools and techniques. Another valid approach is the one that includes some kind of interaction between students and learning tools.

In this paper an intelligent assistant is proposed as an adding learning technology which combines visualization and interaction and brings more focused and interesting way of learning that promotes greater motivation and increased amount of time that students spend learning programming. In order to assess the effectiveness of proposed intelligent assistant a research has been conducted in which students were given intelligent assistant as a part of their programming education about arrays. Students reported that using intelligent assistant increased their motivation to learn, made learning easier and more interesting which resulted in more time spent learning and in increased understanding of programming concepts. Students also stated that they find intelligent assistant to be also potentially beneficial in other areas. Expanding intelligent assistant's scope regarding programming and further testing of its effectiveness on a larger scale as well as research about other potential areas of its application and expanding intelligent assistant with the ability to create answers from various textual information will be a part of future research.

REFERENCES

[1] L. P. Baldwin, and J. Kuljis, "Learning programming using program visualization techniques", In Proceedings of the 34th Annual Hawaii International Conference on System Sciences, pp. 1051-1058, IEEE, Washington, DC, USA, 2001.

[2] S. Bergin, and R. Reilly, "The influence of motivation and comfort-level on learning to program", In Proceedings of the 17th Annual Workshop on the Psychology of Programming Interest Group, pp. 293-304, University of Sussex, Brighton, UK, 2005.

[3] B. Hanks, C. McDowell, D. Draper, and M. Krnjajic, "Program quality with pair programming in CS1", In Proceedings of the 9th annual SIGCSE Conference on Innovation and Technology in Computer Science Education, pp. 176-180, Leeds, United Kingdom, 2004.

[4] T. Jenkins, "On the difficulty of learning to program", In Proceedings of the 3rd Annual Conference of the LTSN Centre for Information and Computer Sciences, pp. 53-58, Loughborough, UK, 2002.

[5] A. Robins, J. Rountree, and N. Rountree, "Learning and teaching programming: A Review and Discussion", Journal of Computer Science Education, vol. 13, no. 2, 2003, pp. 137-172.

[6] S. Alaoutinen, and K. Smolander, "Student self-assessment in a programming course using bloom's revised taxonomy", In Proceedings of the 15th Annual Conference on Innovation and Technology in Computer Science Education, pp. 155-159, ACM, New York, NY, USA, 2010.

[7] T. Jenkins, "On the difficulty of learning to program", In Proceedings of the 3rd Annual Conference of the LTSN Centre for Information and Computer Sciences, pp. 53-58, ICS Subject Centre, Ulster, Ireland, , 2002.

[8] U. Nikula, O. Gotel, and J. Kasurinen, "A motivation guided holistic rehabilitation of the first programming course", ACM Transactions on Computing Education (TOCE), vol. 11, no. 4, 2011, art. no. 24.

[9] A. Yadin, "Reducing the dropout rate in an introductory programming course", ACM Inroads, vol. 2, no. 4, 2011, pp. 71-76.

[10] M. McCracken, V. Almstrum, D. Diaz, M. Guzdial, D. Hagan et al., "A multi-national, multi-institutional study of assessment of programming skills of first-year CS students", ACM SIGCSE Bulletin, vol. 33, no. 4., 2001, pp. 125–140.

[11] M. Konecki, and V. Mrkela, "Algorithmic thinking and animated interactive presentation of sorting algorithms in education of students", In Conference Proceedings of VIVID 2014 (Education in Information Society), pp. 105-112, Faculty of Organizational Sciences, Kranj, Slovenia, 2014.

[12] N. Hawi, "Causal attributions of success and failure made by undergraduate students in an introductory-level computer programming course", Computers & Education, vol. 54, no. 4, 2010, pp. 1127-1136.

[13] A. Gomes, and A. J. Mendes, "An environment to improve programming education", In Proceedings of the 2007 International Conference on Computer Systems and Technologies, pp. 88:1-6, ACM, New York, USA, 2007.

[14] D. A. Norman, and S. W. Draper, "User-Centered System Design: New Perspectives on Human-Computer Interaction", Erlbaum, Hillsdale, NJ, 1986.

[15] M. Konecki, "Problems in programming education and means of their improvement", DAAAM International Scientific Book 2014, pp. 459-470. DAAAM International, Vienna, 2014.

[16] M. Konecki, "Using video lectures in introductory programming courses", In Proceedings of IAC-EIaT 2014, pp. 256-260, Czech Institute of Academic Education z.s., Vestec, Czech Republic, 2014.

[17] M. Hu, "Teaching novices programming with core language and dynamic visualization", In Proceedings of the 17th NACCQ, pp. 94-103, Christchurch, New Zealand, 2004.

[18] J. Stasko, "Animating algorithms with XTANGO", ACM SIGACT News, vol. 23, no. 2, 1992, pp. 67-71.

[19] R. Oechsle, and T. Schmitt, "JAVAVIS: Automatic program visualization with object and sequence diagrams using the java debug interface (JDI)", In Lecture Notes in Computer Science, vol. 2269: Software Visualization, pp. 176-190, Springer-Verlag, Berlin, Germany, 2002.

[20] S. Grissom, M. McNally, and T. Naps, "Algorithm visualization in CS education: Comparing levels of student engagement", In Proceedings of the ACM Symposium on Software Visualization, pp. 87-94, San Diego, California, USA, 2003.

[21] M. Kölling, B. Quig, A. Patterson, and J. Rosenberg, "The BlueJ system and its pedagogy", Journal of Computer Science Education, Special issue on Learning and Teaching Object Technology, vol. 13, no. 4, 2003, pp. 249-268.

[22] A. Moreno, N. Myller, E. Sutinen, and M. Ben-Ari, "Visualizing programs with Jeliot 3", In Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 373-376, Gallipoli, Italy, 2004.

[23] M. C. Carlisle, T. A. Wilson, J. W. Humphries, and S. M. Hadfield, "RAPTOR: A visual programming environment for teaching algorithmic problem solving", In Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education, pp. 176-180, St. Louis, Missouri, USA, 2005.

[24] C. D. Hundhausen, and J. L. Brown, "What you see is what you code: A 'live' algorithm development and visualization environment for novice learners", Journal of Visual Languages and Computing, vol. 18, no. 1, 2007, pp. 22-47.

[25] A. Pears, S. Seidman, L. Malmi, L. Mannila, E. Adam et al., "A survey of literature on the teaching of introductory programming", ACM SIGCSE Bulletin, vol. 39, no. 4, 2007, pp. 204-223.

Data Acquisition Systems - Examples of Laboratory Exercises

Olivera Tasić*, Viša Tasić**, Darko Brodić***, Marijana Pavlov-Kagadejev**, Milena Jevtić****, Marko Anđelić**

*Mechanical and Electrical High School, Bor, Serbia
**Mining and Metallurgy Institute, Bor, Serbia
*** Technical Faculty in Bor, University of Belgrade, Bor, Serbia
e-mail: visa.tasic@irmbor.co.rs

Abstract – In this paper selected examples of laboratory exercises related to data acquisition systems are described. The examples are realized by using PC and microcontroller-based data acquisition systems. Basic elements of the analog signal conditioning, multiplexing, and analog to digital conversion are shown. The illustrated laboratory exercises can be used in the courses linked with the basis of automatic control systems as well as with programming languages. Because of its simplicity, the presented examples can significantly contribute to better understanding of the planned curriculums.

I. INTRODUCTION

Data acquisition is the process of collecting information that describes physical phenomena [1]. PC and microcontroller-based data acquisition systems have very important role in monitoring and control of the production process parameters. The schematic diagram of data acquisition system is shown in Fig. 1.

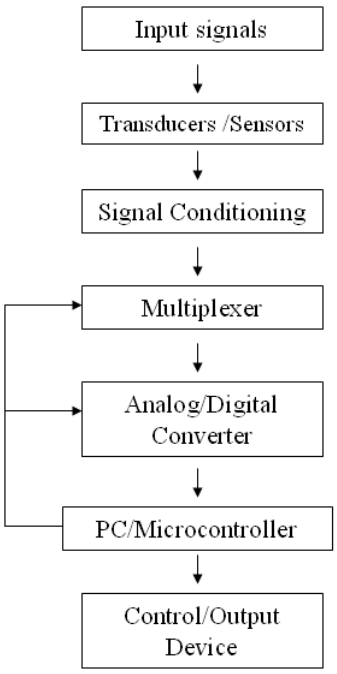


Figure 1. Schematic diagram of data acquisition system

In our examples, the input signal is a physical quantity (temperature, pressure, flow). A transducer or sensor is used to measure the input signal. The multiplexer selects which transducer is sampled at a particular time and directs the proper analog signal to the analog to digital (A/D) converter. The PC or microcontroller controls the multiplexer and A/D converter and receives the digital data from the A/D converter. After that, PC or microcontroller processes the data and sends it to an output device or uses it to control the response of another device. The practical experience that students gain in the laboratory is of particular importance for their understanding of the applicability of theoretical knowledge. Simple, clear, and interesting experiments and exercises are essential to motivate them to expand their knowledge.

Laboratory of Applied Electronics and Computer Science was established at the Institute of Mining and Metallurgy Bor in 2005, as a part of the Industrial Informatics Department [2]. Its primary role is to enable the lab exercises and realization of seminar and diploma papers for students of Computer Science at the Technical Faculty Bor, Serbia. Since the establishment, the laboratory has been used by hundreds of students and professionals. Selected examples of laboratory exercises of microcontroller and PC-based data acquisition systems are presented in the paper. The experiences of students and teachers regarding the work in this laboratory clearly indicate that lab work greatly contributed to a better understanding of the planned educational, professional, as well as the scientific program.

II. DATA ACQUISITION SYSTEMS USED IN LABORATORY EXERCISES

Programmable logic controllers (PLC) are the most popular type of process control computers, used in automation systems. They are controlled by programmed code created on a PC and downloaded to the memory of the controller. Depending on the state of the inputs and results of programmed functions, the controller will activate proper outputs [3]. Flexible solutions for data acquisition and control are also an important tool for measurements in scientific and research laboratories. Very often self-constructed or adapted measuring devices are applied in the laboratories. PLCs are modular devices and this modularity is often used to provide additional

functions, such as an increase the number of available inputs and outputs. In our laboratory exercises we used PLC called MMS [4]. In functional point of view, MMS is an autonomous unit (Fig. 2 and 3). The MMS standard configuration includes CPU module (one-board computer based on MC68HC11E microcontroller with internal eight channels, 8-bit A/D converter), analog input modules (up to 3 modules, each with 20 channels), digital input and output modules (64 + 64 channels), RS232 communication port, LC display and keyboard.

We have developed Supervisory Control And Data Acquisition (SCADA) application for MS Windows. It supports MMS functions in order to perform data transfer as well as to analyze and interpret the results [4]. I/O acquisition modules for PC have been programmed in Pascal, C, or Delphi programming languages. In our experiments, we have used MMS and PC modules (shown in Fig. 4), and measuring equipment, such: digital multimeters, oscilloscopes and signal analyzers.

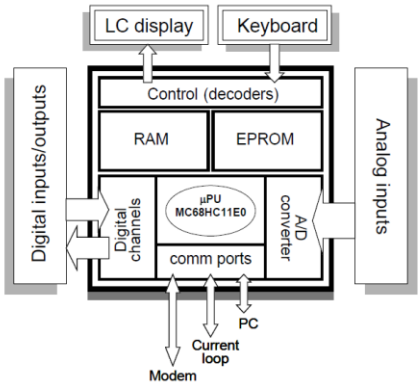


Figure 2. Functional block diagram of MMS

Figure 3. Basic configuration of MMS

Figure 4. Multifunctional data acquisition module for PC

III. LABORATORY EXERCISES

A. Temperature Sensors/Transducers

Transducers convert a physical response or phenomena into an electrical signal, which is proportional to the quantity of the input signal. In most of the lab experiments as sensing elements different types of temperature sensors is used, such as thermocouples, NTC and PTC thermistors (Fig. 5). In the exercise manual students can find basic information about each type of temperature sensors used in the exercise. The goal of this exercise is accomplished by understanding the principle of operation of each type of temperature sensor. Detailed information about the other types of sensors and transducers students can find in the additional literature.

B. Temperature Transmitters/Indicators

In this exercise students learn how to properly connect each given type of temperature sensor on temperature controller or indicator (Fig. 6). In the exercise manual students can find basic information about each type of temperature transmitter or indicator used in the exercise. The educational goal of this exercise is accomplished by fulfilling the technical goal which is setting up a transmitter or indicator parameters to prescribed values.

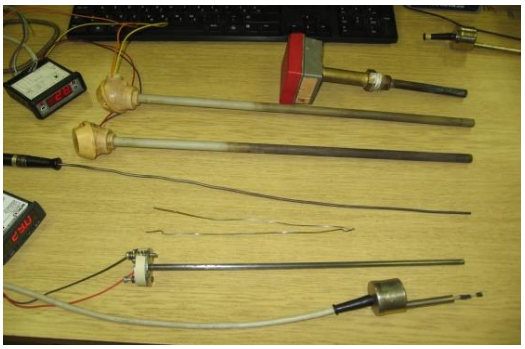


Figure 5. Different type of temperature sensors and transducers

Figure 6. Temperature sensors connected to controllers and indicators

C. Signal conditioning

Signal conditioning is used to modify the output from a sensor/transducer to increase the precision of the measurement. The role of analog signal conditioning is to preserve degradation of the measurement information, despite external influences. This includes signal isolation, amplification, and filtering. In the case, when the control system has no isolated analog inputs, a separate isolator will be needed (shown in Fig. 7).

Figure 7. Loop isolator JK2002. Input 0...4/20mA, output 0...4/20mA (2 separated channels in the same case)

The most of PLC analog input modules provide conversion of high-level signals (such as 0-5 V input to 4-20 mA output) as well as amplification and linearization.

In the exercise manual students can find basic information about the characteristics of MMS analog input modules (shown in Fig. 8). The manual also provides the equations for calculation the required components of the amplifiers. In the first part of the exercise, for a given signal range at the input and required signal range at the output, students determining the missing components of amplifying chain. The aim of this exercise is accomplished by fulfilling the technical goal which is checking the setting of analog channels to required values. The setting is checked in two points (zero and maximum).

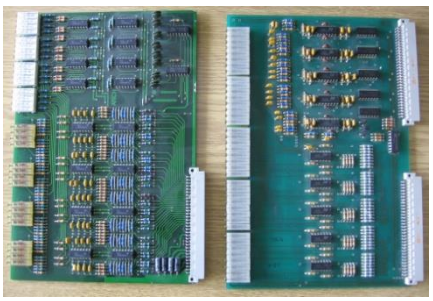


Figure 8. The MMS analog input modules (left – with 16 analog inputs, right – with 20 analog inputs)

Filtering is used to separate desired signals from undesired signals of ‘‘noise.’’ Noise causes the signal to fluctuate, which makes data interpretation very difficult. In the second part of the exercise, for a given level of noise at the input, and desired level of noise at the output of amplifying chain, students determining the components of filters. In the exercise manual, they can find the explanations and equations for the required calculations.

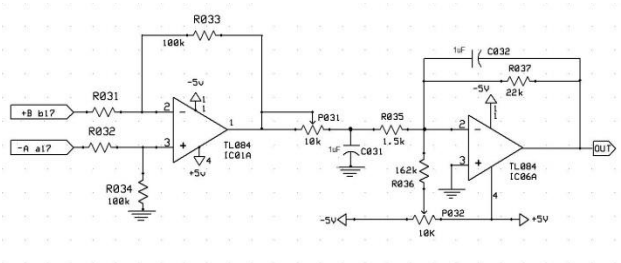


Figure 9. The example of two-stage amplifier (differential amplifier, a passive filter, and inverting amplifier/adder)

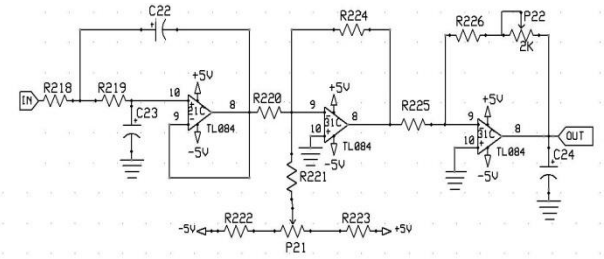


Figure 10. The example of active low-pass filter of the second order, inverting amplifier/adder, and inverting amplifier

The educational goal of this exercise is accomplished by recording the transmission characteristics of the designed filters.

D. Analog multiplexer

The analog multiplexer is a switching device that allows two or more transducers to be connected to one amplifier and one analog to digital converter. Generally, the multiplexer is controlled by acquisition software. The advantage of the multiplexer is that the user does not have to purchase multiple A/D converters if multiple transducers are to be monitored. In this exercise, students learn how to properly connect different types of analog multiplexers/switches (Fig. 11 and 12) and how to control them by acquisition software on MMS and PC.

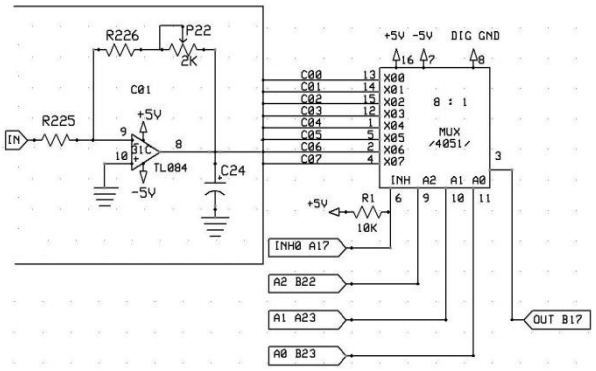


Figure 11. The example of connection the 8-channel analog multiplexer/demultiplexer 74HC4051

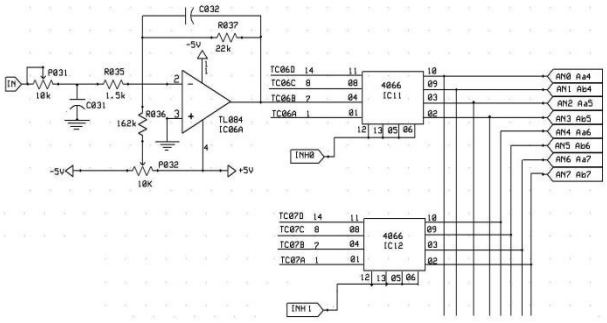


Figure 12. The example of connection the quad bilateral switch CD4066BC

E. Analog toDigital(A/D) converter

The A/D converter is a binary encoder that converts the transducer output into the corresponding binary value which can be further processed. The A/D converter can interface with a PC by direct connection to the internal bus, or through the external communication channels such as the RS-232, USB or Ethernet connection. In the exercise manual students can find basic information about each type of A/D converter used in the exercise.

The program sequence for control the A/D conversion is as follows:

- setting the address of analog channel that will be brought to the input of the A/D converter,
- setting the start of the A/D conversion,
- testing whether the A/D conversion has been completed,
- saving the results of A/D conversion.

The goal of the exercise is accomplished by understanding a short programs written in Turbo Pascal and Turbo C that control the 12-bit A/D converter AD678 (Fig. 13) [2].

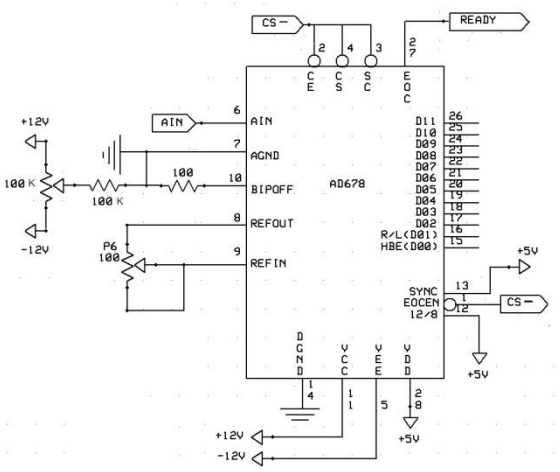


Figure 13. The example of AD678 unipolar input connections with gain and offset trims

F. Data acquisition software

The performance of a data acquisition system is largely dependent on the software controlling the multiplexer, A/D converter, and the output/control devices [1]. The special SCADA real-time application, operating in MS Windows environment (Windows 98/2000/XP), was developed in order to support MMS functions and to perform data transfer, analyses and real-time interpretation of results [4]. It was based on client/server architecture running on both master and remote stations, enabling integration in a complex distributed control and monitoring system. SCADA was developed using Microsoft Visual C++ 6.0 [5]. During the exercise, students learn and practice how to set the monitoring properties of the SCADA application related to the configuration of the input parameters.

Figure 14. The example of data presentation in tabular form in the SCADA application

IV. CONLUSION

The students and teachers experiences regarding the work in this laboratory clearly indicate that it greatly contribute to a better understanding of the planned educational program. In order to labs done correctly some previous knowledge in electronics and electrical engineering are required. It is also necessary that students undertake serious preparations prior performing the exercises. These preparations consist of a detailed study of the instructions for the exercise and recommended literature before coming to the lab. Even though, the majority of students thinks that the exercises are interesting and contribute to a better understanding of the teaching content.

ACKNOWLEDGMENT

This work is supported by a grant from the Ministry of Education, Science and Technological Development of the Republic of Serbia, as a part of the project No. TR33037 “Development and Application of Distributed System for Monitoring and Control of Electrical Energy Consumption for Large Consumers."

REFERENCES

[1] A. Kent, J. G. Williams, Enciclopedia of Microcomputers, vol. 4. New York and Basel, Marcel Dekker, Inc., 1990, pp.197–206.

[2] V.Tasić, D.Brodić, D.Milivojević, M.Pavlov, “Practicing of Lab Exercises at the Laboratory of Applied Electronics and Computer Engineering”, Proceedings of 6th. International Scientific Conference Computer Science 2011, 01.-03.09.2011. Ohrid, FJR Macedonia, pp. 429-434, ISBN 978-954-438-914-7.

[3] R. Buinac, D. Maršić and G. Malčić, “Web-based control of programmable logic controllers in process control computers education,” Proceedings of 37th International Convention on Information and Communication Techology, Electronics and Microelectronics MIPRO 2014, pp. 917-920, Opatija, Croatia, 26.05.-30.05.2014.

[4] V.Tasić, V.Despotović, D. Brodić, M.Pavlov, D.Milivojević, “Twenty years of monitoring and control of electricity consumption in RTB Bor, Serbia,” Proceedings of 36th International Convention, MIPRO 2013, pp. 1134-1138, Opatija, Croatia, 20.5.-24.5.2013.

[5] B.Stroustrup, The C++ Programming Language , Addison-Wesley Pub Co, 3rd edition , 2000.

The Gamification of Education

Sanja Maravić Čisar*, Robert Pinter*, Petar Čisar**
* Subotica Tech-College of Applied Sciences, Subotica, Serbia
** Academy of Criminalistic and Police Studies, Belgrade-Zemun, Serbia
sanjam@vts.su.ac.rs, probi@vts.su.ac.rs, petar.cisar@kpa.edu.rs

Abstract - Programming is a major subject in Computer Science (CS) departments. However, students often face difficulties in the basic programming courses due to several factors that cause these difficulties. The most important reason may be the lack of problem solving abilities that many students show. Learning to code can be made more effective and sustainable if it is perceived as fun by the learner. Code Hunt uses puzzles that players have to explore by means of clues presented as test cases. Players iteratively modify their code to match the functional behavior of secret solutions. The aim of this paper is to present the results of implementation of Code Hunt in learning Java as part of the Object Oriented Programming curriculum. The participants of the research were students of Subotica Tech majoring in computer engineering. The research was conducted over the fall semester in the 2014/2015 academic year.

I. INTRODUCTION

Programming is a complex mental activity that is defined as an abstract process. Understanding and visualizing abstract processes poses a considerable problem for students when learning programming, as well as other fields with similar characteristics. Many students find it frightening, as programming might require thinking in ways of abstraction that has never been practiced before. Learning to code can be made more effective and sustainable if it is perceived as fun by the learner.

Games are an integral component of people's daily lives. Over $20.5 billion was spent on video games by Americans in 2013 [1]. While the games' primary aim is to present entertainment, given how universally applicable they are, games have long-since outgrown their entertainment role. Games are implemented not only to provide entertainment, but also in the fields of defense, education, scientific exploration, health care, emergency management, city planning, engineering, religion, and politics. Apart from entertainment games there are also the so-called serious games whose primary goal is to train, investigate, or advertise.

Code Hunt is a highly effective educational gaming platform that internally leverages fitness values so as to guide the user to successful test/clue generation, furthermore, it also externally provides its users with entertaining learning experiences where search-based test generation is manually emulated. Code Hunt is a new attempt to create a game using serious search-based testing, make it available for a huge group of users, including coders, and especially learning coders [2]. The test cases are constantly changing, since they are built based on the mined data, thus players will enjoy a new experience every time. Learning to code can become increasingly efficient and sustainable if the learner sees it as 'fun' as opposed to 'learning to code'. Code Hunt was launched in May 2014.

II. BACKGROUND-PEX FOR FUN

Code Hunt uses puzzles that players have to explore by means of clues presented as test cases. Code Hunt is based on the test/clue generation of Pex, a white-box test generation tool that uses dynamic symbolic execution.

Pex [3] is an automatic white-box test generation tool for .NET, based on dynamic symbolic execution. This tool is integrated into Microsoft Visual Studio in the form of an add-in. It can generate test inputs which are combined with different unit testing frameworks [4]. Pex have been implemented in classroom teaching at various universities (for example North Carolina State University, University of Illinois at Urbana-Champaign, and University of Texas at Arlington), and also in a variety of tutorials both within Microsoft (such as internal training of Microsoft developers) and outside Microsoft (such as invited tutorials at .NET user groups). Further, the authors of Pex created numerous open source research extensions upon it [5].

One of the most important methodologies that Pex supports is called parameterized unit testing, which broadens the scope of today's industry practice which prefers closed, traditional unit tests (i.e., unit test methods without input parameters) [4].

There are useful characteristics that Pex offers to support for testing. Primarily, there is the option of exploring code and suggesting the tests that should be done. Secondly, assuming that it is a parameterized test, Pex can determine the combination of parameters that has to be tested so as to provide all feasible versions. Lastly, once Code Contracts is being used, Pex uses that information to fine-tune the unit tests that are offered or generated for the user [6].

III. CODE HUNT, AN OVERVIEW

As stated above, Code Hunt is an educational coding game. The play scenario is the following: the student will impersonate the code player, whose task is to detect the missing code fragments. Points are gained by completed level, while extra points are granted for elegant solutions. The game can be implemented for either C# or Java. During the game, the player will make their way to specific sectors and gain knowledge about arithmetic operators, conditional statements, loops, strings, search

algorithms and more. The entire gaming experience has an underlying automated grading engine based on dynamic symbolic execution. The grading engine automatically analyzes the user code and the secret code so that the result table is generated. Most modern browsers including Internet Explorer 9, 10, 11 and recent versions of Chrome and Firefox are suitable for running Code Hunt [2].

Figure 1. The start screen of Code Hunt

Code Hunt resembles most games in the sense that there are sectors and levels. Code is written in an editor window by the player, they use either C# or Java as the programming language.

An initial set of given code examples called puzzles are being displayed in the working area for users (players) to play with. Each puzzle has the main method with the name “Puzzle”. After a puzzle is loaded in the working area, a player could click the “Capture Code” button to compile and run it. After that [2]

(1) the code is sent to a server in the cloud,

(2) the server compiles the code (including an optional Java-to-C# conversion),

(3) the server starts an in-depth analysis of the code, it is compared with the goal algorithm,

(4) the player will then be presented with the results (Figure 2).

Figure 2. The results of the compilation

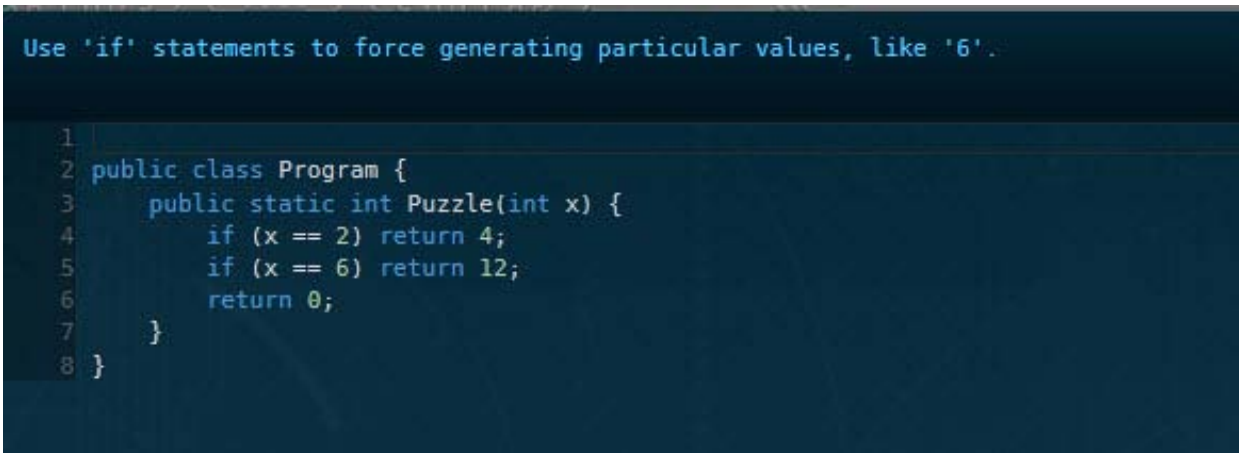


Figure 2.a The secret code

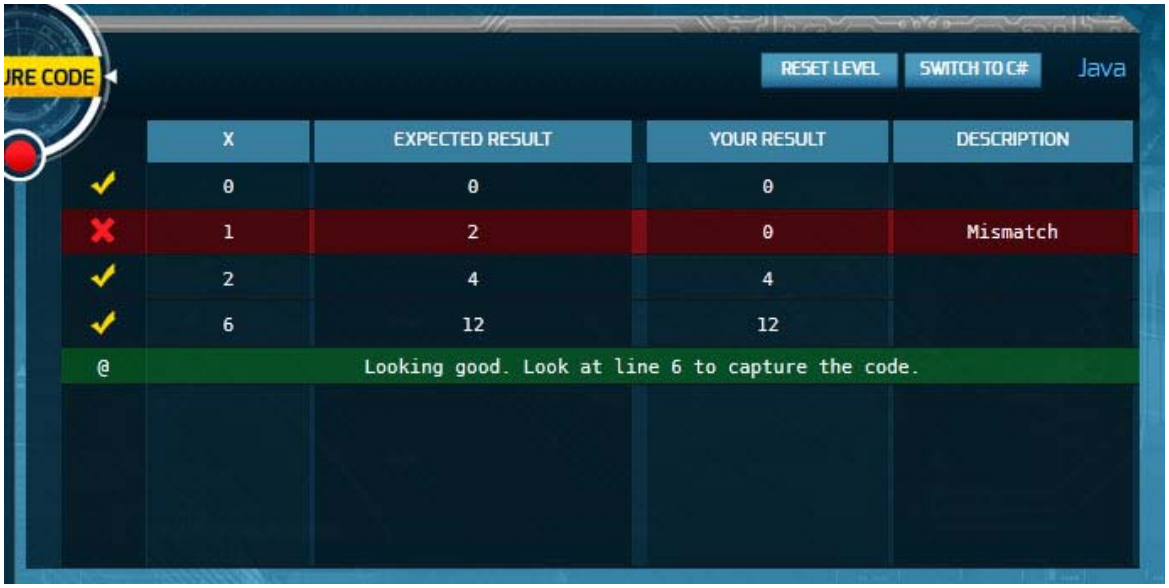


Figure 2.b The result table generated by analyzing the user code and the secret code

The main Puzzle method can take parameters and return values. In order to run such a Puzzle method, someone must provide argument values. Code Hunt automatically finds interesting argument values by analyzing the code. The result will be either a compilation error, with information in the bottom pane, or some mismatches or agreements with the goal algorithm. The code is presented left in Figure 2, while on the right there are the mismatches (red crosses) and agreements (yellow checkmarks).

The player has completed this level successfully if the code has compiled and no mismatches were found, only agreements with the goal algorithm. In the game this is described as the player having “CAPTURED!” the code, here seen in Figure 3.

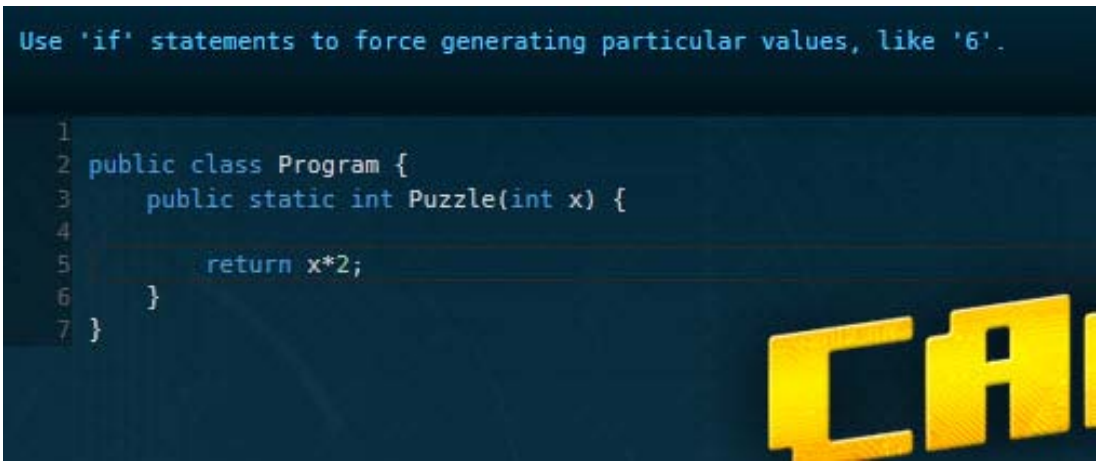


Figure 3. After solving the puzzle, the player receives a score

As a follow-up, each mismatch and agreement with the goal algorithm is returned in the form of a tuple (input, actual result, expected result). In the case when the player's code is in agreement with the goal algorithm, the actual and expected results are identical, while, if a mismatch has occurred, they are different. The player is given a chance to examine the mismatches and figure out how to modify and improve the code so a greater resemblance is achieved with the goal algorithm [2].

Upon the completion of a given level, the player will be prompted to choose a somewhat more difficult level, so as to keep up the flow. Also, all sectors bar the first one are locked as their default setup. Only when the player has completed a satisfactory number of levels in a certain sector does the next sector become unlocked (Figure 4).

Figure 4. The game's sector

IV. RESULTS

The participants of the research were students of Subotica Tech majoring in computer engineering. The research was conducted over the fall semester in the academic year 2014/2015. The sample consists of 52 students from the second year of computer science who attended classes in the subject Object-Oriented Programming II. This course examines object-oriented programming techniques and issues in Java programming language. The student obligations during the semester include passing two midterm exams, doing one home assignment, and completing the final project. From the total numbers of students, 8 were female (15.38%) and 44 male (84.62%) students. Until the end of the semester 4 students had dropped out (they did not attend classes) and did not participate in the research itself.

Students were asked to rate their understanding of various topics that are covered with Code Hunt (arithmetic, loops, conditionals, strings, nested loops, arrays). The scale was from 1 to 5 (1="insufficient", 2="sufficient", 3="good", 4="very good", 5="excellent"). The results are shown in Table 1.

TABLE 1 THE AVERAGE RESULT OF UNDERSTANDING VARIOUS TOPICS

	Average (max of 5.0)
arithmetic	3.71
loops	3.25
conditionals	3.10
strings	3.45
nested loops	3.08
arrays	2.89

As can be seen from the results, students in general were confident in all topics. On the first midterm exam they had an average of 11.23 points out of 15, so it could be concluded that they rated their knowledge and understanding of various topics objectively.

In the second half of the semester students were taught the following topics: classes, objects, encapsulation, virtual functions, inheritance and polymorphism. This was followed by knowledge evaluation. In this test where the students' knowledge was tested, the questions referred to the second part of the semester when CodeHunt was not used in class. In general, students had some difficulties in comprehending and implementing this part of the curriculum. They rated their understanding of topics as shown below, in Table 2. In all categories their knowledge was above "good". On the second midterm exam which

covered these topics they had an average of 7.33 out of 15 points.

TABLE 2 THE AVERAGE RESULT OF UNDERSTANDING OOP TOPICS

	Average (max of 5.0)
classes	2.63
objects	2.54
inheritance	2.47
polymorphism	2.18
virtual functions	2.13

The semester ended with a survey regarding their views on implementing CodeHunt in their studies and if they liked this gaming approach. CodeHunt was used in the first part of the semester. In the second part of the semester the students are quizzed on the following items: classes, objects, encapsulation, virtual functions, inheritance and polymorphism, but the knowledge evaluation was not performed using CodeHunt.

The possible answers were 1=”I hate games”, 2=”it was so-so”, 3=”it was fun”, 4=”it was great“. The value 2.75/5.0 refers to the average grade of 2.75, whereas the maximum was 5.0. The obtained average result of 2.75 indicates that the students liked this way of learning and they marked it as “fun”. Students who played the game, solved an average of 45.7 percent of Code Hunt’s puzzles. The statement “the puzzle aspects of Code Hunt keep me focused on trying to reach solutions,“ showed that the majority of students agreed (“A lot” 37.5% and “Quite a bit” 45.83%) and 16.67% answered “Not at all”. Only 5 students (10.42%) answered that they tried to find an elegant solution and the rest of them (89.58%) replied that they only tried to find a solution that worked (“My solution was a mess but it works”).

V. CONCLUSIONS

There is a considerable difference in learning to code by solving a puzzle and learning to code by writing to a specification. A great number of competitions can be found that have students competing against other students – as well as trying to beat the clock – so that they find a solution matching the given problem. The advantage of this sort of coding is that it simulates real-life situations the students are likely to be up against in their professional lives. Code Hunt is focuses not on the process of learning to code, in fact, it is form of side-effect the problem-solving itself, shown as pattern matching inputs and outputs. The fun part is provided by the challenge of detecting the pattern. Fun has been defined as an indispensable component aimed at increasing the learning efficiency and keeping the students interested in this coding process, which may, at times be seen as tedious and time-consuming process with the result: learning the vital coding skills [2].

Code Hunt is a highly effective educational gaming platform that internally leverages fitness values so as to guide the user to successful test/clue generation furthermore, it also externally provides its users with an entertaining learning experiences where search-based test generation is manually emulated. Code Hunt is a new attempt to create a game using serious search-based testing, make it available for a huge group of users, including coders, and especially learning coders [2]. The test cases are constantly changing, since they are built on the mined data, thus players will enjoy a new experience every time. Learning to code can become increasingly efficient and sustainable if the learner sees it as ‘fun’ as opposed to ‘learning to code’.

REFERENCES

[1] http://www.forbes.com/sites/davidewalt/2013/12/19/americans-will-spend-20-5-billion-on-video-games-in-2013/

[2] N. Tillmann, J. Bishop, R. N. Horspool, D. Perelman, and T. Xie, “Code Hunt: searching for secret code for fun”, ACM ICSE Workshop on Search Based Software Testing 2014,
https://homes.cs.washington.edu/~perelman/publications/sbst14-codehunt.pdf

[3] http://research.microsoft.com/projects/pex/

[4] N. Tillmann, J. de Halleux, and T. Xie, “Pex for fun: engineering an automated testing tool for serious games in computer science”, March 2011, TechReport, MSR-TR-2011-41.
http://research.microsoft.com/pubs/147143/pexforfun-engineering.pdf

[5] http://pexase.codeplex.com/

[6] D. Esposito, “Pex: Microsoft research's unit test generator and evaluator”, November 26, 2012.
http://www.drdobbs.com/testing/pex-microsoft-researchs-unit-test-genera/240009056

Experiences and Practices in Blended Learning Environment

T. Horvat, T. Alajbeg, S. Predanić
University of applied sciences, Department of Electrical Engineering, Zagreb, Croatia
thorvat3@tvz.hr; talajbeg@tvz.hr; spredanic@tvz.hr

Abstract - The teaching materials, laboratory exercises, and the whole assessment system of the Personal Computer Applications (PCA) course, in the professional study of Electrical Engineering at the University of Applied Sciences in Zagreb, are based on the Moodle e-learning system. Generally speaking, there are always pros and cons when e-learning is considered. Students have access to all teaching materials, tasks and direct communication with the teacher at all times. On the other hand, if there are problems, it is hard for the teacher to distinguish the exact cause. Students may encounter problems in understanding the materials or in the use of instruments and computer applications; without direct contact, it is virtually impossible for the teacher to motivate them to continue working despite the difficulties. Another challenge for the teacher is detecting which parts of the subject are most interesting to the students and which parts demand more attention and maybe different methodologies. Using blended learning, combining the advantages of both e-learning and person to person teaching / learning, it is possible to achieve a better effect for the mutual benefit. This paper presents the teachers experience and students perception of blended learning as well as assessment system.

I. INTRODUCTION

New IT technologies are being used in education for years, and they are especially characteristic for higher education. Recently, the combination of classical way of education (face-to-face, brick-a-mortar, chalk-and-talk) and e-learning is becoming more and more popular. That combination is often called blended learning, as well as hybrid learning. There are different definitions of blended learning [1], [2], [3], [4], [5], but generally it can be said that it is a combination of classical and technology mediated learning [2]. Speaking from our own experience [6], [7] and definitions mentioned above; it could be said that the phrase „blended learning“ implies a formal educational system that combines on-line and face-to-face lectures (or knowledge transfer in any possible meaning) with assessment system. All this can be achieved by applying different teaching and assessment methods [5], [6], [7], and by different ways of preparing for the exercises, availability and possibilities for practising and learning.

Based on these experiences and models mentioned in the literature, one can distinguish several forms of blended learning. In the first form the teacher, working in person with the students, notices a strong interest or necessity for additional clarifying, and gives customised additional content for each student (in the laboratory or on-line to work at home). In the second form, the students attend lectures and exercises according to the given schedule, while all of the materials are available online in learning management system (LMS). Different construction assignments are given to each student, through LMS, and they need to solve them using prior gained knowledge and online support. And the third form, in which the students prepare for classes from online materials, and the lectures only increase the level of understanding. In that form students do the exercises independently, online, from home. Although the literature mentions several ways of organizing blended learning, the essence are these three forms. Different combinations and use of each individual component of these three ways are also possible. In the PCA course the second method of blended learning organization is mostly used.

II. ORGANIZATION OF THE PCA COURSE USING BLENDED LEARNING ENVIRONMENT

The teaching and assessment system on the PCA (Personal Computers Applications) course, at the professional study of the Electrical engineering at the University of Applied Sciences in Zagreb, are based on the ideas mentioned in previous chapter. Lectures, laboratory and construction exercises on the PCA course are the parts of blended learning environment which consist of general and personalised instructions, practical and individual work for each student.

Since the course is at the beginning of study, it includes general IT and computing concepts, possibilities, and basic work principles of parts and subsystems in a computer. Also, very important part is the practical work with computer applications which can be of great importance to future engineers. Practical work is individual; each student has their own computer, so it is necessary that they come to the exercises well prepared for the current subject. To assure that, they have constant access to the materials for learning and practising (through Moodle e-learning system) and computer applications (use of free or demo software) [6].

Basic use of computers is not easy to present to the participants who come from different backgrounds. They had different availability of information technology and the Internet, thus varying degrees of knowledge. Their previous interest in the subject can not be ignored as well [6]. The personal contact between the students and the teacher helps to detect students' sphere of interest, their motivation and enables teacher to try to encourage and

motivate students. These are the reasons why the PCA course is designed in the blended learning environment.

First part, of the blended learning environment in the PCA course, are the lectures. Materials – slides for the lectures are available, even before the lectures, in Moodle e-learning system. Web links and a list of additional literature with detailed explanations of terminology, work principles etc. are a part of the slides. Lectures are held, face - to - face, in two relatively large groups (around 100 students in each group), where students can actively participate, ask questions and further explanations.

Second part are the laboratory exercises. Materials needed for the preparation before the exercise, are also available through Moodle. The laboratory exercises are held in smaller groups (around 12 students in each group), so each student works on their own computer. At the beginning of the exercise, the teacher explains the goals, purpose and the way the exercise will be done. Then, the work with a certain application is presented, with the emphasis on the most important parts which the students might need in further work. After that, the students practice working with the application on their own, according to the given instructions. The teacher is in the laboratory the whole time, and is available for all the possible questions – for those students who need extra help, and for those who are interested for more advanced uses of the application.

There are also construction exercises, which are the completely individual student work. Students need to write a simple programming code; using all of the previously gained knowledge and skills; using materials available and sharing experiences with their colleagues.

Blended learning environment has, through the years, proved to be the most efficient way for organizing and administering the course, and assessing students' knowledge. Initially, a lot of time is required for the design, preparation and organization of such a system, the creation and adaptation of necessary teaching materials. Details regarding to certain specific features of material preparation, assuring that the computer applications used in the classes are available for the students to practice at home, as well as system administration and system assessment are outlined and described in [6] and [7].

Creating such a system is never quite finished: over the years, occasional corrections are needed – changes in modalities of organization, administration and ways of processing the topics, all based on empirical indicators. Although there is always a need for regular improvements and modernization (it is on the computer, so constant change is normal), the use of blended learning environment after a few years begins to bear benefit: as for the time, and as far for the resources (space, energy consumption, working hours...) and perhaps most importantly - helps students, in the most simple way, gain necessary knowledge and skills. In the next chapter, perceptions of consumers (students) of blended learning will be discussed. That includes a comparison with the conventional lectures and full e-learning system, and their perceptions of the usefulness of this type of organization of teaching and examination system on the PCA course.

III. STUDENTS' EXPERIENCE IN USING BLENDED LEARNING ENVIRONMENT

Students, during their education, encounter different ways the organization of teaching and assessment. How do students perceive this way the organization of teaching and assessment system on the PCA course? What are the good sides and what are not so good sides of this environment? What do they find extremely important and interesting, what they think is not as important or interesting, and what makes them want to work and know more? Quantitative research on this subject was conducted among the students at the PCA course, at the end of the semester in a form of a survey with 11 multiple choice questions – some single answer and some multiple answers, without ranking. The survey was created through Google forms, which ensured anonymity, and the link on it was available for students through Moodle. There are 216 students enrolled in the PCA course, in the academic year 2014./15., and 153 (randomly chosen) of them completed the survey. The questions regarding pure e-learning and chalk and talk were referring to their experiences in general, while the questions regarding blended learning were referring on their experiences on the PCA course. That way, it can be determined how students perceive blended learning on the PCA course in comparison to chalk and talk methods which are still standard in education. The survey was in Croatian language, so it appears in figures shown in this paper, as it is the mother tongue of almost all of the students.

A. *E-learning – pros and cons*

Question regarding the technologies, which do they find most acceptable for learning and most useful way of transmitting / collecting information and knowledge, was chosen as a starting point. Computer with the Internet connection for 76.47% of the responders represents the most useful and most used medium for this particular purpose. Just a little more then one fifth of them (22.88%) considered books and printed materials the most important and most useful for transmitting and collecting knowledge and information. Interestingly, less than 1% considered the same for TV. Based on this result, it is obvious that it is very important for the students, in information gathering and learning, to have a computer available materials and lessons (lectures). Considering the extremely high percentage, this fact can be further used to design more materials to prepare for the lectures and learning.

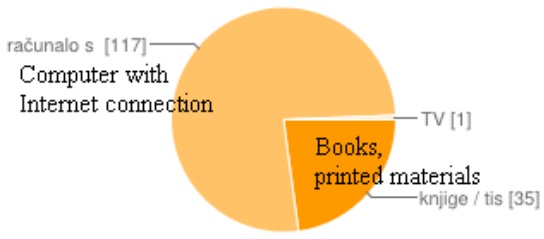


Figure 1 Which technologies do students use for learning?

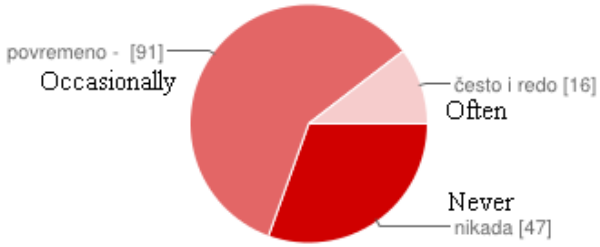


Figure 2 How often did students use pure e-learning?

Therefore, students were asked if they have, and how completely independently and actively, used pure e-learning in any form (lessons, lectures, exercises...). Question considered any subject or area, the important thing was that the materials were only accessible via computer with Internet connection, such as for example: khanacademy.org, tonimilun.com, tesla.carnet.hr or YouTube in terms of tips & tricks how to work with some computer applications. This way of learning is often used by 9.80% of students but 3 times more (30.72%) never used pure e-learning. Pure e - learning was occasionally, therefore rarely, used by 59.48% of responders. Thus, a total of more than 90% of responders never or rarely used computers available materials for independent learning. Seemingly, this is in contrast with the fact that for them almost 77% of them the computer with an Internet connection is the most important and useful access to information and knowledge. However, this is completely independent learning, occasionally or never practised more than 90% of students, and these two data are not to be confused.

As major drawbacks or limitations of pure e-learning system [multiple answers could be checked] those who regularly or occasionally used it (9.80% of them often and 59.48% occasional users, a total of 69.28%) reported the one-way communication (52.29%) and the inability to verify the quantity and quality of learned material (38.56%). Respondents also find sociological component significant, the inability of consultations with fellow students when learning is the limiting factor of e-learning for a third of the population (33.97%). A minority (1.3%) considered material unavailability due to works on the network or slow web link as a big limitation.

However, two-thirds of these students (69.28%) who are periodically or regularly using e-learning system for independent learning, as the biggest advantages of this way of learning lists current and easy access to the requested material (55.56%). The possibility of repeating the lessons (replay) is important to 49.02% of the students. Ability to access materials from anywhere at any time of the day (and night) is considered an advantage for 48.37% of the responders. For a third of respondents (32.68%) it is important that the all of the materials are concentrated in one place - everything needed is accessible within a few clicks.

B. Chalk-and-talk - pros and cons

In a specific time in a specific place, lectures are held ex cathedra, exercises in groups, chalk and black board or marker and white board, tests corrected manually, results available to students in a couple of days. It's hard to keep track of own achievements, especially in several courses with different modes of evaluation, which is not in accordance with the Bologna process in higher education. This is a general and brief factual description of the traditional, classic, way of teaching that still prevails in the professional study programs at the University of applied sciences in Zagreb.

When asked to highlight the most important disadvantages of this, traditional way of teaching organization and assessment system, students have pointed out the murmur and similar distractions in class when the more people are attending the lectures (49.02%). The next greatest shortcoming in students' opinion is that the subjectivity of teachers can affect the grade (48.37%). The fact that they have to come to a specific place at a specific time is a disadvantage for a little more than one third of the population (36.60%), and almost the same percentage (35.29%) feels that the deficiency is the emphasis on aural rather than the visual. Despite the shortcomings, this system of teaching was considered the best for one third of the student population (34.64%) but only if it is performed in small groups.

Of the advantages of the classic ways of organizing teaching and assessment system, students with very high percentage emphasize the interactivity with the teacher (73.20%), therefore, they very highly appreciate and the possibility of direct personal communication with the teacher. Next in significance to them is another sociological component: the possibility of direct and personal communication with their fellow students (59.48%). On the third place in order of importance is also another sociological component: the possibility of the experience transfer (from teachers to the students) in the work or when meeting a certain phenomena (35.29%).

Looking at these results, concerning the classical way of teaching, it can be concluded that it is a way of teaching that fits the students but only if it is a small group. This is concluded on the basis of direct claims of a third of students (34.64%) and based on a high percentage of them who considered direct contact with the teacher (73.20%) and colleagues (59,48%) to be the most important. Of course, this kind of teaching with small groups is not possible to organize in the mandatory courses, especially in first two years of the study where there is a lot of students. Sometimes it is possible to organise classes in this way, for some elective courses in the third year of the study, which enrolled a small number of students.

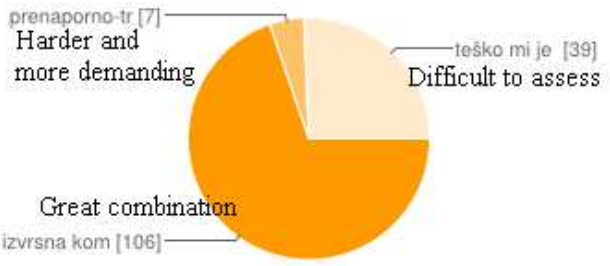


Figure 3 How do students perceive blended learning environment in relation to classic way?

C. *Blended learning*

Comparing the mutual advantages and disadvantages of traditional and e-learning system, it is noted that some benefits of one at the same time disadvantages of another. For example, in the classical way of teaching organization one of the advantages is the interactivity with the teacher and fellow students; while the impossibility of two-way communication is mentioned as a disadvantage in pure e-learning system.

When asked how often they have met with a combination of e-learning and traditional ways of teaching, during their education, regardless of the degree of mutual integration, 7.19% of the respondents answered that it was often. Occasionally (33.99%) or never (58.82%) with such a mode of organization of teaching encountered 92.81% percent of the student population.

How do they, after the entire semester of teaching / learning in blended learning environment, perceive this way the organization in relation to classic way, was the following question. 69,28% of students considered this to be a good combination, while 5.23% estimated blended learning environment harder and more demanding in relation to the classical way of teaching. Quarter of the respondents (25.49%) found it difficult to assess which method is better.

Most of the students (73.20%) highlighted continuous availability of materials and computer applications for learning and practising and all at one place as the most important and most useful characteristic of blended learning system. A large number of respondents (58.17%) considers a direct and personal (eye-to-eye) contact with the teacher as the very significant part. The ability to communicate with the teacher and outside the class schedule or consultation is significant for 44.44% of the students, and almost identical number of them (43.14%) finds direct and personal (eye-to-eye) contact with their colleagues important. Indirect conclusion may be that the social component is one of the key parts in the process of studying and acquiring knowledge and skills.

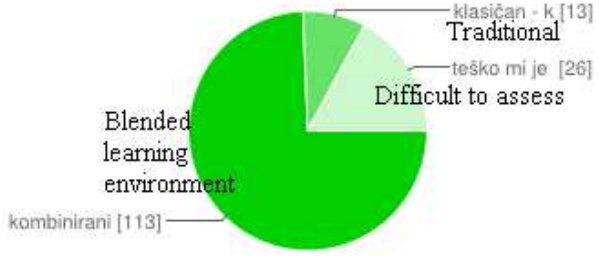


Figure 4 Which manner of organisation motivates students better?

Since the Bologna process requires continuous monitoring of students, the students were asked to evaluate the importance of constant availability of their achievements, but also a plan of lectures and exercises visible in advance so they can plan their tasks and activities. This is, of course, possible to enable only using computers. 55.55% of the respondents considered this important and useful in this kind of organization. For the majority of students, even 83.00%, speed of assessing their tests and the availability of the results as soon as possible is very important and useful.

The attitude of students towards the role and influence of teachers in the evaluation process is particularly interesting: for the 39.87% of them it is an advantage that in tests which are assessed automatically, by the computer, the subjectivity of teachers is eliminated. However, even 47.06% of students considered it important that, despite the fact that the computer evaluates the tests, the teacher can intervene in the assessment at their sole discretion. E.g., if the boundary for the pass is set on 50.00%, the border can be moved to 49.65% by teacher's intervention.

Finally, the study was concluded with a personal impression of students, which manner of class organisation and the assessment system encouraged them to learn more, to achieve better grades and to acquire new skills. Even 73.86% of the population believes it is blended learning environment, compared with 9.15% who thought the traditional manner. Others (16.99%) found it difficult to assess.

IV. CONCLUSION

Blended learning can not be considered as a mere sum of two ways of teaching organization, it can not be regarded as the classic method of teaching with added e-learning component. Or vice versa; e-learning with possible consultations. In the organization, implementation, application and interpretation of blended learning environment is required a holistic approach. Part of the results of the research indicates it, such as the one about the role of computers and teachers in the evaluation process. This research can not be considered exhaustive or final. It has only discovered and opened new questions that deserve to be answered. For example, students (in the part of useful segments in blended learning environment) were asked to assess how important is the ability to communicate with the teacher and outside the class schedule and consultation. This was considered important and useful for 44.44% of students. Without denying that sometimes indeed there is such a need (communication with the teacher outside the class schedule and consultation), one might reasonably ask is this, at least in part, a reflection of students' poor organization of their own time? Or dependence in work and learning when they expect regular incentives by the teacher? To search for answers to these and similar questions, it is needed to carry out extensive research over several generations.

Organisation of the course - lectures, exercises and assessment system is a consequence of the number of students in this course and the subject matter. In the study

of electrical engineering, one of the key segments is the practical work with instruments and a variety of computer applications. This conditions the organization of teaching and assessment system on each professional course. The fact is that the PCA course annually enrols more than 200 students and that they should master practical work with a variety of computer applications. This should be done regardless of their different backgrounds, as for the interests and their own skills to work with computers. Experiences have shown that this type of organization of teaching, exercises and assessment system is very efficient in these situations. The conducted survey confirms this. This regime has been applied for seven years on this course. It would be interesting to accurately calculate and show how much benefit for the institution (University of applied sciences in Zagreb) this brought, how many teaching hours have been saved, as reduced energy consumption.

Results of this study suggest that everything possible in the process of organization of teaching and assessment systems should be organized in the e-learning system. Some of the examples are: preliminary exams and regular exams, records of attendance and other types of administration, materials for attending lectures and preparation for the exercises. Space where students can upload their work (such as seminar papers) should be available as well. Besides the advantages for the teachers, less paperwork being the biggest advantage, the results of research directly indicate the benefits of the online available materials (section III A).Considering that for almost 77% of the respondents computer with an Internet connection is the main way of transmitting information and knowledge, over the next academic years it would be useful to create video clips with instructions for the use of certain tools and software, explanations of some computer terms and procedures. Almost 74% of respondents stated that blended learning encourages them to learn, acquire knowledge and skills and ultimately achieve better-marks which motivates the teachers to continue working and developing in these areas. Research will be repeated in the few next academic years to monitor and compare the results, and after a number of years to make a comparative analysis with a definite conclusion.

LITERATURE

[1] M. Bower, B. Dalgarno, G. Kennedy, M. J. W. Lee, J. Kenney: Blended synchronous learning, a handbook for educators, Australian Government Office for Learning and Teaching, 2014. ISBN: 978-1-74361-686-4 [PDF]

[2] Handbook of research on educational communications and technology, 3rd edition, 2008, Lawrence Erlbaum Associates Taylor & Francis Group New York, ISBN 0-203-88086-2, edited by: J. Michael Spector, M. David Merrill, Jeroen van Merriënboer, and Marcy P. Driscoll

[3] http://www.christenseninstitute.org/blended-learning/ - [15.1.2015]

[4] http://aspirepublicschools.org/media/filer_public/2013/07/22/aspire-blended-learning-handbook-2013.pdf - [15.1.2015.]

[5] C. J. Bonk, C. R. Graham, The handbook of blended learning: global perspectives, local designs, Pfeiffer, San Francisco 2006

[6] T. Alajbeg; T. Horvat; T. Novak: The organization of classes and assessment system using Moodle, MIPRO, 2014

[7] T. Alajbeg; D. Ćika; T. Ražov: Using Moodle to automatically grade laboratory exercises, MIPRO, 2011

Application of ICT in Process of Informatization of Education System - Efficiency of Digital Log Book

Dragana Glušac*, Ivan Tasić*, Biljana Radulović*, Žarko Mušicki**
*University of Novi Sad, Technical Faculty „Mihajlo Pupin“, Zrenjanin, Republic of Serbia
** Primary school „Žarko Zrenjanin“, Novi Sad, Republic of Serbia
glusacdragana.zr@gmail.com
tasici@tfzr.uns.ac.rs
biljana.radulovic66@gmail.com
musickizarko@gmail.com

Abstract - The aim of this study was to evaluate the efficacy of digital i.e. e-Log Book, to determine the opinions of teachers as direct beneficiaries of Log Book about its efficacy, and based on teachers' opinions gain insight into what can be improved in the e- Log Book. The expected result is not to eject the traditional Log Book from the use, but that information in it can be used more effectively, accessible and optimally well archived.

I. INTRODUCTION

The introduction of new services and software in education is part of the overall IT transformation in society, so that global computerization characteristics are applied to this social issue as well. In the basis for the computerization of education some relevant factors are of social, technical and technological nature [1]. On the basis of the explosive development of microelectronics, information technology telematic networks, space for the introduction of information technology is opened. information systems that increase the power in the disposal and handling of information are hurriedly introduced and developed [2]. In developed educational systems records of students and their achievements are largely digitized, and there are studies that have dealt with the qualitative analysis of their application [3] .

In Serbia, there are a number of software solutions for electronic evaluation of pupil achievement. The most famous is just a project called "Electronic Log Book". The project "Electronic Log Book" is applied in 500 schools (which amounts to about 25% of all elementary and secondary schools in the country), the list of schools-users is a trade secret of the author of the project, but it can be checked indirectly through the server: www.e-ocene.org.

When considering the use of software for the electronic evaluation of the success of pupils outside the Republic of Serbia, there are very similar software solutions. On the Croatian territory there are several software packages under the names: "Electronic directory of schools", "School Log Book" and "SMS System". Total applicability is reduced to about 30 schools across the Croatia [4].

.

II. RESEARCH METHODOLOGY

The aim and tasks of research

The aim of the research was to identify attitudes of teachers regarding the effectiveness of electronic log books in elementary school. In particular, perceive the attitudes of teachers in improving administrative tasks as well as enhanced communication with parents. It will be determine whether there is a statistically significant effect- the length of service of teachers on the attitude about e-Log Book. The ultimate goal of this research is the prediction of the success of the implementation of the electronic Log Book on a broader level and proposed measures for its widely popularizing, even legal regulation of its application.

Research Tasks:

1. Examine whether the members of the teaching staff are aware of the possibilities offered by the e-Log Book.
2. Determine if the e-Log Book facilitates the organization of educational work.
3. Determine what are the advantages of e-Log Book in the organization and management of the school.
4. Determine imperfections of the e-Log Book in the organization and management of schools.
5. Explore whether this software facilitates communication with parents.

Research hypothesis

The main hypothesis of this research project is as follows: Elementary school teachers have a positive attitude about the effectiveness of the implementation of e-Log Book.

Auxiliary hypothesis is: Among teachers with different work experience there is a different degree of acceptance of digital i.e. e-Log Book.

Research variables:

Research variables are the following: the effectiveness of e-Log Book, teacher attitudes, years of service.

Research methods and techniques

Research methods are:

•Theoretical methods,
•Descriptive statistics,
•Determine the frequency, percentage, measures of central tendencies and variability measures.

Statistics reasoning - correlation analysis - to determine the correlation coefficient of the independent and dependent variables.

The research technique is questionnaire.

Scale of assessments that is used is the Likert scale. The survey used the Mann-Whitney test and Kruskal-Wallis test.

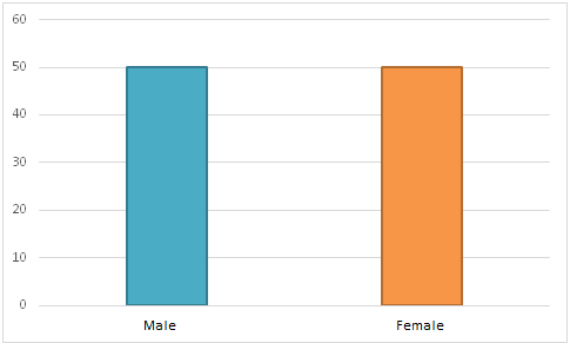
Indicators are written testimonies of teachers on the basis of which will be analyzed for inferring the efficiency of e-Log Book.

Population, location and sample surveys

The experiment was conducted in two elementary schools: Elementary School "Svetozar Markovic Toza" and elementary School "Zarko Zrenjanin" in Novi Sad, within the 2013/2014 school year. The research sample included a total of 100 participants (teachers).

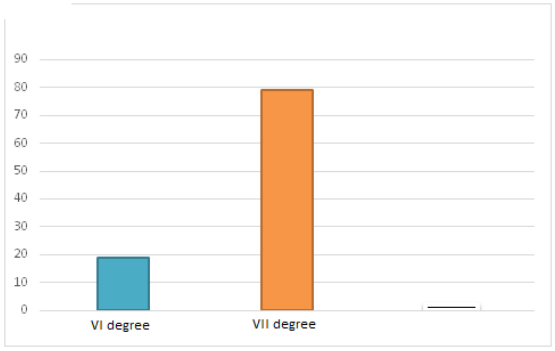
III. RESEARCH RESULTS

The study included 100 patients, with both sexes represented in equal numbers of cases (50 men and 50 women). The results are shown in Graph 1.



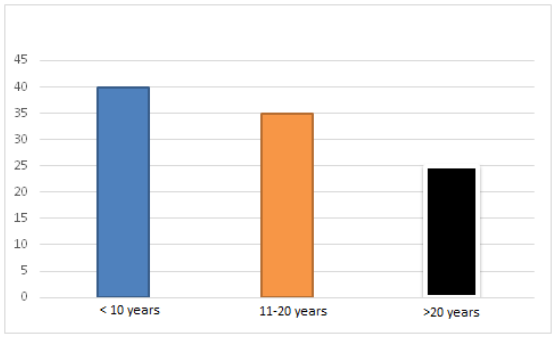
GRAPH 1. THE GENDER STRUCTURE OF THE SAMPLE

In most cases (79%) respondents have seventh or higher level of qualifications, while in a minority of cases (29%) of respondents have higher level of qualifications. In 2% of the respondents did not provide information on the level of educational attainment. The results are shown in Graph no.2.



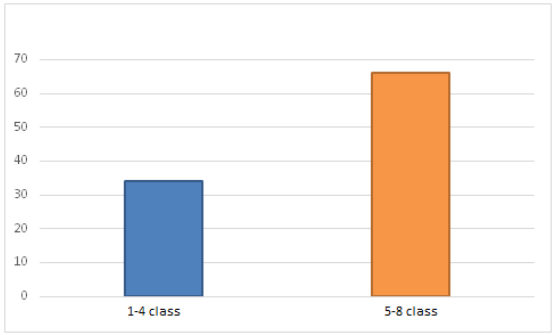
GRAPH 2. STRUCTURE OF THE EDUCATION LEVEL OF RESPONDENTS

In most cases (40%) respondents have up to 10 years of service, then 11 to 20 years of service (35%) and the smallest number of cases over 20 years of service (25%). The results are shown in third Graph.



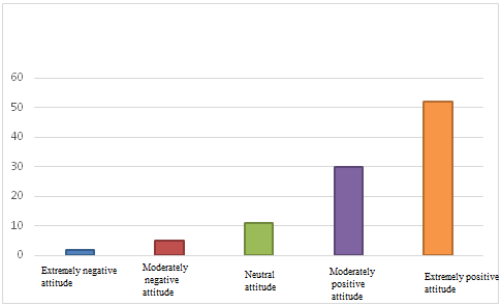
GRAPH 3. STRUCTURE OF RESPONDENTS IN RELATION TO THE YEAR OF SERVICE

In most cases (66%), teachers are from the field of higher grades, while in a minority of cases (36%) of teachers in the area of the lower grades. The results are shown in Graph no.4 in diameter.



GRAPH 4. STRUCTURE OF RESPONDENTS IN RELATION TO THE CURRENT TEACHING AREA

As the assessment of e-Log Book conducted through five-point Likert scale, it can be concluded that 82% of respondents expressed a positive attitude (moderate - 30% of respondents; 52% of respondents strongly) to the efficiency of e-Log Book.

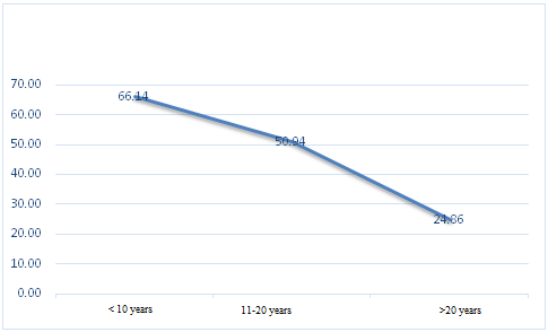


GRAPH 5. STRUCTURE OF THE RESPONSES ON THE EFFECTIVENESS OF THE IMPLEMENTATION OF E-LOG BOOK

In order to test the statistical significance of the observed results, we applied the x^2 goodness of fit test. X^2 test results indicate that the proportion of the distribution of responses obtained in this study is statistically significantly different in relation to the distribution of responses to the same proportions (x^2 (df = 4) = 87.7, $p < 0.001$).

In order to identify differences in the level of all the groups, we applied nonparametric substitute for direct analysis of variance - Kruskal-Wallis test. Results indicate statistically significant differences among all groups (KW (df = 2) = 34.3, $p < 0.001$).

The average rank groups are shown in Graph no.6.



GRAPH 6. DEGREE OF ACCEPTANCE OF E-LOG BOOK IN RELATION TO YEARS OF SERVICE

In order to determine the differences among the groups defined, we applied nonparametric replacement for the T-test for independent samples - Mann-Whitney test. The results of this test show that all three groups are statistically significantly different ($p < 0.01$), which provides teachers with the least seniority show the highest level of acceptance of e-Log Book, and teachers with the most seniority accept the e-book at a minimum. The test results are shown in first Table.

TABLE 1. COMPARISON OF THE GROUPS ACCORDING TO LENGTH OF SERVICE IN THE POSITION OF ACCEPTING E-LOG BOOK

Group*	Ave rank	Group	Average rank	Разлика суме рангова	Mann-Whitney тест	p
I group	43.86	II група	31.30	12.56	465.50	.007
I group	42.78	III група	17.36	25.41	109.00	.000
II group	37.64	III група	20.50	17.14	187.50	.000

** Group I = up to 10 years of service; Group II = from 11 to 20 years of service; Group III = over 20 years of service.*

Appendix 1: Descriptive indicators, measures of central tendency and variability measures for variables: the degree of acceptance of e-Log Book and the degree of efficiency of e-Log Book.

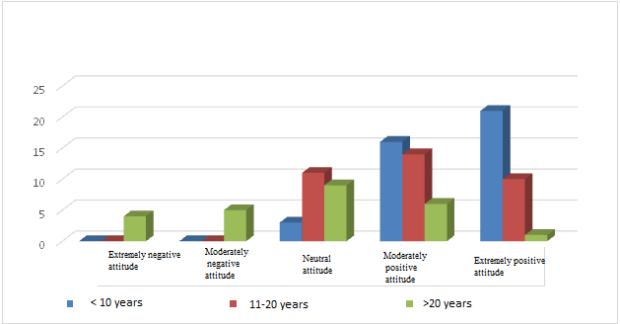
TABLE 2. DESCRIPTIVE INDICATORS, MEASURES OF CENTRAL TENDENCY AND VARIABILITY MEASURES

	Number of respondents	Range of values	Minimum value	Maximum value	Arithmetic middle	Standard deviation	Variance
Efficiency of E-Log Book	100.00	4.00	1.00	5.00	4.25	0.98	.96
Acceptance of E-Log Book	100.00	4.00	1.00	5.00	3.87	1.05	.10

Appendix 2: Frequency and percentage of the degree of acceptance of e-Log Book

TABLE 3. FREQUENCY AND PERCENTAGE OF THE DEGREE OF ACCEPTANCE OF E-LOG BOOK OF TEACHERS WITH DIFFERENT WORK EXPERIENCE

	up to 10 years of service		11 to 20 years of service		Over 20 years of service	
	Frequence	Percent	Frequence	Percent	Frequence	Percent
Extremely negative attitude	0	0.0	0	0.0	4	16.0
Moderately negative attitude	0	0.0	0	0.0	5	20.0
Neutral attitude	3	7.5	11	31.4	9	36.0
Moderately positive attitude	16	40.0	14	40.0	6	24.0
Extremely positive attitude	21	52.5	10	28.6	1	4.0



GRAPH 7. DEGREE OF ACCEPTANCE OF E-LOG BOOK IN RELATION TO YEARS OF SERVICE

Appendix 3: Correlation years of service and questions 3 to 7 (in accordance with the suggestion mentors).

In order to determine the degree of correlation between length of service and monitoring, measurement and evaluation of educational work using the e-Log Book, Spearman's rank correlation coefficient Ro is applied. Results indicate all correlations moderately low (from -0251 to -0319), negative and statistically significant (p <.05). Detailed view of the results is given in Table no.4.

TABLE 4. CORRELATION BETWEEN YEARS OF SERVICE AND VARIOUS ASPECTS OF EDUCATIONAL WORK

	Years of service		
	Spirman Ro coefficient	Level of significance	Sample size
Monitoring of educational - teaching work	-.261	.009	99
Measuring educational - teaching work	-.251	.012	100
Evaluation of educational - teaching work	-.319	.001	99

In order to determine the degree of correlation between length of service and various moguls within the e-Log Book, the Spearman rank correlation coefficient Ro is applied. Results indicate low to moderate correlations (from -0239 to -0.41), negative and statistically significant (p <.05) in the case of Statistics and Teachers modules, while in the case of the remaining modules, correlations are not statistic significant. Detailed view of the results is given in Table 5.

TABLE 5. CORRELATION BETWEEN YEARS OF SERVICE AND DIFFERENT MODULES WITHIN THE E-LOG BOOK

	Years of service		
	Spirman Ro coefficient	Level of significance	Sample size
Administration module	-.083	.415	99
Teaching module	-.410	.000	100
Statistics module	-.239	.017	100
Internet module	.109	.280	100
SMS module	.105	.299	100
Module telephone and answering machine	.097	.336	100

In order to determine the degree of correlation between length of service and improve the organization and management of schools using the e-Log Book, the Spearman rank correlation coefficient Ro is applied. Results indicate all moderate correlation (-0.36 of up to -0365), negative and statistically significant (p <.01). Detailed view of the results is given in Table 6.

TABLE 6. CORRELATION BETWEEN YEARS OF SERVICE AND IMPROVE THE ORGANIZATION AND MANAGEMENT OF SCHOOLS USING THE E-LOG BOOK

	Years of service		
	Spirman Ro coefficient	Level of significance	Sample size
Improvement of the organization	-.365	.000	100
School management	-.360	.000	100

In order to determine the degree of correlation between length of service and various pro-social activities within the school environment Spearman rank correlation coefficient Ro is applied. Results indicate low to moderate correlations (from -0.28 to -0393), negative and statistically significant (p <.01) in the case of creating a positive social climate and improving teaching practice, while in the case of the remaining modules correlations were not statistically significant. Detailed view of the results is given in Table 7.

TABLE 7 CORRELATION BETWEEN YEARS OF SERVICE AND PRO-SOCIAL ACTIVITIES IN THE SCHOOL ENVIRONMENT

	Years of service		
	Spirman Ro coefficient	Level of significance	Sample size
Creating a positive social climate	-.280	.005	100
mprovement of teaching practice	-.393	.000	100
Establishing direct communication with parents	-.186	.064	100
School marketing	-.135	.181	100

In order to determine the degree of correlation between length of service and various interpersonal relationships using the e-Log Book, applied the Spearman rank correlation coefficient Ro. Results indicate low to moderate correlations (from -0235 to -0410), negative and statistically significant (p <.01) in the case of establishing relationships among teachers, teachers and students as well as teachers and parents, while in the case of teacher - Director the correlation was not statistically significant. Detailed view of the results is given in Table 8.

TABLE 8. CORRELATION BETWEEN YEARS OF SERVICE AND VARIOUS INTERPERSONAL RELATIONS, USING THE E-LOG BOOK

	Years of service		
	Spirman Ro coefficient	Level of significance	Sample size
Teacher - teacher	-.410	.000	99
Teacher - pupil	-.345	.000	99
Teacher - parents	-.235	.018	100
Teaher - principle	-.078	.445	99

IV.CONCLUSION

Education must inevitably go up with the times and the needs of social development. So far in about 500 schools in Serbia e-Log Book is applied. There is a need for constant education, improvement of the system and continuity of data entry. This ensures that the mandatory school statistics do not allow errors and far less work. Compatibility with the EIS, and programs for printing certificates leads to the perfect regularity of school documentation and release large amounts of time. In this way, cognition of computers for many, arouse an idea to use computers in teaching.

In this paper there was an investigation of teachers' attitudes about the possibilities of application didactic and IT innovation in the organization and management of schools, where wearing Software Research was e-Log Book. The first application of this software began in 2004, just in elementary school ,,Zarko Zrenjanin". E-Log Book is used in this school for 10 years, but research has shown that there is still room for further improvement, as well as the training of teachers in the field of ICT. Based on the results of performed research, what was stated in the research and the objectives and tasks, can be concluded that main and auxiliary hypotheses are proven. Elementary school teachers have a positive attitude about the effectiveness of the implementation of e-Log Book.

REFERENCES

[1] Salmon, C. (1984). The principles and Pracitce of Cultural Research, London.

[2] Meirink, J. A., Meijer, P. C., & Verloop, N. (2007). A closer look at teachers' individual learning in collaborative settings. Teachers and Teaching: theory and practice, 13(2), 145-164.

[3]Beijaard, D., Korthagen, F., & Verloop, N. (2007). Understanding how teachers learn as a prerequisite for promoting teacher learning.

[4]Radulovic, B., Maric, B. (2012). Comparative analysis of information systems in education in the Western Balkans, Engineering and Informatics in Education, 4th International Conference, Technical Faculty in Cacak

Methodology for online study program development – case study ePITUP

N. Kadoić., N. Begičević Ređep
Faculty of organization and informatics Varaždin, Croatia
{nkadoic, nbegicev}@foi.hr

Abstract - This paper deals with the methodology for online study program development which was constructed to meet the needs of students, teachers and the institution. The proposed methodology consists of four phases: 1.) Problem definition, 2.) Strategic decision making, 3.) Implementation, 4.) Evaluation. In the first phase, situation analysis is performed and central problem is defined. Strategic decision making phase includes making a decision on several issues: choice of learning strategies, decision on delivery formats, choice of appropriate design, etc. In this phase, needs, target and task analysis must be performed. Also, instructional, delivery and evaluation strategy must be created. It is also recommended to perform TOWS and cost benefit analysis. In the implementation phase an online study program is prepared, based on the strategies created in the previous phase, and delivered to students. The fourth phase involves evaluation of student reactions, achievement of objectives and impact of the project on the institution. In addition, we present the case study of developing ePITUP online study at Faculty of Organization and Informatics. The purpose of the paper is to provide a good insight into methodology for online study program development. It will provide an effective guide for developing online courses and establishing an online study program.

I. INTRODUCTION

E-learning encompasses a range of technology for teaching and learning purposes. According to T. Bates' definition [1], e-learning refers to all computer and Internet-based activities that support teaching and learning, both on-campus and at distance; it also includes administrative and academic uses of information and communication technologies that support teaching and learning. E-learning supports skills needed in a knowledge–based society, such as collecting, analyzing and applying information appropriately. E-learning comprises different teaching methods, for example information management, creative thinking, critical thinking, problem solving and collaborative learning [16].

There are different forms of e-learning. E-learning can be used with classroom aids, in computer labs or laptop programs, in mixed mode or in a fully online form [14, 15). In the form of classroom aids, students may be given supplementary online work after class. With computer labs or laptop programs, the use of computing supports traditional Face-to-Face learning (f2f) but it is still within a classroom setting. In a mixed – blended learning mode, classroom time is combined with online teaching and learning. Blended learning means that some of the classes are done Face–to-Face, and other are done online, or include lab and practical work in classroom and online lectures. Fully online learning is a form of learning and teaching where students never have classroom time, only online lectures.

The strategic planning of e-learning includes planning and decision making about implementing E-learning on different levels: state level, university level, faculty level, study program level and individual (student and teacher) level. There are different options for implementing e-learning in teaching and learning process. E-learning can be used as a means of support to the already established systems of education i.e. blended – learning mode. On the other side it can be implemented as an independent form of teaching; in other words as a separate teaching program. It can also be partially introduced (for a single subject, group of subjects, or a study program) [17].

Based on the document "Criteria for the Evaluation of Online Study Programs" prepared by the National Council for Higher Education on 13 February 2013, online study program is a program where minimum 50 percent of courses is done online. A course is done online when minimum 50 percent of teaching hours is done online. Therefore, we can conclude that mixed learning mode can be implemented when preparing and running higher education online study programs in Croatia.

The topic of this paper will be strategic planning and decision making on implementing e-learning at study program level. This paper deals with the methodology for online study program development consisting of four phases: 1) Problem definition, 2) Strategic decision making, 3) Implementation, 4) Evaluation. A case study on developing a vocational online study program, entitled Information Technology in Business Applications at the Faculty of Organization and Informatics (ePITUP), University of Zagreb, will be presented.

II. STATE OF THE ART

Introduction and implementation of the instructional systems design models and methodology, including guidelines for building effective online training or online courses, has received considerable attention in the literature. However, it has rarely been applied in developing online study programs.

Bates and Sangra [1] recommend the best plan for online education. This plan is at program level since a critical area of developing a vision for the use of

technology is at the academic program planning stage. The program planning process must be integrated with the allocation of resources and financial plans.

In his book, Bates [2] presents a case study of how a Polytechnic developed an e-learning strategic plan. It describes the institution's rationale for moving more strongly into e-learning, and the processes followed by the institution which helped it develop a plan and ensure its acceptance throughout the institutional community. It also refers to the factors that facilitated the process. It indicates that paying attention to objectives, core values and principles, faculty development and training, are critical for a successful transition from mainly Face-to-Face teaching to e-learning.

Methodology for strategic planning and decision making on e-learning implementation has been developed and used in the higher education (HE) context in [11]. The authors deal with strategic planning of e-learning implementation at different levels in HEIs.

As we already mentioned, there are many instructional system design models. The most popular is the ADDIE model. The ADDIE model is a framework that lists generic processes used by instructional designers and training developers. It is a guideline for building effective training (courses) and performance support tools in five phases: Analysis, Design, Development, Implementation and Evaluation (ADDIE) [4]. Florida State University initially developed the ADDIE model to explain the processes involved in the formulation of an instructional systems development program for military interservice training. This training adequately trains individuals to do a particular job, and it can also be applied to any interservice curriculum development activity [5]. The idea of the model was to complete each phase before moving on to the next one. Over the years, practitioners revised the steps, and eventually the model became more dynamic and interactive.

Rooij and Wiliams [6] claim that in the digital age instructional designers must possess both a sound instructional design knowledge base and solid project management skills which would enable them to complete courseware projects on time, on budget and in conformance to expectations. Project management skills include the ability to apply repeatable processes, along with interpersonal skills such as communication and leadership.

A broader framework for developing blended learning programs is provided in Harvey's paper (Harvey, 2003). In the paper, Harvey presents Badrul Khan's blended e-learning framework which serves as a guideline to plan, develop, deliver, manage, and evaluate blended learning programs [8]. It is an Octagonal Framework containing eight dimensions: institutional, pedagogical, technological, interface design, evaluation, management, resource support, and ethical dimension. Each dimension in the framework represents a category of issues that need to be addressed to create a meaningful learning environment. Khan (Khan, 2001) claims that the implementation of this framework can help explore strategies for effective learning and performance, as well as ensure effective delivery of learning. Consequently, this will bring a high return on investment.

A review of the most important activities that must be achieved to reach a successful adoption of a blended learning approach is given in [9]. Randy and Kanuka suggest that a successful adoption of a blended learning approach, which would enhance the effectiveness and efficiency of teaching and learning, would require the following: the creation of a clear institutional direction and policy; potential framework, increased awareness and commitment; establishment of a single point of support, quality assurance and project management; the creation of an innovation fund which would provide the financial support and incentives to faculty and departments for initiating blended learning course transformations; investment in establishing a reliable and accessible technological infrastructure; a strategic selection of prototype projects that prove to be exceptionally successful exemplars of effective learning; development of a formal instructional design support available through a blended format; systematic evaluation of satisfaction and success with the teaching, learning, technology and administration of a new course; creating a task group to address issues, challenges and opportunities, as well as communicate and recommend new directions to the University community.

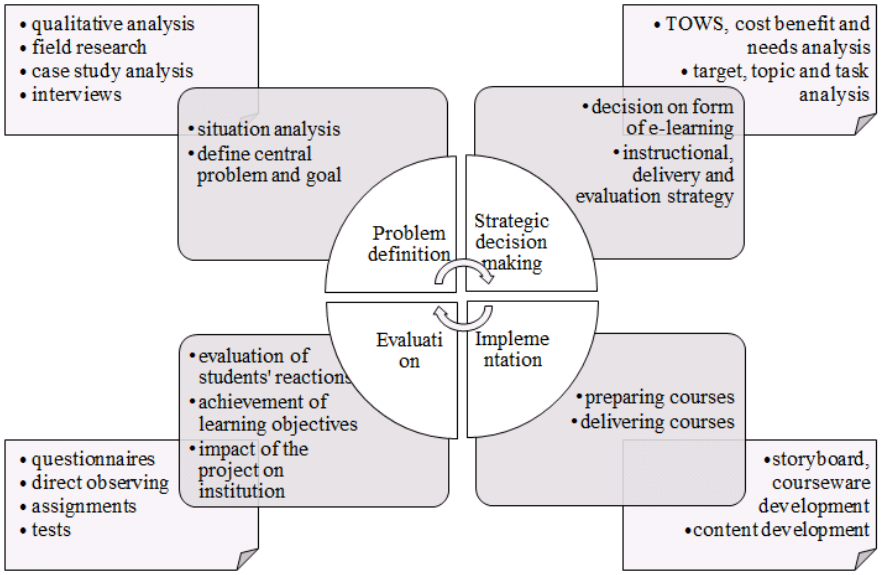
III. METHODOLOGY FOR ONLINE STUDY PROGRAM DEVELOPMENT

Methodology for online study program development presented in this paper consists of four phases: 1.) Problem definition, 2.) Strategic decision making, 3.) Implementation, and 4.) Evaluation.

During the first phase, the situation analysis must be performed and central problem defined. The situation analysis includes a review of key facts and major documents (curriculum, financial report, evaluation report, etc.) and collection of feedback from the stakeholders regarding the problem. The tools used in this phase are qualitative analysis, field research, case study analysis, interviews, etc. Based on the results of situation analysis, the central problem and main goals must be defined.

In the phase of strategic decision making, needs, target, task and topic analysis must be performed; instructional, delivery and evaluation strategy must be created. It is important to conduct needs analysis if there are any indicators for starting a development effort. The results of needs analysis are crucial for making decisions on the e-learning study program form. E-learning study program form can be an ICT supported study program, online program (50% Face-to-Face, 50% online) and fully online study program (100% online). Performance of TOWS and cost benefit analysis is also recommended.

TOWS analysis is a variant of the classic business tool called SWOT analysis. It is an acronym for Threats, Opportunities, Weaknesses and Strengths. It gives answers to the following questions: How to make the most of the strengths?, How to avoid weaknesses?, How to capitalize on opportunities?, and How to manage threats?. TOWS analysis can help decision makers get a better understanding of the strategic choices they face.



Picture 1. Methodology for online study development

Cost–benefit analysis is a systematic process for calculating and comparing benefits and costs of a project or a decision. It has two purposes: to determine if it is a sound investment/decision (justification/feasibility), and to provide a basis for comparing projects. It involves comparing the total expected cost of each option against the total expected benefit, to see whether the benefits outweigh the costs, and by how much [12].

Target analysis is also an important part of this phase because the design and delivery of e-learning materials and preparing online courses must be influenced by characteristics of the students as well as other stakeholders – teachers, administrative staff and management at the institution. Task analysis identifies the tasks that students should learn, and the knowledge and skills they need to develop or reinforce. Topic analysis is carried out to identify and classify the content of study curriculum and course content [3]. The results of topic analysis are identification of course content and classification of content elements. The results of needs, target, task and topic analysis are valuable inputs in the process of decision making on instructional, delivery and evaluation strategy (options).

Decision making on instructional methods includes evaluation of three alternatives: expositive methods, application methods, and collaborative methods concerning the task results and topic analysis. Design of online courses includes the combination of expositive, applications and collaborative methods. Expositive methods emphasize the "absorption" of new information, and include presentations, case studies, worked examples, demonstrations. Application methods emphasize the active processes students use to perform procedural and principle based tasks, and to build new knowledge. Application methods include demonstration-practise method, job aids, case-based or scenario-based exercises, role play, simulations and serious games, guided research, project work. Collaborative methods emphasize a social dimension of learning and engage students sharing knowledge and performing tasks in a collaborative way. They include online guided discussions, collaborative work and peer tutoring [3].

Each method can be delivered in different formats. For example, a presentation can be delivered as a Power Point file or as a video presentation. In preparing a delivery strategy, it is crucial to analyze the following elements: student-related factors (status, time, equipment, delivery channels, technical expertise etc.), technology aspects (infrastructure, connectivity, technical expertise, etc.) and organizational requirements (institution strategy, time, budget, people, etc.).

The aims of evaluation can be different, to check the quality of the courses (program), measure effectiveness of teaching and learning, evaluate present and past state, as well as to compare the two states to see if the aims are reached. Assessments tests are a tool that is often used for evaluation. The tests must be aligned with the learning objectives at study program level and course level.

In the implementation phase an online study program is prepared based on the strategies chosen in the previous phase, and it is delivered to students. The outcome of this phase is an online study program. In this phase, curriculum structure is defined, learning outcomes for each topic, course and study program are prepared; e-learning content is produced with the implementation of instructional, evaluation and delivery strategies. Development of elearning material contains three main modules: content development, storyboard development, and courseware development [3]. Content development is the creation of e-learning materials that integrates the knowledge needed for reaching the learning objectives. Storyboard development is the integration of instructional (pedagogical) methods and media elements. Courseware development is the development of interactive content elements, as well as the integration of these elements into a learning platform accessible by students.

The fourth phase involves the evaluation of student reactions, achievement of learning objectives and impact of the project on the institution. Evaluation of student reactions can be conducted using questionnaires which are usually submitted to students at the end of the course. Evaluation of learning objectives achievement can be done through direct observation, assignments and tests. It is necessary to check if student knowledge and skills are increased, and if the defined learning objectives are

reached. In the evaluation of project impact on the institution, the final results, such as improved quality, decreased costs, increased student interest for study program, must be identified [13].

The proposed methodology for online study program development can be used as an effective guide for developing online courses and establishing an online study program. The framework that was used in developing this methodology was ADDIE instructional design model. The methodology presented in this paper is comprehensive. Therefore, according to the complexity of learning project, project objectives, available resources and expertise, some of the phases and methods described in this methodology can be skipped, simplified or expanded.

IV. CASE STUDY

The methodology presented in this paper was implemented in the development of fully online study program.

The vocational study programme "Information Technology in Business Application" (PITUP) was developed, and is today performed by the Faculty of Organization and Informatics in Varaždin. In the performance of the study program, the Faculty has established cooperation with other institutions of higher education and organizations from private and public sector in the field of applying modern ICT in business organization. The study program "Information Technology in Business Application" is performed during three (3) years and six (6) semesters. Each semester represents one education cycle and lasts 15 weeks.

Till 2008 study "Information Technology in Business Application" was performed Face-2-Face using ICT technology. The Learning Management System called Moodle was used in teaching and learning. However, it was mainly used for uploading teaching materials, providing information about important dates and course deadlines, and for uploading student assignments. Only a small number of teachers used some advanced features of LMS Moodle. This study program is performed at four different locations: Varaždin, Križevci, Sisak and Zabok. In Varaždin, around 50% of students are regular students, and 50% of students are part-time students. In Križevci, Sisak and Zabok, all students are part-time students with full time jobs. The main reason for considering the development of an online study was part-time students who sometimes cannot attend classes due to work obligations. The second reason was time and cost of traveling to cities 2 hours or more away from Varaždin. Still, the decision was made on a more detailed needs analysis, TOWS and cost-benefit analysis. Those methods were used in our methodology for online study program development.

During the first phase (problem definition), situation analysis was conducted and central problem was defined. Based on the conducted situation analysis the following results were gained: the study program PITUP is mostly enrolled by part-time students that are often not able to participate in Face-to-Face activities due to their job; student success is low (low percentage of students who pass exams in the first or second examination period; low grades); high travel costs (sometimes it happens that a teacher travels to Križevci, Zabok or Sisak, and none of students (or only a few of them) attended the lectures).

The situation analysis also showed that there were teachers at Faculty level who have knowledge and experience in e-learning implementation, and were willing to develop an online study program. Also, there is a Center for supporting e-learning activities at the Faculty; so the administrative support is also ensured. In this phase we defined a central strategic goal - the development of an online study program called e-PITUP.

During the second phase, we made a decision on delivery formats that need to be developed in the phase of online course development. We chose delivery formats based on the ADDIE instructional design which was used to be the framework for developing online courses. Some of the delivery formats we defined as mandatory at online course level. These are as follows:

- defined learning outcomes at topic level;
- video lectures for minimum 50% of previous F-2-F lectures;
- video materials for seminar or laboratory activities (mainly tutorials showing how to use a certain tool, application, how to solve a certain problem, etc.);
- different student assignments related to each topic, with aim to include students in the continuous learning process;
- evaluation and self-evaluation tests for each topic.

In developing online courses we decided to use different learning strategies: strategies of repetition, elaboration, organization, understanding and application.

In the second phase, we also performed needs and task analysis. The results show which type of courses can be developed as a role model for further online course development. We choose 5 courses based on the recognized needs of students and teachers and their specific characteristics. We also prepared instructional and delivery model for each of these courses. Also, we planned the resources needed for preparing these 5 pilot online courses (equipment, working hours, etc.). Additionally, time and activity scheduling of implementation process was done.

In the implementation phase, 5 pilot courses were prepared and delivered to the students in Križevci: Quantitative Management, Mathematics, E-Business, Office and Office Management, and Programming.

These courses were performed 50% online and 50% Face-to-Face. In order to achieve positive outcomes, students continuously had assignments, projects, tests, which were previously carefully planned and prepared by teachers. Teachers became instructors who directed students in an active problem solving process. Online

communication became an important part of the teaching process and it was introduced through forums, webinars, chats and direct contact via E-mail or Skype.

In the last phase, evaluation of student reactions, achievements of objectives and impact on the institution was done. In pilot courses, student results were better than the results of only F-2-F teaching: percentage of students who passed exams in the first and second examination period increased. Additionally, average grade point increased. HELAM instrument, upgraded with several open-questions [11], was used for student evaluation. Students evaluated the pilot courses, being a new way of learning, as more convenient and appropriate. In general, the pilot courses got positive feedback. Students were able to compare online learning with Face-to-Face learning. Therefore, the conclusion is that the objectives set by the developed pilot online courses in PITUP study program were achieved. At institutional level, the strategic decision is to develop new online courses at PITUP study and, in a long-term, transform the study into fully online study.

A. Pilot E-course „Quantitative Management"

Quantitative Management is a course at third year of study Information Technology in Business Applications. It is held through 15 hours of lectures and 15 hours of seminars.

There are 5 subjects in the course: analysis of investments, sensitivity analysis, decision making and ranking, decision making in terms of uncertainty and risk, and forecasting. Using the *old* way of learning, in Zabok, Križevci and Sisak, there were 6 weeks with 5 hours of classes per day, one day per week. Each week different subject was covered. The last week a repetition of all subjects, with instructions how to make the project, was done.

In online learning, the goal was to have only three weeks of classes, and the remainder of three weeks switch to fully online activities:

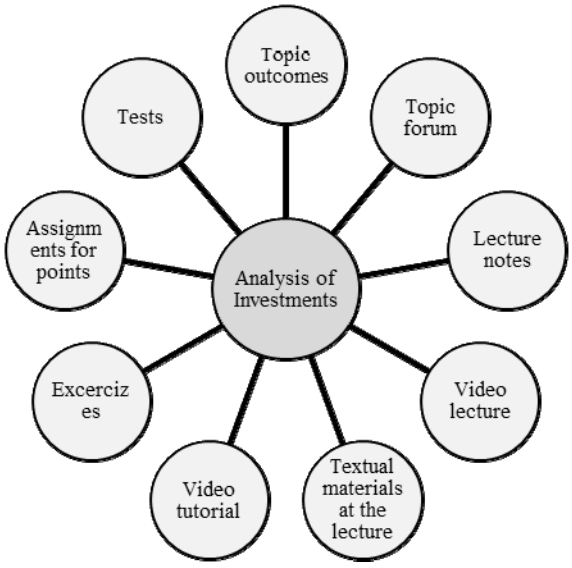
- Week 1 – Face-to-Face learning: introduction into the course, investments analysis;
- Week 2 – online learning: sensitivity analysis;
- Week 3 – Face-to-Face: decision making and ranking;
- Week 4 – online learning: analysis in terms of uncertainty and risks;
- Week 5 – online learning: forecasting;
- Week 6 – repetition, presentation of the project (to be done before final exam)

Pilot course was delivered in Križevci.

Each of topic that was done only online consisted of several elements (according to the delivery formats defined in the methodology):

- Outcomes of each topic;
- Forum for specific topic where students can ask questions and seek for more detailed explanations of how to solve certain assignments, or where teachers can initiate discussions;
- Lecture notes for each topic (cca 20 page document with theoretical base of the topic, examples and exercises);
- Video lecture (simulation of real Face-to-Face lecture; instructor talks about subject matter of topic using a PowerPoint presentation; student can download the presentation separately, print it, and make additional notes during the audio lecture);
- Textual materials at the lecture (if student have strong reading/writing preference for learning according to VARK learning style classification, instead of watching visual video lecture, (s)he can learn about the topic from given textual materials);
- Video lecture/tutorial of how to solve a concrete assignment, use specific software etc., depending on topic;
- Practice exercises (with or without key)
- Assignments for points (grades) which have to be done during the week and uploaded to Moodle;
- Self-evaluation tests and tests for points (grades).

During six weeks students can gather points (grades) by submitting correct solutions on assignments that were previously prepared for each topic, and by solving tests on Moodle. Collected points are added to the project exercise points and to the result on final exam.



Picture 2. Quantitative management: Quantitative management: delivery formats defined by the methodology

Here are some of the results of the new way of learning:

- 80% of students passed the exam at the first examination period, and 100% of students passed the exam, including the first and second examination period;
- Compared to the student results in Križevci, when old way of teaching was implemented,

average grade was increased in online learning for 0.6 of the grade
- 90% of students find online course more attractive than other Face-to-Face courses;
- 90% of students find online course more appropriate for their personal needs considering organization of classes (they encourage the combination of Face-to-Face and online classes because there is flexibility of taking lectures and fulfilling course obligations; also, they can contact the teacher in person).

In general, students are satisfied with the new way of teaching, and find it more convenient for their needs.

V. CONCLUSION

This paper dealt with the development of a methodology for online study program development. The main motive for this topic arose from the current situation at Face-to-Face study called PITUP at Faculty of Organization and Informatics. However, during the performance of the process several issues were identified, mainly related to student needs and impossibility to be present at Face-to-Face classes.

A logical solution for that situation was to create an online study. Firstly, we needed a methodology of how to create an online study. Methodology was created using 4-phase decision making methodology and ADDIE instructional design model. 4-phase decision making methodology combines several important types of analyses (situation analysis, cost-benefit, TOWS, cost-benefit, needs analysis, target, topic, task analysis). Those analyses were primarily used for identifying the context of the problem, as well as for deciding how an online study program would be organized. The basis of online study program is online courses. ADDIE was used to determine how certain course has to be organized to match the study program concept. Finally, we came up with the methodology we applied in our case. Afterwards, we created a framework of e-PITUP pilot study. We chose, prepared and implemented 5 online courses.

The final results of pilot courses were positive: student success was better than before, costs were reduced. We applied the suggested methodology only partially, in future we should work on additional piloting and testing. Nonetheless, the first results are promising.

LITERATURE

[1] Bates, AW Tony, and Albert Sangra, Managing technology in higher education: Strategies for transforming teaching and learning, John Wiley & Sons, 2011.

[2] Bates, A. W., "Strategic Planning for E-learning in a Polytechnic", Making the transition to e-learning: Strategies and issues, 2007, pp. 47-65.

[3] E-learning methodologies, A guide for designing and developing e-learning courses, Food and Agriculture Organization of the United Nation, (FAO), Rome, Italy, 2011.

[4] Morrison, Gary R. Designing Effective Instruction, 6th Edition. John Wiley & Sons, 2010.

[5] Branson, R. K., Rayner, G. T., Cox, J. L., Furman, J. P., King, F. J., Hannum, W. H., Interservice procedures for instructional systems development. (5 vols.) (TRADOC Pam 350-30 NAVEDTRA 106A). Ft. Monroe, VA: U.S. Army Training and Doctrine Command, August 1975, (NTIS No. ADA 019 486 through ADA 019 490)

[6] Van Rooij, Shahron Williams, "Project management in instructional design: ADDIE is not enough." British Journal of Educational Technology 41.5, 2010, pp. 852-864.

[7] Singh, Harvey, "Building effective blended learning programs." EDUCATIONAL TECHNOLOGY-SADDLE BROOK THEN ENGLEWOOD CLIFFS NJ- 43.6, 2003, pp. 51-54.

[8] Khan, Badrul H., "A framework for web-based learning." Web-based training, 2001, pp. 75-98. http://www.academia.edu/259746/A_Framework_for_Web-Based_Learning

[9] Garrison, D. Randy, and Heather Kanuka., "Blended learning: Uncovering its transformative potential in higher education." The internet and higher education 7.2, 2004, pp. 95-105.

[10] Begicevic, N., Divjak, B., Hunjak, T., Development of AHP based model for decision making on e-learning implementation, JIOS, 2007

[11] Ozkan, S., Koseler, R., Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation, Journal "Computer and education", 2009

[12] Boardman, N. E. (2006). Cost-benefit Analysis: Concepts and Practice (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

[13] Kirkpatrick D.L., Kirkpatrick J.D. (2006). Evaluating Training Programs. The Four Levels. San Francisco: Berrett-Koehler Publishers.

[14] Bates, A. and Poole, G., Effective Teaching with Technology in Higher Education San Francisco: Jossey-Bass, 2003

[15] OECD, E-Learning in Tertiary Education: Where Do We Stand?, Paris: OECD, 2005.

[16] Divjak B., Begičević, N.: Imaginative Acquisition of Knowledge - Strategic Planning of E-Learning, ITI 2006 Proceedings of the 28th International Conference on Information Technology Interfaces, University Computing Centre SRCE, University of Zagreb, 2006., p. 47-53

[17] Bubaš G., Balaban I., Begičević N.: Course Evaluation in e-Learning by Comparative Analysis of Two or More e-Learning Courses, 18th International Conference on Information and Intelligent Systems IIS 2007, Varaždin, 2007., p. 65-72.

Effective knowledge exchange with modern didactic concepts

C. Ungermanns, W. Werth
Carinthia University of Applied Sciences (CUAS)
Europastraße 4, 9524 Villach, Austria
Phone: +43 4242 90500
c.ungermanns@fh-kaernten.at
w.werth@fh-kaernten.at

Abstract - Production sciences and manufacturing technologies are worldwide a central element for societies which take profit and welfare by their producing industries. Especially in the highly industrialized countries it has to be considered that these production plants can only be operated with highly qualified staff.

In this contribution we demonstrate a structure and scheme for knowledge transfer from the tertiary to the secondary educational level with examples in the field of Robotics and Flexible Automation.

Thanks to the initiative Leonardo Lifelong Learning – Transfer of Innovation of the EU it is shown that a successful cooperation of Austrian, Croatian, Slovenian and Turkish partners is realized in terms of education and engineering. The essential elements of the project will be presented in this paper.

I. INTRODUCTION

Modern production sciences and manufacturing technologies processes need both flexibility and a high level of automation [1]. Of similar importance is especially in highly industrialized countries that these production processes can only be operated by highly an excellent qualified staff. This is even more important, if these processes have to be optimized.

The practical relevance and the corresponding complexity is the main reason for choosing the topics "Flexible Automation" and "Robotics" as an example for developing a structure and concept for knowledge transfer to different educational levels as well as to SMEs and their staff.

A. *Associated Project*

The LLL-project of "Flexcell" makes a contribution to keep high standard production plants at the industrial location of Europe by offering a modular teaching concept for the educational demands to start operation of a flexible production cell with success. This modular teaching concept considers the levels of demands of the different target groups, e.g. academic staff, students, pupils and workers. Regional close partnership is necessary and supported by skills and technological equipment of cooperating countries and partners in the project [4].

The present paper is therefore based on a collection (summary) of experiences and knowledge of experts who all work in the field of Flexible Automation and robotics. These people work for example as consultants in industry, teachers in technical schools of higher education and lecturers in universities of applied sciences. It is important to mention that they are in close relationship to the needs of industrial applications.

B. *Educational concept*

The main scope of the considered concept is to establish training modules of VET teachers, trainers of practical lessons in schools as well as employees in companies. Furthermore in order to show the effectivity of the concept training sessions have been successfully undertaken. The basic direction of innovation transfer is from Austria and Slovenia to other countries – for example Turkey. Also a good cooperation to the "Industrial and craftsmen school Pula" was established.

Figure 1. Flexcell Project devices of different partners

It is well known that a profound knowledge leads to higher quality, reduced costs, better company reputation and competitiveness.

The developed scheme is a working guideline for finding the needs of industry in the field of education and how to establish a new education system that is based on those needs. Therefore one of the results of this project is for example the preparation of questionnaires, which identify the needs of industry, their expectations for the future, weaknesses, possible potentials for optimization and how they would like to cooperate with schools and universities. Beside the target group further results are adapted training modules for a specific topic, a special dictionary, knowledge transfer in the field of flexible automation, a better cooperation between industry and schools and evaluation sheets for continuous improvement.

Additionally solutions found in a previous project [9] are prepared to be transferred.

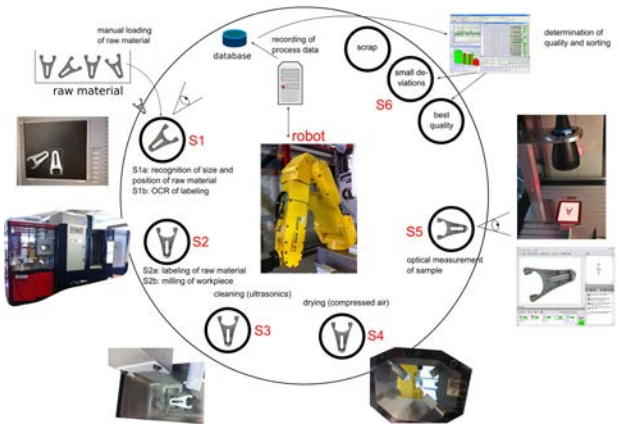


Figure 2. Process Flow FlexCell CUAS

In fig. 1 different flexible cells, which have been developed by project partners and used machines are presented in order to understand the application behind the concept. Basically this can also be seen as a contribution to the implementation of industry 4.0 concept.

II. TRANSFER CONCEPT

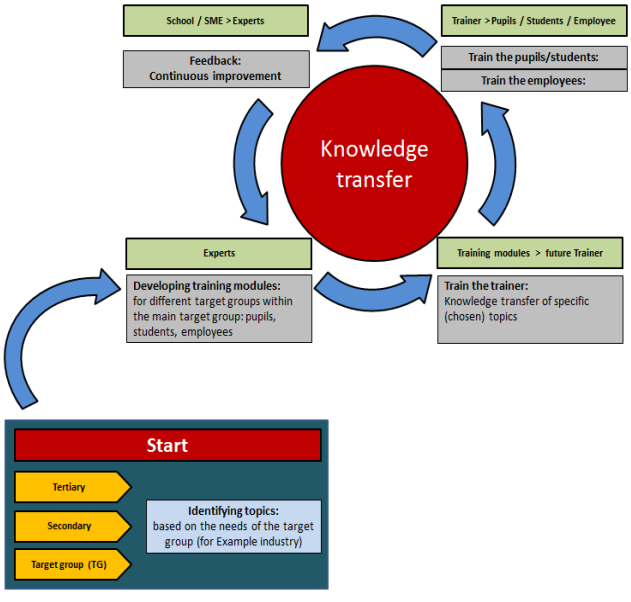


Figure 3. Process of knowledge transfer including continous improvement

The quality of this process is monitored and continuously improved by the Deming circle, which is a four-step method consisting of plan / do / check / act (fig. 3). The last point is essential to assure that the needs of the target groups are met. The combination of theoretical knowledge and practical exercises is also very important for the overall teaching process.

It is worth to mention that an overview of the corresponding school and education system of the partner country is helpful. Fig. 4 gives an example of the Turkish education system from pre-school up to doctorial studies.

Our didactical concept was matched with the current preeducation/knowledge of the students in Turkey. In order to ensure an implementation in the corresponding curriculum it is necessary to understand the overall education structure in Turkey. Finally it could be shown that a smooth integration of the course was successfully performed.

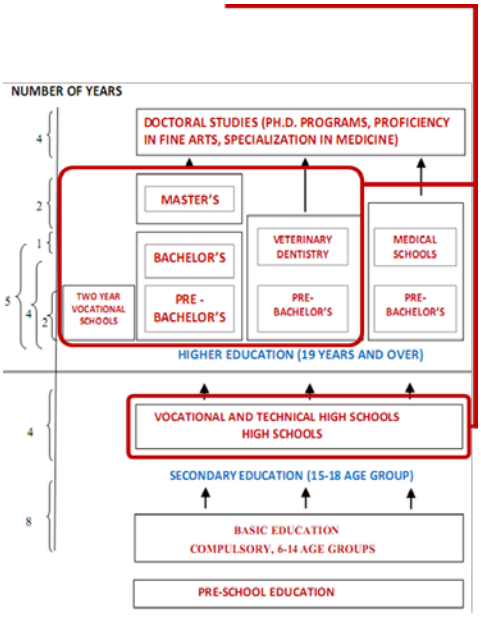


Figure 4. Overall education system in Turkey

III. DIDACTICAL APPROACH

The didactical concept of the project is based on two pillars.

A) Training of trainers in Austria and Slovenia with flexible production cells (see fig. 1).

B) Training of trainers and pupils in partner countries (Turkey and Croatia).

A. Training the trainers

The first part of the training uses the results of the previously reported SIAT project "Koop Flexible Automation" [8], which combines various disciplines of process automation, e.g. robotics, process control for CNC-based manufacturing, image processing and analysis on a high education level (see fig. 2). Due to differently designed cells with various education focus, specific

training modules had been developed earlier and could now be employed in practice (see fig. 1). All prepared courses can be seen in the following fig. 5. Additionally the estimated duration for each part is documented.

Nr.	Module	Hour	Days	Σ
1	CNC Manual Programing (milling + turning)	8	6	48
2	CNC Cam Programing (milling + turning)	8	5,5	44
3	Robotics basic	8	3	24
4	Robotics advanced	6	2	12
5	Quality management	4	1	4
6	Optical measurement systems	4	1	4
7	Process control	6	2	12
8	Machine control and embedded systems b.	6	3	18
9	Machine control and embedded systems a.	6	2	12
10	Safety basic	5	1	5
11	Safety advanced	3	1	3
12	Design I	6	3	18
13	Design II	6	3	18
14	Economics and organization	8	2	16

Figure 5. Training module setup

The modular design has proved to be very advantageous, because it was possible to adapt the training program to the trainees in the short term. Another benefit arises from the fact that the above mentioned modules were presented by the different project partners, which considers their specific skills. This enhances the quality of teaching and offers a wider range of industrial applications. Especially in the field of computer-aided design and material processing our project partners could provide complementary knowhow. Additionally a parallel training of the groups is feasible.

Further details can be found in [9].

B. Training in partner countries

A particular challenge was the training in our partner countries. For this purpose it was essential to perform a local inspection of the sites with special focus on the available infrastructure (see fig. 6).

Figure 6. Local inspection of partner sites

It was immediately recognized that the financial resources of our friends in Turkey and also in Croatia were very limited. Therefore a low-budget concept was absolutely necessary.

An appropriate solution was the implementation of the Open-source learning platform Moodle. It is a learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalized learning environments. Moodle is provided freely as Open Source software, under the GNU General Public License. Anyone can adapt, extend or modify Moodle for both commercial and non-commercial projects without any licensing fees and benefit from the cost-efficiencies, flexibility and other advantages of using Moodle. The system is web-based and so can be accessed from anywhere in the world. With a default mobile-compatible interface and cross-browser compatibility, content on the Moodle platform is easily accessible and consistent across different web browsers and devices [10].

Fig. 7 shows the overall structure of the Moodle course which is divided into different courses. These courses exist for Robotics Basic, Robotics Safety, Quality management, Optical measurements, Cam Programming Turning, Machine Control and Embedded Systems, CNC Manual Programming.

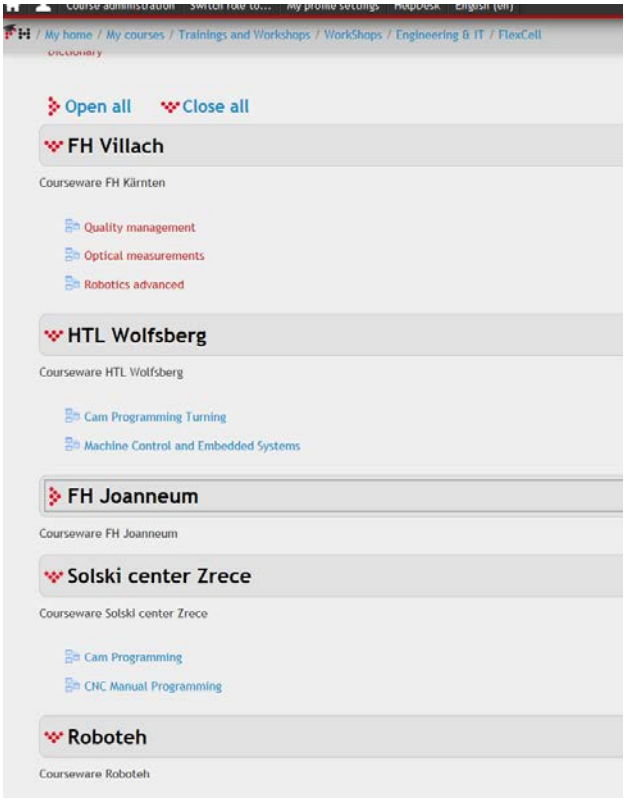


Figure 7. Moodle Training at CUAS

Furthermore a technical dictionary English – Croatian, English – German, English – Slovenian and English – Turkish was added.

C. Software -Tool (Matlab)

Finding solutions for technical problems need first of all physical knowledge of the proposed task and a lot of mathematical relationships. Therefore a successful knowledge (transfer) concept has to cover these topics. Furthermore a modern pedagogical method requires

computational help in solving mathematical problems. It turns out that mathematics is a challenging topic for many pupils and students. On the other hand nowadays there are different pleasant ways to improve mathematical skills. There are for example applications running on a tablet or smartphone, which help to teach and learn mathematical content in a playful way for pupils. For engineers Matlab is a valuable tool - some people would say a "toy" for technicians - for solving mathematical oriented tasks. Matlab and its simulation environment Simulink combine the evaluation of mathematical formulas or complex tasks with the elements of a simple programming language. It is seen as an easy-to-learn tool for solving and understanding mathematical related tasks. Therefore we decided to include an introduction to Matlab in the training module. We are convinced that Matlab is a suitable programming language also for pupils and beginners in a technical field [2].

Due to the fact that our proposed topic is related to robotics the introduction to Matlab is based on the needs of basic robotic calculations.

Furthermore we used the introduced Moodle platform to prepare an interactive "Matlab for robotics" training module [5], [7].

It is clear that users (e.g. pupils involved in the training) need Matlab access when elaborating on the units. We added citrix receiver functionality to the Moodle course, so that the user has access to our Matlab configuration. Thus the student only needs an internet connection but no additional software to perform the training units.

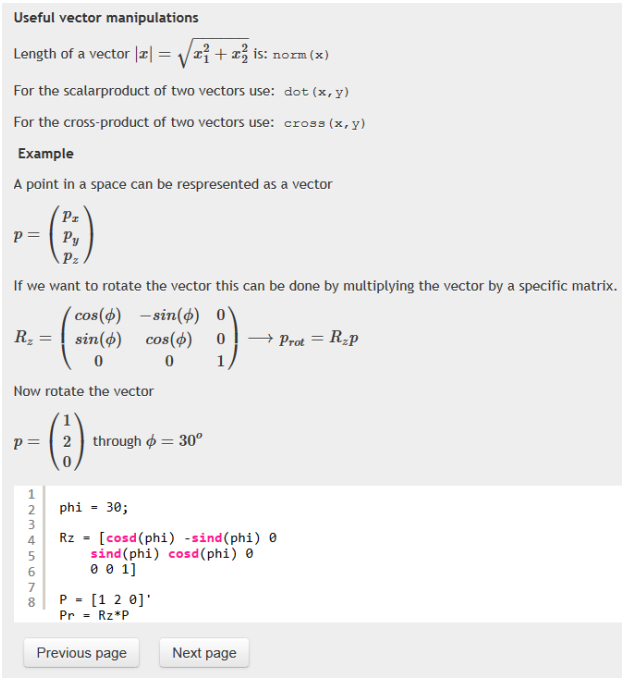


Figure 8. Example of Moodle course

In order to give an impression of the unit the content is summarized. The main task of the module consists of

- Doing simple calculations (scientific calculator)
- Using variables
- Making plots
- Working with vectors and matrices
- Using m-file functionality
- Applying control flow (loops, etc.)
- Trajectory planning

The corresponding pedagogical approach uses three elements. First the considered mathematical background is explained. Then selected examples are presented and solved both by "manual calculation" and Matlab. In addition exercises are included which have to be solved by the user immediately. Of course related solutions can be found in the tutorial so that students are getting direct feedback to their learning progress. A vector/matrix example is presented in fig. 8.

It is clear that there are many printed books and a lot of online material available for Matlab training and Robotics education. Our approach involves comfortable theoretical internet based training, while at the same time Matlab usage is directly integrated in the course.

Moreover it is possible to use the same environment (Moodle-Matlab) for complex robotic tasks. In the case of available budget robust ready-to-use real world components can be integrated. In our labs for example we deal with Quanser elements (e.g. DC-drives) and the Simulink related software "Quarc" to setup robotic experiments. We used this configuration also in the training session for the mentioned project [9], which is integrated in the Moodle course. An additional advantage is that most of the presented Matlab exercises can also be directly solved by free software like octave [11], [12].

D. Low-Cost solution

The training concept that guarantees a successful knowledge transfer and corresponds to the given framework conditions (e.g. long distance training) needs on one hand a web based training platform.

On the other hand practical experiments are still an indispensable pedagogical element in teaching technical subjects. As mentioned above, there are a lot of educational lab products available on the market. Unfortunately this equipment is very expensive.

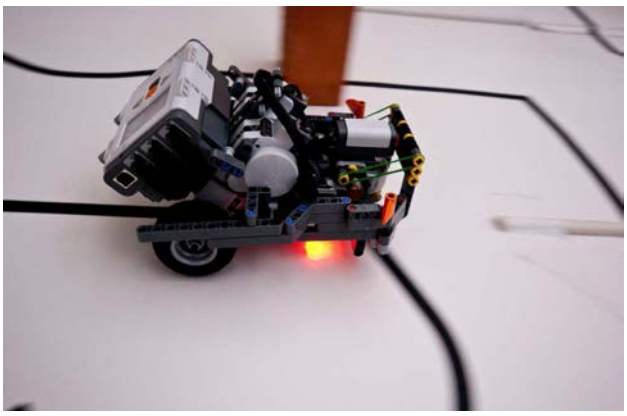


Figure 9. Line follower LEGO

So a setup is needed that takes the particular financial situation of the partners into account. A feasible low-cost hardware solution, which offers especially beginners the possibility to perform robot experiments, is provided by LEGO. The corresponding controller – the LEGO Brick EV3 – and a lot of actuators (DC-Motors) and sensors are used to build a simple robot lab experiment [3]. The specific training module included thereby the following parts:

- Explaining the components of a robot
- Creating a line follower (fig. 9)
- Understanding feedback control [6]

Again a Moodle section was made to provide the theoretical knowledge of the considered hardware. For example the functionality of used sensors is explained at an undergraduate level.

Figure 10. Robotic training of pupils in Turkey

During the local training a line follower was established together with a student group (see fig. 10). It turned out that this was very motivating for both pupils and teachers to get results in a very short time. It has to be mentioned that in this particular session the freely software of LEGO MINDSTORMS was used to create and implement the needed software task. It is a welcome development that the latest controller is now supported by many software producers. Therefore the presented task can basically also be solved by means of Matlab/Simulink.

During the local training basic concept of feedback control was given. The usefulness and mechanism of feedback was explained based on LEGO components. Within one hour pupils without prior knowledge in automated control designed a P-controller for a DC-Motor and performed their first positioning task.

IV. Discussion and Outlook

This paper is a contribution to a modern didactical knowledge transfer concept. It could be shown that a combination of onsite training and web-based learning together with state of the art simulation tools and traditional elements is an excellent method also for long distance education. The feedback of the students and teachers was very positive and a follow-up training was requested.

Beside the technical solution also friendly relationships are established, which leads to an intention of a further collaboration of all partners (fig. 11).

Figure 11. Intention of future collaboration

Acknowledgment

The authors would like to thank the project team for the great collaboration. Without the commitment and good will of all parties, this project could not be realized. Special thanks is given to Mr. Mario Wehr who supported us at any time in case of problems with Moodle.

References

[1] J. Love, Process Automation Handbook: A Guide to Theory and Practice, Springer, 2007

[2] A. Tyagi, MATLAB and SIMULINK for Engineers, Oxford University Press, 2012

[3] D. Benedetteli, The LEGO MINSTRORMS EV3 Laboratory: Buirld, Program and Experiment with five wicked cool robots!, No Starch Press, 2013

[4] J. Zupančič et al, Broschure – Koop Flexible Automation, flexcell.eu, 2014

[5] P. Corke, Robotics, Vision and Control, Fundamental Algorithm in MATLAB, Springer, 2013.

[6] R. Dorf, R. Bishop, Modern Control Systems, Pearson Prenntice Hall, 2005

[7] B. Siciliano, O. Khatib, The Handbook of Robotics, Springer, 2008

[8] C. Madritsch, C. Ungermanns, T. Klinger, W. Werth, An Innovative Industrial Automation System Showcase for Quality Management and Statistical Process Control Lectures, MIPRO 2012, Croatia, 2012

[9] C. Ungermanns, W. Werth, Future-oriented education with a modern flexible production unit, MIPRO 2013, Croatia, 2013

[10] Moodle, https://moodle.org/

[11] J. Schmidt Hansen, GNU Octave Beginner's Guide, Packt Publishing, 2011

[12] Octave, https://www.gnu.org/software/octave/

Implementation of Information Technology in Accounting Course: A Qualitative Analysis

I. Bujan*, A. Flac**
* Polytechnic of Međimurje in Čakovec /Department of Management, Čakovec, Croatia
** Translation Agency / Court Interpreter and Translator for English and German, Čakovec, Croatia
* ibujan@mev.hr, ** info@translatora.net

Abstract - The purpose of this paper is to conduct qualitative analysis of introducing a specific accounting computer program into the teaching process of the course *Foundations* of Accounting held at the Polytechnic of Međimurje of Čakovec. This paper also has a specific purpose of detecting advantages and possible disadvantages of implementing such computer program. A research sample consists of undergraduate students, who are studying management at the Polytechnic of Međimurje. During this research students have been taught accounting in a classical way during the academic year 2013/14 by means of frontal, auditory teaching methods, using whiteboard, but parallel to this, the accounting program was also used. It is hypothesized that students prefer the accounting program and the usage of computer technology over classical teaching methods.

I. INTRODUCTION

Many academic institutions have already built infrastructure which incorporates information technology in the learning process. The main purpose of introducing information technology into the learning process at the academic level is the improvement of learning outcomes and overall success of the students. What is more, technology can enable the effective application of constructive, cognitive, collaborative, and sociocultural models of learning [1]. While analyzing developed countries, especially the United States of America, Western Europe, Great Britain and Australia, classrooms at faculties and other institutions of higher education in these countries are well-equipped with information technology – laptops, computers, iPads, interactive boards and other digital learning media. Nevertheless, computerization of the learning process should be done upon a detailed analysis of needs in the learning process and particularities of each course. It is not to be diminished that also information education of teachers is also a specific issue to analyze. Teachers at institutions of higher education, who are teaching economic group of courses, need to be additionally trained in order to adequately introduce and apply various software solutions to their courses. However, a problem that might arise regarding this situation, is unwillingness of teachers to educate themselves additionally, which is often the case with senior teachers. A national survey on information technology in teacher education in 1999 in the USA reported that most institutions stated that their technology infrastructure is adequate, and faculty IT skills are comparable with students` IT skills [2].

Since the main purpose of this paper is the analysis of information technology introduction into accounting courses, more emphasis shall be put on business schools, institutions of higher education, since accounting is usually taught in such schools.

The second issue to detect, while dealing with the introduction into IT is the ownership of business schools – Faculty or Polytechnic. State-owned schools usually have scarce resources, therefore information technology can not be accessed due to limited funds. On the other hand, private business schools usually invest much more in the learning process, respectively soft skills, but also in information technology – accessibility of computers, iPads, various programs, applications, business simulations and interactive boards. Mentioned issues have been detected in developed and in particularly less developed countries. Special attention shall be paid to as well as more detailed analysis of this issue shall be conducted regarding business schools in Croatia, which can be classified as a post-transition country.

Seeing the importance of IT in education process, Croatian government has adopted e-Croatia program in 2007, which also deals with e-education [3]. However, any action program is useless without a proper strategy. According to the program e-Croatia, some of the detected issues in Croatian institutions of higher education are a relatively high degree of centralization of the Ministry of Education and non-market oriented strategies of the institutions of higher education [4].

The purpose of this paper is to conduct a qualitative analysis among undergraduate students studying at the Polytechnic of Međimurje in Čakovec, Croatia, at the Department of Management and prove the benefits of introducing accounting software into the course Foundations of Accounting. Since the analysis is qualitative, benefits cannot be postulated in this part of the paper but are to be analyzed in the Qualitative Analysis section.

This paper is organized as follows. First section outlines basic ideas on IT in education. Second part analyses information technology and economic schools. Third part encompasses and deals with qualitative analysis

and results and the last section gives basic conclusions, recommendations and limitations of the research.

II. INFORMATION TECHNOLOGY AND ECONOMICS COURSES

Some of the basic economic courses held in business schools in Croatia are statistics, microeconomics, macroeconomics, accounting, finances, marketing, management, organization and other specific courses within a specific field of economics. The above-mentioned courses might be considered as common courses in various business schools. What is more great benefits can be made upon the introduction of information technology – hardware and software solutions. An accounting course has been chosen because a theoretical, "classical" method of teaching significantly differs from the applied software accounting program.

A. Information Technology at the Polytechnic of Međimurje in Čakovec

The Polytechnic of Međimurje in Čakovec is a state-owned institution of higher education, which has been operating since 2008. It performs three main study programs – management in tourism and sports, computing and sustainable development. Since specific study programs, like computing are performed, the Polytechnic is equipped with the latest hardware and software technology necessary for the teaching process. Since it is owned by the Ministry of Education , funds are limited, and hardware is usually purchased through various projects – especially EU funded projects. Software solutions are also funded by EU funded projects, donations and via internal funds obtained by means of market-oriented approach – scholarship from part-time students.

Provided hardware is used for teaching process in the Department of Management. Economic mathematics and statistics, which is taught in the 1st year of the undergraduate study is a good example of IT integration into the learning process. Theoretical content of the courses are well supplemented with Excel. Students use Excel program to attend the course and also pass exam that way. Implementation of IT is well demonstrated in the operational planning course, taught in the 4th semester of the undergraduate program in management. Students take part in classical auditory, frontal lectures but they also do business simulations via CESIM program launched by a Swedish company. The purpose of these simulations is to make business decisions on investments, sales, employments, layoffs and accordingly students can see how their decision turned out – with loss or revenue on a weekly basis. This virtual world gives just a little idea of how different decisions manifest and affect the market.

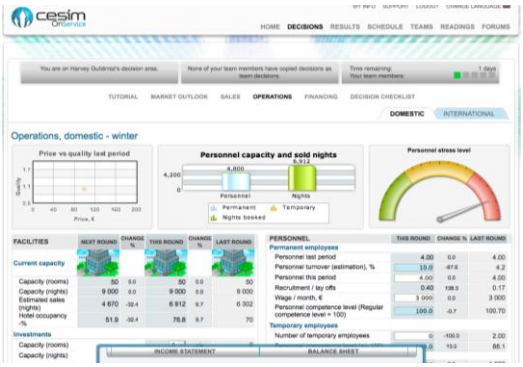


Figure 1. Example of CESIM simulation game

This simulation game is also applicable to other courses – marketing and management, especially to hospitality industry.

The course of a special interest in this paper is the course called Foundations of Accounting taught at the Polytechnic. The mentioned course is taught in the 2nd semester of the undergraduate program for management studies. Teaching methods used in this course can be classified as "classical" and "computer supported" methods. This "classical" way refers to teaching double entry accounting with a "T" account and a "computer "supported" teaching method, which uses various accounting computer programs.

A T account is a graphic representation of a general ledger account consisting of debit and credit side and is a basic training tool in a double entry accounting. Debit entries are depicted to the left of the "T" and credits are shown to the right of the "T". The grand total balance for each "T" account appears at the bottom of the account. A number of T accounts are typically clustered together in order to show all of the accounts affected by accounting transaction. This approach is not used in single entry accounting, since only one account is impacted by each transaction.

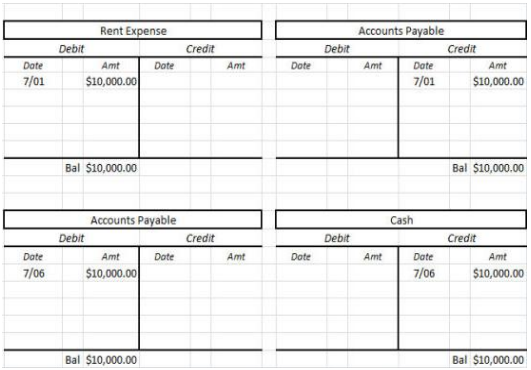


Figure 2. Example of "T" account

The "T" account is very convenient to use for educational purposes, however, it is not used in business practice. Also, it is very useful for compiling more complex accounting transactions. At majority business schools accounting is taught by means of "T" account, not only in Croatia but also in other countries around the world [6]. Professional Croatian magazines, like RIF, RRIF also use "T" accounts, when explaining accounting transactions in the articles. In practice various accounting programs are used [7].

It is more frequently discussed and misunderstood that the usage of an accounting program is possible without passing and learning accounting theory and basic principles of a double entry accounting. The author strongly argues that the implication of accounting software is not possible without any theoretical and "classical" education and that introduction of accounting software gives the right combination of IT and classical teaching.

B. Accounting Programs

There are currently many accounting programs on the market offering various features. Nonetheless, various changes and upgrades of existing accounting programs are quite peculiar to Croatia, due to frequent changes in tax and other laws. Therefore, programmers must provide upgraded versions of programs. Polytechnic of Međimurje has rather limited funds, therefore the accounting program *I.SUSTAV* was obtained from a donation of one Croatian entrepreneur from Čakovec. The main motif of this entrepreneur was to provide the software for students to prepare them for practical work. Also, the students could recommend the program to their future employer since the program offers various other features.

III. QUALITATIVE ANALYSIS

Qualitative analysis has been used as a method of research in this paper since answers of students are necessary to decide what advantages are provided by introducing accounting program (in this case *I.SUSTAV*) into the Foundations of Accounting. Since the method of qualitative analysis shall be used for obtaining non-numerical data, a classical statistical hypothesis testing is not possible. Nevertheless, the author's hypothesis that students understand a basic double entry accounting better when IT, i.e. accounting program is introduced and their basic idea is that unnecessary theory should be excluded from the teaching process and more practical documents should be booked with the accounting program. This hypothesis shall be confirmed by using qualitative questions set in the online survey.

A research sample consists of 70 undergraduate students who have attended Foundations of Accounting. The survey was sent to students` email addresses on 1 February, 2015 and consisted of 9 questions as follows:

Q1. Did you attend an accounting subject in your high school?

Q2. Do you think "T" accounts are more useful than accounting program in the teaching process?

Q3. Do you think that it is possible to learn accounting only by introducing accounting program in the learning process without any theoretical topics?

Q4. Do you think it would be more efficient for learning process, to minimize theory and introduce accounting program?

Q5. Is it easier to learn accounting with the help of accounting program?

Q6. What is, according to your opinion, the most important advantage in using accounting program over "T" accounts?

Q7. What did you dislike about the accounting program?

Q8. Do you think your **computer** literacy is sufficient for using accounting program?

Q9. Would you attend additional classes only to learn working with the accounting program *I.SUSTAV*?

A. Research Results

This section presents a more detailed analysis of research questions. In the e-survey only the above-mentioned questions (from Q1 to Q9) were set, no questions concerning age, gender or any other socio-economic questions were of interest. Concerning Q1. Did you attend an accounting subject in your high school? 70% of the students reported that they have already had some kind of accounting education and 30% did not. At this point it can be concluded that a higher percentage of students, who have some king of accounting knowledge could give better responses in assessing accounting program.

Concerning Q2. Do you think "T" accounts are more useful than accounting program in the teaching process? 45% of students think that "T" accounts are more useful than accounting program, 10% think that "T" accounts could use an additional explanation tool and 45% of students think that it would be more efficient if the accounting program would be used instead of "T" accounts.

Concerning Q3. Do you think that it is possible to learn accounting only by introducing accounting program in the learning process without any theoretical topics? 78% of students think that it is not possible to learn accounting only by introducing accounting program, while on the other hand 22% of students think that it is possible.

Concerning Q4. Do you think it would be more efficient for the learning process, to minimize theory and introduce accounting program? 37% of students think that they would learn accounting more easily without theoretical background. It is interesting to note that a large percentage, i.e. 63% of students think that it is not possible. This question can be correlated with Q8 which investigates students' knowledge of informatics, and it can be concluded that students who dislike informatics also dislike using the accounting program and do not prefer the accounting program.

A.

Concerning Q5. Is it easier to learn accounting with the help of accounting program? It is very interesting to note that 94% of students think that it is easier to learn accounting with the help of accounting program.

Question Q6. What is, according to your opinion, the most important advantage in using accounting program over "T" accounts, was very useful in obtaining primary data. The answers were analyzed and systematized in the Fig. 3.

% of answers	Answer – benefits of the accounting program
37%	I will be able to use the program in the future for my job.
21%	It is quicker and more systematic.
35%	It is more frequently used in practice.
1%	Less paperwork.
3%	I am used of "T" accounts, therefore I don`t like the accounting program.
1%	Easier understanding.
2%	Since computers will be used even more in the future, I think that the accounting program should definitely be used in lectures.

Figure 3. Answers on research Q6.

According to the Fig. 3 it can be concluded that majority of students see the benefits of the accounting program in their future job, think of it as an easier way of learning and understanding the accounting topics.

Q7. What did you dislike about the accounting program? 100% of the students responded that they liked everything about the accounting program.

It is also interesting to point out that regarding Q8. Do you think your information literacy is sufficient for using accounting program, 98% of students consider that their knowledge of informatics is sufficient to use the accounting program without any additional informatics course.

The last questions, Q9 asks students whether they would attend additional classes just to learn to work with the *I.SUSTAV* accounting program. 90% of students reported positively.

IV. CONCLUSION

In this paper a qualitative analysis was conducted on the sample of 70 students studying at the Polytechnic of Međimurje in Čakovec taking the course Foundations of Accounting as part of the bachelor study program Management in tourism and sports. The main goal of this paper was to conduct a qualitative analysis in order to obtain some qualitative data on potentials of using the accounting program *I.SUSTAV* in the future classes. It was hypothesized that students would prefer accounting program, and using computer technology, over classical teaching methods. To prove the hypothesis Q2, Q5, Q7 and Q9 will be discussed.

According to Q2, 45% of students think that it would be more efficient if the accounting program would be used instead of "T" accounts and 10% think that it would be good to use the accounting program with "T" accounts. Furthermore, according to Q5, even 94% of students think it is easier to learn accounting with the help of the accounting program. Q7 was very indicative since 100% of students reported that they have no dislikes regarding the accounting program.

One of the most affirmative questions was Q9 where 90% of students reported that they would attend additional lessons only to learn accounting using the accounting program *I.SUSTAV*.

The questions were analyzed in a qualitative way and according to the percentage of answers the hypothesis could be accepted. Additional support for the hypothesis are benefits of the accounting program set by the Q6 – practical usage of the program, the accounting program is quicker and more systematic, it is more used in practice, less paper work, easier understanding.

On the basis of the conducted research it can be concluded that students generally accept technology in education and that they see many benefits in using the accounting program for their future employment. Limitations of this research lies in the fact that only one generation of students has been tested. For obtaining a more detailed analysis it is recommended to conduct surveys at least five years in a row and start introducing information technology in every accounting course.

REFERENCES

[1] D.E. Leidner, S.L. Jarvenpaa, "The Use of Information Technology to Enhance Management School Education: A Theoretical View", MIS Quarterly, vol. 19, no. 3, Special Issue on IS Curricula and Pedagogy, pp. 265-291, September 1995.

[2] D. Moursund, T. Bielefeldt, "Will New Teachers Be Prepared To Teach in a Digital Age? A National Survey on Information Technology in Teacher Education", International Society for Technology in Education, Eugene, OR, Milken Family Foundation, Beverly Hills, CA., 1999.

[3] G. Sinković, A. Kaluđerčić, "E-učenje – izazov hrvatskom visokom školstvu", Ekonomska istraživanja, vol. 19, no.1, pp.105-113, 2006.

[4] Središnji državni ured za e-Hrvatsku: Operativni plan provedbe Programa e-Hrvatska 2007. za 2004. godinu, Zagreb, 2004.

[5] http://www.cesim.com

[6] National Curriculum, Accouting Guide Grade 11-12, Basic.

[7] V. Brkanić, "Računovodstvo poduzetnika", Knjiga I., RRIG, Zagreb, 2012.

[8] Međimursko veleučilište u Čakovcu, "Studijski program – preediplomski stručni studij menadžmenta u turizmu sportu", Čakovec, 2008.

Evaluating the Quality of Games Designed for Learning Programming by Students with Different Educational Background: An Empirical Study

Tihomir Orehovački* and Snježana Babić **
* University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
tihomir.orehovacki@foi.hr
** Polytechnic of Rijeka, Business Department, Rijeka, Croatia
snjezana.babic@veleri.hr

Abstract - Programming competences have become an important aspect in many different professions. Considering that games give enjoyment to their players and stimulate their curiosity, they represent a convenient tool for gaining programming skills. With an aim to examine the quality of games designed for learning programming from both pragmatic and hedonic perspective, an empirical study was carried out. Study participants were students enrolled in different programmes at two Croatian higher education institutions. Data was collected with subjective measuring instruments. The reported findings uncovered pros and cons of evaluated games meant for leaning programming concepts.

I. INTRODUCTION

In the context of computer and information sciences there is a special emphasis on motivating applicants in acquiring programming skills on all levels of education. Results of the recent study have shown that 60% of engineering students expect the programming courses to be more fun, user friendly and entertaining [1].

One way how students' motivation can be increased is implementation of computer games in the learning process [2]. Games with educational potential in general and serious games for learning programming in particular enable students to acquire certain learning outcomes in an entertaining fashion [3]. Current games meant for learning programming have confirmed their efficiency in enhancing the level of students' knowledge as well as students' overall satisfaction [4].

With an objective to support development of educational games, a number of frameworks and models have been proposed. However, studies related to evaluation and customization of games to certain educational domains are still rather rare [5]. Along with traditional technological features, an important role in the design and evaluation of educational games play specific pedagogical facets that need to be met during the integration of games into educational settings. According to Kazimoglu et al. [3], well-defined learning outcomes are one of the essential motivational aspects of educational games. The same authors claim that pedagogical framework focused on stimulating skills such as problem solving, building algorithms, debugging, simulating, and socializing can be employed as a guideline for the development of educational games meant for growth of logical thinking. Due to the complexity of the educational process and students' characteristics, designers are increasingly integrating students' requirements in different phases of the game lifecycle such as design and evaluation [6].

User experience (UX) represents an outcome of interaction with educational games and is affected by pragmatic and hedonic facets of User-Centred Design (UCD) paradigm [19]. Hassenzahl et al. [8] revealed that UX enables insight in users' perception of software quality, not only its assessment from objective perspective. In other words, UX balances between diverse aspects of software quality including beauty, challenge, stimulation, and self-expression. The perception of user experience is subjective in nature and is based on individual values of particular user and specific context of use (e.g. user's assignment, social environment, organizational characteristics, etc.) [9].

This paper reports results of the empirical study that was carried out with an aim to determine if perception of pragmatic and hedonic dimensions of quality in the context of games designed for learning programming concepts differ among students with different educational background.

The remainder of the paper is structured as follows. The next section provides a brief overview of current advances in the field. Findings of the empirical study are discussed in the third section. The last section offers concluding remarks and guidelines for future work.

II. LITERATURE REVIEW

Considering educational games as a representative sample of slow technology, Orehovački et al. [7] introduced a theoretical framework of hedonic quality that reflects an interplay of following attributes: expectation,

confirmation, perceived usefulness, perceived ease of use, pleasure, aesthetics, playfulness, satisfaction, and loyalty.

With an aim to determine factors that contribute to students' acceptance of educational games, Ibrahim and Jaafar [10] extended a set of constructs introduced by Venkatesh et al. [17] with two additional factors (learning opportunities and enjoyment) and two moderating variables (gender and game experience).

Games with higher level of usability are more appealing and pleasurable to players. Results of the study carried out by Christou [11] indicate that inexperienced players provide more beneficial and precise feedback on usability issues of games than experienced players. The same author also discovered that inexperienced players relate usability with availability of game features whereas the experienced ones relate this concept with the ease of using a game. Since the aforementioned aspects of usability are different but in the same time complementary, both inexperienced and experienced players should be involved in studies related to the quality evaluation of games.

As an outcome of their work, Zaharias et al. [6] confirmed that hedonic and pragmatic facets of quality have equal influence on players' perception of game's attractiveness. The same authors also found that different aspects of perceived usability together with assignments which constitute a particular game significantly affect the motivation of players. Therefore, when developing a game, special attention should be paid to the design of assignments implemented into game as well as interface features.

Barnes and Richter [12] discovered that the awards and penalties as an integral part of in-game's feedback are significantly related to specified learning objectives. Zhang and Lu [13] explored usability of a mobile serious game for learning programming iPlayCode with following attributes: learnability, efficiency, memorability, errors, and satisfaction. The results of their study imply that 61% of students play this game because its interface is easy to manage and use. The set forth usability aspects were operationalized as dimensions of satisfaction that appeared to be a major predictor of students' continuance intentions related to iPlayCode game.

Based on the initial postulates of the UTAUT model [17], Ibrahim et al. [14] confirmed significant influence of performance expectancy, attitude and enjoyment perceived by students on their behavioural intentions to use educational games. On the other hand, it appeared that effort expectancy, self-efficacy and anxiety do not affect students' behavioural intentions in that respect.

III. EMPIRICAL STUDY

A. Measuring Instruments

Drawing on proposed models for evaluating Web 2.0 applications [15][16][20][21][22], a set of attributes aimed for measuring pragmatic and hedonic dimensions of quality in the context of educational games designed for learning programming have been identified. In addition, attribute meant for assessing self-efficiency was adopted from UTAUT model [17] whereas attribute that reflects aspects of challenge was adopted from model introduced by Hainey et al. [18].

Data was collected with two online questionnaires. A total of 60 items constituted the first questionnaire. These items were associated with demographic data of the participants, frequencies in using computer and mobile games, students' anxiety related to programming and playing computer games, their previous experience in playing games, students' perception of programming as a discipline and their desires related to the employment of games in the educational ecosystem. The second questionnaire was composed of 150 items related to the evaluation of 28 different aspects of the perceived quality of the games for learning computer programming. The responses to the items referring to the perceived quality were modulated on a four-point Likert scale (1- strongly agree, 4 – strongly disagree).

B. Participants

A total of 175 subjects took part in the study. The majority of them (70.86%) were students at the Polytechnic of Rijeka (POLYRI) while the remaining 29. 14% studied at University of Zagreb, Faculty of Organization and Informatics (FOI). The sample was mainly composed of full-time students (POLYRI = 96%, FOI = 98%). The average age of FOI students was 20.80 (SD=3.589) years while POLYRI students aged 19.27 years (SD=1.212) on average. At the time the study was carried out, majority of FOI students were enrolled in the Information Systems while only 3.9% of them were enrolled in Business Systems undergraduate study programme. The sample of POLYRI students was composed of 33.9% individuals enrolled in Informatics, 24.2% students of Entrepreneurship and 41.9% of subjects enrolled in Transport undergraduate study programme. The vast majority (28.2%) of POLYRI students do not play computer games, 22.6% of them play computer games less than once a week whereas 19.4% of POLYRI students play computer games between once and twice a week. On the other hand, only 3.9% of FOI students do not play computer games. The majority of study participants from both groups of students (POLYRI=27.4%, FOI =29.4%) play mobile games between once and twice a week. In addition, 41.2% of FOI students and 36.3% of POLYRI students play mobile games more than one hour a week.

C. Procedure

The research was conducted during the winter semester of the academic year 2014/2015 and consisted of two parts. Prior to the implementation of the study, FOI students have completed one course related to the programming. Therefore, they can be referred to as experienced programmers. Several weeks before the study took place, general programming concepts were presented to POLYRI students. In that respect, they can be labelled as novice programmers. At the beginning of the study, the participants were asked to complete first (pre-use) questionnaire. Afterwards, they played each of two educational games (LightBot 2.0 and CodeCombat) designed for learning programming for 30 minutes. When

the time for playing each game has run out, participants were asked to complete second (post-use) questionnaire. Since results of Shapiro-Wilk Tests uncovered that at least one of the variables in a pairwise comparison violates the assumption of normality in data ($p < .05$), differences between evaluated games were explored with Wilcoxon Signed-Rank Tests. Taking into account the aforementioned, reported results are expressed as median values. Considering that data collected from FOI students for the construct "minimal memory load" was normally distributed, differences between games in the context of this construct were examined with parametric T-test.

D. Quality of games designed for learning programming perceived by FOI students

Data that reflect certain hedonic and pragmatic dimensions of quality perceived by FOI students in the context of games meant for learning programming are presented in Table 1. The analysis of collected data revealed that LightBot 2.0 (Mdn=17) fosters significantly more (Z=-2.044, p=.041) self-efficiency among FOI students than CodeCombat (Mdn=18). Namely, it appeared that 47.1% of FOI students had to use a tutorial before solving tasks in LightBot 2.0 whereas 41.2% of them had to do the same before playing the CodeCombat. It was also found that it was significantly easier for FOI students (Z=-2.753, p=.006) to learn to use LightBot 2.0 (Mdn=6) than to employ CodeCombat (Mdn=6). More specifically, the results indicate that 86.3% of FOI students agreed that it was easier for them to become proficient in solving assignments that constitute LightBot 2.0 while 74.5% of them confirmed that the set forth is true for the CodeCombat. Furthermore, 92.2% of FOI students believe that it was easy to learn how to solve the assignments which LightBot 2.0 is composed of where 80.4% of them confirmed the same for CodeCombat. The analysis of data also uncovered that it was significantly easier (Z=-2.901, p=.004) for FOI students to play CodeCombat (Mdn=8) than to use LightBot 2.0 (Mdn=10). Namely, 84.3% of FOI students supported the claim that it was easy to solve assignments implemented into LightBot 2.0 while 94.1% of them agreed with this statement in the context of CodeCombat. Moreover, 98% and 92.2% of FOI students stated that they could alone solve assignments which constitute CodeCombat and LightBot 2.0, respectively. According to the results of data analysis, FOI students had to employ significantly less physical activity (Z=-4.380, p=.000) while solving the assignments integrated in CodeCombat (Mdn=14) than when dealing with assignments implemented in LightBot 2.0 (Mdn=13). Majority (90.2%) of FOI students confirmed they had to use mouse frequently when they were playing LightBot 2.0. On the other hand, most of FOI students reported they had to employ keyboard (86.3%) and exchange work with keyboard and mouse (64.7%) continually when they were solving assignments integrated into CodeCombat.

Study results imply that FOI students had to invest significantly less (T=-3.789, p=.000) mental effort while dealing with the assignments of CodeCombat (M=17.53,

TABLE I. HEDONIC AND PRAGMATIC QUALITY OF GAMES DESIGNED FOR LEARNING PROGRAMMING PERCEIVED BY FOI STUDENTS

CATEGORY / ITEMS	Percentage of the participants responded with "Strongly Agree" or "Agree"	
	LightBot 2.0	***CodeCombat***
SELF-EFFICACY		
SLF5. Before I was able to solve assignmenta that constitute the game for learning programming on my own, I had to watch the tutorial.	47.1	41.2
LEARNABILITY		
LRN1. It is easy to become proficient in solving assignments integrated in the game meant for learning programming.	86.3	74.5
LRN2. It is easy to learn how to solve assignments that are part of the game designed for learning programming.	92.2	80.4
EASE OF USE		
EOU2. It is easy to solve assignments that constitute the game aimed for learning programming.	84.3	94.1
EOU4. It is possible to solve assignments implemented in the game for teaching programming without any help.	92.2	98.0
MINIMAL ACTIVITY		
MAC1. Solving assignments integrated in the game for learning programming requires a great deal of mouse usage.	90.2	33.4
MAC2. Solving assignments integrated in the game for learning programming requires a great deal of keyboard usage.	13.7	86.3
MAC3. While solving assignments that constitute the game for learning programming, it is necessary to frequently switch between the mouse and the keyboard.	15.7	64.7
MINIMAL MEMORY LOAD		
MML1. Using the game for learning programming requires a great deal of mental and perceptive activity.	72.5	49.0
MML2. Solving assignments that are part of the game for learning programming requires a great deal of thinking and decision making.	90.2	58.8
AESTHETICS		
AES1. The design of the interface of the game for learning programming is visually appealing.	78.5	86.3
AES4. The game for learning programming has an attractive interface.	64.7	80.4
AES5. The game for learning programming looks nice.	76.5	86.3
RELIABILITY		
RLB1. While using the game for learning programming, there was no operational interruptions.	91.1	72.5
RLB2. While solving the assignments that constitute the game for learning programming, there were no errors.	80.4	62.8
RLB4. The game for learning programming is stable.	94.1	78.5
RLB5. The game for learning programming is bug-free.	90.2	80.4
CHALLENGE		
CHL2. The game for learning programming increases the difficulty level of assignments with an acceptable speed rate.	74.5	86.2
CHL3. The assignments that are implemented in the game for learning programming are not challenging enough.	17.6	47.1

SD=2.969) than when solving those of LightBot 2.0 (M=16.27, SD=3.007). More detailed, 72.5% of FOI students believe they had to undertake a lot of mental and perceptive activities while playing LightBot 2.0 while 58.8% of them agree they had to think a lot and make too many decisions when solving assignments integrated into CodeCombat. It was also discovered that compared to LightBot 2.0 (Mdn=17), FOI students perceived CodeCombat (Mdn=16) as significantly more attractive (Z=-2.264, p=.024) game for learning programming. Namely, majority of FOI students agree that CodeCombat is visually appealing (86.3%), has attractive interface (80.4%), and looks nice (86.3%). On contrary, results of the analysis suggest that in comparison with LightBot 2.0 (Mdn=8), CodeCombat (Mdn=10) was perceived by FOI students as significantly less reliable (Z=-3.346, p=.001) game meant for learning programming concepts. More specifically, 90.2% of FOI students agree that LightBot 2.0 does not have bugs and is therefore perceived as dependable game for learning programming by 94.1% FOI students. On the other hand, 27.5% of FOI students experienced interruptions when playing CodeCombat whereas 37.2% of them reported errors while solving assignments implemented in CodeCombat. Finally, it appeared that assignments which constitute LightBot 2.0 (Mdn=7) represent a significantly greater challenge (Z=-2.924, p=.003) for FOI students than those integrated in CodeCombat (Mdn=6). While 74.5% of FOI students agree that difficulty of assignments implemented into LightBot 2.0 increases at acceptable rate, only 17.6% of them thinks that assignments which constitute LightBot 2.0 are not challenging enough.

E. Quality of games designed for learning programming perceived by POLYRI students

Results of the data analysis that indicate extent to which evaluated games aimed for learning programming have met particular facets of hedonic and pragmatic quality from the viewpoint of POLYRI students are illustrated in Table 2.

According to study results, it was significantly easier (Z=-2.271, p=.023) for POLYRI students to learn how to use CodeCombat (Mdn=6) than to learn how to employ LightBot 2.0 (Mdn=6). More specifically, it appeared that 83.1% of POLYRI students agree that it is easy to become proficient in playing CodeCombat whereas 88.7% of them support the claim that it is easy to learn how to solve assignments which are included in this game for learning programming. The analysis of collected data also uncovered that POLYRI students had to employ significantly more physical activity (Z=-8.109, p=.000) when solving assignments in LightBot 2.0 (Mdn=13) than when doing the same in Code Combat (Mdn=15). Namely, 89.9% of POLYRI students reported they had to use mouse very often while solving assignments in LightBot 2.0. On the other hand, 87.9% and 73.4% of POLYRI students argued they had to frequently use keyboard and switch between mouse and keyboard, respectively, when dealing with the assignments that constitute CodeCombat. It appeared that interaction with LightBot 2.0 (Mdn=16) required significantly more mental load (Z=-3.165, p=.002) of POLYRI students than it demands playing the CodeCombat (Mdn=17) game. More detailed, 78.2% and 88.7% of POLYRI students claimed that solving assignments implemented in LightBot 2.0 required a lot of mental and perceptive effort as well as a great amount of thinking and making decisions, respectively. The results of data analysis also revealed that compared to LightBot 2.0 (Mdn=7), POLYRI students perceived CodeCombat (Mdn=8) as significantly less reliable (Z=-5.879, p=.000) game. More specifically, 88.7% of POLYRI students have not experienced operational interruption when playing LightBot 2.0 while 83.1% of them claimed there were no errors when solving assignments in LightBot 2.0, Moreover, 96.8%, 96%, and 92.8% of POLYRI students agreed that LightBot 2.0 is reliable, stable, and bug-free game designed for learning programming concepts, respectively.

POLYRI students also determined that CodeCombat (Mdn=5) is significantly more usable (Z=-3.323, p=.001) to individuals with wide variety of characteristics than LightBot 2.0 (Mdn=5) is. Namely, 86.3% of POLYRI students agree that CodeCombat displays labels that describe functionality of particular interface functionalities when mouse cursor hoovers over them. According to the results of data analysis, help resources offered by CodeCombat (Mdn=10) are of significantly higher quality (Z=-2.786, p=.005) than those provided by LightBot 2.0 (Mdn=11). More specifically, 93.6% of POLYRI students think that CodeCombat has diverse forms of user help, 95.1% of them agree that help resources can be accessed to equal extent from different parts of CodeCombat game, while 92% of POLYRI students believe that solutions to assignments that constitute CodeCombat can be found in its help materials. It was also found that quality of feedback provided by LightBot 2.0 (Mdn=12) was of significantly lower quality (Z=-2.392, p=.017) than those displayed by CodeCombat (Mdn=14). Namely, 90.3% of POLYRI students support the claim that LightBot 2.0 displays appropriate messages while they are solving the assignments, 85.4% of them agree that LightBot 2.0 offers information to what extent particular assignment has been solved whereas only 51.4% of POLYRI students think that LightBot 2.0 provides feedback on accuracy of solution to certain assignment.

On the other hand, POLYRI students believe they were significantly more effective (z=-2.317, p=.020) in solving assignments implemented in LightBot 2.0 (Mdn=13) than when doing the same in CodeCombat (Mdn=13). More detailed, 70.3% of POLYRI students reported that playing LightBot 2.0 improved their productivity in learning programming concepts whereas 67.4% of them stated that solving assignments which constitute this game resulted in an increase of their understanding of programming concepts. It also turned out that playing LightBot 2.0 (Mdn=7) represented a significantly greater challenge (Z =- 3.557, p=.000) for POLYRI students than solving assignments that are part of CodeCombat (Mdn=6). Namely, the results of data analysis indicate that 86.3% of POLYRI students were pleased with speed rate at which LightBot 2.0 increases difficulty of its assignments. Moreover, 76.6% of POLYRI students agree that assignments implemented in LightBot 2.0 are challenging enough.

The study results also indicate that LightBot 2.0 (Mdn=14) held the attention of POLYRI students to the significantly greater extent (Z=-2.775, p=.006) than CodeCombat (Mdn=15). More specifically, 72.6% of POLYRI students stated that they have not noticed what was happening around them when they were playing LightBot 2.0. In addition, 89.5%, 84.7%, and 88.7% of POLYRI students reported that solving assignments that constitute LightBot 2.0 encourages their creativity, imagination, and curiosity, respectively. Finally, compared to CodeCombat (Mdn=8.5), LightBot 2.0 (Mdn=7) has met expectations of POLYRI students to significantly greater level (Z=-2.172, p=.030). Namely, 65.1% of POLYRI students reported they were pleased with the assignments implemented in LightBot 2.0.

IV. CONCLUSION

The contribution of this paper is twofold. First, a set of fourteen different attributes which reflect pragmatic and hedonic aspects of quality in the context of games intended for learning programming concepts was

TABLE II HEDONIC AND PRAGMATIC QUALITY OF GAMES DESIGNED FOR LEARNING PROGRAMMING PERCEIVED BY POLYRI STUDENTS

CATEGORY / ITEMS	Percentage of the participants responded with "Strongly Agree" or "Agree'	
	LightBot 2.0	***CodeCombat***
LEARNABILITY		
LRN1. It is easy to become proficient in solving assignments integrated in the game meant for learning programming.	76.7	83.1
LRN2. It is easy to learn how to solve assignments that are part of the game designed for learning programming.	85.5	88.7
MINIMAL ACTIVITY		
MAC1. Solving assignments integrated in the game for learning programming requires a great deal of mouse usage.	87.9	38.7
MAC2. Solving assignments integrated in the game for learning programming requires a great deal of keyboard usage.	13.8	87.9
MAC3. While solving assignments that constitute the game for learning programming, it is necessary to frequently switch between the mouse and the keyboard.	21.0	73.4
MINIMAL MEMORY LOAD		
MML1. Using the game for learning programming requires a great deal of mental and perceptive activity.	78.2	68.5
MML2. Solving assignments that are part of the game for learning programming requires a great deal of thinking and decision making.	88.7	71.8
RELIABILITY		
RLB1. While using the game for learning programming, there was no operational interruptions.	88.7	67.0
RLB2. While solving the assignments that constitute the game for learning programming, there were no errors.	83.1	69.4
RLB3. The game for learning programming is reliable.	96.8	90.3
RLB4. The game for learning programming is stable.	96.0	87.9
RLB5. The game for learning programming is bug-free.	92.8	85.5
ACCESSIBILITY		
ACS1. When the mouse hovers over a particular interface feature, a label that describes its functionality is displayed.	71.8	86.3
HELPFULNESS		
HLP1. The game for learning programming contains various forms of user help.	85.5	93.6
HLP2. User help is accessible from any part of the game for learning programming.	85.5	95.1
HLP3. In a situation when the user does not know how to solve a specific assignment, the solution can be found in help resources of the game for learning programming.	78.2	92.0
FEEDBACK		
FDB1. While solving assignments, the game for learning programming displays appropriate feedback messages.	90.3	91.9
FDB7. The game for learning programming provides feedback on the extent to which an individual assignment is solved.	85.4	91.1
FDB8. The game for learning programming offers feedback on the extent to which the solution to an assignment is correct.	51.4	89.5
EFFECTIVENESS		
EFE1. The game increases students' productivity in learning programming concepts.	70.3	64.6
EFE2. Solving assignments that constitute the game enhances students' understanding of programming concepts.	67.4	65.1
CHALLENGE		
CHL2. The game for learning programming increases the difficulty level of assignments with an acceptable speed rate.	86.3	89.5
CHL3. The assignments that are implemented in the game for learning programming are not challenging enough.	23.4	46.0
PLAYFULNESS		
PLY3. While solving assignments integrated in the game for learning programming, I am not aware what is happening around me.	72.6	67.0
PLY6. Solving assignments implemented in the game for learning programming stimulates my creativity.	89.5	78.3
PLY7. Solving assignments that constitute the game for learning programming arouses my imagination.	84.7	75.0
PLY8. Solving assignments integrated in the game for learning programming fosters my curiosity.	88.7	80.6
SATISFACTION		
STF2. I am satisfied with the assignments that constitute the game for learning programming.	65.1	62.8

identified. Second, relevance which students with different educational background assign to quality attributes has been determined. More specifically, it appeared that both experienced programmers (FOI students) and programming novices (POLYRI students) found differences between evaluated games in terms of quality attributes intended for measuring the extent to which was easy to learn how to solve assignment that constitute particular game, levels to which playing a certain game requires minimal action and memory load, and the extent to which a certain game represent a challenge. The analysis of data collected from FOI students uncovered that they also found differences between LightBot 2.0 and CodeCombat related to the level to which playing particular game improves their self-efficacy in learning programming concepts, the extent to which is easy to solve assignments implemented in specific game, the degree to which a game is visually appealing, and the level to which a game is reliable. On the other hand, data gathered from POLYRI students indicate that games involved in empirical study differ also in terms of the degree to which a game is usable to people with wide variety of capabilities, the extent to which a game contains diverse help materials, the level to which a game provides useful feedback information, the degree to which playing a game enhances students' effectiveness in learning programming concepts, the extent to which a game successfully holds students' attention for a long period of time, and the level to which a game has fulfilled students' expectations. Third, based on the aforementioned study results, advantages and disadvantages of games meant for learning programming have been discovered. Taking the above into consideration, findings reported in this paper can be used as a backbone for future advances in the field. In addition, identified set of attributes can be employed for the assessment of programming-oriented educational games. However, since only two games were evaluated in the empirical study, results cannot be generalized to all genres of games designed for learning programming. With an objective to examine the robustness of presented results, in our future work we will carry out several additional studies that will be focused on exploring interrelations among quality attributes as well as their contribution to students' continuance intentions related to games designed for learning programming.

REFERENCES

[1] A. Yacob, and M. Y. M. Saman, "Assessing level of motivation in learning programming among engineering students". In: The International Conference on Informatics and Applications (ICIA2012). The Society of Digital Information and Wireless Communication, pp. 425-432, 2012.

[2] F. Zhang, D. Kaufman, and S. Fraser, "Using video games in computer science education." European Scientific Journal, August, vol. 10, No 22, pp. 37-52, 2014.

[3] C. Kazimoglu, M. Kiernan, L. Bacon, and L. MacKinnon, "Learning programming at the computational thinking level via digital game-play". Procedia Computer Science, 9, pp. 522-531, 2012.

[4] M. A. Khenissi, F. Essalmi, and M. Jemni, "Comparison between serious games and learning version of existing games". In Proceedings of the 6th World Conference on Educational Sciences, pp. 06-09, 2014.

[5] C. Malliarakis, M. Satratzemi, and S. Xinogalos, "Designing Educational Games for Computer Programming: A holistic Framework". Electronic Journal of e-Learning, vol. 12(3), 2014.

[6] P. Zaharias, C. Gatzoulis, and Y. Chrysanthou, "Exploring user experience while playing educational games: Focus on temporality and attractiveness". International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), 4(4), 19-32, 2012.

[7] T. Orehovački, A. A. Al Sokkar, J. Derboven, and A. Khan, "Exploring the Hedonic Quality of Slow Technology". In: CHI 2013 workshop on Changing Perspectives of Time in HCI, http://bib.irb.hr/datoteka/617623.workshop_paper_final3.pdf, 2013.

[8] M. Hassenzahl, E. L. C. Law, and E. T. Hvannberg. "User Experience-Towards a unified view." Ux Ws Nordichi 6, pp. 1-3, 2006.

[9] H. C. Jetter, and J. A. Gerken, "Simplified Model of User Experience for Practical Application". Ux Ws Nordichi 6, pp. 1-3, 2006.

[10] R. Ibrahim, and A. Jaafar, "User acceptance of educational games: A revised unified theory of acceptance and use of technology (UTAUT)". World Academy of Science, Engineering and Technology, 77, pp. 551-557, 2011.

[11] G. Christou, "A comparison between experienced and inexperienced video game players' perceptions", Human-centric Computing and Information Sciences, vol. 3(1), pp.1-15., 2013.

[12] T. Barnes, A. Chaffin, A. Godwin, E. Powell, and H. Richter, "The role of feedback in Game2Learn". pp. 1-5, In CHI, vol. 2007, 2007.

[13] J. Zhang, and J. Lu, "Using Mobile Serious Games for Learning Programming". In INFOCOMP 2014, The Fourth International Conference on Advanced Communications and Computation, pp. 24-29, 2014.

[14] R. Ibrahim, A. Jaafar and K. Khalil, "Habits and factors affecting undergraduates acceptance of educational computer games: a case study in a malaysian university", In The International Conference on Informatics and Applications (ICIA2012) (pp. 404-418). The Society of Digital Information and Wireless Communication, 2012.

[15] T. Orehovački, A. Granić, and D. Kermek, "Evaluating the Perceived and Estimated Quality in Use of Web 2.0 Applications", pp. 3039-3059, Journal of Systems and Software, vol. 86(12), 2013.

[16] T. Orehovački, S. Babić, and M. Jadrić, "Exploring the Validity of an Instrument to Measure the Perceived Quality in Use of Web 2.0 Applications with Educational Potential", In: P. Zaphiris, A. Ioannou (eds.) LCT, Part I, HCII 2014. LNCS, vol. 8523, pp. 192-203. Springer, Heidelberg, 2014.

[17] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View", 425-478, MIS Quarterly, vol. 27(3), 2003.

[18] T. Hainey, W. Westera, T. M. Connolly, L. Boyle, G. Baxter, R. B. Beeby and M. Soflano,"Students' attitudes toward playing games and using games in education: Comparing Scotland and the Netherlands", Computers & Education, 69, 474-484, 2013.

[19] T. Orehovački, "Development of a Methodology for Evaluating the Quality in Use of Web 2.0 Applications". In: P. Campos et al. (eds.) INTERACT 2011, Part IV. LNCS, vol. 6949, pp. 382–385. Springer, Heidelberg, 2011.

[20] T. Orehovački, "Perceived Quality of Cloud Based Applications for Collaborative Writing", In: J. Pokorny et al. (eds.) Information Systems Development – Business Systems and Services: Modeling and Development, pp. 575–586. Springer, Heidelberg, 2011.

[21] T. Orehovački, "Proposal for a Set of Quality Attributes Relevant for Web 2.0 Application Success". In: Proceedings of the 32nd International Conference on Information Technology Interfaces, pp. 319–326, IEEE Press, Cavtat, 2010.

[22] T. Orehovački, A. Granić, and D. Kermek, "Exploring the Quality in Use of Web 2.0 Applications: The Case of Mind Mapping Services". In: A. Harth, N. Koch (eds.) ICWE 2011 Workshops, LNCS, vol. 7059, pp. 266–277, Springer, Heidelberg, 2012.

Approaches for Implementation of Distance Learning in Higher Education

T.A. Hryhorova*, V.P. Lyashenko* and D.M. Kalinichenko*
* Kremenchuk Mykhailo Ostohradskyi National University, Kremenchuk, Ukraine
gasta1@yandex.ru

Abstract - Life requires an integration of the on-line distance learning into higher educational institutions of Ukraine. It can be used as an element of distance learning, as a separate form of learning, and as an additional instrument in the traditional form of learning. The introduction of this form of training revealed a number of problems like a lack of motivation among teachers, a preparation of teaching material, and the learning model of the majority of distance learning systems. An approach for integration and spreading of the on-line distance learning in the higher education institutions has been proposed in this work. It will allow solving the problems by using special software that takes into account the organization of the educational process management in Ukrainian universities and enables a preparation of the multimedia training courses.

I. INTRODUCTION

The use of the Internet in the education system allows a significant enrichment of the higher education institutions activities by improving the quality of education and expanding its availability. Life demands the new approaches to the learning process in the Ukrainian universities to be introduced in order to ensure the development of the communicative, creative and professional skills of students as well as an improvement of the educational process management. All of this significantly expands the possibilities of the existing traditional learning technologies. The development of the distance education in Ukraine is included in the National Program of Information Development.

A. Advantages of Distance Learning

- For students: it provides the most favorable conditions for using technological capabilities of modern computers and communications. Distance learning improves searching and obtaining for information, helps in developing of cognitive and communicative abilities, an ability to quickly make informative decisions in complex situations, etc.
- For teachers: it provides the release from teaching formal knowledge, the freedom to choose a form of interaction with students. It gives an, opportunity to improve approaches to studying a discipline based on individual abilities and needs of students by learning through discussions, group planning and critical analysis of the obtaining results, and unconventional approaches to soling problems.
- For institutions: it adds a new content to the traditional lecture-form study format.

Platforms for distance learning (E-learning platforms) are already used in some universities Ukraine, and they have many years of practical use in other countries.

That said, we can confidently assert that the on-line distance learning can be used as a separate form of education, as a part of a distance learning, and as an aid in the traditional form of education. But despite the all advantages it has still not became a widespread form of learning in Ukrainian universities.

B. Reasons

- Teachers are not motivated.
- There is a need to create new methods and lectures for new method of learning. All of this can be done using e-learning platforms, as well as using special editor's software. However, the use of editors is not a very convenient approach it because this requires additional set of skills. Also, editor's software environments do not have expanded system of recommendations on how to create a multimedia educational material.
- Most of systems are focused on the Western model of learning (focus on the creation of courses). They have an excessive commercialization, a surplus universality, and complicated set of settings. Also, they often have the bulky and unclear user interface that requires additional development time.

II. RESULTS AND DISCUSSION

A purpose of this work is to develop approaches that will facilitate the implementation of the on-line distance learning for various forms of training in the higher education system, as well as the creation of an information system based on these approaches.

In order to better implement new forms of learning, each institution is required an adaptation period (learning curve). During this period it is better to leave the familiar form of organization and learning management.

A learning management system is a combination of the several components including educational and methodological support, organizational of educational processes, a maintenance of computer and telecommunications systems, a training and professional development of teaching staff, etc.

Institutions, faculties, and departments are basic structural units of the higher education institutions of the accreditation levels III-IV. Department is the basic structural unit of the teaching and research institution of the higher education system that provides educational, methodical, scientific, and research work with the relevant specialty training. The mechanism of the educational and teaching process at the level of the department staff is essential. The educational process should be organized in such a way that provides a teacher with a freedom of choice in selecting methods and forms of organizing the learning process, and it should give a student an academic freedom of choice to select his(or her) own education "trajectory" [1].

If we consider the department from a system's perspective, its main elements are the staff people: a head of the department, teachers, facilitators, laboratory assistants, and students. Each of the system's elements performs a specific task, but it is the interconnection of these elements that provides an achieving of the main task of the system: students learning. All elements of the system are interconnected and form a whole structure; the system simply does not exist in the absence of even one element.

To some extent, it can be argued that the task of the automatic management of the educational process of the higher educational institution is reduced to the task of the automatic management of the educational process within the department. Data on the learning process of various departments are poorly linked. Implementation of the curriculum, monitoring progress, and other elements of the educational process within a specific department faculty do not affect the learning process in other departments of the faculty.

Main Functions of the Learning Management System are:

- the user management system;
- the management of roles (administrator, teacher, student, etc.);
- the management of access rights;
- learning management;
- providing of different types of communication between students and teachers;
- the creation, storage, and delivery of e-learning materials;
- the analysis of learning processes;
- monitoring of the success of training;
- building of the reports, etc.

A research has shown that by using such existing distance learning platforms as Moodle, Claronline, ATutor, SharePointLMS, Live @ EDU, eFront, you can effectively solve the problem of management training activities in the environment of the designing of the adaptive learning courses, its development, and individual control and analysis of the learning process. However, there is no such system structural unit as the Department, there is no role of the "Head of the Department", there is no such a term as 'semester' and no concept of the training course within the semester, here is nothing exist in terms of academic group in these systems. Also, the ability to operate with the grading of students' achievements within semester is not provided in these systems. As a result, there is no way to make a final statement, for instance, of all disciplines taught in the semester. For these reasons, the management of the Kremenchuk Mykhailo Ostohradskyi National University in 2013, refused to use the system Moodle and the like. Therefore, the need has arisen for the information and organization system of educational process of universities that should be oriented primarily on the use of educational groups. Such system should take in the consideration the distribution of learning courses within the semester and support the organizational control of the educational process. This will help to introduce new forms of learning safely and also help monitoring the success of training in the study groups within a separate department.

Another problem that has a serious influence on providing a distant learning is difficulty in preparation of the training materials. Most of teachers are not trained and are not motivated professionally and psychologically to use remote technologies in the educational process [2, 3]. New technologies differ from traditional forms of education; therefore, teachers often have no clear idea of how to develop remote courses in such a way that material is better acquired by students [4]. Also, teachers often have no experience with operating in the electronic educational systems. Despite the fact that special pedagogical technologies for new forms of education are developed for training of teachers of Higher Education Institutions (HEI), the readiness of teachers to use the distance education system increased insignificantly [5].

The offered approach to preparation of the training materials for distance learning system and its program realization will help solving a problem of psychological training of teachers by reducing a time spent on the lecture material preparation and on learning new technologies [6]. This approach is based on the taking into account pedagogical and psychological components of the learning process. It includes recommendations developed in the text editor on the training material preparation. The editor is a component of information system and has an option to generate a package of material automatically. The package includes lectures or recommendations concerning practical and laboratory works developed on the basis of the text document prepared by teacher. All this will allow increasing an interest in use of control systems of the electronic training and will improve a quality of training in different forms of student education.

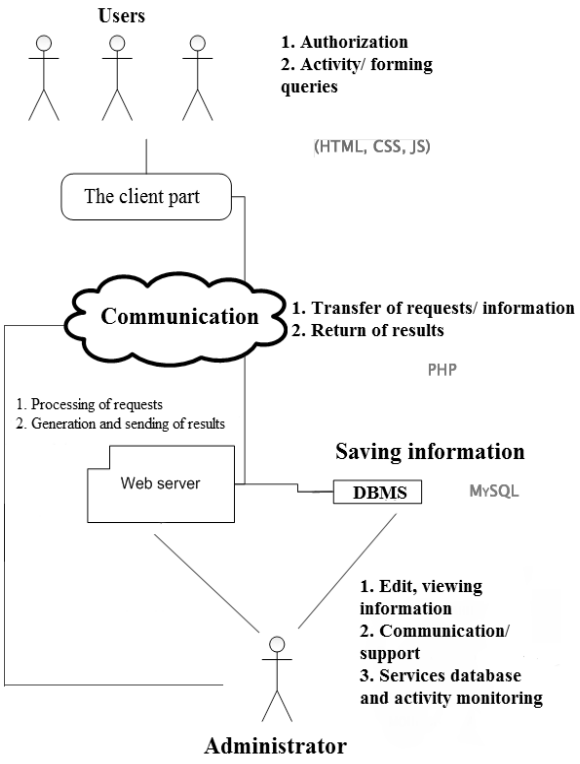


Figure 1. The structure and the interdependence of system components

Three level client-server architecture (Fig. 1) and component model were used for development of information system.

The distance learning system contains such a structural unit as Chair, and new chairs of the HEI can be added, and by adding new Chairs the system increase could be provided. Thus, the system supports a possibility of creating an unlimited number of subsystems, each of which at the same time is an independent link in system. So, each subsystem in its turn has client-server architecture as well.

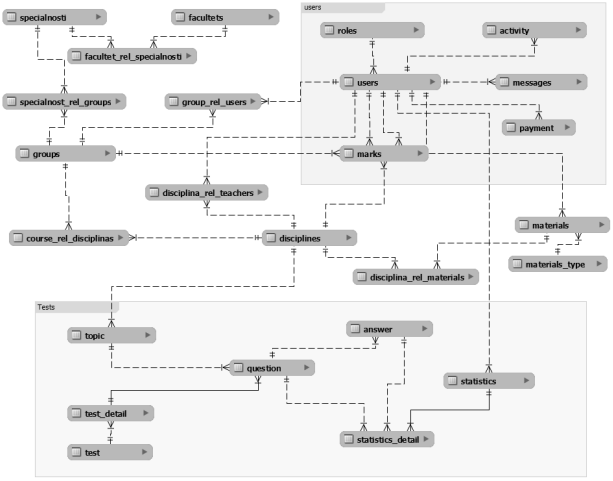


Figure 2. Database of Information System

During the data model development a method of building multi-sites was used with a possibility of creating an unlimited number of databases for each created subsystem with the identical list of metadata.

A database learning management system meets all requirements of the institution and follows the three basic principles, such as scalability, availability and, reliability. Fig. 2 shows the relational database that consists of the tables of information entities.

The data model is divided into three parts: the "Users", "Learning Processes", and "Test". Part "Users" contains information about the users of the system, which includes three main groups: students, teachers and administrators. A head of the department is a part of a group of teachers with elevated privileges that allow access to more information. Part "Learning Process" contains information about the department, study groups, disciplines, courses, and teaching materials. Part "Test" contains information for screening tests in various academic disciplines.

The system consists of four components that are presented as separate modules: Student, Teacher, Administrator and Editor of the system. The module of the Administrator of a control system of educational process gives the user with the permissions of the Administrator a full access to viewing and editing information placed in the MySQL database that is necessary for correct functioning of the system.

The following permissions are available to Administrator:

- to add and delete users, to edit their data, and to define a role of the user;
- to connect students with groups;
- to connect teachers with disciplines;
- to view activity history of users;
- to add faculties, departments, and groups;
- to review information about progress of students with possibility of export;
- to review and edit information contained in a MySQL database by means of the library Adminer.

The module of the Administrator includes the following pages:

- page that displays the main information about users (first name, middle initial, last name, group, login, and payment details);
- page that displays information about activity of users (time of visit, activity on the site);
- page that displays information about progress of the student (first name, middle initial, last name, group, a subject, an assessment, first and last name of the teacher);
- page that displays information on the training materials (available lectures, links of access, and so on);
- page-form for registration of new users;

- page for editing information about users (first name, middle initial, last name, group, login, the password, details of payment, and other);
- page for access to a database;
- page for sending messages to other users of system.

The module of the Teacher allows the user to perform the following actions:

- to add and create materials;
- to delete materials;
- to post and review grades;
- to communicate with other users and administrator;
- to add information about the discipline;
- to view students and teachers activities history;
- to add tests to the system.

Head of the Chair has all the permissions of the Teacher and some additional ones:

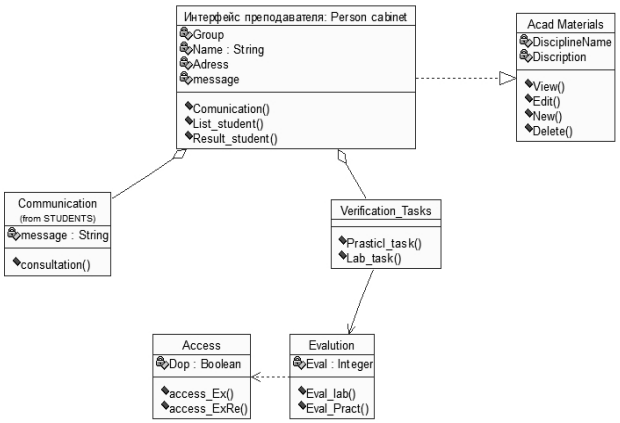


Figure 3. Chart of classes of the module of the teacher

- to view and revise grades of all groups of students in all disciplines of the chair;
- to view the activity history of students and teachers;
- to connect teachers with disciplines, and add in new disciplines;
- to connect student groups with teachers.

After registration or authorization Teacher has an access to his private page where all options/activities are available to him.

The chart of classes of the module is shown in Fig. 3.

The following options are available to Student:

- to view materials on disciplines;

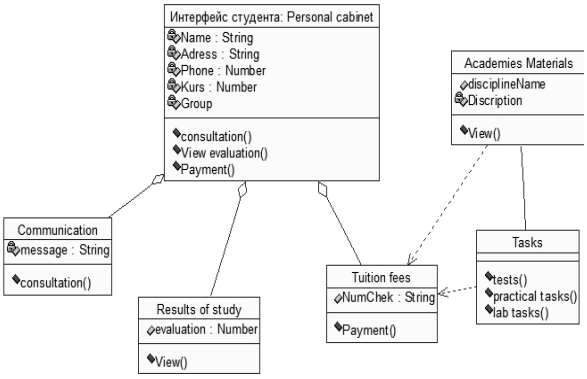


Figure 4. The chart of classes for the student

- to view grades;
- to communicate with other users and the administrator;

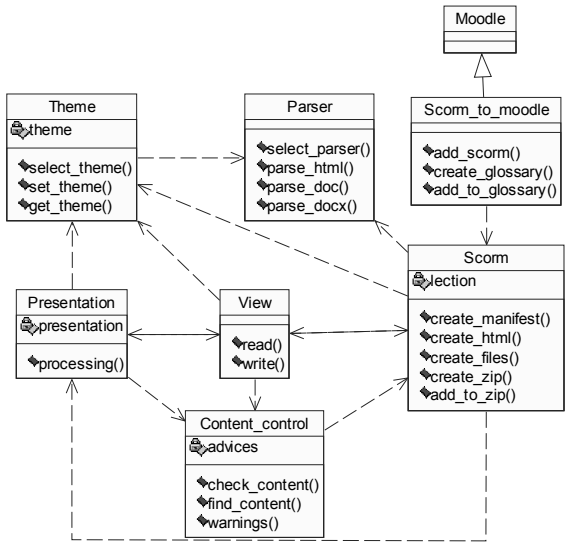


Figure 5. Class diagram of the module editor

- to send works for grading to the teacher;
- to take tests.

After registration or authorization student has to go enter his private page in order to be able to perform all allowed operations.

A chart of classes of the Student module is shown in Fig. 4.

The Editor module allows creating a training material in a form of html document or a SCORM package that includes a lecture text, presentation, and methodical materials with support of the glossary.

A chart of classes of the Editor module is shown in Fig. 5.

The module Editor allows analyzing a lecture material text and creates a template of electronic lecture or presentation on the basis of the rules which are described in the external xml-file. The template is created using keywords found in the text, based on the specifics of a

subject, the amount of available information including such supplemental information like schedules, tables, drawings, video, etc.

The template usually is *.css file. It contains information on color combinations, font size, font style and rules of positions of elements on pages. Separate templates are created for lectures and presentations that account for difference in the features of each lecture types.

A PHP language and the MySQL database server were chosen for implementation of the project. WAMP web server was used for local testing. Macromedia Dreamweaver was chosen as the development environment for creating of web-knots. An interface was created by means of the markup languages HTML and CSS. Interactivity is provided using JavaScript. The database was designed in the MySQL Workbench environment. All necessary statistical information is displayed in tables. Queries in the Database are carried out on the server side with the help of the PHP.

Only a user registered as Administrator in this database has an access to a Chair database. The database and the module are protected from external intervention by mandatory authorization.

A hash function SHA-512 was used to ensure password security. Also, a SSL technology and HTTPS protocols are used for increasing a level of security of the data transfers within system modules. A key (certificate) received by the client from the server is used for encrypting /decrypting of a traffic.

The system supports the 100-poins scale system of grading. The teacher independently determines a type of control for each discipline and number of points reserved for each type. The teacher can grade tasks/types of activities only in those disciplines for which he is assigned. The system automatically calculates number of points by each type of tasks, students rating in group, and estimates both on a national scale, and on the ECTS scale. The system allows preparing and analyzing midterm and final grade sheets for groups on the chosen discipline. For each student, the page with all his/her grades for all types of classes in this discipline is created. An additional module is developed at this stage; this module allows creating structured courses based on the methodical laboratory works prepared in a word-processor Microsoft Word. The courses account for specialty of the learning material design.

III. CONLUSION

The custom system was created at Department of Informatics and Higher mathematics that can be used not only for the organizing distance learning, but can also help increase efficiency of the classical in-class education. The system ensures functioning of the following groups of users: students, lecturers, managers, and administrators of educational institutions, and system administrator. For each user the system supports the authorized access to the resources according to its permissions. The system gives to the management of the educational organization a chance to realize its policy in the field of Internet training, namely: to organize process of training according to curricula, to keep records of academic performance in study groups, to keep records of the working time teachers and students in the system, to keep records and review prepared textbooks and training material of the courses, to connect accounting institutions for tuition payment records keeping if necessary. This control system of educational process is available to public and targets a wide audience.

REFERENCES

[1] V. N. Maximov, Interdisciplinary communication in the learning process. - M .: Education, 1988, 192 p.

[2] S. P. Kudryavtseva, V. V. Colossus, International information: Teach. User. - K. : Publishing House "Word", 2005, 400 p.

[3] T. V. Tkachenko, L. S. Shevchenko, Multimedia technology for distance learning. Teaching Discourse: Collected works. Institute of Education Sciences of Ukraine, Khmelnytsky Humanitarian and Pedagogical Academy - Khmelnitsky: HHPA, vol. 2, 2007, pp. 171-174.

[4] E. V. Burmistrova, Designing the learning process (using remote technologies) in high school. Dissertation of the Ph.D. Omsk, 2005.

[5] L. L. Skvortsova, Psychological aspects of multimedia support lectures. Innovative technologies training in the preparation of economists and managers: a collection of scientific papers. State Grodno University of Kupala. - Grodno: GRNU, 2012, pp. 42-51.

[6] T. A. Hryhorova, A. I. Mandryka, Automation of the preparation of lectures and presentations for distance learning systems. Journal of Kharkiv National University Karazin, vol. 1105. Kharkiv, 2014, pp. 66–74.

Usage of Databases and the Information Literacy of Students of the Faculty of Electrical Engineering and Computing

N. Jelača*
* University of Zagreb, Faculty of Electrical Engineering and Computing / Central Library,
Unska 3, 10000 Zagreb, Croatia
nina.jelaca@fer.hr

Abstract - Given the area of their research, the students of the Faculty of Electrical Engineering and Computing (FER) need the latest scientific information from both domestic and foreign journals. That is why the Croatian Ministry of Science, Education and Sports has provided them with the access to various databases of scientific and technical journals. Furthermore, the Faculty of Electrical Engineering and Computing is a provider of one of the most important databases for engineers – IEEE Explore. Students can access these databases either from their homes or from the Faculty premises; they can also attend seminars on information retrieval. The goal of this article is presenting the research on students of the Faculty of Electrical Engineering and Computing, their usage of available databases and their familiarity with the search tools for those databases.

I. INTRODUCTION

The main interest of this article is finding out whether students use online databases, whether they seek information independently and whether they evaluate information they find online. Students can gather information for their research from many sources, such as libraries, textbooks, journals and articles, their colleagues, etc. The question is – do they look for that information properly? There is a need for students to attend classes and courses on information seeking. This would help them work on their information literacy and means of looking for new information, thus helping them complete their papers, research, seminars and assignments more quickly and efficiently.

The term "information literacy" was first used by Paul Zurkowski in 1974. There are numerous definitions that strive to explain what the term means. Information literacy is a set of skills and knowledge an individual uses to locate, retrieve, analyze and use effectively the information that is needed. The Association of College and Research states that colleges need to incorporate information literacy into their programs and promote it to students, because information-literate people are the ones who have learned how to learn. They are prepared for lifelong learning and have the ability to find and evaluate the information they need. However, the fact that students are not aware of what problems information illiteracy causes is a growing issue [1].

The research that Project Information Literacy has conducted on information-seeking behavior of college students shows that students mostly use Google and a few similar information sources for their research and are not aware of how search engines work, which keeps them from conducting quality research [2]. Their information seeking behavior is more random than organized and they rarely evaluate information they receive as long as it answers the question given in the assignment [3].

The Faculty of Electrical Engineering and Computing (FER) and its Central Library provide students with information on databases they can access and courses they can take up in order to learn more about information retrieval. For example, on FER's Central Library's web page, one can find news about how and where students and other members of the academic society can access databases such as EBSCO, Elsevier, Thomson Reuters and the Web of Knowledge [4],[5]. The Central Library's web page also shows which catalogues and bibliographies, such as PERO and the NSK catalogue, are available for students [6]. There are also news about online training webinars from IEEE Xplore that students can attend in order to adopt the best practices in searching IEEE Xplore [7].

The Faculty of Electrical Engineering and Computing also organizes seminars for the acquisition of research, academic and business generic skills for PhD students. One of the workshops is focused on scientific information retrieval and publishing research results. The students who attend these seminars can gain basic knowledge about relevant bibliographic citation databases, methods of effective searching and retrieval of information and the basic elements of writing scientific articles and their submission process, review and publication in journals [8]. Students may also choose an elective humanistic course that teaches the basic principles of scientific ways of thinking and conducting of scientific research. [9].

Croatian students are given access to many quality online databases from a wide range of interest fields. Access is provided through the Online Database Center, a project of the Croatian Ministry of Education and Sports, CARNet and the Library of Ruđer Bošković Institute [10]. Databases can be used via an AAI@EduHr account – electronic academic identity.

The Online Database Center provides network access to 63 commercial bibliographic and open access databases to all members of scientific and academic communities in Croatia. All databases are acquired after consulting scientists, teachers, users and librarians that are well acquainted with the needs of their users. The Online Database Center provides access to:

- Bibliographic databases from different scientific fields (Medline, INSPEC, MathSciNet, Agricola, GeoRef, PsycINFO, FSTA, Compendex).
- Databases covering all scientific fields (Current Contents, Web of Science, Academic Search Premier, Scopus).
- Full text databases (Blackwell Synergy, Cambridge Journals Online, Springer Link, Science Direct, Oxford Journals, Wiley Interscience, Top 100 Lippincott Journals, Evidence Based Medicine Reviews, Business Source Premier, HeinOnline, Ovid Core Biomedical Collection).
- Open access databases (ArXiv, CiteSeer, DOAJ, Google Scholar, OAIster, PLoS, PubMed Central, Scirus, spiers, etc.) [11].

The Center also provides educational courses for users that teach practical searching skills. Courses are held online and at universities [12].

FER students can also access articles from IEEE Xplore. IEEE Xplore is one of the most valuable databases which covers literature on electrical engineering, computer science and ICT, as well as many interdisciplinary subject areas [13]. It provides access to scientific and technical content and offers more than 3 million full text articles published by IEEE and its publishing partners [14]. In 2012, FER established institutional approach to the IEEE Xplore online database. FER and IEEE singed a 3 year contract, from July 1st 2012 to June 30th 2015, which means that all employees and students can now access the database from the IP address of the University of Zagreb, including student dormitories [15].

II. THE SURVEY

The goal of this study is determining how FER students carry out their research for assignments and papers, what online resources they use and whether they seek help from librarians.

A survey that was distributed to a random selection of 100 students from the Faculty of Electrical Engineering and Computing had nine questions and was anonymous. It was conducted in a three day period during the February of 2015. All 100 surveys where filled. The study addressed following questions:

- What information resources do FER students use to conduct their research?
- Do students use WEB 2.0 applications to cooperate with their colleagues while writing papers?
- Who do students ask for help in finding relevant information resources for their research?
- Do students use advanced options while searching databases and the web?
- Have they ever attended seminars or courses on information retrieval?
- Does the Faculty of Electrical Engineering and Computing need to add more courses on that subject?
- Which databases do students mostly use?

III. RESULTS

The survey was mostly filled by male students (75%, 75 out of 100 students), from the bachelor programme (78%, 78 out of 100 students) and the master programme (22%, 22 out of 100 students).

Table 1. Survey respondents gender

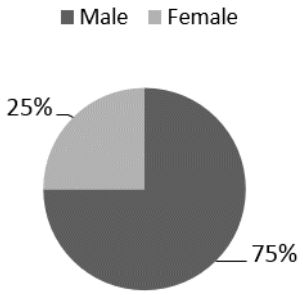
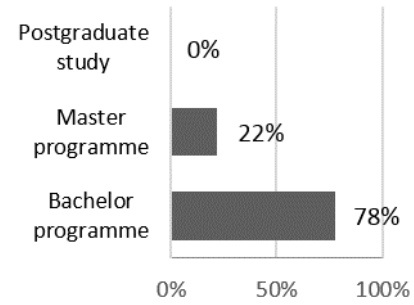
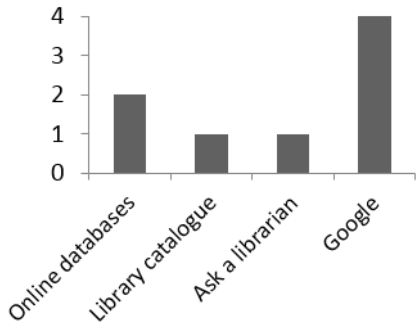


Table 2. Survey respondents college programme



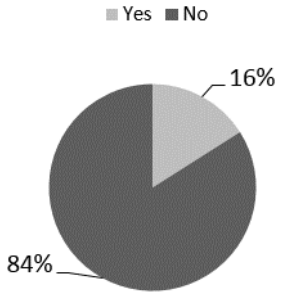
The results indicate that students usually turn to Google for their research and rarely use anything else, such as databases or the library catalogue. They never use library services such as Ask a Librarian.

Table 3. Use of Web sources



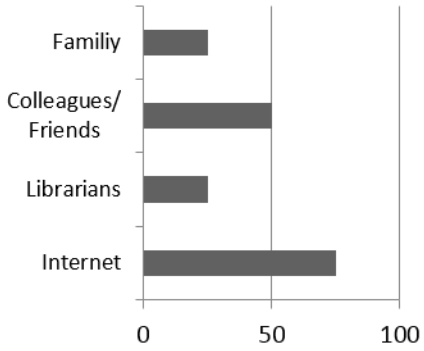
84% (84 out of 100 students) of FER students do not use WEB 2.0 applications (such as Mendeley or Endnote) to help them in bibliography organization and collaboration with colleagues.

Table 4. Use of WEB 2.0 applications



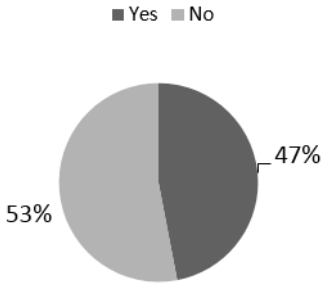
They always use the Internet if they have problems finding the right information and also often turn to their colleagues and friends. They ask their parents or librarians for help rarely.

Table 5. Asking for help



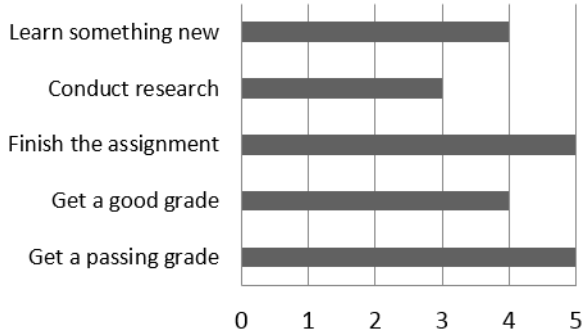
Only 47% (47 out of 100 students) of students use advanced options while searching through databases or the Internet.

Table 6. Use of advanced search options



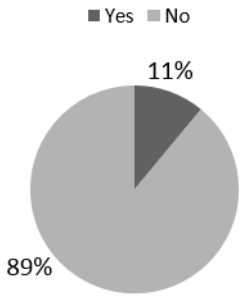
The most important thing to FER students is getting a passing grade and finishing the assignment while getting a good grade. They do not consider learning something new or conducting research to be very important.

Table 7. What is important to FER students



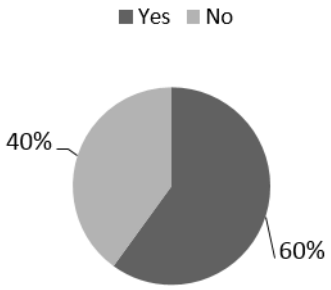
There are few students who attended seminars, courses or lectures on information retrieval and information literacy, only 11% (11 out of 100 students). Seminars where mostly attended by graduate students on FER, in secondary school and Algebra [15].

Table 8. Information literacy and information searching seminars



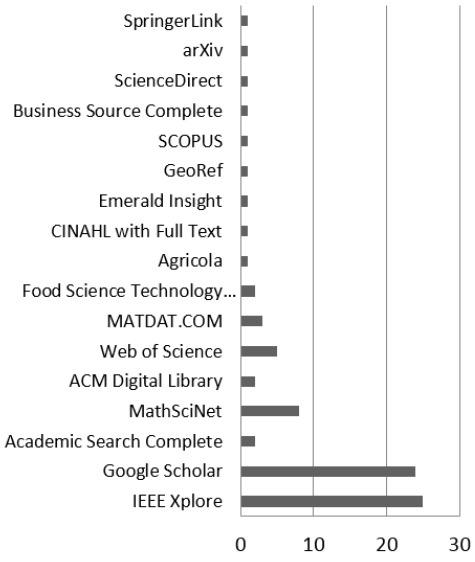
Only 40% (40 out of 100 students) of students think FER should include more of said courses for all students into its programme.

Table 9. Need for more seminars about information literacy



Students where asked which databases they use to conduct research. 51% (51 out of 100 students) of students responded to this question. Databases they use the most are IEEE Xplore, Google Scholar, MathSciNet and Web of Science.

Table 10. Most used online databases



IV. DISCUSSION

The technological advances of today make access to information much easier, yet students still limit themselves to searching for information on Google. They are not information literate and do not have the necessary skills to find, retrieve and evaluate the information they need. Nowadays, students are not acquainted with different types of information sources, such as databases or library catalogues.

Students are either unacquainted with the value of these sources or are uninformed about them and means of using them properly. Students find that getting a passing grade and finishing the given assignment is most important. They are much less concerned with getting a good grade or learning something new. With that mindset, the students' information seeking is reduced to a minimum required to finish the given assignment. Virtually none of the students have extra knowledge and skills in information retrieval because only few had attended seminars on the subject and merely half of them feel that FER should include that type of education in their programme.

Almost all students use the Web as the only resource of information and rely on the Internet for writing papers and finishing assignments. They do not tend to use reliable information sources such as library catalogues or databases which they are welcome to use freely. The results indicate that students either lack the knowledge about certain information resources or about their usage. More than half of the respondents say they never use advanced options, which would help them find relevant information more easily and more quickly, while researching. FER students never use library services or databases the access to which is provided by Croatian Ministry of Education and Sports or FER.

Although master programme students have more access to courses on information literacy on FER than bachelor programme students this research has shown that there is not a big difference between their attendances. From 11% of students who attended those kind of seminars 50% are bachelor and 50% are master programme students.

All of this indicates that libraries and FER have a big task ahead of them. The fact is that acquiring access to databases such as IEEE Xplore means spending a huge amount of money. This research, on the other hand, makes it obvious that students do not use these databases frequently. The college and its library need to work together on promoting and educating students about relevant information sources and how to use them, so that students could start trusting and using them with the skills needed for conducting proper research that would give relevant results.

V. CONCLUSION

Most of the findings from this article indicate that libraries as well as professors have a big problem to solve. Students from bachelor and master programme search for information only on the Web, very rarely consulting relevant information sources such as the library catalogue or scientific databases. There is a need to engage students in critical thinking and to work on their information literacy. Although students are offered different options as to where and how to learn about conducting scientific research, they do not apply to those courses or attend seminars and workshops. Perhaps making those courses mandatory would improve how students see the importance of information literacy. Professors need to help solve this problem by requiring students to use relevant sources for writing their papers and assignments. The Faculty of Electrical Engineering and Computing's educational programme should

incorporate more information literacy and research methodology courses that would be mandatory for students at least at some point of their education. The Library also needs to be included in the implementation of these courses by supporting the teachers and the learning process with its expertise. The Central Library has to promote various quality information resources and help students understand the important role of the library in a student's life. This would make students more independent and critical in their research and provide them with the skills required for lifelong learning – the one that extends beyond their college education.

REFERENCES

[1] Americam library association. Presidential committee on information literacy: final report. 1989. Available at: http://www.ala.org/acrl/publications/whitepapers/presidential [cited 2015-02-05].

[2] Head, A. J. „Project Information Literacy: What can be learned about the information-seeking behavior of today's college students?" Available at: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/2013/papers/Head_Project.pdf [cited 2015-02-05].

[3] J. Lasić-Lazić, S. Špiranec, M. Banek Zorica and Sanja Žunić, "Discipline–related infromation behaviour research: educational policy implications" in Proc. 30th Jubilee International Convention: Computers in Education = Zbornik radova 30. Jubilarnog međunarodnog skupa MIPRO 2007 : Računala u obrazovanju : savjetovanje, CE, Rijeka, 2007, pp. 21–26.

[4] Fakultet elektrotehnike i računarstva. Pristup online bazama podataka EBSCO, Elsevier i Thomson Reuters izvan lokacija matične ustanove. Available at: http://www.fer.unizg.hr/knjiznica?@=2fb3k [cited 2015-02-02].

[5] Fakultet elektrotehnike i računarstva. Udaljeni pristup Web of Knowledge. Available at: http://www.fer.unizg.hr/knjiznica?@=2c5cq#news_9300 [cited 2015-02-02].

[6] Fakultet elektrotehnike i računarstva. Katalozi i bibliografije. Available at: http://www.fer.unizg.hr/knjiznica/baze_podataka/pero [cited 2015-02-02].

[7] Fakultet elektrotehnike i računarstva. EBSCO - IEEE Xplore Digital Library Training. Available at: http://www.fer.unizg.hr/knjiznica?@=2eeq8#news_9300 [cited 2015-02-03].

[8] Fakultet elektrotehnike i računarstva. Osnovne informacije. Available at: http://www.fer.unizg.hr/doktorski/gv/pzipri [cited 2015-02-03].

[9] Fakultet elektrotehnike i računarstva. Istraživanja, publikacije i odgovornost u znanosti. Available at: http://www.fer.unizg.hr/predmet/ipouz [cited 2015-02-16].

[10] Centar za online baze podataka. Available at: http://www.online-baze.hr/o-centru [cited 2015-01-27].

[11] Ministarstvo znanosti, obrazovanja i sporta. Centar za on-line baze podataka. Available at: http://public.mzos.hr/Default.aspx?art=8925&sec=3139 [cited 2015-01-27].

[12] Centar za online baze podataka. Raspored tečajeva pretraživanja online baza. Available at: http://www.online-baze.hr/izobrazba [cited 2015-01-27].

[13] Nacionalna i sveučilišna knjižnica. Omogućen slobodan pristup do više od 450 e-knjiga iz područja računarstva. Available at: http://www.nsk.hr/omogucen-slobodan-pristup-do-vise-od-450-e-knjiga-iz-podrucja-racunarstva/ [cited 2015-01-27].

[14] IEEE Explore Digital Library. About IEEE Explore Digital Library. Available at: http://ieeexplore.ieee.org/xpl/aboutUs.jsp [cited 2015-01-27].

[15] Središnja knjižnica. Available at: http://lib.fer.hr/ [cited 2015-01-27].

[16] Algebra : otvoreno učilište. Available at: http://www.algebra.hr/ [cited 2015-02-10].

Comparing MOOCs in m-learning and e-learning settings

Antonia Bralić, Maja Ćukušić and Mario Jadrić
University of Split
Faculty of Economics
Address: Cvite Fiskovića 5, 21000 Split, Croatia
Phone: +385 21 430758
E-mail: {maja.cukusic, mario.jadric}@efst.hr

Summary

Massive Open Online Courses (MOOCs) are known as free online courses available for hundreds or thousands of participants worldwide. MOOCs have been a hot and current topic within the field of e-learning, with numerous platforms developing and growing, each of which is specialized for courses in different areas. Due to the technology development and its pervasion in everyday activities, platforms that offer MOOCs are developing courses and all related activities in a way that they are available for usage on different platforms and devices, adapting to the trend of ubiquitous learning. MOOCs are being adjusted for mobile and desktop usage and a need to create a frame for evaluating and comparing mobile and desktop instances of MOOCs arose. This paper reviews literature, questions mobile and desktop versions of MOOCs, and provides a frame for their comparison.

I. Introduction

E-learning is a widespread way of disseminating and acquiring knowledge, in all its forms and appearances. It could be defined as a usage of information and communication technologies in order to leverage and enhance process of learning and teaching [1]. One of the most popular e-learning instances are MOOCs, Massive Open Online Courses, which are basically free online courses available for hundreds or thousands of participants worldwide [2].

Furthermore, MOOCs have become very popular because publication and participation are easy to manage, students do not have to pay for a MOOC which is an advantage [3], and the technology has become an essential part of everyday life. Naturally, an extreme role technology has in everyday personal and professional activities raises the question of adapting learning in every situation and in all times. MOOCs are not excluded from this issue.

The paper starts with differing m-learning from e-learning and considering access to MOOCs in mobile and desktop environments. The chapter is followed by presenting potential criteria for comparing MOOCs in m-learning and e-learning.

II. MOOCs In Mobile And Desktop Environment

A. *E-learning, m-learning, t-learning and u-learning*

E-learning is defined in many ways, and one of most popular definitions is that it is the use of technologies in the process of gaining and disseminating knowledge. Similarly, mobile learning (m-learning) is considered as constructing every kind of knowledge and skills by using mobile technologies [4]. m-learning is also defined as learning through multiple contexts, social interactions and interactions of content, by using personal electronic devices [5].

M-learning can be viewed as a natural "extension" of e-learning and, as such, as another component/module of more complex ICT solutions for hybrid learning. In particular, e-learning is a macro-concept that includes the environment for online and mobile learning. E-learning is also a "subset" of distance learning, which is also part of the flexible learning concept [6].

As with e-learning, some of the most important advantages of m-learning (illustrated in Figure 1) are that the study can be done by anyone, in any place, at any time when the user is ready to learn thus taking advantage of mobile phones and wifi enabled PDAs [7]. In general, teachers believe that learners' participation and engagement increases when using mobile devices for class activities outside of class, but does not positively affects their participation and engagement while in class [8]. Obviously, there should be some concern with regards to this type of access to learning content. At the same time, there are a number of advantages that usage of mobile devices brings forth elaborated hereinafter.

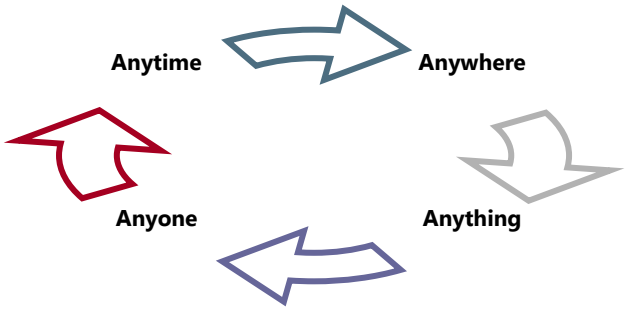


Figure 1: Advantages of m-learning, adapted from [7]

E-learning producers tend to adjust their content and products to make use of the new opportunities enabled by technology development accordingly. Consequently, end users interact with each other using ubiquitous devices because of the increased availability of free wireless network access points, and in this way they are "extending traditional e-Learning into a new phenomenon named Ubiquitous Learning"[9]. U-Learning can be defined as "a function of different parameters: u-Learning = *{u-Environment, u-Contents, u-Behaviour, u-Interface, u-Service}*" [10]. Despite the ubiquitous character of learning, there are some areas for which m-learning is more appropriate and have been emphasized as convenient for m-learning initiatives. These include any form of scan-and-learn project, just-in-time guides, initiatives that can include location, refresher learning [11] and support for on-demand learning in everyday business activities. A research [12] on proliferation of mobile learning in business settings (with 538 respondents from 481 organisations) shows little difference in uptake of mobile learning between not-for-profit (74%), public (73%), and private (69%) sector. As for industries with the highest use of mobile devices commercial training providers (90%), manufacturing, science and engineering (79%), and professional and technical services (72%) are among the leading adopters. Similarly, BYOD (*Bring Your Own Device*) movement contributes to u-learning adoption. It is more frequent in large organizations (20000+ employees), multinational companies and companies with over 10 years´ experience of using e-learning (ibid). As great effort is needed to integrate BYOD properly (especially from security perspective), it is clear that large companies would be at the forefront.

Popularity of tablet devices contributes to u-learning movement as well. Tablet learning can be perceived as a type of mobile learning, concerning portability and personalization of the device used. Still, a lot of theoreticians and experienced professionals emphasize that tablets are not as portable and ubiquitous as mobile devices. Also, t-learning offers many more opportunities for delivering and disseminating content. In all cases, course design is one of the most important factors that influence performing activities and finally completing course assignments.

Presence of mobile internet access grew rapidly in previous years, and a proliferation of smartphones (later, tablets as well) became one of the most relevant game-changers in area of online learning. Due to high availability of devices and learning material, learners should always be able to retrieve the material without the technology becoming a deterrent [7]. An example of successful m-learning project is Merrill Lynch. Their compliance training courses were delivered on BlackBerry mobile phones [13] to: (i) enable learning outside the offices, (ii) make use of existing mobile device usage habits, and (iii) allow people to consult learning content in chunks at time and place they prefer. Reported results are exemplary, namely, the company reported a 1.21% increase in average competency score (comparing to control groups) and a 12% higher completion rate in the set timeframe. Overall, it took less time for employees to complete the course (on average 45% less time in training). Similar superiority of m-learning was identified in the previously mentioned study [12]: as high as 70% improvement in productivity, 47% improvement in manager perception of additional business benefit, and 25% improvement in revenue was identified in some cases when m-learning was used to increase the numbers of participants in learning and development programmes. Similarly, the eLearning Guild in their Mobile learning report [13] emphasise that some of their members who have implemented m-Learning report a 52% improvement in user performance and an 83% increase in making learning available to users.

There are a number of studies comparing mobile and desktop versions of same course. An empirical study [15] inspected delivery of an existing e-learning course within Department of Defense (USA). In the original (e-learning) version of the course, it took around 40 to 60 minutes to complete the course. For the mobile version of the course, completion time dropped to 25 to 45 minutes. In addition, subjective preference was measured – 38% rated the mobile version as 5/5 (being very satisfied with it), 46% as 4/5, and 13% as 3/5. Around 70% of the participants that previously used the e-learning version preferred the m-Learning course, while 26% of the participants preferred the e-Learning course.

Conversely, another study [16] found no difference in learning performance across the three modes of workforce training: classroom face-to-face instruction, laboratory-based e-training, and mobile learning. This supports the commendation that knowledge acquisition can be equally good regardles of the way material is delivered – via desktops or mobile devices (depending on the context of course).

Based on the above, a profound consideration of devices used to access learning materials in MOOCs is needed. Courses need to be created and developed in accordance with instructional design principles, bearing mobile, tablet and desktop computer usage in mind. Companies "push" the use of mobile devices, tablets and own computers and so according to the reported study [12] 52% of organisations that participated in the study provided learners/employees with smartphones (or as high as 60% in "top learning" companies) while 48% provide learners with tablets and 41% have a policy of BYOD (or as high as 66% in "top learning" companies.

Another comparison of e-learning and m-learning/t-learning in the context of use that illustrates why it is important to support both ways is given in Table 1 based on [7]. Collaborative character of e-learning is related to learners behavior during knowledge acquiring: learners can exchange opinions and build ideas and frameworks, as well as challenge their ideas. Further on, networked learning is "learning in which information and communications (ICT) is used to promote connections: between one learner and other learners, between learners and tutors; between a learning community and its learning resources" [17]. With regards to access using mobile devices it is important to note that it is mostly used for

informal learning, which is a type of learning in which learner sets his own goals and ways to achieve these goals, as well as the moment in which they are achieved [18]. In short, m-learning characterizes accesing adapted material that it stripped of excess media content, i.e. is more lightweight.

TABLE I. DIFFERENCE BETWEEN E-LEARNING AND M-LEARNING ACCESS

E-learning	M-learning
Collaborative	Networked
Media rich	Lightweight
Distance learning	Situated learning
Formal	Informal

B. MOOCs

MOOCs have been described as the new form of online learning, while their popularity is growing with academics and within business circles. The term was first used to describe an online open course *'Connectivism and Connective Knowledge (CCK08)'* developed by George Siemens and Stephen Downes [19]. Early MOOCs have been known as cMOOCs, which are based on the *connectivism* theory of learning, which favours networks of learners evolving informally [20]. In 2011 a number of universities in the United States started offering MOOCs via commercial platforms, and these MOOCs were called xMOOCs. xMOOCs have been "focused around a set of short, modularised video-lectures, followed by automated, multiple-choice testing of learners' understanding of the content" [19], and are used to define and describe MOOC concept we use today.

As already stated, what mostly differentiaties an xMOOC from a cMOOC is who provides a MOOC. Namely, cMOOCs are mostly provided by a group of individuals while xMOOCs are usually provided by one or more higher education institutions or lately more often for-profit companies [21].

The most popular MOOC providers are presented in Figure 2. In 2013 Coursera was the largest MOOC provider by far, with almost half (47%) of all the open online courses offered [22]. Its share of courses is slowly being reduced by new MOOC providers.

In 2014 Coursera's share of courses was 36% enrolling around 10.5 million students [23]. Nonetheless, Coursera's share is still twice as large as competitors. Namely, edX holds 16% of the provided courses and has around 3 million registered students.

Based on the same data source, other providers in the "top 5" list per number of students are Udacity (1.5 million students), MiriadaX (1 million students) and FutureLearn (800.000 students), while in the "top 5" list per number of courses besides Coursera and edX are Canvas Network (8.4%), Miriada (5.3%) and FutureLearn (4.7%).

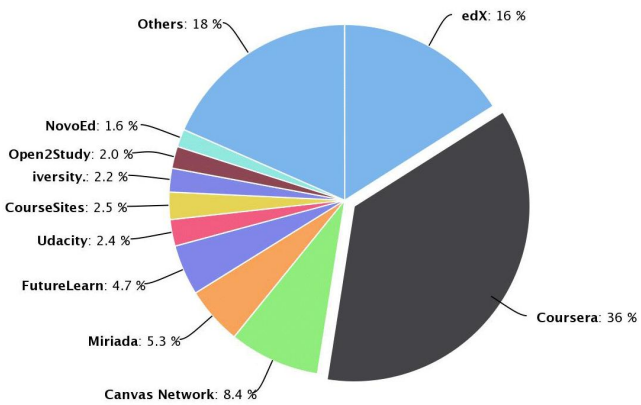


Figure 2: Course distribution by provider in 2014 [23]

C. Accessing MOOCs in mobile and desktop environments

Many of the listed providers maintain and upgrade MOOC learning platforms that offer responsive websites or (native) mobile apps (tablet and smartphone versions). Mobile and desktop versions are often different. When it comes to considering MOOC design, creators have different options for adapting the course for usage on different types of devices.

First of all, it is possible to implement responsive design. By doing this, the web app (e-learning website or a platform) is automatically resized and adjusted for usage on mobile and/or tablet devices, i.e. onscreen content automatically changes layout in response to the size of the screen it is being displayed on [24]. Responsive design is a good idea if creators of a MOOC want to enable mobile learning without dealing with mobile apps. Naturally, responsive design and development is a more complicated and resource-consuming task than e-learning development.

Access to a MOOC titled "E-learning and Digital Cultures" provided by The University of Edinburgh available from Coursera using different devices is illustrated below. The course is not optimized for access from mobile phones or tablets using responsive design principles (compare Figure 3 and Figure 4) but is available through Coursera native mobile app for iOS (Figure 5).

MOOC designers also have an option of developing mobile apps in addition to desktop version of the online course. Majority of MOOC providers do post their mobile apps in Windows, Android and Apple stores for download. Apps are attractive because they are, among other features, "formatted perfectly for the devices on which they are based, tend not to tempt the user away from the job in hand with links, pop-ups and extraneous clutter and they can usually be accessed with a single touch" [25].

Both mobile apps and responsive web apps have advantages and disadvantages that should be taken into consideration when evaluating and comparing MOOCs in m-learning and e-learning environments. Such criteria is presented and analysed in the following chapter of the paper.

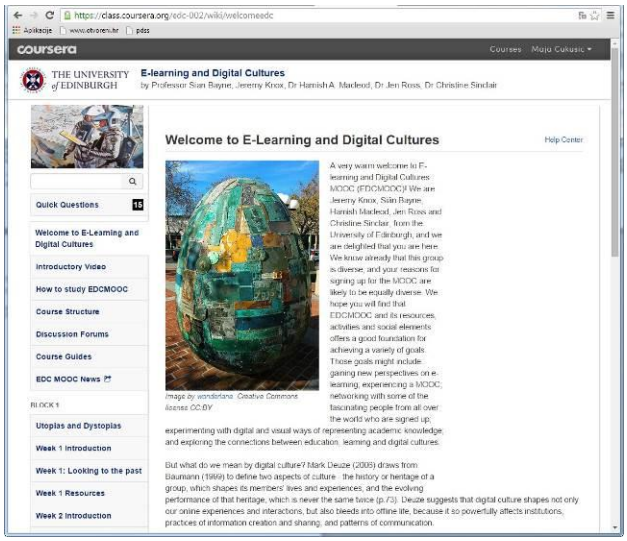


Figure 3: Accessing a MOOC using browser on a desktop computer

Figure 4: Accessing a MOOC using browser on a mobile phone

III. Criteria For Comparing MOOCs In M-Learning And E-Learning Settings

In order to inspect thoroughly MOOCs in different learning contexts it is important to consider sets of criteria for comparison. This paper brings forth the preliminary criteria based on the available body of literature and as such presents a starting point for our future empirical research/comparison of major MOOC platforms (such as Coursera, Khan Academy, Udacity, edX). The aim of the research is to provide a methodological and eventually and empirical overview of similarities and differences of MOOCs design in mobile and desktop environments, supported by comparison of the courses provided by major MOOC platforms.

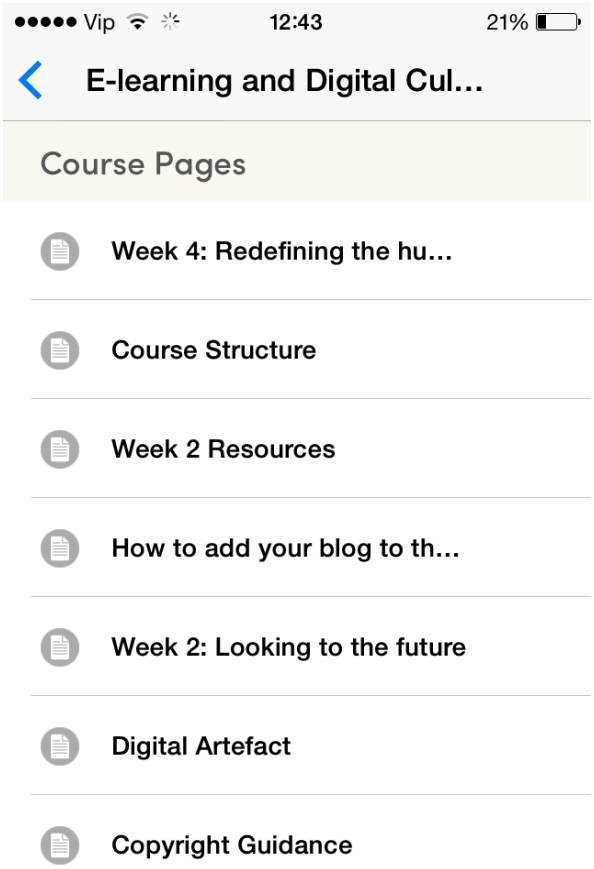


Figure 5: Accessing a MOOC using mobile app

First of all, MOOCs can be observed and compared in relation to popular learning theories. For example, [26] states that "popular MOOCs employ predominantly instructivist approaches to teaching". Courses from the likes of Coursera and Udacity tend to follow a highly instructivist model but, because of their size, spawn networks and study groups of learners who meet face-to-face, and through various social media such as Facebook, to enhance and support one another using quite different and more connectivist approaches [27].

Some common criteria for comparing MOOCs in mobile and desktop environment are presented in Table 2. In their paper, [28] list elements of effective traditional learning, that should be connected to particular characteristics of online learning, to ensure that a successful learning context will be created. The same characteristic should be used to evaluate MOOCs as an instance of e-learning. Other criteria are based on previous considerations of quality e-learning courses (in [29]; [30]; [31]) and common platform features (in [32]).

With regards to devices used to access MOOCs it is important to consider course design and ease of use. Perceived ease of use is one of key variables that influence learning flow in online learning process [33]. It is related to course design, which depends on series of issues. Screen size, time and frequency of use, connectivity, internet connection quality, type of content, target group, all are variables that influence course design.

When it comes to online learning design, content must be broken into smaller chunks of content, called learning

objects [34]. By creating and combining learning objects, online courses are developed in an easier way, and they are more manageable. Also, thanks to granularity and the way that learning object is programmed, marked and saved, possibility of sharing and re-use is increased. One of the questions set before instructional designers is: „How to envision and create learning objects in order to incorporate them in different methodological and technological predispositions connected to different environments".

Differences between e-learning and m-learning in terms of content and parts of instructional design are fading in MOOCs. Idea of MOOCs in mobile devices is to provide the same resources, and slightly altered experience to the users (considered in Table 3 under two different categories of criteria). The experience is the crucial difference between these versions.

TABLE II. CRITERIA FOR COMPARING MOOCS IN MOBILE AND DESKTOP ENVIRONMENT – MOOC CHARACTERISTICS

Quality and expected MOOC features	Author(s)
MOOC should be interactive	Wild, Griggs, & Downing, 2002
MOOC should provide an opportunity to repeat and practice	Wild, Griggs, & Downing, 2002
MOOC should provide a choice between different presentation styles	Wild, Griggs, & Downing, 2002
MOOC content should be relevant and practical	Wild, Griggs, & Downing, 2002
Goals of MOOC are defined	Ćukušić & Jadrić, 2012
Working plan of MOOC activities (*syllabus*) should be created	Ćukušić & Jadrić, 2012
Educational content should be created, its course (sequence) and structure defined	Ćukušić & Jadrić, 2012
Students are being motivated	Johnson & Aragon, 2003
Social interaction is encouraged	Johnson & Aragon, 2003 Gillani & Eynon, 2014 Dron & Anderson, 2014 Ozuorcun & Tabak, 2012
Additional types of activities are provided	Johnson & Aragon, 2003
Acquired knowledge is assessed	Garača, 2008
Platform provides auxiliary functions for instructors (logs, enrollment options…)	Garača, 2008
Course is scheduled (the start and end dates are defined)	Ćukušić & Jadrić, 2012

TABLE III. CRITERIA FOR COMPARING MOOCS IN MOBILE AND DESKTOP ENVIRONMENT – ACCESS TO MOOC FEATURES

Mobile MOOC features	Author(s)
Allows offline viewing (mobile apps)	Korucu & Alkan, 2011 Elucidat, 2013
Allows download only on Wi-Fi	
Content is broken into small pieces	Govindasamy, 2002
Connection of a mobile learning tool to an individual	Ozuorcun & Tabak, 2012
Able to watch in vertical in horizontal position of device	
Responsive MOOC features	**Author(s)**
Supports a range of devices and screens	Elucidat , 2013
Over 90% of website content is delivered through mobile version of the responsive web page as well	

Access to MOOC features using: **a) Desktop computer b) Mobile device c) Mobile App**
Badges overview
Live chat (collaboration)
Access to transcripts
Browse by language
Browse categories of content
Browse by skill level
Browse by author company
Browse by availability
Browse by programming language
Search bar
Chapter list
Progress tracker
Estimated effort for one course shown
Student reviews of a course (other people's feedback)
Verified certificate access
Unverified certificate access
Overview of currently active courses
Communication with authors and authoring organization
Settings area (my profile…)
Syllabus overview
Schedule overview
Access to exercises and assignments
Access to forum
List of known content errors in course
Feedback – survey possibilities
Notifications area

Another important feature that mobile environments provide is offline/online access to learning material [35], [36]. Online access provides up-to-date information and access to data, while offline status ensures speed and no communication costs.

Mobile setting is different than the desktop one, and particular rules should be applied when designing an m-learning course. According to [7], there are several requirements when it comes to designing an m-learning tool: portability, connection to an individual, retrieving knowledge without the technology becoming a deterrent, communication, adaptation to learners development, persistency, usefulness and user-friendliness.

The second part of the table 3 is prepared based on the review of common MOOC features that should be available in both desktop and mobile versions. The content, as already suggested should be delivered in an appropriate way. The representation of both the content, and the features should be different when accessing from different devices. The list is non-exhaustive and will be amended by additional features based on extensive research of 76 randomly selected MOOCs conducted by [19].

IV. CONCLUSION

In the years to come MOOC market is expected to consolidate. It is expected that only largest platform that provide highest quality content will survive thus achieving even greater integration and exploitation of the massive potential of this type of learning, in all learning environments (such as academic or business).

Platforms that offer MOOCs are developing courses and all activities related to those courses in a way that they are available for usage on different platforms and devices, adapting to the trend of ubiquitous learning.

The aim of this paper was to investigate the potential of adjusting MOOCs for desktop and mobile environment and to propose a set of criteria for comparison of MOOC design and implementation tasks and challenges in both surroundings, having methodological and technological terms in mind, moreover, questioning them in MOOC's mobile and desktop environment.

So far, studies focused on mostly methodological issues in MOOC learning since some of the instructional design principles may pose a challenge due to large numbers of users (such as achieving expert human feedback). A number of key areas in which instructional design of these courses should be improved were identified (see for example [19] for a through comparison of over 70 MOOC courses from Coursera, EdX, Udacity, Codeacademy, Canvas Network, FutureLearn, and other websites). Nevertheless, other than rethinking "soft" aspects of MOOCs (some of which are presented earlier in table 2), technological adequacy of the MOOCs should be considered as well. "Pedagogy first, technology second" underpins the criteria presented earlier. Seamless transfer from one device to another, continuing with learning material should be supported by technological solutions and that is the base for criteria presented in table 3.

Future research will aim to validate and further expand the proposed criteria. Also, it will be reviewed whether top 5 MOOC platforms have responsive websites or mobile apps (tablet and smartphone versions), and how are they designed, i.e. what are similarities and patterns followed by more platforms. Additionally, it will be reviewed if mobile and desktop version follows methodological and technological ground of e-learning and how do they (if they do) differ from each other.

REFERENCES

[1] R. Koper, Open source and open standards. In J.M. Spector, M.D. Merrill, J.V. Jeroen van Merrienboer & M.P. Driscoll, eds. Handbook of research on educational communications and technology. Mahwah, NJ: Lawrence Erlbaum, 2007, pp.355-65.

[2] A. Watters, Are MOOCs the Future of Online Learning? [Online] Available at: http://blogs.kqed.org/mindshift/2011/08/are-moocs-the-future-of-online-learning/ [Accessed 12/092014].

[3] W. Bernhard, N. Bittel, S. Van Der Vlies, M. Bettoni, and N. Roth, The MOOCs business model. Procedia - Social and Behavioral Sciences 106, 2013, pp.2931-2937.

[4] C. D. Grosso, Progress Report – Change Manager, 2003, [Online] Available at: http://www.flexiblelearning.net.au/leaders/fl_leaders/fll03/interim/del_grosso_cinthia_progreport.pdf [Accessed 20/08/2014].

[5] H. Crompton, Mobile Learning. New approach, new theory.. In Z.L. Berge & L.Y. Muilenburg, eds. Handbook of Mobile Learning. New York: Taylor & Francis, 2013, pp.47-57.

[6] S. C.-Y. Yuen, Mobile Learning with PDAs. Invited lectures at Beijing Normal University, Beijing; Southwest China Normal University, Chongqing, China; Beijing Language and Culture University, Beijing, China, 2004, [Online] Available at: nic.gdut.edu.cn/GCCCE2007/02.ppt [Accessed 15/02/2015].

[7] N.C. Ozuorcun, and F. Tabak, Is M-learning versus E-learning or are they supporting each other? Procedia - Social and Behavioral Sciences, 46, 2012, pp.299-305.

[8] P. Pollara, Mobile Learning in Higher Education: A Glimpse and a Comparison of Student and Faculty Readiness, Attitudes and Perception, 2011, [Online] Available at: http://etd.lsu.edu/docs/available/etd-11042011-105812/unrestricted/PollaraFinalDissertation.pdf [Accessed 14/04/2015].

[9] M. Ebner, C. Stickel, N. Scerbakov and A. Holzinger, A Study on the Compatibility of Ubiquitous Learning (u-Learning) Systems at University Level. U C. Stephanidis (Ur.), Universal Access in HCI, Part III, HCII 2009, LNCS 5616, Berlin: Springer-Verlag Berlin Heidelberg, 2009, pp. 34-43.

[10] G. Zhang, and Q. Jin, Research on Collaborative Service Solution in Ubiquitous Learning Environment. 6th International Conference on Parallel and Distributed Computing. Applications and Technologies (PDCAT 2005), 2005, pp. 804–806.

[11] M. Harrison, C. Bean, Designing Mobile Learning, 2012, [Online] Available at: https://www.kineo.com/m/0/designing-20mobile-20learning-p1.pdf [Accessed 14/04/2015].

[12] Towards Maturity Community Interest Company, Mobile learning in the Workplace, 2014, [Online] Available at: http://towardsmaturity.org/shop/wp-content/uploads/2014/06/In-Focus-2014-Report-Mobile-Learning-in-the-Workplace.pdf [Accessed 14/04/2015].

[13] S. Wexler, J. Brown, D. Metcalf, D. Rogers & E. Wagner, Mobile Learning, Santa Rosa, CA: The eLearning Guild, 2008.

[14] K. Swanson, Merrill Lynch: Bullish on Mobile Learning, 2008, [Online] Available at: http://www.clomedia.com/articles/merrill_lynch_bullish_on_mobile_learning [Accessed 14/04/2015].

[15] J. Haag, From eLearning to mLearning: The Effectiveness of Mobile Course Delivery, Proceedings of Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando FL, USA, 3-13.

[16] T. V. Paul, An Evaluation of the Effectiveness of E-Learning, Mobile Learning, and Instructor-Led Training in Organizational Training and Development. The Journal of Human Resource and Adult Learning, Vol. 10 No. 1, 2014., pp. 1-13.

[17] P. Goodyear, M. Asensio, C. Jones, V. Hodgson & C. Smith, Relationships between conceptions of learning approaches to study and students' judgements about the value of their experiences of networked learning. ALT-J, Vol. 11, No. 1, 2003, pp. 17-27. DOI: 10.1080/0968776030110103

[18] M. Driscoll, Hype versus reality in the boardroom: Why e-learning hasn't lived up to its initial projections for penetrating the corporate environment. In: S. Caliner & P. Shank (Eds.), The e-learning handbook. Past promises, present challenges, San Francisco: Wiley. 2008, pp. 29-55.

[19] A. Margaryan, M. Bianco and A. Littlejohn, Instructional quality of Massive Open Online Courses (MOOCs), Computers & Education, Vol. 80, Jan 2015, pp. 77–83. DOI:10.1016/j.compedu.2014.08.005

[20] D. El-Hmoudova, MOOCs Motivation and Communication in the Cyber Learning Environment. Procedia - Social and Behavioral Sciences 131, 2014, pp.29-34.

[21] D. Morrison, The Ultimate Student Guide to xMOOCs and cMOOCs, April 2013. Available at http://moocnewsandreviews.com/ultimate-guide-to-xmoocs-and-cmoocso/, [Accessed 20/02/2015].

[22] D. Shah, MOOCs in 2013: Breaking Down the Numbers, December 2013. Available at: https://www.edsurge.com/n/2013-12-22-moocs-in-2013-breaking-down-the-numbers, [Accessed 20/02/2015].

[23] D. Shah, MOOCs in 2014: Breaking Down the Numbers, December 2014. Available at: https://www.edsurge.com/n/2014-12-26-moocs-in-2014-breaking-down-the-numbers, [Accessed 20/02/2015].

[24] S. Ivec, Is Responsive Design Right for Your e-Learning Course? 2014, [Online] Available at: http://elearningindustry.com/responsive-design-right-e-learning-course [Accessed 12/092014].

[25] C. Shepherd, Are apps the future of e-learning? 2010, [Online] Available at: http://clive-shepherd.blogspot.com/2010/11/are-apps-future-of-e-learning.html [Accessed 12/092014].

[26] C. Severance, Teaching the World: Daphne Koller and Coursera. Computer , 45(8), 2012, pp. 8-9.

[27] J. Dron, and T. Anderson, Teaching Crowds: Learning and Social Media. Edmonton: AU Press, Athabasca University, 2014.

[28] R. H. Wild, K. A. Griggs, and T. Downing, A framework for e-learning as a tool for knowledge management. Industrial Management and Data Systems, 102(7), 2002, pp. 371-380.

[29] M. Ćukušić, and M. Jadrić, e-učenje Koncept i primjena. Zagreb: Školska knjiga, 2012.

[30] S. D. Johnson, and S. R. Aragon, An Instructional Strategy Framework for Online Learning Environments. U New Directions for Adult and Continuing Education, No. 100, Wiley Periodicals, Inc., 2003, pp. 31–43.

[31] N. Gillani, and R. Eynon, Communication patterns in massively open online courses. Internet and Higher Education, 23, 2014, pp. 18–26.

[32] Ž. Garača, Poslovni informacijski sustavi. Split: Ekonomski fakultet u Splitu, 2008.

[33] Y. J. Joo, K. Y Lim, and S. M. Kim, A Model for Predicting Learning Flow and Achievement in Corporate e-Learning. Educational Technology and Services, 15(1), 2012, pp. 313-25.

[34] T. Govindasamy, Successful implementation of e-Learning Pedagogical considerations. Internet and Higher Education, (4), 2002, pp. 287–99.

[35] A. T. Korucu and A. Alkan, Differences between m-learning (mobile learning) and e-learning, basic terminology and usage of m-learning in education. Procedia Social and Behavioral Sciences, 15, 2011, pp. 1925–1930.

[36] Elucidat, So - What is Responsive Design and Why is it Important for e-Learning? Available at: https://www.elucidat.com/blog/5-byod-and-responsive-learning-tips, [Accessed 20/02/2015].

Mining Association Rules in Learning Management Systems

M. Perušić Hrženjak, M. Matetić* and M. Brkić Bakarić*
*Department of Informatics, University of Rijeka, Rijeka, Croatia
mhperusic@gmail.com, majam@inf.uniri.hr, mbrkic@inf.uniri.hr

Abstract – Learning management systems collect huge amounts of data that can later be analysed. The University of Rijeka uses MudRi e-learning system, which is based on the Moodle open source software. This paper focuses on the Programming course, for which data over several years are available. The data can be interpreted and valuable knowledge can be obtained and used for improving the quality of lectures, as well as making the lectures more suitable for students based on the actions and material deemed the most popular. Since the MudRi database contains many facts that might affect each other (e.g. homework might affect the final grade), association rule mining, which discovers regularities in data, is the most suitable data mining method. Apriori algorithm for the discovery of association rules is used for finding connections between various actions and final grades. Many interesting rules and information are discovered, which lead to conclusions on actions that seem to be in relation with the course success.

I. INTRODUCTION

Many companies accumulate huge amounts of data from their daily operations. For instance, shops gather a lot of data about its customers based solely on products they buy. All the transactions are stored in databases and data warehouses and they are available for further analysis, which helps understanding customers' behaviour better, which in turn helps managing the store, its promotions etc. Similar application can be seen in education since it nowadays relies on e-learning systems which collect a lot of information about students and their actions.

This paper focuses on the data from the Programming course in the MudRi e-learning system, with the purpose of getting information which help better understanding students' behaviour, improving the quality of course and e-learning system, and establishing connection between students' usage of the system and achieved success.

There are many kinds of knowledge that can be found in data. This paper focuses on the discovery of association rules, as well as their explanation. The tool that we use is R, the free software environment for statistical computing and graphics.

The remaining sections are as follows:

Section 2: Lists some related work

Section 3: Gives a description of the dataset and explains the process of mining association rules

Section 4: Discusses the results of this analysis

Section 5: Gives the conclusion of this research

II. RELATED WORK

Educational data mining is a promising area of research, so there are many works in this field [1]. Although methods and goals vary, all these works are trying to improve education systems.

The authors in [2] use data mining algorithms to identify behaviour of failing students in order to warn students who behave in such a way prior to the final exam, as well as to help teachers with managing their class.

The research in [3] predicts students' final grades using classification algorithms based on features extracted from the data logged in the e-learning system, while the one in [4] uses data mining to help engineering students to select an appropriate major.

In [5], the authors propose a novel Critical Relative Support to deal with the criticality or significant level of least association rules on the examination result of computer science students from the University of Malaysia Terengganu. The authors in [6] use data mining techniques to try to connect students' behaviour and their final success.

The work in [7] presents a case study on the data from the subjects which are common at graduation and post-graduation level to predict the factors that explain student's success or failure by using association rules discovery.

Several case studies related to video lectures and their impact on the learning process, along with their advantages and disadvantages are presented in [8, 9, 10].

III. ASSOCIATION RULE MINING

The process of association rule mining could be divided into five most common phases: selection, pre-processing, transformation, data mining and the evaluation of rules [11].

A. Selection

The data used in this research come from the Programming course in the Moodle e-learning system called MudRi, and contain five years of continuous system usage in the form of transactions. Every action that students take is recorded, no matter if it is just a click on some resource or upload/download of some file, along with the date and time of the action. Students of the course are required to attend at least 70% of the lectures. Additional activities and course content files are at their disposal within the MudRi system. The two activities specially observed in this paper are demonstrator classes and video lectures, the latter being a new element in the course.

The actions that the system records are joined with students' final grades at the end of the selection process, since the main objective of the paper is to see whether there is a correlation between those two activities and students' final grades.

B. Pre-processing

The first thing that needs to be done on this set of transactions is the anonymisation of students' personal information. Everyone is given a unique ID to enable keeping track of their actions. Furthermore, as five years is a long span, the format of some fields might have changed, so it is crucial to make sure everything is consistent.

Another important pre-processing step is the use of taxonomy. Taxonomy makes it possible to define groups of items based on their shared characteristics [12]. In this case it is used to group points, so all the grade points between 80 and 100 become grade A, points between 70 and 80 become grade B, etc. Most of these point totals are transformed to its letter counterpart directly by the system, but some of them are wrongly aligned, possibly due to some missing values, and as such present a noise that needs to be cleaned. An example can be seen in Fig. 1 where numeric grade 71.33 matches grade F instead of B.

C. Data transformation and data mining

Since R needs special data formats to deal with data mining, our set of transactions needs to be transformed into a data frame which is a list of vectors of equal length.

Fig. 2 represents an excerpt of the dataset used in this research in the form of a data frame, with its fields and matching values shown.

Association rules could be defined as an expression of the form X -> Y, where X and Y are disjoint and non-empty item sets [13]. The interpretation of the expression is that when all the items from set X are in a transaction, the items from Y are in the transaction too.

There are many algorithms that could be used for mining association rules, and in this research the Apriori algorithm is used.

Apriori is a classic algorithm for association rule discovery. It can be used on any database that contains transactions, where for every item set it tries to find a subset with at least a minimum number of item sets. Apriori uses a bottom-up approach [14], so it adds one item at a time to frequent subsets (candidate generation step) and tests the new group against the data. When there are no further successful extensions found, the algorithm terminates, as seen in Fig. 3.

For counting of items, Apriori uses a breadth-first search and a tree structure [14]. It first generates item sets of length *k* from item sets of *k-1*, and then prunes the candidates with infrequent sub pattern. Although the size of item sets grows exponentially, efficient search is possible since it uses the downward-closure property of *support* [15] (anti-monotonicity [16]), which guarantees that for an infrequent item set, all its supersets must also be infrequent.

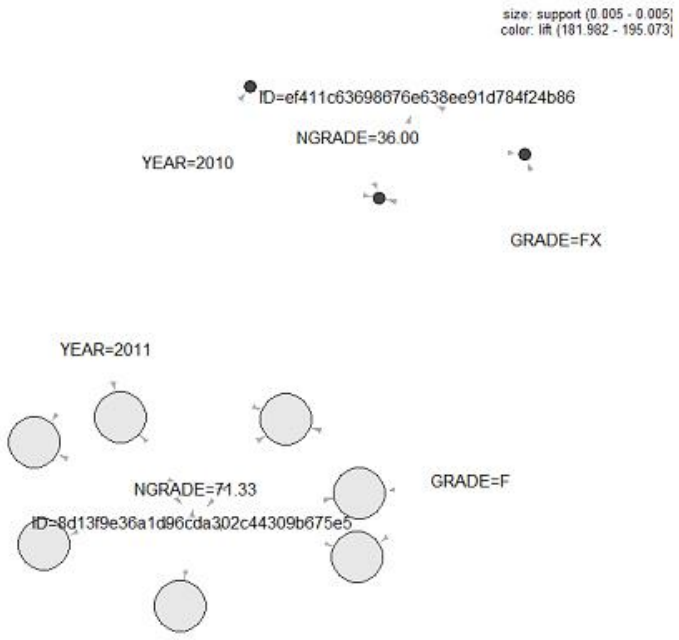


Figure 1. Noise present in the system

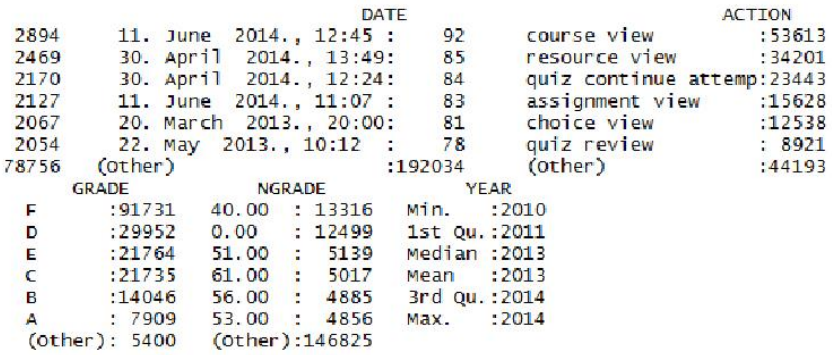


Figure 2. Data frame

$$
\begin{aligned}
&\textit{Apriori}\,(T, \varepsilon)\\
&\quad L_1 \leftarrow \{\textit{large 1-itemsets that appear in more than } \varepsilon \textit{ transactions}\}\\
&\quad k \leftarrow 2\\
&\quad \textit{while}\ L_{k-1} \neq \varnothing\\
&\quad C_k \leftarrow_{\textit{Generate}} (L_{k-1})\\
&\quad \textit{for transactions}\ t \in T\\
&\quad C_t \leftarrow_{\textit{Subset}} (C_k, t)\\
&\quad \textit{for candidates}\ c \in C_t\\
&\quad \operatorname{count}[c] \leftarrow \operatorname{count}[c] + 1\\
&\quad L_k \leftarrow \{c \in C_k |\ \operatorname{count}[c] \geq \varepsilon\}\\
&\quad k \leftarrow k + 1\\
&\quad \textit{return}\ \bigcup_k L_k
\end{aligned}
$$

Figure 3. Apriori algorithm [15]

D. *Rules validation*

For rules validation, a few metrics of interest are used.

- Support

Support is first introduced in [15]. It is defined as:

$$\text{supp}(Z) = P(Z) \tag{1}$$

Support is defined on a dataset and gives a proportion of the number of rules that contain *Z*. It is used as a measure of importance of a dataset. As it uses the count of transactions, it is often called frequency.

Support is the main measure of significance (importance) of an item set. An item set is called large or frequent only if it has *support* greater than the set minimum.

The main feature of *support* is the downward-closure property or anti-monotonicity. It means that all subsets of a frequent set are also frequent, while all supersets of an infrequent set are also infrequent. That feature is used for tree pruning in level-wise algorithms, such as the Apriori algorithm.

The main weakness of *support* is the rare item problem. In some systems, due to uneven distributions, many transactions that could be significant get pruned away because of low *support*. However, it is also possible that transactions with low *support* happen simply by chance.

- Confidence

Like *support*, it is first introduced in [15]. Its definition is:

$$\text{conf}(X \rightarrow Y) = \text{supp}(X \rightarrow Y)/\text{supp}(X) =$$

$$P(X \text{ and } Y)/P(X) = P(Y \mid X) \tag{2}$$

That means that *confidence* is the probability of seeing the consequent of a rule under the condition that the transactions also contain the antecedent. *Confidence* is directed and gives different values for the rules X -> Y and Y -> X [17]. It does not have the downward-closure property.

Confidence, like *support*, is developed by Agrawal et al. (the so-called support-confidence framework). It is used after pruning of the search space for producing rules from frequent item sets which satisfy the minimum confidence that is set.

The biggest problem with *confidence* is its sensitivity towards the frequency of the consequent, which means that consequents with higher *support* will automatically produce higher *confidence* values.

- Lift (Interest)

Introduced in [18], its definition is:

$$\text{lift}(X \text{ -> } Y) = \text{lift}(Y \text{ -> } X) = \text{conf}(X \text{ -> } Y)/\text{supp}(Y) =$$

$$\text{conf}(Y \text{ -> } X)/\text{supp}(X) = P(X \text{ and } Y)/(P(X)P(Y)) \tag{2}$$

Lift measures how many times X and Y occur together more often than expected if they were statistically independent [18].

As *confidence*, *lift* does not have the downward-closure property, but it does not suffer from the rare item problem. However, in small databases it is susceptible to noise. Rare item sets with low counts (low probability) which occur together only few times (or only once) completely by chance can produce enormous *lift* values [17].

IV. Experimental Results

Unfortunately, for every useful piece of information obtained, there is a lot more trivial information being generated [19]. General rules that could be mined from our data do not have great value for this research, since there are many unique values which produce specific rules that describe specific situations which are based either on student ids or academic years. However, after eliminating the data that cause such rules (i.e. student ID), and defining what we want to mine, associations discovered become a little more interesting.

The subject of our research are student demonstrator classes, and the visualization of data in Fig. 4 shows us that there are some rules that stand out with their high *confidence*.

Upon further inspection, the rules with the highest *confidence* about student demonstrator classes show that most of the students interested in these classes try to find information on the course forum. The majority of related transactions happened in year 2012. Rules that support this fact can be seen in Fig. 5.

Another element we find interesting for research are video lectures. Programming 2 video lectures were recorded out of the class. In total 13 short 4-15 minutes lectures were captured covering about 40% of the course topics. The video lecture activities are represented in the analysis by the total number of clicks on each particular resource. Since watching video lectures does not have a direct influence on the total course score but serves solely as a supportive activity, i.e. students do not earn points by watching video lectures, our hypothesis is that watching video lectures nevertheless impacts the learning process. Video lectures are of interest since this activity is newly introduced and since it is based on modern technology which increases learning motivation. The latter is particularly important for introductory programming courses. Mining on transactions that contain video lectures produces several interesting rules, as seen in Fig. 6.

Since this is a fairly new element in the big dataset, it is not plausible to rely only on the frequency of rules, which is why *support* and *lift* have greater impact on making conclusions.

Ordered by *support*, the most important rule is that students who watch Video lecture 1 are also students graded E, which is concluded with the values of *confidence* and *lift* 0.284 and 1.268, respectively. However, the same lecture has grades C and D as consequences, as can be seen in Table 1. A similar situation happens with other lectures. Although the rules can be read in this form, it is easier to find the real knowledge in them by visualisation, in our case with R package ArulesViz, as in Fig. 6.

The visualisation in Fig. 6 makes it evident that some of the students graded FX watch all the videos, but most of them watch only the first one, as indicated by the size of the circle which represents *support* of the rule. These facts are somewhat more connected than it would be expected if they were independent, as indicated by the darker shade which represents *lift*. That can be explained by the fact that many students open the first video out of curiosity, but then decide that they dislike learning from videos.

Grade E is quite similar to FX in a way, since most of the students watch the first lecture, but only few of them watch other lectures, probably those that they have some special interest in.

C and D students watch most of the lectures, but it is interesting that there are no C students that watch the last two lectures.

Grade B students watch most of the lectures, but they are especially interested in lecture number 5, while A students watch only lectures 3 and 4. It is interesting that those two videos are absent from grade B, which makes it the biggest difference between the two grades. The colour of circles in grade A indicates that there is a strong connection between lectures 3 and 4 and the grade, with *lift* values of 1.6 and 1.7 being some of the biggest *lift* values among the rules.

The most interesting rule in this research is actually the one that is absent. Not even one student that watched video lectures got grade F, which means that those students did not fail the class.

It can be concluded that video lectures really raise the quality of the class and that they show connection to the final grades.

Table I. Part of Video Lecture Rules

Rule left-hand side	Rule right-hand side	support	confidence	lift
{SUBJECT = Video lekcija 1 - Sortiranje umetanjem}	{GRADE=E}	0.050632911	0.2834646	1.2675680
{SUBJECT = Video lekcija 1 - Sortiranje umetanjem}	{GRADE=C}	0.030942335	0.1732283	0.8378596
{SUBJECT = Video lekcija 1 - Sortiranje umetanjem}	{GRADE=D}	0.023909986	0.1338583	0.6846995

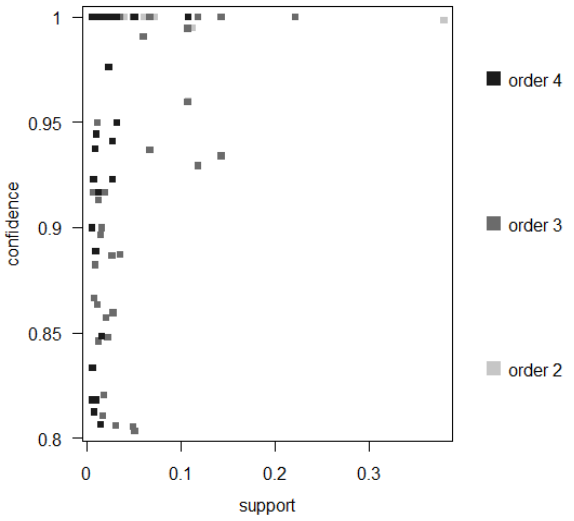


Figure 4. Visualisation of rules

```
>  rules.sorted2 <- sort(rulesd, by="support")
>  inspect(rules.sorted2)
     lhs                                                    rhs
1    {SUBJECT=Demonstrature       }                     => {ACTION=forum view discussion}
2    {YEAR=2012}                                        => {SUBJECT=Demonstrature       }
3    {YEAR=2012}                                        => {ACTION=forum view discussion}
4    {SUBJECT=Demonstrature       ,
      YEAR=2012}                                        => {ACTION=forum view discussion}
5    {ACTION=forum view discussion,
      YEAR=2012}                                        => {SUBJECT=Demonstrature       }
```

Figure 5. Association rules mined

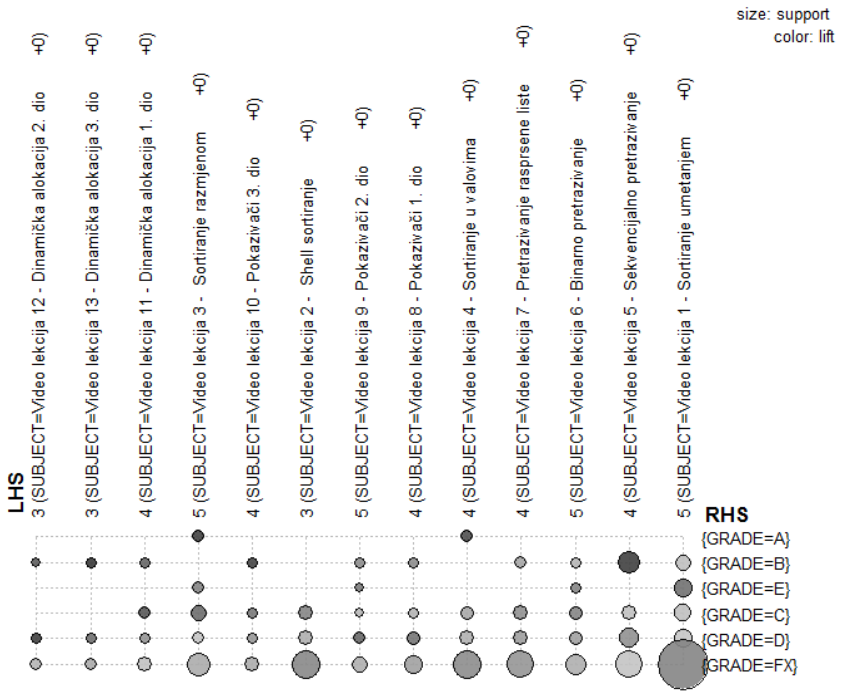


Figure 6. Visualisation of rules regarding video lectures

V. CONCLUSION

Mining association rules is a very valuable tool in data mining, since rules made from patterns might be strange or unexpected at first, but are realistic picture of the system. Although the most famous usage of association rule mining is "market basket analysis", it is valuable for educational purposes as well.

In this work we use students' data from the MudRi e-learning system database, and apply the data mining technique of association rule mining in order to discover knowledge. We find some rules that support the fact that students have better success in the course when using video materials, and we find out which lessons seem to have greater connection to the final grades. Discovered knowledge could be used in raising the quality of the course and checking whether video lectures can be used for improving the overall success of students.

ACKNOWLEDGMENT

This research has been supported under the Grant No. 13.13.1.3.03 of the University of Rijeka.

REFERENCES

[1] C. Romero and S. Ventura, "Data mining in education," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3(1), pp. 12–27, 2013.

[2] A. Merceron and K. Yacef, "Educational Data Mining: a Case Study," In Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED, 2005, pp. 467–474.

[3] B. Minaei-Bidgoli, D. Kashy, G. Kortemeyer, and W. Punch, "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System," In Proceedings of the 33rd ASEE/IEEE Conference of Frontiers in Education, 2003, pp. 13–18.

[4] K. Waiyamai, Improving Quality of Graduate Students by Data Mining. Thailand: Department of Computer Engineering, Faculty of Engineering, Kasetsart University, 2003.

[5] Z. Abdullaha, T. Herawani, N. Ahmadb, and M. Mat Derisc, "Mining significant association rules from educational data using critical relative support approach", Procedia-Social and Behavioral Sciences, vol. 28, pp. 97–101, 2011.

[6] El-Halees, A. Mining Students Data to Analyze Learning Behaviour: A Case Study. Palestine: Department of Computer Science, Islamic University of Gaza, 2008.

[7] V. Kumar and A. Chadha, "Mining Association Rules in Student's Assessment Data," IJCSI International Journal of Computer Science Issues, vol. 9(3), pp. 211–216, September 2012.

[8] E. D. Manley and T. M. Urness, "Video-based instruction for introductory computer programming," Journal of Computing Sciences in Colleges, vol. 29(5), pp. 221–227, 2014.

[9] J. H. Sharp and L.A. Schultz, "An exploratory study of the use of video as an instructional tool in an introductory C# programming course," Information Systems Education Journal, vol. 11(6), pp. 33–39, 2013.

[10] A. Matthíasdóttir, "How to teach programming languages to novice students? Lecturing or not," in Proceedings of the International Conference on Computer Systems and Technologies-CompSysTech, vol. 6, 2006, pp. 15–16.

[11] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI magazine, vol. 17(3), pp. 37–54, 2006.

[12] R. Agrawal and R. Srikant, Mining Generalized Association Rules. CA: IBM Research Division, 1995.

[13] H. Toivonen, "Sampling Large Databases for Association Rules", in Proceedings of the 22th International Conference on Very Large Data Bases (VLDB), vol. 96, 1996, pp. 134–145.

[14] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Procedings of the 20th International Conference on Very Large Data Bases (VLDB), vol. 1215, 1994, pp. 487–499.

[15] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, vol. 22(2), 1993, pp. 207–2016.

[16] J. Pei, J. Han, and L. V. S. Lakshmanan, "Mining frequent itemsets with convertible constraints," in Proceedings of the 17th International Conference on Data Engineering, 2001, pp. 433–442.

[17] M. Hahsler, K. Hornik, and T. Reutterer, "Implications of probabilistic data modeling for mining association rules," in From Data and Information Analysis to Knowledge Engineering, M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, Eds. Springer Berlin Heidelberg, 2006, pp. 598–605.

[18] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in Proceedings ACM SIGMOD International Conference on Management of Data, vol. 6(2), 1997, pp. 255–264.

[19] S. D. Bay and M. J. Pazzani, "Discovering and Describing Category Differences: What makes a discovered difference insightful?," Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society, 2000, pp. 603–608.

Integrated Solution for On-line Environments Users' Learning and Support

V. Riabtsev*, K. Sinitsa** and O. Voychenko**
* The Interregional Academy of Personnel Management, Kyiv, Ukraine
** IRTC ITS, Kyiv, Ukraine
alexp_p@mail.ru

Abstract - With increasing complexity of modern software, including on-line environments, it is important to organize user training and support, a significant part of which could be implemented through the ICT-based learning activities.

During the learning process a learner goes through several stages. Starting from the understanding the basics of the given software, through training course the learner finally comes to a practical work, which still requires learner to refer to a user manual or contact the customer support to resolve some complicated issues. Within a traditional approach the content blocks used at every stage of the learning process are not interconnected, despite they are related to the same subject domain.

The approach offered in the work is aimed to increase the results of such learning process, by establishing an effective management of all available learning content within particular subject domain. The approach is based on integration of four main blocks of the content: one used to build the learning course, content from its built-in help system, user guides and information resources used by the customer support team.

The use of this approach allows keeping the balance between sets of the content as well as optimize the learning process at all stages.

I. INTRODUCTION

Recently growing numbers of distributed multi-user online systems are brought to life. With their increasing complexity and extending functional features we witness the escalation of demands to the qualifications and skills of their users. The users should know the whole spectrum of the system's functionality, be able to solve complex, sometimes ill-defined, problems, cope with time pressure and permanent changes. As an example, one can consider the requirements to the operators of the modern on-line call center, which need to know supporting computer system in detail to be able to quickly resolve client's issues, be aware of current products, services, and changes, customer bases, corporate policies and procedures [1].

Approaches and methods of personnel training for multi-user online systems do not follow software lifecycle trends quickly enough and still mostly rely on traditional professional training, use of manuals and distribution of information about recent changes through communication channels (corporate email). As a result, the users are insufficiently prepared for their professional activity. Moreover, a detailed analysis of operator's activity demonstrated the need for careful planning of learning and training, which should penetrate business and work processes to ensure efficiency of the human-computer system as a whole. The design of human-centered systems should envisage potential needs of their users in problem solving and interaction support to provide individual and personalized assistance and help. Ideally, these systems should adaptively be able to help situated problem solving or assist the user to move to the next action of activity where needed. However, a detailed analysis of users activity needed to organize intelligent support, such as described in [2, 3] is so far rather laborious so there is a need for some efficient solution for existing systems support.

II. PROBLEM STATEMENT

Observation and analysis of user activities during their work with some online systems, in-depth study of help requests from customers who successfully finished their training made it clear that training on specific system functions and distribution of reference materials is sometimes insufficient for practical problem solving. Significant amount of requests to the system's help desk was caused by the lack of understanding how the task may be completed with the system rather than related to technical failures.

To simplify on-line system management and support by the staff we focused on better automatic assistance for the users. In case of limited resources the purpose was to enhance self-training and performance support of the users by better organization and structuring of training and reference content. Another idea was to address the most critical issues first, at the same time collecting information for further enhancement. This course of action corresponds to the process quality approach, focusing on major issues and thus enhancing quality without delay for detailed analysis and reengineering.

III. ANALYSIS

In the first phase of the study we have analyzed the main components of the user training and support system. As a result, it was possible to identify four basic sources of information which constitute knowledge about system functionality and its use:

- Distance training course that provides general information about the system and its functions. Successful completion of it is a prerequisite for an operator to obtain the access to the system.
- User Guide – a thick book (or its electronic version) which describes all functions of the system. It is assumed that the reference information from the Guide is used as an additional information source by the staff that has already completed a training distance course and is working with the system.
- inline help – small reference blocks, built directly into the system interface. It is usually implemented as an icon with a question mark and simply by clicking or hovering the mouse on it user can get a concise reference about the function of this interface element. Such reference blocks are small, but there are a lot of them and they are present in almost all sections of the system.
- Customer support service, contacted by users in case of problems. Customer Support is contacted not only when there are technical problems, but also in the case a user does not know how to perform a particular operation. In this situation, the support service should solve this problem and explain to the user how he or she may achieve the desired result.

It should be noted that in this study we have not considered the mechanisms of informal learning, when people learn from each other and communicate with each other to obtain an advice or hint, which could be useful for performing their professional tasks.

Undoubtedly, this learning mechanism also exists and can be quite effective, but it has been excluded from the study because, in general, we could not control the communication of the system users. Although it is possible to use social services to support a community of the users [4, 5], we could not control how people communicate with each other, for example, outside the office and we could not influence this process.

Therefore, in this study, we have started from the assumption that each user was working alone.

General scheme of the user activity is as follows:

- The user passes the distance training course successfully and after that gains access to the system.

In case of problems user's activity goes the following way:

- To solve a problem, the user can click on inline help elements;
- If there are no clues, user refers to the User Guide;
- If still there is no solution, a call or an email to customer support service follows.

General scheme is provided on figure 1.

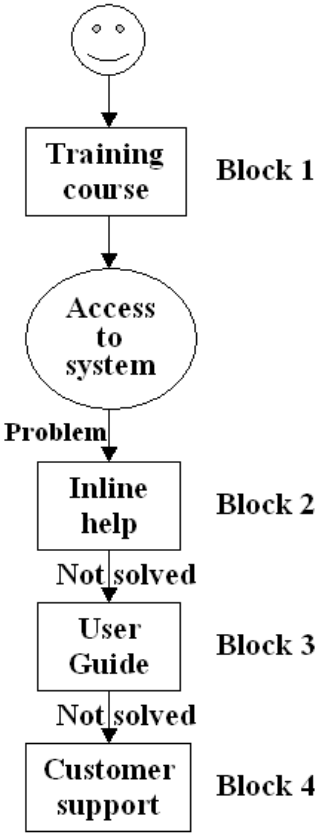


Figure 1. General scheme of the user activity in case of problem.

Further study of the reference and training information brought the following basic questions:

- What are the information sources for the four blocks mentioned above?
- Who manages this information?
- On what procedure and set of criteria the information is updated?

After analysis, we concluded that these four blocks are not directly linked. They are produced and managed independently. Inline help is produced by the developers. Tutors of distance training course get basic reference materials and then build the course and make changes on their own, sometimes with no one consulting them on the system features. Technical customer support service operates with its own database. The User Guide is written by technical experts who are not always directly linked with the developers, training course tutors and customer support service.

However, we managed to identify a single source of all information initially used in all four blocks: the system project documentation. There is no other information at the initial stage of user training and support system creation. On the basis of this project documentation all four blocks are subsequently built.

IV. SOLUTION

The proposed solution, aimed to increase overall efficiency of users' training and reference system, consists of several stages.

The first stage - the initial processing of the raw project information. The totality of the information that may be useful should be extracted from the project documentation and cut into some elementary units (chunks) that may be used to describe a single function of the system, system interface element, a system task, a problem situations, a solution of a typical problem, etc.

Further, these chunks are used to compose the content of the main four blocks (distance training course, User Guide, inline help and Customer support service, as shown on the figure 1). In this way, they, first, do not contradict each other in the details, and secondly, all the information is formulated in a uniform manner, stylistically and logically. This improves the user's perception of the main four information blocks used further for users' training and support. The first stage is presented on figure 2.

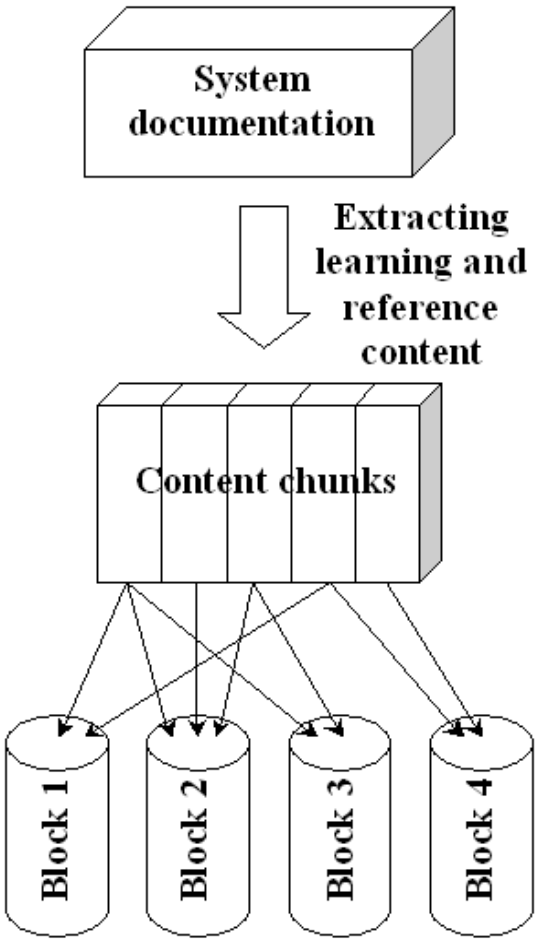


Figure 2. The first stage.

The second stage is the collection of statistical data (logs) of the users' activities. In order to efficiently collect data about users' activity in the system, all four blocks have to be online. In particular, a User Guide should be implemented as a web resource that is not downloadable to a local drive. All users' queries to each of the four blocks should be logged. Every request to customer support service, every session of User Guide use and every click on an element of inline help should be logged. All information recorded should be stored in separate databases for all four blocks.

The third stage is the analysis and processing of the collected data. Information stored in each database passes statistical processing and analytical reports are submitted for analysis. Then it is possible to locate situations where users have problems most often, find what functions caused the greatest number of questions and what interface elements are problematic for use.

As a result of the analysis it is possible to identify common problem situations that arise during users' activity from the entry into distance training course to the routine work with the system.

The fourth stage is the modification of learning and reference information. After analyzing the information gathered a set of decisions should be made about entire information update. As a result, content chunks are updated. At this stage, not only the content of four main blocks of information is updated and modified, but also their structure. For example, the number of chapters in the User Guide may be changed, as well as the structure of distance training course may be significantly updated, not only the particular course elements.

In the inline help new items may appear or, on the contrary, some unclaimed ones are excluded. The database of problem situations operated by the customer support service is also a subject of modifications. Stages 2-4 are presented at figure 3.

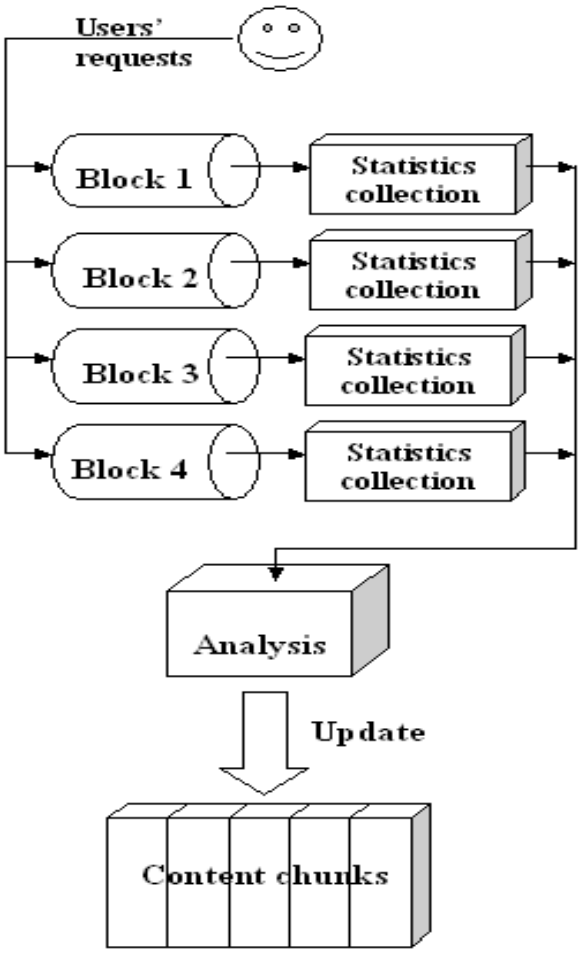


Figure 3. Stages 2-4.

The fifth stage is a deployment of updated learning and reference components within existing information system.

At the last stage the updated blocks should be included back into the system and run. The described approach can be effective for small information systems or for the releases of the new versions of already used information system.

V. CONCLUSION

The proposed approach initiates ongoing user support enhancement process. Monitoring of user activities can provide valuable information about potential causes of user's mistakes and difficulties and thus indicate how the issue could be addressed. For multiuser online systems with hundreds of users the cycles of updates should take place on a regular basis, without delay. All registered issues should be analyzed, processed according to the suggested procedure and resolved through changes in an online user guide and inline help. If updates of both learning and reference content are managed by the same team, the changes are easily coordinated and could be done whenever necessary without synchronizing them with updates in the system itself. In other words, it enables permanent enhancement of learning and performance support for the user.

It should be mentioned that information regarding user actions and problems encountered while working with the system is an excellent source of information for developers who could use it for making system interface more efficient and friendly. It was also recognized that many users do not feel comfortable switching between several windows but need to have user guide at hand when working with the system.

Therefore, it was decided to prepare a trial version of user guide for mobile devices to assist them.

REFERENCES

[1] K. Synytsya, and H. von Brevern, "Information processing and holistic learning and training in an organization: a systemic-structural activity theoretical approach," in Human-Computer Interaction and Operators Performance: Optimization of work design, Activity Theory approach, Taylor & Francis, 2010, pp. 385-409.

[2] H. von Brevern, and K. Synytsya, "A systemic activity based approach for holistic learning & training systems," Educational Technology & Society, vol.9 (3), 2006, pp. 100-111.

[3] H. von Brevern, and K. Synytsya, "Systemic-structural theory of activity: a model for holistic learning technology systems," in Proc. Int. Conf. on Advanced Learning Technologies ICALT-2005, Kaohsiung, Taiwan R.O.C., 2005, IEEE Computer Society Press, pp. 745-749.

[4] R. Van Noorden, "Online collaboration: scientists and the social network", Nature, vol. 512, issue 7513, http://www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711

[5] O. Voychenko, and K. Synytsya, "Knowledge sharing via Web 2.0 for diverse student groups in distance learning", Global Engineering Education Conference (EDUCON), IEEE, 2011, pp. 933-936.

Istraživanje
Mišljenje učenika, roditelja, učitelja i uprave škole o primjeni e-Dnevnika, unutar III. osnovne škole Čakovec

Nataša Boj*
* III. osnovna škola Čakovec, Čakovec, Republika Hrvatska
natasha.boj@gmail.com

Sažetak – u III. osnovnoj školi Čakovec je provedeno istraživanje o primjeni e-Dnevnika na inicijativu administratora. Ispitanici su bili učenici, roditelji učenika, učitelji i uprava škole s ciljem da se sazna kakva su njihova razmišljanja i zaključci o primjeni istog. Očekivano, većina ispitanika je dala pozitivne odgovore i pohvale aplikaciji. Bilo je nekoliko primjedbi i prijedloga od strane svih ispitanika. Jedan od ciljeva ovog istraživanja bio je da se prezentacijom rezultata na različitim skupovima i objavom u medijima potaknu svi oni koji ne koriste e-Dnevnik da započnu s njegovom primjenom, tim više ukoliko raspolažu potrebnom opremom i imaju brzi pristup Internetu u cijeloj školi.

I. UVOD

E-dnevnik je u III. OŠ Čakovec u primjeni od šk. god. 2013./2014.

S obzirom da je proteklo određeno vrijeme i svi su dionici odgojno-obrazovnog procesa imali priliku koristiti se e-Dnevnikom, provedeno je istraživanje o njihovim iskustvima u radu s istim kako bi se dobile informacije o tome jesu li korisnici zadovoljni u potpunosti ili bi se eventualno još nešto moglo poboljšati u samoj aplikaciji ili načinu rada samih učitelja a na zadovoljstvo učitelja s jedne strane te učenika i roditelja s druge strane.

Istraživanjem je bilo obuhvaćeno 24 učitelja, 153 roditelja i 130 učenika i 3 člana uprave škole (ravnateljica, psihologinja, tajnica).

Ankete su napravljene u Google Disk okruženju.

U glavnom dijelu će biti prikazani rezultati istraživanja.

II. GLAVNI DIO

A. Rezultati istraživanja vezanog za učitelje

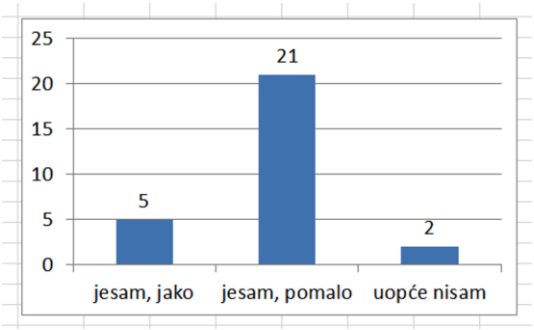
Anketna pitanja:

- „Ti si - učitelj RN, učitelj PN, učitelj PN i razrednik?“
- „Godine radnog staža?“
- „Jesi li nakon najave uvođenja e-dnevnika a prije edukacije osjećao strah i nelagodu?“
- „Jesi li nakon edukacije osjećao strah i nelagodu?“
- „Koji način vođenja dnevnika/imenika ti se više sviđa?„
- „Prednost daješ e-Dnevniku zbog: pristupa svim podacima od bilo kud i bilo kad, zato što vidim samo svoj predmet, zato što sve važne podatke imam na jednom mjestu, zato što mogu biti u e-komunikaciji s roditeljima“
- „Je li slučajni odabir učenika u redu?“
- „Jesi li u e-komunikaciji s roditeljima?“
- „ Koristiš li mogućnost upisivanja bilježaka bez obzira na ocjene?“
- „Kada upisuješ sat i izostanke i ocjene?“
- „Jesi li zadovoljan opremom koja ti je na raspolaganju?“
- „Koju ocjenu daješ adminu e-Dnevnika?“
- „Poruka za kraj?“

ANALIZA ODGOVORA UČITELJA

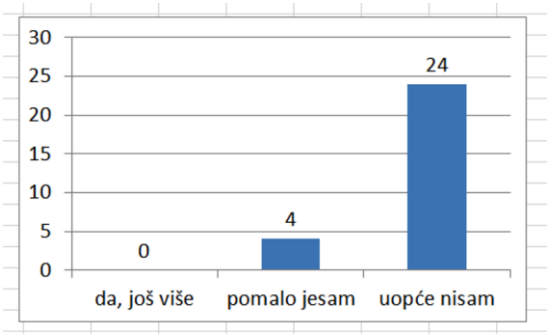
- Status učitelja: 8 učitelja razredne nastave, 8 učitelja predmetne nastave koji su ujedno i razrednici te 8 učitelja predmetne nastave koji nisu razrednici.
- Godine radnog staža: različite, od 1 do 31 i više
- Većina učitelja je nakon najave uvođenja e-Dnevnika osjećala pomalo strah i nelagodu, ali nije bilo otpora.
- Nakon provedene edukacije u školi samo manji broj njih i dalje osjeća strah i nelagodu.
- Pri kraju školske godine većina njih se izjasnila kako im se više sviđa elektronski način vođenja evidencije (22).

Jesi li nakon najave uvođenja e-Dnevnika a prije edukacije osjećao/la strah i nelagodu?



Slika 1. Osjećaj straha i nelagode kod učitelja, bez otpora

Jesi li nakon edukacije osjećao/la strah i nelagodu? - učitelji



Slika 2. Osjećaj straha i nelagode kod učitelja nestaje nakon edukacije

- Najveću prednost učitelji vide u tome što imaju pristup svim podacima u bilo koje vrijeme i s bilo kojeg mjesta te što nisu ograničeni prostorom gdje mogu upisivati bilješke. Nadalje, navode kako je sve uredno i vrlo pregledno.
- Zadovoljni su mogućnošću slučajnog odabira učenika.
- Koriste mogućnost e-komunikacije sa roditeljima.
- Redovito uz ocjene upisuju bilješke ali i samo bilješke.
- Većina njih upisuje nastavne sata na početku sata, no, ima i onih koji to rade pri kraju sata, u nekoj pauzi ili kod kuće.
- Većina njih ocjene upisuje na satu. Većina je u potpunosti zadovoljna opremom koja im stoji na raspolaganju.
- Svi anketirani učitelji su zadovoljni načinom na koji je bila provedena edukacija u školi kao i sa suradnjom administratora e-Dnevnika.
- Učitelji imaju primjedbe što nemaju više vremena za korekcije grešaka već za sve moraju kontaktirati administratora.
- Nekim učiteljima nedostaje neposredni kontakt s roditeljima kojeg je prije uvođenja e-Dnevnika bilo više, a nekima je baš suprotno.

B. Rezultati istraživanja vezanog za roditelje/staratelje

Anketna pitanja:

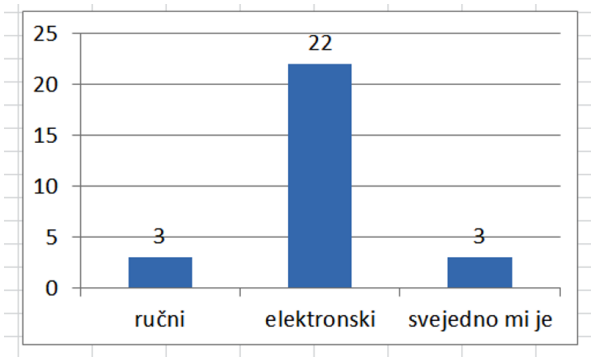
- „Jeste li upoznati s primjenom e-Dnevnika u školi?“
- „Možete li sami do web adrese e-Dnevnika?“
- „Kojem dnevniku dajete prednost – papirnatom ili elektronskom?“
- „Zašto elektronskom?“ – zato što: a)“imam pristup svim podacima bilo kad i bilo gdje“, b) „imam apsolutnu kontrolu nad djetetom“, c) „mogu biti u e komunikaciji s razrednikom“
- „Komunicirate li s razrednikom preko e-maila?“
- „Dolazite li i dalje na primanja zbog ocjena?“
- „Provjeravate li ocjene prije nego li dijete dođe iz škole?“
- Jeste li zadovoljni evidencijama koje vode predmetni učitelji?“
- „Poruka za kraj?“

U istraživanje je bilo uključeno 153 roditelja, od toga 67 roditelja učenika razredne nastave i 86 roditelja učenika predmetne nastave.
54% roditelja ima srednju školsku spremu, 13% višu i 33% visoku školsku spremu.

ANALIZA ODGOVORA RODITELJA

- Svi roditelji su upoznati s primjenom e Dnevnika u školi.
- Većina roditelja sama pristupa aplikaciji, samo manjem broju njih pomognu njihova djeca.
- 89% roditelja daje prednost e-dnevniku, 6% ručnom načinu a 5% njih se izjasnilo kako im je svejedno.
- Polovica roditelja je u povremenoj komunikaciji s razrednikom/com putem e-pošte.
- Većina roditelja sada manje dolazi na primanja roditelja, ili nemaju potrebu ili idu na primanje samo kada nastane neki problem
- Polovina roditelja nema potrebu provjeravati ocjene prije nego li dijete dođe iz škole, dok 16 % njih to ipak radi redovito.
- Većina roditelja je zadovoljna načinom na koji učitelji vode evidencije koje se odnose na njihovu djecu u e-Dnevniku, manji postotak njih je zadovoljan samo s pojedinim učiteljima.

Kojem načinu vođenja evidencije daješ prednost?



Slika 3. Odgovori učitelja

Pozitivne primjedbe roditelja:

- „Pohvala e-Dnevniku općenito, jednostavan pristup ocjenama bilo kad i bilo gdje.“
- „E-dnevnik – super stvar, oduševljeni smo e-dnevnikom, svaka čast na uvođenju e-Dnevnika.“
- „Zadovoljni smo sa svime, sve pohvale, ne može bolje.“
- „Pohvala razrednici na upisanim komentarima i objašnjenjima kod svake ocjene.“
- „Vrlo praktično, brzo i efikasno.“
- „Sve je vrlo pregledno i razumljivo.“

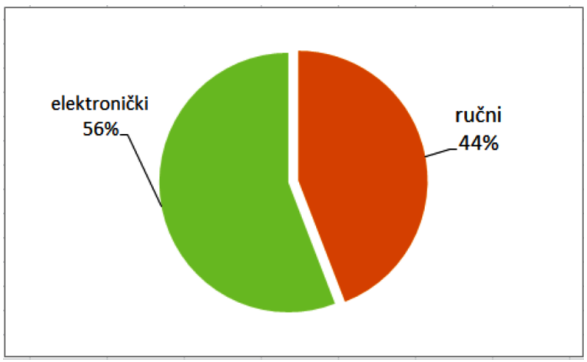
Negativne primjedbe roditelja:

- „Pojedini učitelji upišu ocjene tek nakon nekoliko dana.“
- „Otuđenost od razrednika, škole.“
- „Bilo bi dobro da se neke općenite obavijesti vide u e-dnevniku a ne samo u e-mailu.“

Neutralne primjedbe roditelja:

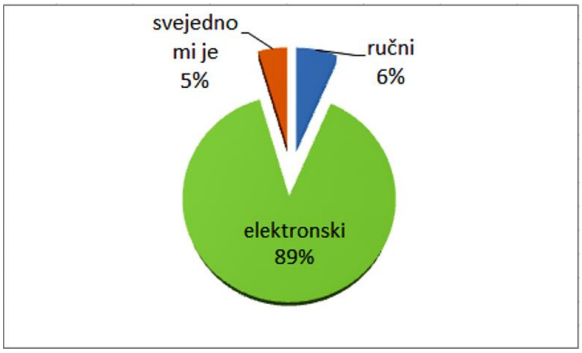
- „Podržavam e-dnevnik ali isto tako volim doći u školu i čuti se s učiteljem.“
- „Nadamo se da ćemo uskoro imati i e-udžbenike.“

Kojem načinu vođenja evidencije daješ prednost?



Slika 4. Odgovori učenika

Kojem načinu vođenja evidencije dajete prednost?



Slika 5. Odgovori roditelja

C. Rezultati istraživanja vezanog za učenike

Anketna pitanja:

- „Znaš li što je e-Dnevnik?“
- „Znaš li pristupne podatke?“
- „Pratiš li svoje ocjene i bilješke i koliko često?“
- „Kojem načinu vođenja daješ prednost – ručnom ili elektronskom? „
- „Jesi li zadovoljan što tvoji roditelji imaju dnevni pristup ocjenama, bilješkama i izostancima? „
- „Koju ocjenu bi dao svojim učiteljima o vezi načina korištenja e-Dnevnika? „
- „Što misliš o mogućnosti slučajnog odabira učenika?“
- „Poruka za kraj“

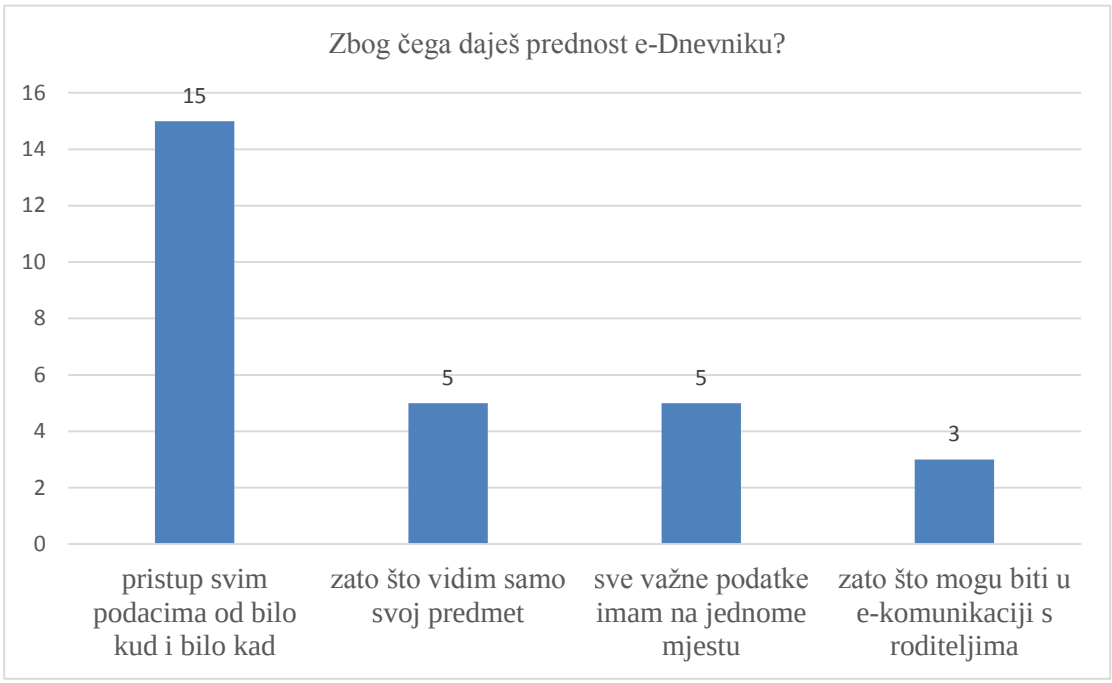
U istraživanje je bilo uključeno 130 učenika predmetne nastave. Svi su upoznati s e-Dnevnikom i većina zna napamet pristupne podatke za ulaz.

ANALIZA ODGOVORA UČENIKA

- 46 % učenika redovito prati ocjene i bilješke
- 57% učenika daje prednost e- načinu vođenja evidencije, 19 % učenika ručnom vođenju.
- Razlog - učenici nisu zadovoljni što roditelji imaju svakodnevni pristup ocjenama, bilješkama i izostancima.
- 44% učenika je dalo učiteljima ocjenu vrlo dobar za način na koji koriste e-dnevnik, a 5% njih ocjenu nedovoljan.

Primjedbe učenika:

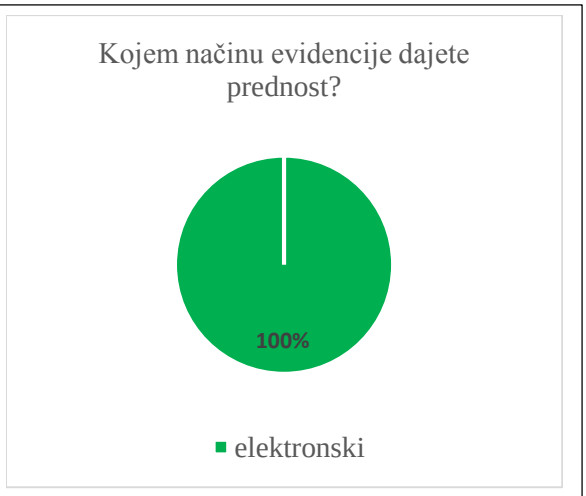
- „Super stvar, trebale bi ga uvesti sve škole.“
- „Dobro je što učitelji ne vide ocjene drugih predmeta, dobar dizajn.“
- „e-Dnevnik je fer prema učenicima, super je jer mogu vidjeti sve ocjene kod kuće.“
- „Slučajni odabir mi se ne sviđa.“
- „Slučajni odabir je pravedan.“
- „Nije dobro to što roditelji mogu vidjeti dnevno sve ocjene.“



Slika 6. Odgovori učitelja

D. Rezultati istraživanja vezanog za upravu škole

U upravu škole ulaze ravnateljica, tajnica, psihologinja i knjižničar kao stručni suradnici te računovotkinja.
U svom radu e-Dnevnikom se služe ravnateljica, tajnica i psihologinja.
Na osnovu rezultata istraživanja može se zaključiti kako oni u cijelosti daju prednost e-Dnevniku u odnosu na papirnati.



Slika 7. Odgovori uprave škole

III. ZAKLJUČAK

Na osnovu rezultata provedenog istraživanja može se sa sigurnošću zaključiti kako su primjenom e–Dnevnika svi zadovoljni, čak oduševljeni, što se može vidjeti iz odgovora ispitanika i primjedbi koje su naveli na kraju anketa.

Najviše su zadovoljstva izrazili roditelji, učitelji i uprava škole.

Učiteljima je ovakav način evidencije znatno opširniji s obzirom da nisu ograničeni prostorom za upis kao što su bili u papirnatom dnevniku/imeniku i zato imaju bolji pregled. Nadaje, svi važni dokumenti su na jednom mjestu i kao takvi uvijek dostupni i lako se pronađu. Veliku prednost navode to što imaju pristup svim podacima u bilo koje vrijeme i s bilo kojeg mjesta i što na kraju nastavne godine mogu vrlo brzo zbog prijenosa podatka iz e-Dnevnika u e-Maticu, ispisati svjedodžbe i zaključiti nastavnu godinu.

Roditelji su zadovoljni što ne trebaju više tako često dolaziti u školu zbog ocjena, već im je sve dostupno od kuće a većina u školu dolazi samo ako nastane neki problem.

Zadovoljna je i većina učenika, posebno oni koji nemaju problema s učenjem i ocjenama, dok oni drugi nisu baš oduševljeni uvođenjem e-Dnevnika zbog toga što se svaka loša ocjena i negativna bilješka ne može sakriti.

Zadovoljna je i uprava škole.

Rezultati provedenog istraživanja su do sada prezentirani Učiteljskom vijeću III. OŠ Čakovec, Vijeću roditelja III. OŠ Čakovec, Vijeću učenika III. OŠ Čakovec, nadalje na Županijskom stručnom vijeću učitelja informatike Međimurske županije, Županijskom stručnom vijeću ravnatelja Međimurske županije i na Carnetovoj korisničkoj konferenciji prošle godine.

LITERATURA

[1]. Pikić, A., „Metodologija anketnog istraživanja“, Gradska knjižnica i čitaonica Pula, 2011., dostupno u pdf formatu na http://www.academia.edu

Digitalne kompetencije za nastavnike

V. Anđelić*, H. Laušić Ramljak **

* Agencija za strukovno obrazovanje i obrazovanje odraslih/Odjel za razvoj i osiguravanje kvalitete sustava strukovnog obrazovanja, Zagreb, Republika Hrvatska

** Agencija za strukovno obrazovanje i obrazovanje odraslih/Odjel za međunarodnu suradnju i projekte, Zagreb, Republika Hrvatska

vesna.andelilc@asoo.hr

hrvojka.lausic.ramljak@asoo.hr

Sažetak - Vrijeme brzih promjena i razvoja tehnologije, potreba praćenja gospodarskih i društvenih promjena nameće cjeloživotno učenje kao potrebu, a ne kao izbor. Da bi odgovorili izazovu vremena, nastavnici trebaju pratiti napredak tehnologije i suvremene tijekove u obrazovanju i kroz uporabu informacijsko komunikacijske tehnologije (IKT). Promjene utječu na način učenja i poučavanja. Zato se nameće potreba za usavršavanjem nastavnika za sigurnu i kritičku uporabu IKT- a u poučavanju odnosno za postizanje digitalne kompetentnosti. Razvoj digitalne kompetentnosti postiže se kroz sudjelovanje i komuniciranje pomoću IKT-a te uporabom alata za izradu e-sadržaja.

Ključne riječi: informacijsko-komunikacijska tehnologija, stručno usavršavanje nastavnika, digitalna kompetencija

.

I. UVOD

Digitalna kompetencija je jedna od osam ključnih kompetencija za cjeloživotno učenje (Europski referentni okvir, 2004):

1. Komuniciranje na materinjem jeziku
2. Komuniciranje na stranom jeziku
3. Matematička pismenost i osnovna znanja iz znanosti i tehnologije
4. Digitalna kompetencija
5. Učiti kako se uči
6. Međuljudska i građanska kompetencija
7. Poduzetništvo
8. Kulturno izražavanje

Europska komisija potvrdila je razvoj digitalne kompetencije kao relevantan prioritet politika Europske unije (European Commission, 2010a, 2010b). Sudjelovanje u društvu danas zahtijeva skup kompetencija povezanih s tehnologijom koje se prepoznaju i kao „životne vještine". Potrebna je digitalna inkluzija koja ovisi, naravno, o pristupačnosti i mogućnosti, ali i o znanju, vještinama i stavovima. A to se uči i razvija kroz digitalnu kompetenciju.

II. DIGITALNA KOMPETENCIJA - POJAM, OKVIR

Gotovo sve europske zemlje imaju posebnu nacionalnu strategiju vezanu za digitalnu kompetenciju. *Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje Republike Hrvatske* definira digitalnu kompetenciju na sljedeći način:

"Digitalna kompetencija - odnosi se na osposobljenost za sigurnu i kritičku upotrebu informacijsko-komunikacijske tehnologije za rad, u osobnomu i društvenomu životu te u komunikaciji. Njezini su ključni elementi osnovne informacijsko-komunikacijske vještine i sposobnosti: upotreba računala za pronalaženje, procjenu, pohranjivanje, stvaranje, prikazivanje i razmjenu informacija te razvijanje suradničkih mreža putem interneta."

Digitalna kompetencija je najčešće korišten koncept za opis tehnološki povezane vještine. Posljednjih godina, nekoliko naziva se koristilo za opis kompetencije korištenja digitalnih tehnologija: IKT vještine, tehnološke vještine, vještine 21. stoljeća, digitalne vještine, digitalna pismenost. Promjene u društvu i kulturi temeljene na novoj tehnologiji, imaju utjecaj i na nazivlje. Digitalna kompetencija sastoji se ne samo od digitalnih vještina već i od socijalnih i emotivnih aspekata uporabe i razumijevanja digitalnih uređaja.

Prema okviru za razvoj i razumijevanje digitalne kompetencije (*A Framework for Developing and Understanding Digital Competence in Europe, 2013*), područja digitalne kompetencije su:

- Informacija: pronaći, pohraniti, organizirati i analizirati digitalne informacije, prema njihovoj važnosti i svrsi.
- Komunikacija: komunicirati u digitalnome okruženju, dijeliti resurse pomoću *online* alata, surađivati kroz digitalne medije, interakcija na mreži, svijest o međukulturalnosti.
- Izrada sadržaja: izraditi i urediti novi sadržaj (tekst, slike, video), integrirati prethodna znanja i sadržaje, stvarati kreativne multimedijske uratke, programirati, primjenjivati prava intelektualnog vlasništva i licence.
- Sigurnost: osobna zaštita, zaštita podataka, zaštita digitalnog identiteta, mjere sigurnosti, sigurno i održivo korištenje.
- Rješavanje problema: utvrditi digitalne potrebe i resurse, odabirati odgovarajuće digitalne alate s obzirom na namjenu ili potrebu, rješavati konceptualne probleme putem digitalnih sredstava, krea-

tivno koristiti tehnologiju, rješavati tehničke probleme.

Okvir digitalne kompetencije strukturiran je u pet dimenzija:

1. Identificirano područje kompetencija
2. Kompetencije primjerene svakom području
3. Razina stručnosti (osnovna, srednja, napredna)
4. Primjeri znanja, vještina i stavova za svaku kompetenciju
5. Primjeri o primjenjivosti u različite svrhe (učenje, zapošljavanje, slobodno vrijeme, kupnja i prodaja, ..)

Okvir digitalne kompetencije u prve dvije dimenzije:

Područje kompetencije Dimenzija 1	Kompetencija Dimenzija 2
Informacija	**1.1 Pregledavanje, pretraživanje i filtriranje informacija** Pretraživati i pristupati informacijama *online*, iskazati potrebu za informacijom, učinkovito pronaći resurse, kretati se između raznih *online* izvora. **1.2 Vrednovanje informacija** Sakupiti, obraditi, razumjeti i kritički vrednovati informacije. **1.3 Pohrana i pronalaženje informacija** Manipulirati i pohranjivati informacije i sadržaje za lakše pretraživanje, organizirati informacije i podatke.
Komunikacija	**2.1 Interakcija putem tehnologije** Komunicirati kroz različite digitalne uređaje i aplikacije, razumjeti kako se ostvaruje digitalna komunikacije te načine komuniciranja putem digitalnih sredstava; razlikovati komunikacijske formate, prilagoditi načine komunikacije i strategije određenoj publici. **2.2 Razmjena informacija i sadržaja** Podijeliti s drugima mjesto i sadržaj pronađene informacije, spremnost i sposobnost za razmjenu znanja, sadržaja i resursa, biti proaktivan u širenju vijesti, sadržaja i resursa, znati kako se citiraju informacije, integrirati nove podatke u postojeća znanja. **2.3 Uključenje u *online* društvo** Sudjelovati u društvu kroz online angažman, tražiti mogućnosti za vlastiti razvoj i osnaživanje u korištenju tehnologije i digitalnog okružja, biti svjestan potencijala za sudjelovanje pojedinca pomoću tehnologije. **2.4 Suradnja putem digitalnih kanala** Koristiti tehnologiju i medije za timski rad i suradnju, zajednička izgradnja i stvaranje resursa, znanja i sadržaja. **2.5 Pravila lijepog ponašanja na internetu** Znati i primjenjivati norme ponašanja u *online*/virtualom okruženju, biti svjestan kulturoloških raznolikosti, zaštititi sebe i druge od mogućih *online* opasnosti (npr. zlostavljanje putem interneta), razviti aktivne strategije otkrivanja neprimjerenog ponašanja. **2.6 Upravljanje digitalnim identitetom** Stvoriti, prilagoditi i upravljati jednim ili više digitalnih identiteta, zaštititi svoj e-ugled, raditi s podatcima kreiranim kroz više osobnih računa i aplikacija.
Izrada sadržaja	**3.1 Razvoj sadržaja** Stvarati sadržaje u različitim formatima uključujući multimediju, urediti i poboljšati razvijene sadržaje, kreativno se izraziti putem digitalnih medija i tehnologije. **3.2 Integriranje i ponovno razrađivanje** Mijenjati, poboljšavati i kombinirati postojeće resurse za stvaranje novih, izvornih i relevantnih sadržaja i znanja. **3.3 Autorska prava i dozvole** Primjenjivati prava intelektualnog vlasništva i licence na informacije i sadržaj. **3.4 Programiranje** Primjenjivati postavke, modificirati programe, programirati aplikacije i uređaje, razumjeti načela programiranja.
Sigurnost	**4.1 Zaštita uređaja** Zaštititi svoje uređaje i razumjeti *online* rizike i prijetnje, razumjeti potrebu provođenja sigurnosnih mjera. **4.2 Zaštita osobnih podataka** Razumjeti uobičajene uvjete korištenja/pružanja usluge, zaštititi osobne podatke, poštivati privatnost drugih, zaštititi se od *online* prijevara, prijetnji i virtualnog zlostavljanja. **4.3 Zaštita zdravlja** Poznavati zdravstvene rizike povezane s korištenjem tehnologije u smislu fizičkog i psihičkog zdravlja. **4.4 Zaštita okoliša** Biti svjestan utjecaja IKT-a na okoliš.

Rješavanje problema

5.1 Rješavanje tehničkih problema

Identificirati moguće probleme (od jednostavnih problema u radu do rješavanja složenijih problema) i riješavati ih pomoću digitalnog sredstva.

5.2 Prepoznavanje potreba

Procijeniti vlastite potrebe u smislu resursa, alata i razvoja kompetencija, povezati potrebe s mogućim rješenjima, prilagođavati alate za osobne potrebe, kritički procijeniti moguća rješenja i digitalne alate.

5.3 Inovacije i kreativno korištenje tehnologije

Koristiti tehnologiju kreativno/ inovativno, surađivati u izradi digitalnih i multimedijskih sadržaja, kreativno se izraziti kroz digitalne medije i tehnologiju, stvarati znanje i rješavati konceptualne probleme pomoću digitalnih alata.

5.4 Prepoznavanje potrebe unaprjeđivanja digitalne kompetencije

Prepoznati potrebu unaprjeđivanja vlastite razine digitalne kompetencije u skladu s novim potrebama i tehnologijama, podržati razvoj digitalne kompetencije drugih pojedinaca.

III. MOGUĆI MODEL UČENJA DIGITALNE KOMPETENCIJE

Učenje informacijsko-komunikacijske tehnologije od najranije dobi kroz jednostavne, intuitivne aplikacije ili društvene mreže dovodi do toga da se čini da su pojedinci toliko naviknuli na tehnologiju da poučavanje informacijsko-komunikacijske tehnologije više nije potrebno. Međutim, međunarodna istraživanja i akademska literatura upozoravaju na nedostatak digitalne osposobljenosti pojedinaca.

Jačanje digitalnih kompetencija je jedan od prioriteta u okviru strategije Europa 2020. Razvoj digitalnih vještina u svim područjima obrazovanja može doprinijeti razvoju inovativnih pedagoških pristupa učenicima. Povećanjem udjela IKT-a u nastavi i u učenju, otvara se prostor korištenja i stvaranja otvorenih obrazovnih sadržaja (Open Education Resources, OER). Potrebno je ustanoviti sustav trajnog usavršavanja nastavnika u području digitalnih kompetencija te procjenu njihova znanja.

Jedan od načina usavršavanja nastavnika u području digitalnih komptencija može biti MOOC model (Massive Open Online Course). Danas je MOOC model učenja snažno prisutan u svijetu. Međutim, prilikom razvijanja *online* tečajeva treba imati na umu da je sadržaj namijenjen isključivo za učenje na daljinu. Koriste se različita digitalna sredstva (audio, video, igre, grafikoni, appleti) jer je drugačiji način donošenja informacija (pripovijedanje, uvodni video, postavljanje problemske situacije, primjer). Uvode se interaktivni elementi kao i procjena/priznavanje stečenih kompetencija. U MOOC modelu učenja, kada može biti veliki broj sudionika, teško je imati dovoljan broj mentora koji će pomagati i voditi sudionike do cilja – uspješnog svladavanja stručnog usavršavanja. Dobra ideja može biti uvođenje peer-mentorstva kada svi pomažu svima. Važnu ulogu u učenju tada imaju i društvene mreže. Također, za procjenu usvojenosti digitalne kompetencije može se razviti peer-review model. Osim procjene znanja pomoću kvizova s raznim tipovima pitanja, tada se uvodi „završni uradak“ kojeg procjenjuju barem još dva sudionika stručnog usavršavanja. Uspješno svladavanje kvizova znanja, pozitivna procjena završnog uratka te izvršena procjena barem još dva završna uratka definirat će prolaznost/uspješnost sudionika stručnog usavršavanja. Tako se može dobiti održiv sustav stručnog usavršavanja.

Europska komisija u izvješću iz 2012 *Digital Agenda for Europe, Survey of Schools, ICT in Education (2)* navodi da je Hrvatska jedina država EU u kojoj se digitalna kompetencija ne stječe kroz kurikulume raznih nastavnih predmeta već kroz zaseban nastavni predmet *Informatika/Računalstvo*. Treba napomenuti i da nastavni predmet *Informatika* nije obvezan u osnovnom obrazovanju. Također, prema *Eurydice Report: Developing Key Competence in school in Europe*, *2012* u hrvatskim strukovnim školama više od 50% nastavnika strukovnih predmeta koristi IKT u manje od 25% nastavnih sati. To se svakako može objasniti nedostatkom digitalnih kompetencija kod nastavnika. Nastavnici trebaju biti digitalno sigurni da bi koristili digitalnu tehnologiju u učionici, i za učenje i za poučavanje.

IV. ZAKLJUČAK

Analize pokazuju kako današnji učenici provode vrijeme na internetu najviše za druženje u virtualnim okruženjima i društvenim mrežama, komunikaciju putem elektroničke pošte, pretraživanje interneta za zabavu te za izradu školskih projekata. Njihovi prioriteti su prijatelji, zabava, interakcija u realnom vremenu. Ne vole složenost i loš dizajn. Stoga, da bi nastavnici odgovorili zahtjevu vremena i tehnologiji trebaju steći digitalnu kompetentnost. Trebaju koristiti i stvarati digitalne nastavne materijale (kvalitetne i primjerene prezentacije, multimedijske i interaktivne plakate, konceptualne mape, stripove, ..) te koristiti IKT za učenje i poučavanje. Samo tako će održati pažnju svojih učenika i ostvariti zadane ishode učenja svojih nastavnih predmeta. Potrebno je sveobuhvatno i kontinuirano stručno usavršavanje.

Literatura

[1] L. Ilomaki, A. Kantosalo, M. Lakkala, “What is digital competence”, University of Helsinki, Finland, March 2011.

[2] M. Lakkala, L. Ilomaki, A. Kantosalo, “Which areas of digital competence are important for teacher?” University of Helsinki, Finland, March 2011.

[3] A. Ferrari, „DIGCOMP: A Framework for Developing and Understanding Digital competence in Europe, Report EUR 26035 EN, 2013.

[4] „Digital Agenda for Europe, Survey of Schools: ICT in Education“, Final Report

[5] „Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje“, MZOS, Zagreb, 2011.

Primjena internetskog foruma u prezentaciji i obrani seminarskih radova – studija slučaja kolegija Planiranje informacijskih sustava

Adela Sočković*, Alen Jakupović*
*Veleučilište u Rijeci/Poslovni odjel, Rijeka, Republika Hrvatska
asockovi@veleri.hr
alen.jakupovic@veleri.hr

Sažetak – Seminarska nastava je oblik nastave tijekom koje studenti izrađuju seminarski rad – stručni rad kojim se detaljnije obrađuje neka tematika iz domene nastavnoga programa, te ga javno prezentiraju i brane. Jedan od ciljeva javne prezentacije i obrane je poticanje rasprave o obrađenoj temi. Ponekad do rasprave ne dolazi, a neki od uzroka toga su ograničeno vrijeme trajanja prezentacije i obrane, te nemotiviranost studenata za raspravom. Jedno od rješenja navedenog problema je primjena internetskog foruma u prezentaciji i obrani seminarskoga rada. Članak opisuje jednu takvu primjenu kroz analizu studije slučaja seminarske nastave iz kolegija Planiranje informacijskih sustava koja je održana na specijalističkom diplomskom stručnom studiju Informacijske tehnologije u poslovnim sustavima na Veleučilištu u Rijeci. U članku je prikazana metodika seminarske nastave koja je temeljena na internetskome forumu, kao i korištena informacijsko-komunikacijska tehnologija. Analizom 28 obrana seminarskih radova opisana su iskustva studenata koji su sudjelovali u nastavi bilo kao autori seminarskoga rada ili kao njegovi čitatelji. Ukazano je na prednosti i nedostatke primjene internetskoga foruma u seminarskoj nastavi, te su predložena poboljšanja.

I. UVOD

Članak se bavi problemom nedovoljne aktiviranosti studenata slušatelja tijekom javne prezentacije i obrane seminarskih radova. Nedostatak rasprave tijekom seminarske obrane i javne prezentacije, kojom bi i autor i slušatelji došli do novih spoznaja i bili ponukani na šire razmišljanje o temi iz nastavnog programa, čest je slučaj u nastavi. Brojni su razlozi uzrok tom problemu, a želja je u što većoj mjeri nadići taj problem i pronaći adekvatno rješenje koje bi bilo primjereno današnjim studentima, njihovom tempu i načinu života, što za posljedicu može imati i osuvremenjivanje održavanja nastave.

Nastava je prvenstveno interakcijski stvaralački napor u savladavanju propisanih sadržaja i stjecanju znanja. Važno je da je ta interakcija kvalitetno podržana i ostvarena [1]. S druge pak strane, internet tehnologije danas gotovo ne možemo zaobići u svakodnevnom životu, a njihova prednost u ostvarivanju interakcije je već neko vrijeme prepoznata. Brojne prednosti internet tehnologije su svakim danom sve više korištene u obrazovanju i ona je, k tome, bliska mladim ljudima [2].

Člankom se želi utvrditi opravdanost uvođenja internetskog foruma u seminarski dio nastave. Svrha je prikazati jednu takvu primjenu suvremene metode prezentacije i obrane seminarskih radova u edukativne svrhe. Ciljevi su usmjereni na opisivanje primjene metode internetskog foruma, odnosno uvođenje računala i individualnog pristupa u taj nastavni dio na Veleučilištu u Rijeci, njenih prednosti i nedostataka, dosadašnjih ostvarenja, iskustva samih korisnika, kao i mogućnosti za poboljšanjem i širom primjenom.

Članak se osim uvoda sastoji iz još četiri poglavlja. U drugome poglavlju se kratko opisuje povijest e-učenja. Treće poglavlje detaljno opisuje metodu primjene internetskoga foruma u obrani seminarskih radova te rezultate provedene analize. Četvrto poglavlje opisuje mogućnost primjene ove metode u rješavanju problema dugoga studiranja i male prolaznosti. Peto i zadnje poglavlje – zaključak – donosi sintezu prednosti, nedostataka i mogućih poboljšanja prikazane metode obrane seminarskih radova.

II. E-UČENJE KROZ POVIJEST

Kada govorimo o primjeni internet tehnologije u obrazovanju i nastavi ne možemo se ne susresti s pojmom e-učenja. *„E-učenje (engl. e-learning) je proces obrazovanja (proces učenja i podučavanja) koji se izvodi uz uporabu nekog oblika informacijske i komunikacijske tehnologije, a s ciljem unapređenja kvalitete toga procesa i kvalitete ishoda obrazovanja"* [3]. Primjena internetskog foruma svakako spada u oblik e-učenja. To je pojam s kojim se sve češće svakodnevno susrećemo, a sam početak obrazovanja na daljinu veže se uz Sjedinjene Američke Države krajem 19. stoljeća, kada je pokrenuto više inicijativa za obrazovanje putem dopisnih studija. Ubrzo je početkom 20. stoljeća pozornost usmjerena na kreiranje novih pedagoških modela za studije tog tipa i standarda kvalitete za njihovo provođenje.

Kod suvremenog načina učenja naglasak je svakako na računalnoj tehnologiji, koja je tada trebala osigurati interaktivno poučavanje u obrazovanju na daljinu. Kako bi se takvo podučavanje ostvarilo, 1960. godine pojavljuju se posebno dizajnirani obrazovni softveri (engl. *courseware*). Razvoj razmjene podataka, odnosno *World Wide Weba* je izazvao pak posebnu tehnološku i pedagošku revoluciju u obrazovanju na daljinu.

S vremenom su se za navedene oblike poučavanja pojavili pojmovi koji ih definiraju, pa su neki od primjera

računalom podržano učenje (engl. *computer assisted learning*, CAL), kao i na računalu zasnovano podučavanje (engl. *computer based instruction*, CBI). Ovi pojmovi su već više od trideset godina u uporabi u jeziku, no još uvijek nedovoljno primijenjeni u obrazovanju [4].

III. METODA INTERNETSKOG FORUMA U PREZENTACIJI I OBRANI SEMINARSKIH RADOVA

Više no ikada prije nužno je uskladiti obrazovanje s potrebama tržišta. U vremenu inovacija i rasta društva iznimno su važne vještine rješavanja složenih problema, inovativnost i kreativni napor. Te vještine se ne stječu dovoljno kvalitetno (ili ponekad uopće) u školama i na fakultetima [5].

U cjelokupnom obrazovanju izostaje kritičko mišljenje, koje je ključ za kreativnost i inovativnost za kojima tržište vapi. S tim na umu krenulo se u primjenu metode internetskog foruma u seminarskom dijelu nastave kolegija Planiranje informacijskih sustava. Uloga internetskog foruma je pokrenuti raspravu među studentima o temama iz struke. Pored toga su omogućene razne ostale prednosti karakteristične za uporabu internet tehnologija u obrazovanju, među kojima se ističu individualnost u pristupu i podjednaka angažiranost svih studenata tijekom ovog oblika nastave, odnosno sudjelovanje u učenju i kritičkom razmišljanju i od strane čitatelja, kao i od strane autora.

A. Izbor i obrada teme seminarskog rada metodom instrukcijskog dizajna

Tema seminarskog rada je samostalno izabrana s ponuđenog popisa tema unutar kolegija Planiranje informacijskih sustava te obrađena metodom instrukcijskoga dizajna od strane studenta, autora seminarskog rada. To znači da je za izabranu temu student otvorio novu diskusijsku temu unutar foruma „Seminarski radovi“ u kojoj je definirao ishode učenja. Oni u metodi instrukcijskog dizajna zapravo predstavljaju uspostavljanje budućeg sadržaja jer sadržaj seminara treba biti upravo takav i volumenski toliki da pokriva postavljene ishode učenja. Ishodi učenja su „kostur“ seminarskog rada i ujedno vodilja autoru u oblikovanju izlaganja tematike, a predstavljaju i temelj za izradu zadataka za provjeru kojima čitatelji mogu provjeriti jesu li pročitanim sadržajem uspjeli savladati definirane ishode, postavljene od strane autora rada.

Unutar istoga foruma se također navodi i literatura koja će se koristiti u izradi seminara te se može diskutirati i postavljati pitanja profesoru u vezi izabrane teme prilikom izrade seminarskog rada. Treba primijetiti da nitko od 30 studenata nije iskoristio mogućnost postavljanja pitanja profesoru u vezi izabrane teme. Više je mogućih razloga tome od kojih neki mogu biti da su studenti navikli samostalno izrađivati radove ovog tipa i pretraživati literaturu, da su neki taj dio obavili putem konzultacija ili elektroničke pošte, da su se studenti ustručavali pokazati da im nešto nije jasno i slično. S obzirom da su se studenti prvi put susreli s ovakvim načinom izrade seminarskih radova i njihove obrane, postoji mogućnost da će se ta opcija više koristiti u daljnjem radu s internet forumom, nakon što im taj način održavanja seminarskog dijela nastave postane bliskiji.

Tijekom izrade seminara autor se koristi dokumentima „Sadržaj seminarskoga rada i kriterij ocjenjivanja“ i „Bloomova taksonomija - Kognitivna domena“. To su materijali na internet stranicama kolegija koji objašnjavaju koje elemente treba rad sadržavati i na koji način se isti vrednuju te kako što bolje oblikovati rad metodom instrukcijskog dizajna kako bi se uspješno poučilo kolege o prezentiranoj tematici. S obzirom da studenti do sada nisu koristili taj način poučavanja, Bloomova taksonomija je bila od velike pomoći u oblikovanju rada na način da autori spoznaju čemu žele naučiti svoje kolege i time izdvoje bitno od nebitnog te oblikuju kvalitetan materijal za učenje na internet forumu. Čitatelji su ishodima učenja dobili sliku onoga što se od njih očekuje i na što trebaju obratiti posebnu pažnju tijekom proučavanja sadržaja [6].

B. Prezentacija i obrana seminarskog rada

Svaki student je bio dužan izraditi i obraniti seminarski rad, jer je to preduvjet za izlazak na završni/popravni ispit. Po završetku izrade seminarskog rada je tako napravljen materijal u formatu za elektronsko izdavaštvo, odnosno PDF (engl. *Portable Document Format*) obliku poslan na čitanje studentima (njih 28). Studenti čitatelji su potom trebali u roku od tri dana od objave seminara unutar posebnog foruma „Obrane seminarskih radova“ postaviti netrivijalno pitanje (ono pitanje čiji odgovor ne slijedi direktno iz seminara) otvaranjem diskusijske teme. Autor seminarskoga rada, u roku od sedam dana od objave rada, postavlja objavu na otvorenu diskusijsku temu svakog pojedinog čitatelja i na taj način odgovara na postavljeno pitanje.

Studenti su dodatno motivirani bodovanjem za postavljanje pitanja i odgovaranje na ista. Za svaki seminarski rad student može postaviti najviše jedno pitanje i za svako postavljeno netrivijalno pitanje student čitatelj dobiva jedan bod. Za svaki odgovor na netrivijalno pitanje autor dobiva jedan bod. Također se dodjeljuju i negativni bodovi, odnosno student koji ne postavi netrivijalno pitanje dobiva -1 bod, kao što i autor koji ne odgovori na netrivijalno pitanje dobiva -1 bod.

C. Rezultati metode

Prilikom korištenja metode internetskog foruma pojavila su se razna pitanja od strane čitatelja vezana uz teme seminara, ali su i autori na kraju prezentacije i obrane svojih seminarskih radova iznijeli unutar objave „Zaključak“ svoje dojmove o provedenoj edukaciji koju su oni predvodili, njenim korisnicima, odnosno studentima čitateljima, kao i o mogućim poboljšanjima. Na taj su način veoma efikasno, brzo i „iz prve ruke“ empirijskom metodom prikupljeni dojmovi samih korisnika metode internetskog foruma.

Primjeri najčešće postavljanih pitanja odnose se na više informacija o temi koje nisu spomenute u samom seminaru, tj. prednosti, nedostaci, gdje su bile primijenjene, softveri koji podržavaju metodologiju ili barem neke njene dijelove, mogućnost preskakanja faza i sl.:

„Koji su nedostaci Event-driven process chain tehnike za modeliranje?“

„Da li si možda tijekom istraživanja za seminarski rad naišla na informaciju koja su poznatija poduzeća razvijala svoj informacijski sustav po Nolan Norton metodologiji?“

„Navela si u radu računalno podržane dijagrame, neki primjer možda softvera u kojem se izrađuju?“

Također bi u drugoj pak skupini bila pitanja koja su povezana uz dodatne informacije u vezi napisanog u seminarskom radu, odnosno ona koja iskazuju potrebu za detaljnijim objašnjenjem nekog dijela, pojma i sl. Ona su obično prikazana poput poveznice na tekst, citirajući autora:

„U smjernicama za upravljanje procesima razvoja piše: "potrudite se da upravni odbor radi svoj posao". Na koje načine bi se to trebalo postići? Koliko upravni odbor utječe na daljnji razvoj?“

Trećoj skupini bi pripadala pitanja koja se odnose na mišljenje autora:

„Da li imaš ideju kojem projektu/tvrtki iz realnog života možeš danas savjetovanjem temeljem prethodno stečenog znanje iz Nolan - Norton metodologije pomoći?“

„Imaš li ideju kako bi se uspješno mogao primijeniti ovaj model na primjeru Veleučilišta u Rijeci?“

Ova se pitanja smatraju veoma važnima jer pokreću kritičko mišljenje kod studenta, čime je svakako primjenom internetskog foruma učinjen korak prema kvalitetnijem obrazovanju u kojem je studentu omogućeno da s kolegama diskutira i kreativno odgovara na izazove koji se pred njega postavljaju.

Kad je riječ o zaključku na kraju prezentacije i obrane seminarskih radova, u kojem su se iznosili dojmovi autora o provedenoj edukaciji koju su predvodili, korisnicima, odnosno čitateljima i mogućim poboljšanjima, valja napomenuti da su se autori često osim o dojmu o primjeni internetskog foruma s gledišta autora, izjasnili i o svom dojmu u primjeni kao čitatelji. Tu su dobivene korisne informacije o stajalištu korisnika ove metode učenja.

„...Što se tiče same obrane ovog rada, bila je na drugačiji način, zanimljivija, svatko od nas je „prisiljen“ pročitati rad ostalih kolega i postavljati pitanja te se na taj način pokušava više zainteresirati za usvajanjem novog znanja iz svakog seminarskog rada. Na taj način, postavljanjem pitanja se dolazi i do dodatne literature i proširuje se sam seminarski rad. Ovo je prvi put da smo imali ovakvu obranu seminarskog rada, zanimljivija je, više se sazna o temi nego na klasičan način i mislim da bi trebali imati što više obrana seminara na ovaj način.“

„S obzirom da sam se prvi puta susreo s online obranom završnog rada moram priznati da mi se jako dopalo iako sam bio malo skeptičan na početku jer ni ja kao i većina kolega nismo imali mogućnost online obrane seminarskog rada. Sama online obrana seminarskog rada je po meni dobra jer autor seminarskog rada konstantno kroz vrijeme obrane proširuje svoje znanje o metodologiji te tako još više proučava i nadopunjuje znanje vezano uz tematiku svojega rada, a usput kroz proučavanje radova koje su napravile kolege susrećemo se i upoznajemo s velikim brojem različitih metodologija što je svakako korisno u budućem radu i školovanju. Kroz eventualna pitanja kolega kroz samu obranu seminarskog rada se najbolje mogu uočiti moguća poboljšanja ili nedostaci samog seminara i definitivno mislim da je ovaj način obrane rada daleko kvalitetniji od klasične obrane jer i od autora rada i od čitatelja zahtijeva veće shvaćanje i proučavanje tematike seminarskog rada jer dolazi do veće diskusije a samim time i do većeg saznanja svih sudionika diskusije. “

„Što se tiče provedene edukacije, smatram da smo ovim načinom obrane i kolege i ja naučili nešto više. Oni su imali priliku seminarom se dodatno zainteresirati za metodologiju i njenu primjenu, te također izraziti eventualne nedoumice u vezi pročitanog. To je pak mene poticalo na dublje istraživanje, i također „šire“ razmišljanje. Svi smo sa svojim iskustvima i znanjem (što tijekom obrazovanja, što iz obavljanja naših poslova) razmjenjivali mišljenja te konačno sudjelovali i u vrednovanju metodologije.

Smatram da je ovakav način dobar za učitelje, profesore, izlagače, kada žele dobiti povratnu informaciju o tome koliko su pažljivo čitatelji pročitali lekciju, predavanje i sl., te koliko su istu shvatili. Dodatno bi se, ukoliko je to potrebno, provjera mogla „pojačati“ na način da čitatelji pošalju svoje odgovore na Zadatke za provjeru ostvarenih ishoda učenja.

Meni osobno su pitanja svakako pomogla da provjerim koliko sam naučila metodologiju i to znanje primijenim u dodatnom istraživanju i razmišljanju o mogućnostima i načinima primjene. Po pitanjima svakako vidim da su kolege gradivo seminara nastojali iskoristiti kombinirajući postojeća znanja, stečena na fakultetu, što je i bila svrha seminara.“

„Provedena edukacija mi se je jako svidjela iz razloga što postoje studenti koji ne mogu kvalitetno izlagati pred publikom, a drugi razlog je to što se tema obradi znatno kvalitetnije. Kada se izlaže na predavanju nismo u mogućnosti odgovoriti na neka pitanja vezana za temu jer nam je "promakao" taj dio u istraživanju, a ovdje ako nam i promakne možemo nastaviti istraživanje i naći odgovor te produbiti istraživanje o temi.

Bilo bi jako dobro kada bi profesori na Veleučilištu prihvatili ovakav načina izlaganja seminara jer će se bolje i kvalitetnije obraditi seminari. ...“

Prema rezultatima provedene metode, odnosno napisanim „Zaključcima“ može se primijetiti da je primjena metode internetskog foruma u nastavi polučila dobre rezultate i dobila i više no prolazne ocijene od strane studenata koji su je uz vodstvo profesora primijenili kako bi prezentirali i obranili svoje seminare, ali i sudjelovali u edukaciji postavljanjem pitanja. Korisnici su kao prednosti u svojim komentarima navodili barem kvalitetnu izmjenu mišljenja i proširenje sadržaja seminara, kao i da je pristup zanimljiv jer predstavlja za njih nešto novo i nepoznato, dok je jako velik broj iznimno zadovoljan povećanjem kvalitete u obrani jer se dobiva na vremenu za dodatno proučavanje literature i detaljnije i točnije odgovaranje na pitanja. Svakako ne treba zanemariti kako su pojedini studenti istakli i prednost samog izlaganja bez publike, koja mnogima stvara stres te u tim situacijama često gube bodove jer ne

mogu pokazati svoje pravo znanje radi velike treme i sličnih uzroka. Time svakako dobivamo motiv i za nekim budućim istraživanjem primjene ove metode na kolegijima gdje javna usmena obrana nije ključna.

Osim po pozitivnim komentarima, prema kojima se vidi da je svatko našao nešto za sebe u ovoj metodi, i prema broju onih koji su napravili seminarski rad i branili ga na ovaj način možemo zaključiti da je metoda jako dobro prihvaćena jer je 28 studenata od njih ukupno 30 (93,33 %) sudjelovalo u ovom načinu obrane i prezentacije seminarskog rada.

D. Prednosti primjene metode

Prednost za slušatelje, odnosno u ovom slučaju čitatelje je da samostalno biraju vrijeme i mjesto u kojem će posvetiti svoju pažnju čitanju sadržaja i učenju nečeg novog. To svakako nudi kvalitetniji i individualniji pristup učenju. Ovakva metoda obrane i prezentacije seminarskog rada se posebno preporučuje za kolegije koji su vremenski kasnije po rasporedu, a imaju u sklopu nastavnog plana predviđene seminare. U tim situacijama pažnja studenata obično popušta i svakako izlagani sadržaj ne bi zadobio dovoljno koncentracije i pažnje ciljane publike.

Ovakav način obrane seminara je više nego dobrodošao za obranu i prezentaciju tema koje pokrivaju opširnije gradivo, kao i u slučaju da seminarski dio nastave ima velik broj sličnih tema koje je potrebno prezentirati. On omogućava čitateljima da se posvete tom seminaru s punom pažnjom, da sami iz udobnosti svoga doma, uz pomoć Interneta, istraže razlike i sličnosti u odnosu na neke prethodne teme seminara te se dodatno zaintrigiraju tematikom. Time se dobivaju pitanja studenata koja su postavljena na „višoj razini“ te oni imaju mogućnost sudjelovati u raspravama i pokretati iste, koristeći dostupnu literaturu na Internetu ili kod kuće. Internet je svakako najveći zbir podataka i uvelike skraćuje pretraživanje te ubrzava postupak dolaženja do novih spoznaja koji je nekada trajao tjednima, mjesecima u pretraživanju i obilaženju knjižnica i sličnih izvora.

S druge strane autor, koji je najviše upoznat s temom, ima priliku proširiti svoje znanje dodatnim istraživanjem te kvalitetno odgovoriti na pitanja. Autor ponukan postavljenim pitanjem kolega kreće u dublje istraživanje i pretraživanje literature i Interneta. Odgovaranjem nadopunjava svoje znanje, ali i sam sadržaj postojećeg seminara. Moguće je da neke dijelove autor nije dovoljno detaljno i razumljivo objasnio, da je neke bitne činjenice ili zanimljivosti propustio u svom radu, a na to će mu sve ukazati pitanja kolega. Svakako se vidi da iz ovog načina proizlaze kvalitetni rezultati za obje strane, a sama obrana je podignuta u svojoj razini kvalitete.

Čest je problem prilikom javne usmene prezentacije nemogućnost autora da provjeri svoj odgovor za neko postavljeno pitanje te kvalitetna pitanja često ostanu bez kvalitetnih odgovora. To je i razumljivo jer autor seminara, koliko god da je upoznat s temom, ne može znati odgovor na sva pitanja „iz glave“ jer obrađena tema seminara nije njegov „izum“, „otkriće“ i sl. On je temu obradio proučavajući dostupnu literaturu, no on zato ima jako dobre predispozicije da u jako kratkom vremenu pronađe ispravan odgovor, detaljnije proučavajući već korištenu literaturu ili neku njoj sličnu. Ovim se načinom osigurava interakcija studenata po principu brzog, ali i kvalitetnog i provjerenog odgovaranja na postavljena pitanja s ciljem dobivanja trenutne slike o tome jesu li studenti razumjeli što im se izlaže [2].

E. Nedostaci primjene metode

Autor svakako jako puno nauči o svojoj temi, no to se ne može reći i za čitatelje. Svakako nije jednaka razina znanja o temi u oba slučaja. Velik broj pročitanih seminara ne može rezultirati jednakom kvalitetom naučenog za čitatelja, odnosno on neće jednako dobro naučiti sve teme, a posebno ih sve ne može ni približno dobro razumjeti kao autor pojedine teme.

Studenti često ne provjeravaju svoje znanje o naučenom prema zadacima za provjeru ostvarenih ishoda učenja. Ishodi učenja su većinom primijenjeni samo od strane autora, koji prema njima stvara sadržaj seminara, no od strane čitatelja su slabo ili nikako korišteni.

Ovaj način obrane seminara ne preporučuje se za primjenu na svim vrstama kolegija jer je važno da studenti i dalje imaju javnu obranu, usmeno izlaganje ideja, mišljenja i zaključaka koje trebaju znati prezentirati pred publikom.

F. Prijedlozi za poboljšanje metode

Bilo bi veoma korisno omogućiti da student može sudjelovati (npr. u obliku ankete i sl.) u odgovaranju na pitanja zadataka za provjeru ostvarenih ishoda, a odgovori se odmah analiziraju i dobiva se jasna predodžba koliko je studenta razumjelo obrađeno gradivo. Taj dio može se vrednovati kao aktivnost na nastavi ili biti dijelom kolokvija. Ovime se osigurava i iskorištenost zadataka za provjeru ishoda učenja od strane čitatelja, koji tako ponavljaju pročitano te dobivaju povratnu informaciju o naučenom, i da isti budu ponukani bolje razumjeti pojedinu temu, bolje je proučiti i više o njoj zapamtiti. Time se popravlja i razina znanja čitatelja teme u odnosu na autora.

Mogućnost jačeg fokusiranja na raspravu, također, bi mogla doprinijeti boljem pamćenju sadržaja pojedinih seminara. Tijekom primjene metode internetskog foruma čitatelji su većinom postavljali pitanja na koja je autor odgovarao, a iznimno je malen broj čitatelja koji su nakon odgovora ulazili u diskusiju i ako su istu željeli pokrenuti nisu dobivali odgovor. To se može popraviti uvođenjem obveznog pitanja od strane autora prilikom odgovaranja, npr. autor odgovori na pitanje, ali da bi dobio jedan bod treba postaviti i protupitanje, koje bi preporučljivo moglo biti vezano uz čitateljevo pitanje, a odgovor se može naći u samom sadržaju seminara. I u tom slučaju i čitatelj dobiva jedan bod tek kad i sam odgovori na protupitanje, a u protivnom oba mogu dobiti samo po pola boda (ako u autorovu odgovoru nije bilo pitanja i na to se nije reagiralo). U slučaju da čitatelj zamijeti da nije dobio protupitanje, ima pravo sam postaviti komentar na odgovor (protuizjavu – onu koja se suprotstavlja autorovoj) ili još jedno netrivijalno pitanje za autora. On tada može dobiti jedan bod, a autor će isti ostvariti odgovaranjem na protuizjavu ili drugo postavljeno pitanje svojim argumentima (ako ne odgovori ostaje mu pola boda).

Ovime rasprava može dobiti svoj intenzitet i postati zanimljivija, no vrlo je vjerojatno da će samo čitanje i odgovaranje u tom slučaju zahtijevati produljenje potrebnog roka koje se mora uzeti u obzir. Ipak, produljenjem roka i pojačanjem fokusa na raspravu će se uistinu morati od strane čitatelja pratiti autorovi odgovori, a pitanja se neće postavljati „radi reda“ te će se time i puno više naučiti.

Zornije prikazivanje povezanosti sadržaja s ishodima učenja, također, bi moglo biti od pomoći u boljem postizanju istih, pa bi se kroz primjerice kviz prilikom netočnog odgovora mogao prikazati dio teksta u kojem se nalazi odgovor. To bi pak uključivalo uvođenje dodatnih tehnologija i ulaganje dodatnog vremena pa bi se u daljnjim istraživanjima mogla provjeriti korisnost i isplativost takve mogućnosti.

Klasična nastava ne treba izgubiti svoj značaj. Na nastavi se zajedno s profesorom može komentirati prezentirano gradivo i tu se treba očekivati puna aktivnost studenata [2]. Tada studenti mogu usmeno raspraviti o temi i dobiti korisne povratne informacije od profesora.

IV. BUDUĆNOST METODE

Smatra se da bi se na ovaj način čak i na „neinformatičkim“ studijima mogli postići kvalitetni rezultati. Neki studiji, poput Prava, Ekonomije i sl. bi se u seminarskom dijelu nastave nekih kolegija mogli veoma učinkovito odrađivati na ovaj način. Do ideje se došlo proučavanjem istraživanja o koncentraciji studenata, gdje je jedan projekt na tu temu proveden na Pravnom fakultetu u Zagrebu. Zbog problema dugog studiranja i jako male prolaznosti, proveden je spomenuti projekt u svrhu unaprjeđivanja kvalitete studiranja, a unutar njega je provedeno i istraživanje sa zanimljivim rezultatima.

Okolnosti koje su studenti naveli kao otežavajuće za studij su najvećim dijelom bile nemotiviranost i nezainteresiranost (36 %), lijenost i pomanjkanje upornosti (31 %) i sl., te na kraju i manja sposobnost usvajanja gradiva i loša koncentracija (5,5 %) [7]. Na takvim studijima studenti imaju seminarski dio nastave i često „barataju“ s poprilično opsežnim gradivom, sličnim pojmovima i puno informacija odjednom. Ovakav način, u kojem bi svatko imao kod kuće, u knjižnici, na tabletu u prirodi i sl. mogućnost proučavanja gradiva te samostalnog istraživanja, djelovao bi više motivirajuće i potaknuo veću zainteresiranost. Bitno je da studenti sami upravljaju svojim vremenom te da nisu nigdje fizički prisiljeni biti (npr. na predavanju).

V. ZAKLJUČAK

Članak opisuje iskustva u primjeni internetskog foruma u obrani seminarskih radova, do kojih se došlo analizom provedene seminarske nastave u akademskoj 2014./2015. godini na kolegiju Planiranje informacijskih sustava koji se održava u sklopu specijalističkog diplomskog stručnog studija Informacijske tehnologije u poslovnim sustavima na Veleučilištu u Rijeci.

Rezultati analize, koja se temeljila na deskriptivnoj povratnoj informaciji studenata – sudionika prezentacije i obrane seminarskih radova, pokazuju izrazito dobru prihvaćenost ovakvoga oblika nastave. Neke od uočenih prednosti su: slobodan izbor vremena i mjesta kada će se seminar pročitati, kada će se postaviti pitanje i dati odgovor (uz poštivanje početno zadanoga vremenskog intervala od sedam dana), postavljena pitanja potiču autora na produbljivanje znanja, otklonjen je pritisak izlaganja pred publikom itd.

Pored navedenih prednosti, uočene su i neke mane ovoga pristupa: razina usvojenosti znanja autora o pojedinoj temi je zbog dodatnih pitanja veća nego kod čitatelja, čitatelja se ne prisiljava da iskoristi zadatke za provjeru dostignuća ishoda učenja, nije dobro apsolutno ukloniti javno izlaganje itd.

Nakon provedene analize iskristalizirale su se ideje za dodatnim poboljšanjem primjene internet foruma u obrani seminarskih radova. Neke od ideja su: obavezno uključivanje provjere dostignuća ishoda učenja čitatelja, uključivanje pitanja autora čitatelju, zornije povezivanje ishoda učenja sa sadržajem seminarskoga rada itd.

LITERATURA

[1] A. Milovac, "Osnove andragogije, sociologije i psihologije za potrebe ECDL predavača i ispitivača, priručnik," Hrvatski informatički zbor, Zagreb, 2003.

[2] I. Domitrović, "Studenti, ovo biste TREBALI ZNATI ako na predavanja nosite laptop", http://dnevnik.hr/vijesti/hrvatska/studenti-ovo-biste-trebali-znati-ako-na-predavanja-nosite-laptop---303990.html, 07.02.2015.

[3] E-učenje, http://www.srce.unizg.hr/ceu/o-centru/e-ucenje/ 07.02.2015.

[4] Povijesni razvoj obrazovanja na daljinu, http://www.carnet.hr/referalni/obrazovni/mkod/naslovnica/zastomet/pov.html, 07.02.2015.

[5] A. Poce, "Content Analysis and critical thinking: an assessment study",https://connect.srce.hr/p5wok8k7k25/?launcher=false&fcsContent=true&pbMode=normal, 07.02.2015.

[6] Metodika, http://www.carnet.hr/referalni/obrazovni/mkod/metodika.html, 07.02.2015.

[7] M. Ajduković, V. Branica, L. Rajhvajn, Đ. Gardašević, "Mišljenje studenata IV. godine Pravnog fakulteta o mogućnostima unaprjeđivanja kvalitete studiranja i reformi obrazovnog sustava," projekt "Razvoj modela unaprjeđivanja kvaliteta studija na Pravnom fakultetu Sveučilišta u Zagrebu," neobjavljen

Upravljanje Arduino platformom pomoću programskog jezika Processing

Andrej Lacković
Visoko učilište Algebra – visoka škola za primijenjeno računarstvo, Zagreb, Hrvatska
andrej.lackovic@racunarstvo.hr

Sažetak - Arduino platforma je jednostavna i univerzalna platforma zasnovana pomoću mikrokontrolera na ATmel tehnologiji. Idealna je za razvoj upravljačke elektronike i robotike. Jedan od mogućih načina upravljanja Arduino platformom je pomoću računala kroz programski jezik Processing. Processing je vrsta programskog jezika baziranog na simplificiranoj varijanti Java programskog jezika i integriranog razvojnog okruženja. Rad opisuje osnovno programiranje Arduino mikrokontrolera, te upravljanje pomoću računala kroz programski jezik Processing kroz realan primjer upravljanja RGB LED trakom.

I. UVOD

Mikrokontroler je digitalna elektronička naprava u obliku integriranog kruga. Namjena mikrokontrolera je upravljanje uređajima i procesima, pa u sebi ima integrirani mikroprocesor, memoriju, digitalne i analogne ulaze i izlaze, komunikacijske sklopove i druge dodatke. Mikrokontroler normalno radi u kontroliranoj petlji, dakle očitava ulaze i zatim podešava izlaze sukladno sa svojim programom. Petlja se stalno ponavlja dok traje kontrola procesa.

Programi za mikrokontroler se pišu na računalu te pomoću programatora prebacuju u njihovu memoriju. Najpoznatiji proizvođači mikrokontrolera, uglavnom za hobiste, su Atmel i Microchip.

II. ARDUINO

Arduino je platforma otvorenog koda za kreiranje elektroničkih prototipova baziran na sklopovlju i programskom paketu koji je fleksibilan i jednostavan za korištenje. Namijenjen je umjetnicima, dizajnerima, hobistima, elektroničarima i svima koji su zainteresirani za kreiranje interaktivnih objekata i okruženja.

Slika 1. Arduino Uno mikrokotroler

Arduino je skup elektroničkih i softverskih komponenti koje se mogu jednostavno povezivati u složenije cjeline s ciljem izrade zabavnih i poučnih elektroničkih sklopova. Srce Arduina jesu mikrokontroleri. Najčešće koristi 8 bitne mikrokotrolere koje proizvodi tvrtka Atmel.

A. *Verzije Arduina*

Arduino dolazi u nekoliko različitih verzija. Uglavnom se razlikuju po mikroprocesoru, količini memorije, te broju analognih i digitalnih ulaza/izlaza.

Pa tako razlikujemo:

- Arduino Uno (ATmega 328, I/O: 14) (slika 1.)
- Arduino Mega (ATmega 2560, I/O: 54)
- Arduino Micro
- Arduino Nano

B. *Moguće nadogradenje*

Arduino je moguće nadograditi sa njemu kompatibilnim modulima za različite potrebe. Moguće nadogradnje možemo svesti na module i dodatke.

Neki od mogućih modula:

- GSM Shield (GSM modul)
- Ethernet Shield (mrežni modul)
- WiFi Shield (WiFi modul)

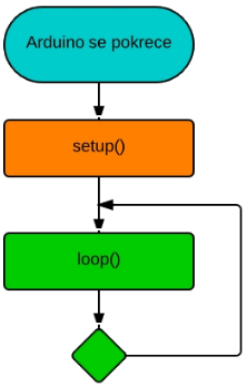
Neki od mogućih dodataka:

- TFT LCD screen (LCD monitor)
- SD card adapter (dodatna memorija)
- Mini USB/Serial adapter

C. *Programska podrška i programiranje*

Programiranje Arduino platforme radi se kroz posebno dizajnirano sučelje nazvano Arduino IDE (slika 2.) Inače sam programski jezik baziran je na AVR C programskom jeziku.

Slika 2. Arduino IDE



Slika 3. Dijelovi programskog koda

Arduino IDE ujedno služi za pisanje programa i za prenošenje samog programa na mikrokontroler. Dok kod većine drugih mikrokontrolera je potreban fizički programator, kod Arduina to nije potrebno. Na samom Arduinu nalazi se Arduino „bootloader“. Mali program koji služi za prenošenje programa na sam mikrokontroler bez nekakvog dodatnog hardvera.

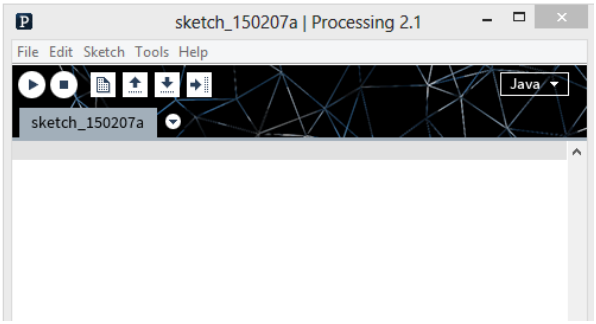
Sam program se sastoji od dva osnovna dijela(slika 3):

- Prvog podešavanja – setup()
- Program koji se izvršava stalno – loop()

Dio prvog podešavanja (setup()) predstavlja dio koda kojim se vrši podešavanje Arduino kontrolera (ulazno izlazne pinovi, definiranje vrste komunikacija kako sa računalom tako i sa ostalim komponentama). Drugi dio (loop()) je dio koda koji Arduino stalno ponavlja. Ovim ponavljanjem može se simulirati „pametan“ uređaj koji stalno osluškuje komande koje mu dolaze kroz ulazne komponente, te na taj način prati ponašanje okoline.

III. PROCESSING

Okruženje Processing jezika je Java. Programi napisani u Processing-u se prevode u Java programski jezik i pokreću kao Java aplikacija. Najčešće će se pokretati brže od programa baziranih na skriptim jezicima kao što je ActionScript, što je posebno važno kod mnogih grafičkih aplikacija.
Najveća razlika između Processing-a i Java programskog jezika je što Processing nudi jednostavno sučelje za programiranje koje ne zahtjeva napredne koncepte programiranja kao što su klase, objekti ili animacije, no ujedno omogućuje naprednim korisnicima njihovo korištenje.



Slika 4. Programsko okruženje Processing

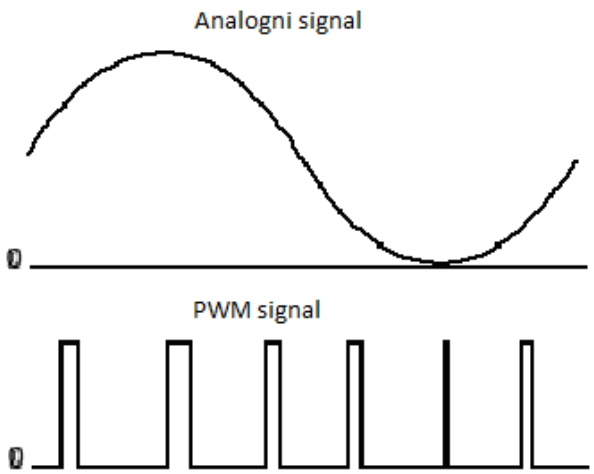
Zbog toga su programi kraći i čitljiviji. Processing koristi razne pristupe renderiranja tipografije uključujući i rasterska i vektorska rješenja. Povijesno gledajući, grafički dizajneri su tipografiju na Webu mogli oblikovati pomoću standardnih i sveprisutnih fontova (koji se mogu naći na računalima svih korisnika) ili koristeći kompliciranija, memorijski zahtjevnija te nepredvidiva rješenja. Processing omogućuje programiranje i prikazivanje tipografije na Web-u neovisno o korisničkom Web pregledniku i bazi fontova.

A. *Definicija Processinga*

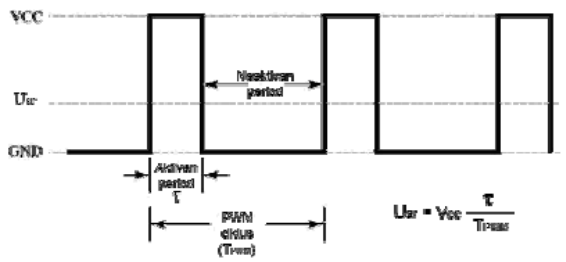
Processing povezuje koncepte programiranja sa principima vizualne forme, pokreta i interakcije. On integrira programski jezik i razvojno okruženje u jedinstveni sistem. Kreiran je za učenje osnova kompjuterskog programiranja unutar vizualnog konteksta. Jezik Processinga je specifično namijenjen za generiranje i modificiranje slika. Cilj mu je dostići ravnotežu između jasnoće (jednostavnosti) i naprednih mogućnosti, tako da je prilagođen i amaterima i naprednim programerima. Sistem omogućava korištenje i izradu kompjutorske grafike i interakcije uključujući vektorske/rasterske crteže, procesiranje slika, boje, događaje, mrežnu komunikaciju i objektno programiranje. Također ima mogućnost generiranja zvuka, slanja i primanja podataka u raznim formatima te slanja i primanja 2D i 3D datoteka. Processing je otvorenog formata što omogućuje svima da se njime bave i da ga razvijaju bez obzira na financijske mogućnosti.
Osnova jezika je Java koji također dijeli elemente sa C programskim jezikom te time omogućava brži razvitak i razumljivost programerima upoznatim tim programskim jezicima.

IV. PWM – PULSE WIDTH MODULATION

Impulsno širinska modulacija je tehnika kontroliranja analognih strujnih krugova preko digitalnih izlaza (najčešće mikrokotrolera). Napon, odnosno struja se dovode do analognog primaoca kao serija impulsa, a informacija o amplitudi analognog signala se predstavlja širinom (trajanjem) impulsa PWM signala (slika 5).



Slika 5. Analogni signal i PWM signal



Slika 6. Zavisnost srednjeg napona u PWM signalu

Analogni naponi se koriste za direktnu kontrolu uređaja, te se taj način pokazuje kao jako jednostavan i jasan način kontrole, ali nije uvijek praktičan i izvediv. Analogna kontrola ima razne mane kao što su promjenjive karakteristike komponenti tokom vremena, veća potrošnja, grijanje, osjetljivost na šum itd. Digitalnim kontroliranjem analognih signala smanjuje se potrošnja energije, imun na šum način upravljanja, te energetski efikasan način. Šum može utjecati na digitalni signal samo ako je toliko jak da promijeni logičku jedinicu na nulu ili obrnuto. Također mnogi današnji mikrokontroleri imaju ugrađene periferije za PWM modulacije što uvelike olakšava implementaciju.

Glavna prednost korištenja PWM signala je što signal ostaje digitalan cijelim putem od kontrolera do primaoca, tako da konverzija nije potrebna. PWM se smatra digitalnim signalom, jer u svakom trenutku vremena nivo signala je ili na nuli ili na maksimumu.

Zavisnost i računanje srednjeg napona (napona koji dolazi do uređaja kojim se upravlja) vidljiva je na slici 6.

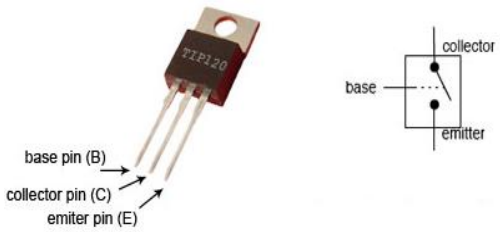
V. PROJEKT RGB TRAKA

Ideja projekta je pomoću računala, odnosno programa napisanog u programskom jeziku Processing upravljati Arduino platformom, odnosno RGB LED trakom spojenog na kontroler. Komunikacija računala i Arduina ostvarena je pomoću serijske veze.

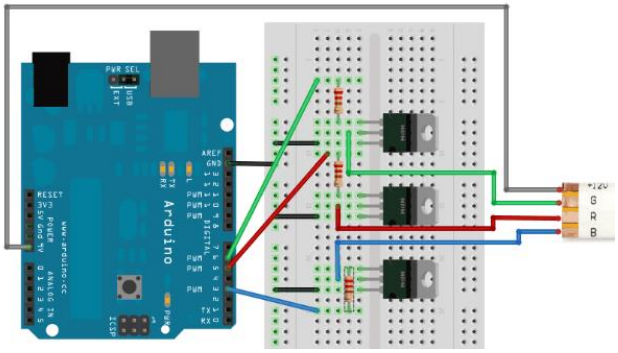
Komponente korištene u projektu:

- Arduino Uno (ATmega 328)
- RGB LED traka (12V DC, 14,4 W/m)
- Tranzistor TIP 120 (slika 8)
- Računalo (Windows 8)

Izlazni pinovi na Arduinu imaju ograničenje na 5V, a RGB traka je na 12V, te je korišten tranzistor TIP120



Slika 8. Tranzistor TIP 120



Slika 9. Shema spajanja

Shema spajanja RGB LED trake prikazana je na slici 9. Traka je direktno spojena na plus 12V DC, a upravljanje pojedinom bojom (crvena/zelena/plava) vrši se tako da se minus pol dovede preko Arduina (pomoću tranzistorske sklopke). Za tranzistorsku sklopku potreban je i otpornik vrijednosti 10 kΩ. Osim tranzistora moguće je koristiti i MOSFET.

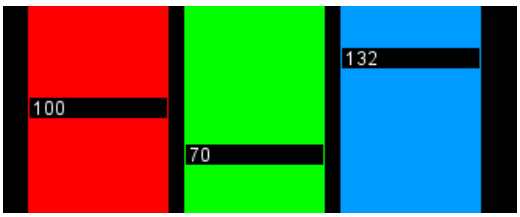
A. *Arduino programski kod*

```
void setup(){
  // deklariranje serijske veze na 9600
        Serial.begin(9600);

  // izlazni pinovi
        pinMode(5, OUTPUT); // crvena
        pinMode(6, OUTPUT); // zelena
        pinMode(3, OUTPUT); // plava

  // još jedan izlazni pin kao GND
        pinMode(8, OUTPUT); // GND
        digitalWrite(8, LOW);
}
void loop() {
  // poziv funkcije GetFromSerial()
        switch(GetFromSerial()){
         case 'R':
                analogWrite(5, GetFromSerial());
                break;
          case 'G':
                analogWrite(6, GetFromSerial());
                break;
          case 'B':
                analogWrite(3, GetFromSerial());
                break;
       }
}
// čitanje serijskog porta
int GetFromSerial(){
       while (Serial.available()<=0) {
}
return Serial.read();
}
```

Izlazni pinovi koji se koriste su PWM izlazi (Pulse Width Modulation). To nisu klasični digitalni pinovi već digitalni pinovi koji mogu poprimiti 255 mogućih stanja i ovisno o tome imaju analognu vrijednost napona.



Slika 10. Pprosessing - upravljanje

Vrijednost 0 je 0V, dok je 255 jednako 5V. Vrijednosti između 0 i 255 daju različite vrijednosti napona. To se postiže širinom 0 ili 1 intervala u određenim vremenskim razmacima.

B. Processing programski kod

```
import processing.serial.*;
Serial port;
sliderV sV1, sV2, sV3;
color cor;

void setup() {
      size(500, 500);
      println("Available serial ports:");
      println(Serial.list());

 // provjera otvorenih portova na računalu
      port = new Serial(this, Serial.list()[1], 9600);

 // kreiranje 3 instance sliderV klase
       sV1 = new sliderV(100, 100, 90, 255,
       #FF0000);
       sV2 = new sliderV(200, 100, 90, 255,
       #03FF00);
       sV3 = new sliderV(300, 100, 90, 255,
       #009BFF);
}

//crtanje
void draw() {
       background(0);
       sV1.render();
       sV2.render();
       sV3.render();

  // slanje „sync“ znaka
  // slanje željene vrijednosti
      port.write('R');
      port.write(sV1.p);
      port.write('G');
      port.write(sV2.p);
      port.write('B');
      port.write(sV3.p);
}

//Slider klasa
class sliderV {
      int x, y, w, h, p;
      color cor;
      boolean slide;

      sliderV (int _x, int _y, int _w, int _h,
      color _cor) {
      x = _x;   y = _y;
      w = _w;  h = _h;
      p = 90;   cor = _cor;
      slide = true;
 }
 void render() {
      fill(cor);
      rect(x-1, y-4, w, h+10);
      noStroke();
      fill(0);
      rect(x, h-p+y-5, w-2, 13);
      fill(255);
      text(p, x+2, h-p+y+6);
      if(slide==true&&mousePressed==true&&mouse
        X<x+w && mouseX>x)
      {
      if((mouseY<=y+h+150) && (mouseY>=y-150))
      {
            p = h-(mouseY-y);
            if (p<0) {
                p=0;
            }
            else if (p>h) {
                p=h;
            }
      }
     }
   }
 }
```

Rezultat izvršavanja cijelog processing koda vidljiv je na slici 10. Pomicanjem klizača sa vrijednošću na pojedinoj boji šaljemo naredbe Arduinu za pojedinu boju. Osim 3 osnovne boje, moguće je dobiti paletu boja tako da kombiniramo intenzitete 3 osnovne (crvena/plava/zelena).

VI. ZAKLJUČAK

Dovršenje jednog projekta je u biti gotovo uvijek početak nekog sljedećeg u smislu nadogradnje postojećeg novim idejama i mogućnostima. Komunikacija mikrokotrolerima putem mreže nudi mogućnost upravljanja putem udaljenih računala, mobilnih ili tablet uređaja što i ovom opisanom projektu izgleda kao super mogućnost i vrlo vjerojatno sljedeći korak. Rad sa mikrokontrolerima je uvijek zanimljiv i dinamičan, te je samo mašta granica.

VII. LITERATURA

[1] P. Zenzerović, Arduino kroz jednostavne primjere, Hrvatska zajednica tehničke kulture, Zagreb, 2014.

[2] M. Benzi, Getting Started with Arduino, 2nd Edition, O' Reily Media, 2011.

[3] Brock Craft, Arduino projects for dummies, Wiley, 2013

[4] http://www.arduino.cc/, 2015.

[5] https://processing.org/

UČENJE PROGRAMIRANJA KROZ ROBOTIKU

Petar Dobrić* i Maja Mačinko Kovač**
*OŠ Julija Klovića/Nova cesta 133, Zagreb, Hrvatska
**OŠ Eugen Kvaternik/Školska 4, Velika Gorica, Hrvatska
*e-mail: dobripero@gmail.com
**e-mail: xjamax@gmail.com

Sažetak – Učenje programiranja danas djeci često predstavlja nešto strano i nerazumljivo. Programiranje je budućnost i jedan od načina kako ga učiniti zanimljivim je učenje kroz igru. Ovim smo radom pokazali kako učiti programiranje s praktičnom primjenom u robotici pomoću programskog jezika RoboPro te ga usporedili s programiranjem u jeziku Basic koji učenici već poznaju. Pokazali smo praktične primjere pomoću kojih odmah vide učinak napisanih programa i automatizaciju istih. Bez unapređivanja načina kako mlade privući programiranju, nema ni digitalne revolucije u obrazovanju. A učenje robotike kroz igru je upravo ono što nam je potrebno.

I. UVOD

Učenje programiranja u osnovnoj školi predstavlja izazov za učitelje zbog činjenice da današnji učenici imaju velikih problema u usvajanju osnovnih konceptualnih i problemskih načina razmišljanja prilikom algoritamskog načina rješavanja zadataka. Usvajanje programskih jezika (LOGO, BASIC, Phyton) tijekom osnovnoškolskog razdoblja kod velike većine učenika stvara animozitet i otpor. Do pozitivnih načina i metoda pri učenju programiranja, kako bi prve programerske korake učenici lakše usvojili, došli smo kroz izvannastavnu aktivnost robotike. Programiranje u programu RoboPro razotkriva uobičajeno mišljenje o teškoći istog. Objektno orijentirani princip učenja je konceptualno jednostavan način usvajanja algoritamskog razmišljanja tijekom rješavanja problemskih zadataka te kao takav olakšava proces učenja programiranja. Učenici odmah vide rezultat svog programa te ga u slučaju pogreške u kratkom vremenu lako korigiraju. Izvođenje zadanog problemskog zadatka odvija se pomoću sučelja koje je povezano s računalom i ulazno-izlaznim elementima (žaruljice, motori, različiti senzori...). Naše iskustvo je da učenje programiranja kroz igru daje odlične rezultate i olakšava rad učenicima i učiteljima na zadovoljstvo svih sudionika obrazovnog procesa. Ovakav način poučavanja kod učenika razvija vještinu samostalnog spajanja i programiranja modela robota te poboljšava učenikove psihomotoričke vještine. Pobuđivanje interesa za istraživanjem i rješavanjem problemskih zadataka, potiče algoritamski način razmišljanja.

II. ZAŠTO BAŠ ROBOTIKA I KOJI JEZIK IZABRATI

Mnogi se pitaju zašto učiti osnove programiranja pomoću robota umjesto tradicionalnih metoda učenja. U posljednjih nekoliko godina vidjeli smo ubrzano širenje uređaja koji sakupljaju i upravljaju informacije u realnom vremenu putem nekog sučelja i komponente među njima, kao što su roveri na Marsu, ABS sustavi, umjetni udovi i operacije, pa čak i IPodi. Kako ti uređaji postaju sveprisutni, slobodoumno obrazovana osoba trebala bi biti upoznata s načinima kako takvi uređaji rade, kao i s njihovim mogućnostima i ograničenjima.

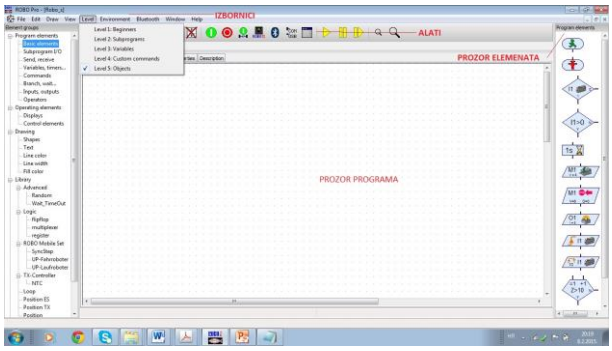
Odgovarajući jezik, jezične strukture i arhitekture za razvoj softvera robota oduvijek su tema rasprava još od prvih računalno upravljanih robotskih sustava. Činjenica da ne postoji univerzalni dogovor ne čudi s obzirom na probleme kojima se jezik susreće. Prvi dio problema odnosi se na činjenicu da se robot stvara na nestrukturiranom, dinamičnom svijetu koji se brzo mijenja i isto tako mora biti u mogućnosti efektivno reagirati na njega. Na osnovnoj razini to implicira da bi jezik trebao olakšati upravljanje algoritama koji se baziraju na radu senzora. Isto tako, bilo koji takav algoritam mora moći biti prekinut u bilo koje vrijeme. Idealno, dobar robotski program trebao bi omogućiti jednostavan skup naredbi za opisivanje i kombiniranje kontroliranog ponašanja. Drugi dio problema proizlazi iz prirode zadataka koji bi taj robot izvršavao, a posljednji problem javlja se oko praktičnih problema razvoja robotskog softvera. Ono s čime se i sami učenici često susreću jest činjenica da je programiranje robota veoma eksperimentalno u prirodi, često ne znamo kako u početku najbolje namjestiti senzore, pogone i ostalo. Često je potrebno puno testiranja, dorade i naknadnog programiranja da bi robot idealno radio. Potrebno je pronaći jezik koji bi dovoljno brzo izvršavao zadane algoritme, upravljao protokom istih, a da sve odvija dovoljno dinamično i zadani zadaci se izvrše.

RoboPro je vrlo jednostavan programski softver. Čak i početnici mogu jednostavno početi programiranje korištenjem dijagrama toka i raznih građevnih blokova. Osim toga, razmjena informacija koristi se varijablama i grafičkim vezama. To omogućava prikaz programske funkcije na jasan i razumljiv način. Grafički programski jezik RoboPro pruža sve elemente modernog programskog jezika, kao što su polja, funkcije, objekti, rekurzije što je koristan alat čak i za profesionalne programere. Programi su prevedeni izravno u strojni jezik za učinkovito izvršenje čak i najsloženijih programa. Programski jezik ne poznaje

granice i lako je pisati programe ili razmjenjivati podatke s ostalim Windows softverima.

III. ROBOPRO SUČELJE

Automatizacija i upravljačka tehnologija prisutna u svim područjima ljudskog djelovanja olakšava rad i ubrzava napredak naše civilizacije. RoboPro sučelje omogućava potpunu kontrolu pri upravljanju Fishertechnick modelima. Intuitivnost i jednostavnost omogućava brzo učenje logike programiranja kroz kontrolne programe ili dijagrame tijeka. Takvi programi za kontrolu i upravljanje robotima mogu se sastaviti i pomoću programskih jezika opće namjene kao što je npr. Qbasic ili ROBOPro. Grafičko korisničko sučelje omogućava odabir ikona pomoću miša gdje slažemo program u obliku dijagrama tijeka. Kako bismo pomoću računala mogli upravljati robotom, potreban je elektronički sklop - sučelje, koje povezuje računalo i robota. Napisani program prenosimo putem međusklopa - sučelja koje prenosi naredbe modelu robota. Sučelje na računalo spajamo pomoću USB kabela koji prenosi programe s računala u memoriju robota. Osim RoboPro programa, na računalo moramo instalirati i upravljačke programe koji omogućavaju dvosmjernu komunikaciju između računala i modela robota. Sučelje napajamo akumulatorskom baterijom napona 9V ili ispravljačem.

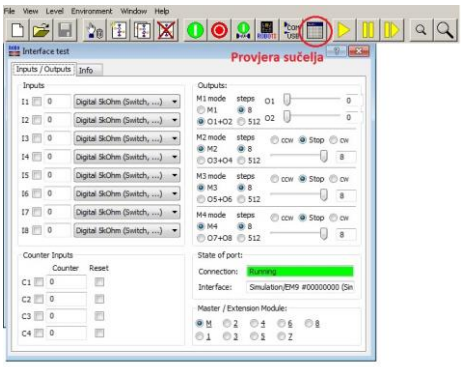


Slika 1. Sučelje programa RoboPro

Podešavanjem sučelja definiramo način rada i odabir vrste sučelja. Program podržava rad sa spojenim elementima (lampice, motori, tipkala i ostali senzori) na sučelje i bez sučelja – simulacija. Ovaj način rada je pogodan ako nemamo vremena za spajanje elemenata ili ako programiramo izvan škole.

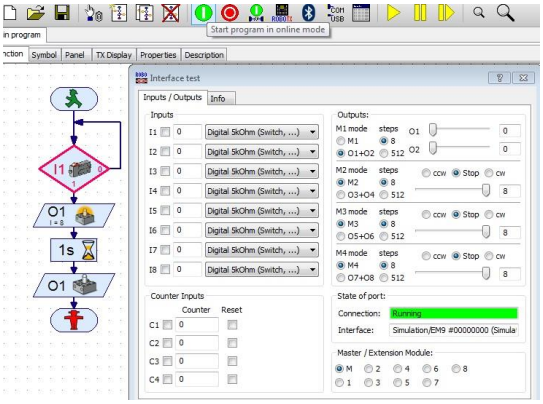
Slika 2. Postavke izlaznog sučelja

Provjera rada izlaznog sučelja i veze s računalom omogućava pravovremeno otklanjanje grešaka (nepravilno spojeni elementi, neispravan USB kabel, pad napona na bateriji ispod 9V) na početku rada. Zelena traka u dnu prozora pokazuje da je sučelje pravilno spojeno.



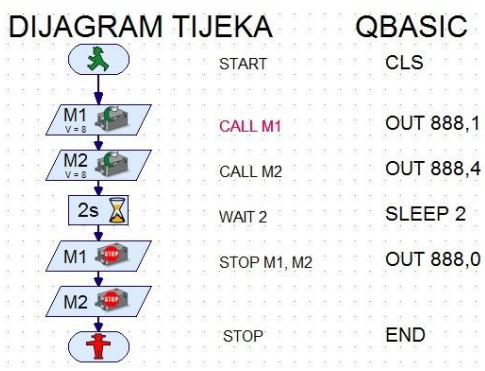
Slika 3. Provjera ispravnosti veze sučelja i računala

Dijagram tijeka sastoji se od programskih elemenata koji se umeću i povezuju u prozoru programa pomoću miša (tehnika povuci – pusti). Grafički prikaz tijeka programa olakšava praćenje tijeka programa i učenje logičkog pristupa programiranju.



Slika 4. Dijagram tijeka i provjera rada programa

Primjer pokretanja robotskog vozila u programima Qbasic i RoboPro zorno prikazuje jednostavnost logike izrade programa. Pokretanjem programa motori M1 i M2 se pokrenu i vozilo se kreće naprijed dvije sekunde. Motori se zaustavljaju i slijedi završetak programa. Učenici vide učinak napisanih naredbi i njihovu ulogu u automatiziranom procesu. Spremanjem programa u memoriju sučelja vozilo pokrećemo neovisno o fizičkoj vezi računala i robota. Grafičko sučelje olakšava početno zanimanje za programiranje i rješavanje problemskih zadataka već kod učenika nižih razreda.

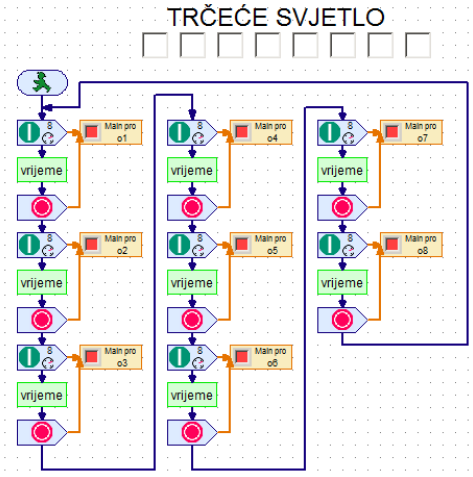


Slika 5. Program pokretanja robotskog vozila

IV. PRAKTIČNA PRIMJENA

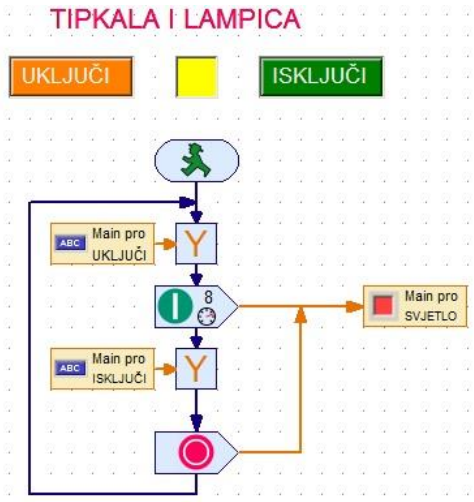
Rješavanje problemskih situacija ostvarivo je bez opreme koja je jako skupa i nedostižna većini škola. Podešavanje programa na modul simulacije donosi novu dimenziju u učenju i razumijevanju programiranja. RoboPro podržava i rad s podprogramima, čime je dobivena preglednost i lakše snalaženje u praćenju tijeka programa i pronalaženju mogućih grešaka.

Primjer zadatka „trčeće svjetlo“ pokazuje jedno od mogućih rješenja. Lampice je potrebno paliti i gasiti zadanim redoslijedom (O1 do O8), u pravilnim vremenskim razmacima. Podprogram „vrijeme“ pojednostavljuje promjenu vremena paljenja i gašenja lampica.



Slika 6. Simulacija – program trčeće svjetlo

Primjer uključivanja i isključivanja svjetla pomoću tipkala (uključi, isključi) riješen je pomoću simulacije. Tipkalo „uključi“ pali svjetlo, a „isključi“ gasi svjetlo. Tipkala su povezana ulaznim panelima, čime je omogućena jednostavna kontrola zadanog modela. Jedan panel je pridružen svakom programu ili podprogramu. Prvo moramo kreirati panel sa svim elementima i tek tada možemo povezati panel s panel programskim elementom.



Slika 7. Simulacija – program uključi, isključi svjetlo

V. ZAKLJUČAK

Ovim načinom rada učenicima je omogućeno da vide praktičnu primjenu svih naredbi koje u nastavi informatike uče i koja im je jedna od mogućih uloga. Samim programiranjem bez dodatnih stimulacija, učenicima učenje zna postati zamorno i besmisleno, a na ovaj smo im način omogućili da kroz zabavan način zavole programiranje i sve mogućnosti koje ono pruža. RoboPro sučelje nije jedini mogući način kojime se uči kroz igru, no svakako je jedan od jednostavnijih i dostupnijih načina kako to postići. Logika programiranja vrlo je slična programskom jeziku Basic kojeg uče u nastavi informatike, i samim time učenicima je lakše pokazati jedan od mogućih načina praktične primjene istoga.

Infomatika je budućnost, i ako ne pronađemo načine kako učenicima približiti i pojednostavniti njeno učenje, teško ćemo zadobiti njihovu pažnju i volju za učenjem. Programiranje nije samo za računalne fanatike ili one koji se opredjeljuju za karijeru u IKT sektoru. Cijela paleta kreativnih i zanimljivih poslova ovisi o poznavanju programiranja, kao što je filmska industrija, analiza medicinskih podataka, izrada sigurnosnih softvera… Programiranje je nit koja se provlači kroz sve buduće profesije, i moramo pronaći načine kako ga učenicima učiniti zabavnim i korisnim.

LITERATURA

[1] F. Pavlović, Elektronika i robotika, Profil, Zagreb, 2008.

[2] Z. Kovačić, S. Bogdan, V. Krajči, Osnove robotike, Graphis, Zagreb, 2002.

[3] H. Vrhovski, I. Kolarić, Priručnik robotike za mentore i učenike, Hrvatska zajednica tehničke kulture, Zagreb, 2010.

[4] P. Crookall, Programiranje za početnike, Niro „Tehnička knjiga“, Beograd, 1983.

[5] Lj. Vidanović, Zbirka zadataka za natjecanje iz programiranja u Qbasicu, Alfa, 2007.

Primjer implementacije e – učenja u nastavi Informatike

Jozica Pažanin*, Marko Rosić**, Ani Grubišić**
*Ekonomsko – birotehnička škola, Split, Hrvatska
**Prirodoslovno-matematički fakultet, Split, Hrvatska
jozica.pazanin@gmail.com

Sažetak – U cilju unapređenja kvalitete procesa poučavanja, upoznavanja suvremenih modela učenja važnih za cijeloživotno obrazovanje te razvoja digitalnih kompetencija učenika, kroz nastavu predmeta Informatika određeni nastavni sadržaji realizirani su konceptom e-učenja. E-učenje svojim klasifikacijama pruža različite načine implementacije u nastavni proces. U ovom radu prikazana je implementacija e-učenja kombiniranim (hibridnim) modelom nastave putem dva e-kolegija: „Sigurnost na Internetu“ i „Matematičke osnove računala“. E-kolegij „Matematičke osnove računala“ korišten je kao dopuna klasičnoj nastavi i predstavlja popratni nastavni materijal obrađenoj nastavnoj jedinici u učionici. E-kolegij „Sigurnost na Internetu“ sadrži cjelovitu nastavnu jedinicu, te je njegova realizacija koncipirana na način da se učenike tijekom sata u učionici upućuje u nastavni sadržaj koji oni potom u potpunosti usvajaju u online okruženju. U radu su navedeni rezultati istraživanja provedenog eksperimenta u kojem jedna skupina učenika nastavnu jedinicu usvaja modelom e-učenja putem sustava Loomen, a druga skupina putem klasične nastave u učionici. Rezultati dobiveni eksperimentom pokazuju usporedbu postignutih obrazovnih ciljeva zadane lekcije u klasičnom i online modelu učenja. Usporedba ostvarenih rezultata obrađene lekcije, pokazuje zadovoljava li koncept e-učenja postizanje obrazovnih ciljeva. U radu su prikazani i rezultati anketnog upitnika koji pokazuju stavove učenika o e-učenju, te njihovu spremnost i motiviranost za prihvaćanjem novih suvremenih modela nastave.

I. UVOD

E-učenje kao proces obrazovanja (proces učenja i podučavanja) koji se izvodi uz uporabu nekog oblika informacijske i komunikacijske tehnologije, s ciljem unapređenja kvalitete toga procesa i kvalitete ishoda obrazovanja [1] omogućava različite načine integriranja u redovni nastavni proces. Kako e-učenje podrazumijeva različite aspekte korištenja informacijsko-komunikacijske tehnologije ovisno o razini zastupljenosti, intenzitetu i načinu korištenja tehnologije u nastavi proizlazi klasifikacija e-učenja. [2]

Tablica 1. prikazuje proširenu klasifikaciju e-učenja koja podrazumijeva korištenje tehnologije u nastavi.

U Ekonomsko-birotehničkoj školi u Splitu kroz nastavu Informatike u hibridnom modelu nastave implementirana su dva e-kolegija: „Matematičke osnove računala“ i „Sigurnost na Internetu“.

TABLICA 1. KLASIFIKACIJA E-UČENJA

Kategorija	**Opis**	**Način korištenja ICT-a**
Klasična nastava (face2face – f2f)	Klasična nastava – nastava u učionici	Tehnologija se uglavnom koristi na razini upotrebe programa za obradu teksta (npr. pisanja priprema u Word-u).
Nastava uz pomoć ICT-a	Tehnologija se uglavnom koristi da bi se poboljšala klasična nastava	Primjer korištenja ICT-a: Power Point prezentacije, programi za testiranje (kvizovi za samoprovjeru i provjeru znanja), društvene mreže, Web stranice za kolegije
Hibridna (kombinirana) nastava	Kombinacija klasične nastave u učionici i nastave uz pomoć tehnologije (online nastave)	LMS (*eng. Learning Management Systems*) - sustavi za upravljanje učenjem
Online nastava	Nastava se odvija isključivo u online okruženju (na f2f nastave)	E-kolegiji se „dostavljaju“ putem LMS-a

U navedenim kolegijima hibridni model nastave predstavlja kombinaciju klasične nastave u učionici i online nastave putem sustava za upravljanje učenjem Loomen (hrvatska inačica sustava Moodle).

E-kolegij „Matematičke osnove računala“ koristi se u prvom razredu ekonomske gimnazije, kao dodatni nastavni sadržaj koji upotpunjuje klasičnu nastavu u učionici. Kolegij se realizira na način da učenici tijekom nastave u učionici usvajaju nove nastavne sadržaje a potom u online okruženju e-kolegija koriste popratne sadržaje (prezentacije i digitalne knjižice sa obrađenim sadržajem nastavne jedinice, primjere riješenih zadataka i kvizove za samoprovjeru znanja) te ponavljaju i utvrđuju gradivo obrađene nastavne jedinice. Osim kao dopuna klasičnoj nastavi u učionici, e-kolegij se koristi i kao samostalna cjelina u online modelu nastave te u tom obliku služi učenicima za pripremanje državne mature iz Informatike i natjecanja „Osnove Informatike“.

E-kolegij „Sigurnost na Internetu" se provodi već tri godine među učenicima prvog razreda ekonomskog usmjerenja. Kolegij se provodi na način da se učenici na

satu u učionici samo upućuje u nastavni sadržaj koji potom samostalno usvajaju u online okruženju. E-kolegij sadrži sljedeći niz aktivnosti i resursa za usvajanje nastavnog sadržaja:

- digitalnu knjigu s objašnjenim sadržajem lekcije, unutar koje je inkorporirano niz zadataka, pdf datoteka i kvizova za ponavljanje
- rječnik koji sadrži nepoznate pojmove iz lekcije
- interaktivne igre (milijunaš, križaljka, vješalo) za ponavljanje gradiva
- forum za komunikaciju sa ostalim polaznicima i komunikaciju s mentorom
- linkove na dodatne izvore znanja.

Detaljno obrazložen sadržaj e-kolegija „Sigurnost na Internetu“, sa definiranim ciljevima i detaljno opisanim zadacima prikazan je u radu [3] a dalje u ovom radu, ovaj e-kolegij biti će izložen u kontekstu provedenog eksperimentalnog istraživanje o efikasnosti e-učenja putem LMS sustava.

II. ISTRAŽIVANJE EKSPERIMENTOM

Tijekom veljače 2015. god. provedeno je eksperimentalno istraživanje u kojem je tema „Sigurnost na Internetu“ među dvije skupine učenika (dva razreda) obrađena različitim modelima nastave. Jedna skupina učenika (eksperimentalna skupina – 18 učenika) nastavnu jedinicu usvajala je modelom e-učenja putem sustava Loomen, a druga skupina (kontrolna skupina – 18 učenika) putem klasične nastave u učionici. U nastavku su izloženi cilj i zadaci ovog istraživanja kao i analiza postignutih rezultata.

Cilj istraživanja

Ispitati, s obzirom na upotrebu različitih modela učenja istog nastavnog sadržaja, postoji li razlika u postizanju obrazovnih ciljeva obrađene nastavne jedinice tj. postoji li razlika u postignutim rezultatima između klasične nastave u učionici i online nastave putem sustava Loomen.

Zadaci istraživanja

- Formirati eksperimentalnu i kontrolnu skupinu učenika. Ove dvije skupine moraju biti ekvivalentne u razini predznanja njihovih članova vezano za predloženu temu. Stoga je potrebno provesti i inicijalni test predznanja o predloženoj temi kako bi se formirale prikladne skupine učenika.
- Realizirati nastavu iz preložene teme. Nastavu realizirati putem klasičnog načina u učionici (kontrolna skupina) i pomoću sustava za e-učenje – Loomen (eksperimentalna skupina).
- Provedba završnog testa kojim se ispituju ostvareni obrazovni ciljevi obrađene lekcije u eksperimentalnoj i kontrolnoj skupini.

Tijek istraživanja

Nakon provedenog inicijalnog testa, učenici su upućeni u obradu novih nastavnih sadržaja na način da je eksperimentalna skupina obrazovne sadržaje usvajala u kombiniranom modelu nastave, a kontrolna skupina putem klasične nastave u učionici. U nastavku se ukratko izlaže realizacija nastave u obje skupine.

Realizacija nastave u eksperimentalnoj skupini učenika

Učenici eksperimentalne skupine usvajaju nastavne sadržaje na način da ih tijekom nastave u učionici nastavnik upućuje u nastavni sadržaj te im zadaje tjedne zadatke koje oni potom samostalno rješavaju u online okruženju putem e-kolegija „Sigurnost na Internetu“. Učenici unutar e-kolegija pristupaju različitim aktivnostima putem kojih pregledavaju nastavni sadržaj koristeći prezentaciju, proučavaju sadržaj lekcije putem digitalne knjige i rječnika te ponavljaju nastavno gradivo putem interaktivnih igara.

Slika 1. Okruženje za realizaciju nastave u eksperimentalnoj skupini

Realizacija nastave u kontrolnoj skupini učenika

Učenicima kontrolne skupine je tijekom klasične nastave u učionici putem PowerPoint prezentacije prezentiran sadržaj nastavne jedinice. U zadnjem dijelu sata obrađeni sadržaj je ponovljen putem interaktivne križaljke. Da bi nastavni materijal za učenje (detaljno razrađen sadržaj nastavne jedinice) bio odgovarajući kao i u eksperimentalnoj skupini, učenicima iz kontrolne skupine podijeljeni su nastavni materijali za učenje identični materijalima iz eksperimentalne skupine. Svi dodatni materijali za učenje (linkovi na kvizove i dodatne izvore znanja) učenicima kontrolne skupine objavljeni su na web stranici škole.

Nakon obrađenog nastavnog sadržaja među učenicima eksperimentalne i kontrolne skupine provedena je pisana provjera znanja testom kojim se provjerava usvojenost obrazovnih sadržaja.

Prikaz i analiza rezultata

Rezultati inicijalnog testa i testa provjere znanja obrađeni su u programi Microsoft Office Excel 2007.

Uspoređujući rezultate inicijalnog testa T-testom (t = - 0,6896) se dobilo da se eksperimentalna i kontrolna grupa statistički značajno ne razlikuju, tj. da su ekvivalentni jer je $|t| < 1,684$ (granična vrijednost). Interpretacija rezultata napravljena je u skladu s uputama [4] i [6].

U tablici 2. prikazani su rezultati obrade nad razlikom rezultata završnog i inicijalnog testa.

TABLICA 2. REZULTATI STATISTIČKE OBRADE

TEST (= Test provjere (završni test) – Inicijalni test)		
	Kontrolna grupa	Eksperimentalna grupa
Broj učenika	18	18
Aritmetička sredina	31,94	38,47
Standardna devijacija	13,3822	13,4803
T-test	-1,458	

Izračun standardizirane veličine učinka inicijalnog i završnog testa proveden je prema Glass-u[5], tj. razlika aritmetičkih sredina eksperimentalne i kontrolne grupe je podijeljena sa standardnom devijacijom kontrolne grupe. Rezultat razlike između veličine učinka inicijalnog i završnog testa predstavlja da je eksperimentalna grupa bolja od kontrolne grupe za 0,4878 standardne devijacije.

Obradom podataka, kojom se utvrđivalo postoji li statistički značajna razlika između eksperimentalne i kontrolne grupe pokazalo se da ovim eksperimentom nema statistički značajne razlike jer je $|t| <$ granične vrijednosti ($|-1,458| < 1,684$) .

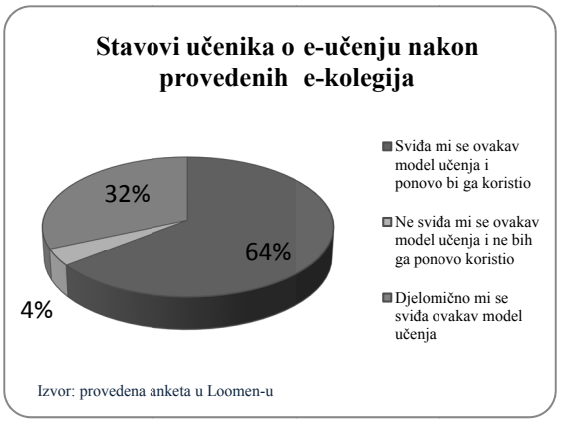
III. ISTRAŽIVANJE ANKETNIM UPITNIKOM

Anketnim upitnikom su ispitani stavovi učenika o e-učenju. Upitniku koji je proveden u sklopu e-kolegija „Sigurnost na Internetu“ pristupilo je 20 učenika. Tablica 3. prikazuje izdvojena tri pitanja specifična za tematiku ovog rada.

TABLICA 3. REZULTATI ANKETNOG UPITNIKA ROVEDENOG U EKSPERIMENTALNOJ SKUPINI

Lakoća korištenja sustava Loomen s obzirom na kompleksnost sustava i informatičko predznanje Skala od 1-5 gdje je 1-izuzetno teško, 5 –izuzetno lako.		4,2
Karakteristika koji smatrate najkorisnijom za učenje putem Loomen-a	Mogućnost ponavljanja gradiva uz pomoć interaktivnih igara (milijunaš, križaljka, vješalo)	70%
	Komunikacija sa ostalim polaznicima i nastavnikom	20%
	Lako su dostupni izvori za učenje	10%
Kad biste mogli birati koji način učenja bi odabrali	E-učenje bez nastavnika	10%
	Tradicionalnu nastavu s nastavnikom	10%
	Kombinaciju e-učenja i klasične nastave s nastavnikom	80%

Tijekom školske godine 2013./2014. proveden je slični anketni upitnik o e-učenju na uzorku od 152 učenika Ekonomsko-birotehničke škole Split koji su koristili e-učenje u svojem obrazovanju. Od ukupnog broja anketiranih učenika njih 103 su bili sudionici e-kolegija „Sigurnost na Internetu (učenici prvih razreda ekonomskog usmjerenja), 32 učenika su sudionici e-kolegija „Izrada marketing plana“ (učenici maturalnih razreda ekonomskog usmjerenja) i 17 učenika su bili sudionici e-kolegija „Matematičke osnove računala“ (učenici prvog razreda ekonomske gimnazije). Rezultati provedene ankete bile su osnova za izradu priloženog grafikona (Slika 2.) koji prikazuje zainteresiranost učenika za ovakav model učenja.



Slika 2. Prikaz razmišljanja učenika o e-učenju

IV. ZAKLJUČAK

Rezultati provedenog eksperimenta pokazuju da ne postoji statistički značajna razlika u postizanju obrazovnih ciljeva klasičnim modelom učenja u učionici u odnosu na online model e-učenja putem LMS-a. Iz toga možemo zaključiti da se konceptom e-učenja mogu u potpunosti postignuti planirani obrazovni ciljevi. Tumačenjem ovih rezultata dolazi se do zaključka da e-učenje:

- u pojedenim nastavnim sadržajima može biti i zamjena klasičnoj nastavi
- može predstavljati izvrsnu dopunu klasičnoj nastavi u učionici.

Brojne su prednosti koje u takvom modelu učenja doprinose kvaliteti procesa obrazovanja npr. mogućnosti ponavljanja gradiva samo-provjerama putem različitih interaktivnih kvizova čime učenici mogu pratiti vlastiti napredak, dostupnost drugih izvora znanja, postizanje interakcije putem komunikacijskih alata itd.

Rezultati anketnog upitnika pokazuju da su učenici spremni za učenje uz pomoć informacijske i komunikacijske tehnologije te da im je ovakav pristup nastavi ne samo prihvatljiv nego i zanimljiv. Izuzetno veliki postotak učenika (80%) izabralo bi obrazovanje kombiniranim modelom.

Učenici kroz kombinirani program e-učenja i klasične nastave osim što uspješno prate planirani nastavni plan i program pojedinog predmeta, unapređuju informatičku pismenost te razvijaju vlastite digitalne kompetencije. Ovakvim modelom nastave učenici upoznaju e-učenje kao izvrstan način za cijeloživotno obrazovanje i neprestano stručno usavršavanje neophodno na današnjem tržištu rada.

LITERATURA

[1] E-učenje, (pregledano 23.02.2015.) Dostupno na < http://www.srce.unizg.hr/ceu/o-centru/e-ucenje/ >

[2] Xxx, Klasifikacija e-učenja, (pregledano 23.02. 2015.) Dostupno na < http://ahyco.uniri.hr/metodika/e_learning.htm >

[3] J. Pažanin, S. Mikulić, Implementacija e-učenja u strukovnom obrazovanju, zbornik radova - CUC 2014

[4] A. Grubišić (2006) Vrednovanje učinka inteligentnih sustava e-učenja, Fakultet elektrotehnike i računalstva, Magistarski rad

[5] G.V. Glass, B.McGaw, and M.L.Smith, (1981) Meta-Analysis in Social Research. London: Sage.

[6] B. Petz, Osnovne statističke metode za nematematičare , Naklada Slap

Digitalna zbirka zadataka s evaluatorom programskih rješenja

Vinkoslav Galešev, Robert Šipek
Udžbenik.hr, Zagreb, Hrvatska
vgalesev@enter.bug.hr, robert@bug.hr

Jedna od glavnih prednosti on-line digitalne zbirke zadataka je njena fleksibilnost, odnosno mogućnost dorade i prilagodbe sadržaja potrebama nastavnog procesa te interaktivnost s korisnikom. Također omogućava i neke nove mogućnosti posebice važne u didaktičke svrhe, a koje nije moguće realizirati klasičnom zbirkom zadataka u tiskanom obliku.

Za ostvarenje didaktičkih ciljeva digitalna zbirka zadataka treba sadržavati i mogućnost evaluacije programskih rješenja. Iako već postoje aplikacije za evaluaciju programskog koda, gotovo u pravilu nisu namjenski kreirane za potrebe nastave. Zato smo razvili vlastitu aplikaciju koja objedinjuje digitalnu zbirku zadataka i evaluator programskih rješenja pisanih u programskom jeziku Python. Aplikacija *Eva* koju opisujemo u ovome radu namijenjena je kako samostalnom vježbanju učenika, tako i podrški nastavnicima u organiziranju i provođenju nastavnog procesa.

I. UVOD

Unazad par godina programski jezik Python počeo se uvoditi u nastavne programe pojedinih srednjih škola u nas. U skladu s time, izdani su i prvi udžbenici koji obrađuju ovaj programski jezik. Slijedom toga, ukazala se potreba i za izradom odgovarajuće zbirke zadataka. Budući da se radi o predmetu informatike, svrsishodno je postaviti temeljno koncepcijsko pitanje - u kojem medijskom obliku realizirati takvu zbirku?

Nedostatak je klasične zbirke zadataka u tiskanom obliku što je statična, tj. nepromjenjiva, barem do sljedećeg reizdanja. Nasuprot tome, jedna od glavnih prednosti on-line digitalne zbirke je njena fleksibilnost, odnosno razmjerno jednostavna mogućnost dorade i prilagodbe njenog sadržaja i oblika. Osim spomenute prednosti, on-line digitalna zbirka zadataka omogućava i neke nove didaktičke elemente primjene u nastavnom procesu, o čemu ćemo nešto više reći u nastavku.

U klasičnoj tiskanoj zbirci, pod pretpostavkom da sadrži i rješenja zadataka, povratna se informacija većinom svodi na dihotomni ishod – rješenje je točno ili nije. Analiziramo li funkcionalnost zbirke zadataka iz programiranja, potrebno je uzeti u obzir da kod programskih rješenja postoji dodatna specifičnost – zadatak se gotovo u pravilu može riješiti na više, manje ili više učinkovitih, načina. Pronalaženje ne samo točnog nego i optimalnog programskog rješenja od izuzetne je važnosti za usvajanje vještine programiranja. U tom pogledu digitalna zbirka ima nesumnjive prednosti nad "analognom", posebice ako je u on-line obliku, odnosno ako je dostupna preko Internet mreže. Primjerice, tehnologija on-line digitalne zbirke omogućuje mjerenje brzine izvođenja programskog koda, pamćenje svih pokušaja rješavanja zadatka, praćenje rada učenika (rješavanje zadataka) u stvarnom vremenu itd.

Za pretpostaviti je da spomenuti didaktički elementi, koji nisu ostvarivi uporabom klasične tiskane zbirke, doprinose kvaliteti učenja, a isto tako i poučavanja, kao što su potvrdila neka dosadašnja istraživanja [1].

Osim rješavanja zadataka u užem smislu, digitalna zbirka omogućava i e-ispitivanje (eng. *e-assessment*), pojam koji se odnosi na ciljane aktivnosti u kojima se informacijsko-komunikacijska tehnologija (IKT) koristi za procjenu, odnosno provjeru znanja i vještina. Jedna od ključnih prednosti takvog načina provjere znanja je automatsko vrednovanje rezultata, odnosno računalom potpomognuta evaluacija programskih rješenja.

Iako se kod nas još uvijek rijetko koristi u nastavnom procesu, automatska evaluacija ima čitav niz prednosti. Primjerice, bitno skraćuje vrijeme ispravljanja zadataka odnosno smanjuje potrebne ljudske resurse, omogućuje brzu i sadržajnu povratnu informaciju ne samo osobama čije znanje se provjerava, nego i sastavljaču ispita o svakom pojedinom ispitnom zadatku i/ili provjeri u cjelini [2].

Navedene prednosti bile su motiv da se upustimo u izradu on-line digitalne zbirke zadataka s automatskom evaluacijom koja će poslužiti kao prateće dodatno nastavno sredstvo udžbeniku iz Informatike za 2. razred prirodoslovno-matematičkih gimnazija u izdanju nakladničke kuće SysPrint. Taj smo udžbenik odabrali budući da sadržajno pokriva uvod u algoritamsko rješavanje problema uporabom programskog jezika Python. Naime, polazište nam je bilo napraviti takvu zbirku koja će učenicima osigurati podršku u rješavanju problema programiranjem, a nastavnicima omogućiti vođenje i praćenje učenika kroz taj proces.

II. PLANIRANJE I RAZVOJ APLIKACIJE

Na Internetu se može pronaći nekoliko aplikacija za evaluaciju programskih rješenja u programskom jeziku Python. Primjerice, *Python-evaluator*, *Pylint*, *Taskeval*, *MIOC evaluator* i dr., Međutim, nijedna od tih aplikacija nije namjenski kreirana za uporabu u nastavi, odnosno kao specijalizirana podrška nastavnom procesu. Stoga smo se odlučili razviti vlastitu, koja će se poglavito rukovoditi obrazovnim i didaktičkim ciljevima.

Jedno od pitanja koje treba riješiti već na samom početku koncipiranja digitalne zbirke jest način evaluiranja programskih rješenja. Naime, postoje dva pristupa vrednovanju programa:

- analizatori programskog koda (engl. *Source Code Analyzer*) koji ispituju različite aspekte programa: sintaksu, semantiku, tijek programa, njegovu učinkovitost (npr. uporabom metode *Dead Code Elimination*), sigurnost i slično,
- evaluatori programskog rješenja koji ispituju izlazne vrijednosti (*output*) programa i uspoređuju ih s unaprijed zadanim, odnosno referentnim kriterijem.

Budući da je aplikacija primarno (digitalna) zbirka zadataka pomoću koje želimo provjeravati točnost rješenja što ga generira programski kôd, odlučili smo se za potonji pristup, tj. *black-box* evaluaciju kakva se koristi i na međunarodnim informatičkim natjecanjima.

Drugo je važno pitanje kako postaviti i organizirati on-line digitalnu zbirku na Internet, tako da bude stalno dostupna svim korisnicima, ali da količina korisnika i zadataka ne utječe na pouzdanost i sigurnost procesa evaluacije.

A. Arhitektura sustava

Aplikacija se sastoji od dvije nezavisne komponente: web sučelja (*front end*) i jedne ili više radilica (*back end*). Web sučelje je prezentacijski a radilica funkcionalni dio. Korisnik ima interakciju isključivo s web sučeljem. Pomoću njega unosi svoje programsko rješenje i prima rezultate evaluacije. Zadatak web sučelja je i briga o autentifikaciji korisnika te pohranjivanje zadataka i korisničkih podataka u jedinstvenu bazu podataka.

Radilica je pozadinski program koji od web sučelja prima zahtjeve za evaluacijom. Zahtjevi se zatim asinkrono izvršavaju i kad se pojedini zahtjev izvrši, rezultat evaluacije vraća se web sučelju

Kako se radi o procesorski intenzivnoj operaciji, radi poboljšanja performansi u slučaju velikog broja zahtjeva, moguće je pokrenuti više radilica na više računala. Raspodjela posla na više radilica je za korisnika transparentna. U slučaju da je pokrenuto više radilica, Web sučelje na temelju opterećenosti radilica odlučuje kojoj će poslati sljedeći zahtjev. Opterećenje radilice prati se na temelju trenutnog prosječnog zauzeća procesora. Zahtjev se šalje radilici s najmanjim prosječnim opterećenjem. Računala na kojima se izvršavaju radilice mogu biti u lokalnoj mreži ali i bilo kojoj lokaciji povezanoj Internetom.

Komunikacija između web sučelja i radilice odvija se preko standardnih TCP/IP socketa. Radilice međusobno ne komuniciraju.

B. Radilica

Radilica od web sučelja prima zadatke u obliku zahtjeva. Zahtjev za evaluacijom sadrži sljedeće podatke: ID zahtjeva, naziv zadatka, timeout, programsko rješenje zadataka (programski kôd u Pythonu), testne ulazne podatke i referentne izlazne podatke.

Za svaki primljeni zahtjev radilica u zasebnom *threadu* pokreće Python interpreter s proslijeđenim programskim rješenjem. Standardni izlaz se preusmjerava u datoteku koja se zatim uspoređuje s referentnim skupom izlaznih podataka. Rezultat komparacije vraća se u web sučelje. U slučaju greške ili prekoračenja dozvoljenog vremena, vraća se poruka o grešci.

Radilica se sastoji od više funkcionalnih modula koji se pokreću na dvije razine sigurnosti: 1. razina – visoka sigurnost, 2. razina – niska sigurnost.

Na razini visoke sigurnosti izvršavaju se sljedeći funkcionalni moduli:

- upravljački modul
- modul za nadzor ispravnosti, pokretanje i zaustavljanja sustava
- modul za komunikaciju s web sučeljem
- modul za praćenje statistike opterećenja
- modul za udaljeni nadzor i upravljanje radilicom
- modul za usporedbu rezultata izvršavanja testnog programa s referentnim rezultatima

Na razini niske sigurnosti izvršavaju se Python interpreter i korisničko programsko rješenje u Pythonu.

Procesi koji se pokreću na razini niske sigurnosti izvršavaju se u izoliranoj i zaštićenoj virtualiziranoj okolini (*sandbox*). Svaki od ovih procesa tako ima vlastiti zaštićeni memorijski prostor i zasebni zaštićeni i virtualizirani datotečni sustav. Program koji se izvršava unutar *sandboxa* nema izravan pristup do memorije, datotečnog sustava niti mrežnih funkcija računala. Svaki program se pokreće u zasebnom sandboxu a nakon izvršavanja upravljački proces briše cijelu virtualnu okolinu. Ukoliko se testni program ne završi u predviđenom vremenu (*timeout*), kontrolni proces će ga zaustaviti i pobrisati njegov *sandbox*. Valja napomenuti da se u virtualiziranoj okolini ne pokreće samo testni program već i Python interpreter tako da čak i u slučaju sigurnosnog propusta u samom interpreteru programi koji se istovremeno izvršavaju niti na koji način ne mogu utjecati jedni na druge ili na sam sustav na kojem se izvršavaju.

Korištenje *sandboxa* je nužno i osigurava zaštitu od malicioznog koda, mrtvih petlji i beskonačnih rekurzija. Ovakav način rada ograničava mogućnost evaluacije programa samo na programe koji ne upravljaju direktno hardverskim aspektima računala, no to je uvjet za stabilan rad.

Kako bi više korisnika moglo istovremeno koristiti evaluator, radilica pokreće svaki *sandbox* proces u zasebnom *threadu*. *Threadovi* se izvršavaju paralelno i imaju isti prioritet izvršavanja.

Funkcionalnost radilice pojednostavljeno se može opisati pseudojezikom na sljedeći način:

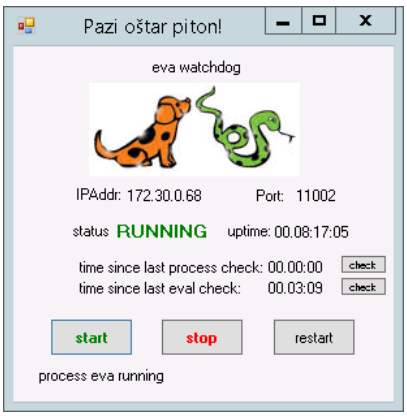
```
početak
    zaprimanje zahtjeva od web sučelja
    ako je zahtjev naredba onda:
        pozivanje upravljačkog modula
        izvršavanje naredbe (pokretanje, zaustavljanje, restart, statistika)
    ako je zahtjev program za evaluaciju onda:
```

pokretanje novog *threada*
formiranje novog *sandboxa*
pokretanje Python interpretera u *sandboxu*
pokretanje testnog programa u Python interpreteru
redirekcija standardnog izlaza u datoteku
dok proces ne završi radi

ako je prošao timeout onda
prekidanje procesa
terminiranje *threada* i prijava greške
ako je proces završio bez greške onda
formalna kontrola izlaznih podataka
usporedba izlaznih podataka s referentnim podacima
brisanje *sandboxa*
vraćanje rezultata i terminiranje *threada*
inače
brisanje *sandboxa*
vraćanje greške i terminiranje *threada*
kraj

C. *Raspoloživost sustava*

Kako bi se postigla visoka razina raspoloživosti sustava zamišljeno je da se radilice izvršavaju na računalima u udaljenim podatkovnim centrima, bez izravnog pristupa korisnika. Rad radilice stoga nadzire zasebni nadzorni program (*watchdog*). Nadzorni program provjerava je li na nivou operacijskog sustava aktivan proces radilice te mu u predodređenim vremenskim razmacima šalje na evaluaciju testni program i na taj način provjerava funkcionalnost radilice.



Slika 1. Nadzorni program Eva Watchdog.

Ukoliko nadzorni program ne dobije odgovarajući odgovor u unaprijed određenom vremenu, prekida izvršavanje radilice i ponovno ju pokreće. Nadzorni program može se pokrenuti i na nekom drugom računalu kako bi se detektirao privremeni prekid rada pojedinog računala te uslijed toga pokrenula rezervna radilica na drugom računalu.

Izravna interakcija administratora s radilicom nije predviđena, nego se radilicom upravlja putem administratorskog web sučelja. Tako administrator može istovremeno upravljati s više radilica, nadzirati njihov rad i mijenjati im parametre.

III. WEB SUČELJE I ZNAČAJKE APLIKACIJE

Budući da je aplikacija (u daljnjem tekstu Eva) zamišljena kao pomoćno nastavno sredstvo, bilo je potrebno ugraditi elemente koji prate nastavni sadržaj iz udžbenika. Međutim, da bi on-line digitalna zbirka zadataka bila dostupna svima, mora omogućiti uporabu i korisnicima izvan nastavnog procesa. To smo riješili tako što se Eva sastoji od dva, relativno nezavisna modula:

- *Fitness* način rada namijenjen je za vježbanje i (samo)provjeru vlastitoga znanja. Učenici ga mogu prakticirati samostalno, kod kuće ili bilo gdje drugdje gdje imaju pristup Internetu. Ovaj način ne zahtijeva povezivanje podataka sa školom, razredom ili nastavnikom pa je potpuno neovisan o njima.
- *Stress* način koristi se za organiziranje i provođenje školskih testova, što zahtijeva određene korake i predradnje. Primjerice, nastavnik najprije mora odabrati zadatke za test i zakazati datum i vrijeme provođenja testa, za svaki razred zasebno. Da bi učenik pristupio testu, mora se prijaviti u taj razred i mora znati pristupu šifru razreda (dodjeljuje ju nastavnik).

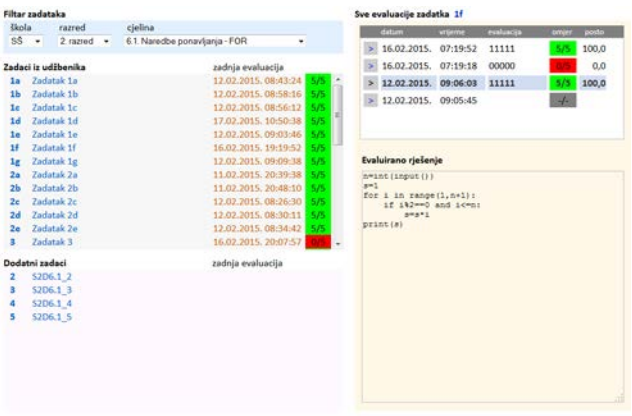
Spomenuli smo u uvodu da je aplikacija Eva namijenjena kako učenicima, tako i njihovim nastavnicima. Zato je web sučelje organizacijski i funkcionalno podijeljeno u dva modula, ovisno o vrsti prijavljenog korisnika (učenik ili nastavnik).

Još je jedna dvojakost značajna za aplikaciju Eva. Pored zadataka koji su inicijalno već ugrađeni i sastavni su dio digitalne zbirke, nastavnicima je omogućeno da u zbirku dodaju (*uploadaju*) svoje zadatke.

A. *Modul za učenike*

Nakon što se učenik prijavi na uvodnoj web stranici, odabire način rada – fitness ili stress (kasnije se uvijek može vratiti na taj izbornik i promijeniti svoj odabir).

Na uvodnoj stranici u fitness način (slika 2.) učenik ima pregled nad svojim dosadašnjim evaluacijama fitness zadataka. Za svaki zadatak koji je učenik rješavao prikazani su datum i vrijeme te rezultat zadnje evaluacije u obliku omjera točnih i ukupnog broja izlaznih parova zadataka. Naime, za svaki zadatak u zbirci priređeno je 5 do 10 parova referentnih ulazno/izlaznih podataka s kojima se uspoređuje izlaz kojeg generira programsko rješenje učenika. Osim što učenik ima pregled nad svim evaluacijama, može provjeriti i koje je programsko rješenje poslao na evaluaciju.



Slika 2. Uvodna stranica *fitness* načina rada.

Klikom mišem na naziv pojedinog zadatka otvara se nova stranica u kojoj se programski kôd tj. programsko rješenje zadatka šalje na evaluaciju (slika 3.).

Slika 3. Stranica za slanje programskog rješenja na evaluaciju.

Učenik može učitati programsko rješenje iz .py datoteke ili ga neposredno upisati u tekstni okvir. Svijetlo-plavo dugme "U/I podaci" služi za prikaz referentnih ulazno/izlaznih podataka s kojima će se usporediti izlaz (*output*) učenikova programa i tako evaluirati njegovo programsko rješenje zadatka.

Pomoću dugmeta "→" učenik može učitati jedno od svojih prijašnjih rješenja, promijeniti tj. popraviti ga i ponovo poslati na evaluaciju, što se u aplikaciji naziva postupak reevaluacije. Ova mogućnost ima dodatnu didaktičku vrijednost jer potiče učenike da popravljaju i dorađuju svoj programski kôd, bez da se to evidentira kao nova evaluacija.

Od posebnog didaktičkog značaja je funkcija dugmeta "| |" koja učeniku omogućava usporedbu referentnih izlaznih podataka i podataka koje generira njegovo programsko rješenje (slika 4.).

Procedura za usporedbu pronalazi outpute učenika koji se razlikuju od referentnih i oboja ih crvenom bojom. Na taj način pomaže učeniku u otkrivanju grešaka, koje ponekad mogu biti teško zamjetne, a ipak dovoljne da rješenje vrednuje kao netočno. Na učeniku je da sam otkrije grešku u svom programu i ispravi je pa mu u tome modul ne može pomoći.

Slika 4. Usporedba izlaznih podataka.

Drugi način korištenja aplikacije Eva je stress način rada, poglavito namijenjen pisanju školskih testova. Da bi učenik sudjelovao u pisanju testa, aplikacija traži da se najprije prijavi u svoje razredno odjeljenje (više o tome u opisu modula za nastavnike).

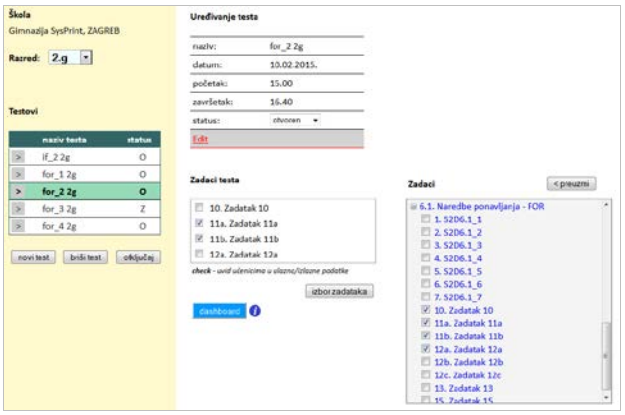
Ako je test otvoren, učenik mu može pristupiti i rješavati zadatke koje je nastavnik prije toga odabrao. Nakon što nastavnik zaključi rješavanje testa, učenik može vidjeti svoje rezultate, tj. ishod evaluacije svojih programskih rješenja za svaki zadatak.

Iako je postupak prijave rješenja i njegove evaluacije u mnogočemu sličan fitness načinu, zbog prirode školskog teksta postoje i određene razlike. Primjerice, u stress načinu učeniku nije omogućena reevaluacija rješenja kako bi nastavnik imao potpuni uvid u sve njegove pokušaje rješavanja pojedinog zadataka. Naime, s pedagoškog i dokimološkog aspekta nije svejedno je li učenik riješio zadatak u dva ili u deset pokušaja.

B. Modul za nastavnike

Ovaj je modul opremljen znatno većim brojem mogućnosti nego modul za učenike. To je iz razloga što omogućava nastavniku da:

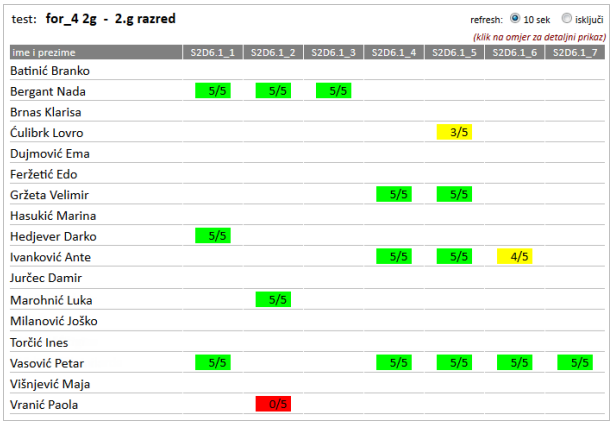
- organizira razredna odjeljenja škole, uređuje i ažurira njihove podatke te podatke učenika, komunicira s učenicima elektroničkom poštom i sl.
- pregledava i nadopunjuje zadatke iz zbirke,
- priprema i provodi školske testove,
- nadzire rad učenika tijekom rješavanja školskog testa i ispisuje završne popise njihova postignuća,
- nadzire rad učenika tijekom fitness načina rada.



Slika 5. Kreiranje i uređivanje školskih testova.

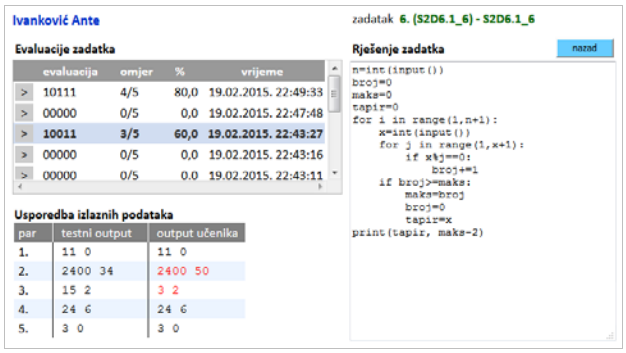
Na slici 5. prikazana je stranica na kojoj nastavnik kreira i uređuje školske testove, što podrazumijeva odabir zadataka pojedinog testa te određivanje statusa svakog zadatka (npr. nastavnik ovdje bira hoće li učenici imati uvid u referentne ulazno/izlazne podatke).

Tijekom rješavanja testa nastavnik na svojem računalu ima pregled nad uradcima učenika – koje zadatke su riješili i koje rezultate su postigli. To mu omogućava stranica pod nazivom dashboard prikazana na slici 6.



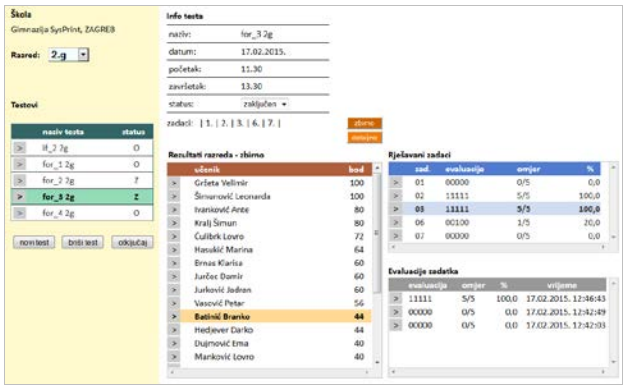
Slika 6. Pogled na rezultate tijekom riješavanja testa.

Osim "panoramskog" pogleda na rezultate čitavog razrednog odjeljenja, klikom na pojedini rezultat nastavnik može detaljno analizirati postignuća učenika tijekom rješavanja tog zadatka – koliko je učenik imao pokušaja i s kojim ishodom, koje programsko rješenje je poslao na evaluaciju, pri čemu mu uvid olakšava komparator outputa (slika 7.).



Slika 7. Detaljna analiza rješavanja zadatka učenika.

Nakon što nastavnik zaključi test, stranica za školske testove (na slici 5. prikazana dok je test još bio otvoren) poprima drugačiji oblik. Nastavnik sada ima na raspolaganju rang-listu ukupnih rezultata učenika na testu, te može analizirati kompletan tijek rješavanja svakog pojedinog zadatka (slika 8.), što mu pomaže u donošenju konačne ocjene. Osim toga, nastavniku su na raspolaganju i dvije završne rang-liste priređene za ispis na pisač.



Slika 8. Pregled ukupnih i pojedinačnih rezultata.

Uz rješavanje školskih testova i provjeru znanja neminovno se vezuje i pitanje prepisivanja tuđih rješenja. Zasad u aplikaciju nisu ugrađeni moduli koji bi to sprječavali, ali učitelju je na raspolaganju mogućnost da prirediti A i B formu testa.

Za kraj ovog letimičnog prikaza aplikacije Eva ostavili smo mogućnost za koju vjerujemo da ima veliki značaj za unaprjeđenje kvalitete rada nastavnika.

Učenje i vježbanje učenika kod kuće u klasičnoj je nastavi velika nepoznanica svakom nastavniku. Međutim, on-line uporaba nastavnog sredstva otvara nove mogućnosti komunikacije, pa smo to pokušali ugraditi u ovu digitalnu zbirku. Pomoću modula *Vježbe* nastavnik ima mogućnost uvida u fitness način rada učenika, i to u stvarnom vremenu. Primjer uporabe tog modula prikazan je na slici 9.

Slika 9. Uvid nastavnika u rad učenika kod kuće (fitness).

Princip rada je sličan *dashboardu* za školski test, s tom razlikom što ovdje nastavnik može birati razredno odjeljenje te bilo koji dan za koji želi dobiti uvid u rad učenika kod kuće (fitness način rada).

IV. TESTIRANJE APLIKACIJE

Kako bismo aplikaciju Eva što više, odnosno što bolje, uskladili s potrebama učenika, nastavnika i općenito nastavnog procesa, najprije smo je probno aplicirali u nekoliko škola. Učenici i nastavnici testirali su rad aplikacije i slali svoje primjedbe i prijedloge (popis škola koje su sudjelovale u testiranju naveden je u Zahvali). Na kraju testne faze organizirali smo za učenike on-line anketu s desetak pitanja o njihovim dojmovima.

U trenutku zaključivanja pisanja ovoga rada još se prikupljaju i analiziraju rezultati on-line ankete iz testne faze. Iako raspolažemo s nepotpunim podacima, preliminarna obrada na uzorku 61 učenika ukazuje na dobru prihvaćenost ovakve online digitalne zbirke.

Prije prikaza rezultata nekoliko čestica iz online anketnog upitnika, spomenimo da je ispunjavanje ankete bilo na dragovoljnoj osnovi.

Na pitanje "*Da možeš birati, što bi odabrao za vježbanje zadataka iz informatike?*" od tri ponuđena odgovora 88,5% učenika odgovorilo je online digitalnu zbirku kao što je ova, 8,2% učenika odabralo je digitalnu zbirku, a samo 3,3% učenika odabralo bi tiskanu zbirku.

Tablica 1. Odgovori učenika na pitanja iz ankete.

čestica	da	ne
Bi li koristio/la zbirku i bez da nastavnik to traži?	87%	13%
Dobro - Odmah vidim da mi je rješenje točno.	85%	15%
Loše - Nastavnik ima previše uvida u moj rad i vježbanje.	10%	90%
Potiče li te zbirka da još više vježbaš zadatke?	79%	21%
Potiče li te zbirka da više učiš gradivo iz udžbenika?	61%	39%
Pomaže li ti zbirka da brže i/ili bolje naučiš gradivo iz informatike?	95%	5%
Bi li ovu zbirku preporučio/la drugima za vježbanje/učenje?	98%	2%

Iz tablice 1. zamjetno je da velika većina učenika pozitivno vrednuje on-line digitalnu zbiru zadataka (kao što je aplikacija Eva), da bi je koristili i bez da nastavnik to traži (87%) te da bi je preporučili i drugim učenicima/korisnicima za vježbanje i učenje (98%). Za daljnji razvoj aplikacije svakako je poticajno što zbirka potiče učenike da još više vježbaju (79%) i uče gradivo (61%) te da mi pomaže da brže i/ili bolje nauče gradivo iz informatike (95%).

Ako se ovi preliminarni rezultati pokažu istinitim, odnosno da vrijede za značajnu većinu učenika, moći ćemo zaključiti da smo postigli cilj koji smo si postavili na početku – načiniti digitalnu zbirku zadataka koja će poslužiti kao upotrebljivo nastavno sredstvo.

V. ZAKLJUČAK

Povratne informacije tijekom probne primjene aplikacije Eva te preliminarni rezultati ankete za učenike upućuju na zaključak da je on-line digitalna zbirka zadataka upotrebljivo i, štoviše, poželjno nastavno sredstvo. Ne samo što je kao medij bolja od tiskane zbirke, nego pruža i posve nove mogućnosti primjene u didaktičke svrhe. Primjerice, nastavnik može pratiti rad učenika kod kuće i to u stvarnom vremenu.

No, dosadašnja iskustva ukazuju da je potrebna određena doza opreza i pažljive pripreme u koncipiranju i realizaciji digitalne zbirke zadataka, posebice u pogledu načina evaluacije programskih rješenja. U uvodu smo već spomenuli da smo u svrhu evaluacije primijenili *black-box* metodu. Pojedini autori koji su analizirali učinkovitost ove metode u primjeni na informatičkim olimpijadama ističu nedostatke takvog načina evaluacije [3], sve do kritika da u nekim slučajevima postoje zadaci za koje se uopće ne mogu napisati odgovarajući referentni ulazno/izlazni podaci [4].

Iako se na srednjoškolskoj razini učenici neće susretati s tako zahtjevnim zadacima, svejedno je potrebno iznimno pažljivo sastavljati referentne ulazno/izlazne podatke te ih provjeriti na više (što više) načina. Prema dosadašnjem iskustvu, to je ujedno i jedino osjetljivo mjesto opisane aplikacije.

Na osnovu početnih povoljnih dojmova i iskustava korisnika, daljnji razvoj aplikacije planiramo u nekoliko segmenata:

- mogućnost evaluacije programskih rješenja koja zahtijevaju pristup datotečnom sustavu,
- implementacija metrike trajanja izvršavanja programa,
- mogućnost zadavanja referentnih rezultata u slobodnijoj formi, uporabom regularnih izraza,
- izvršavanje radilica na Unix, odnosno Linux operacijskim sustavima,
- implementacija modela dinamičkog skaliranja pokretanjem dodatnih radilica u slučaju većeg opterećenja sustava,
- implementacija komunikacije web sučelja s radilicama putem SSL protokola.

Nova inačica aplikacije koja je trenutno u fazi testiranja u školama nudi mogućnost evaluacije programskih rješenja i u drugim programskim jezicima osim Pythona. U toj je inačici realizirana podrška za C, C++, C#, Pascal, QuickBasic, SmallBasic te Java programski jezik.

ZAHVALA

U probnoj fazi primjene aplikacije sudjelovalo je nekoliko škola koje su svojim primjedbama i prijedlozima doprinijele njenom konačnom izgledu i funkcionalnosti: XV. gimnazija (Zagreb), III. gimnazija (Split), Gimnazija Lucijana Vranjanina (Zagreb) i Srednja škola Krapina (Krapina).

LITERATURA

[1] J. M. Noyes and K. J. Garland, "Computer- vs. paper-based tasks: Are they equivalent?", *Ergonomics*, vol. 51, no. 9, Sep. 2008, pp. 1352-1375.

[2] J. Petrović, Primjena računala u provjeri znanja, kvalifikacijski doktorski ispit, 2012., https://www.fer.unizg.hr/_download/repository/Juraj_Petrovic_,KDI.pdf, veljača 2015.

[3] M. Forišek, "On the Suitability of Programming Tasks for Automated Evaluation", *Informatics in Education*, Vol. 5, No. 1, 2006, pp. 63–76.

[4] W. T. van Leeuwen, A Critical Analysis of the IOI Grading Process with an Application of Algorithm Taxonomies, Master's Thesis, Technische Universiteit Eindhoven, Faculty of Mathematics and Computing Science, 2005. http://www.win.tue.nl/~wstomv/misc/ioi-analysis/thesis-final.pdf, veljača 2015.

Primjena računalne animacije u nastavi općeg obveznog tehničkog odgoja i obrazovanja

Damir Purković*
* Filozofski fakultet u Rijeci/Odsjek za politehniku, Rijeka, Hrvatska
damir@ffri.hr

Sažetak - Suvremeno temeljno tehničko obrazovanje zahtijeva sve učinkovitije metode, postupke, strategije i pristupe razvoju tehničko-tehnoloških kompetencija učenika. Problem razvoja tih kompetencija izražen je zbog nedostatnih materijalno-tehničkih i organizacijskih uvjeta, ali i zbog iskustvenih ograničenja učenika, povezanih s razinom obrazovanja i dobi učenika. Konstruktivistička teorija i kontekstualni pristup nastavi i učenju predstavljaju temelj za razvoj potrebnih kompetencija učenika, ali je osmišljavanje prikladnog nastavnog konteksta često složeno i neizvjesno. Prikladna uporaba računalnih animacija bi mogla doprinijeti oplemenjivanju kontekstualne nastave, kao segment koji može upotpuniti prazninu između teorijske podloge i praktičnih tehničko-tehnoloških vještina. U tom smislu u radu se iznose uvjeti i ograničenja uporabe animacije u metodičkom i tehničkom smislu, te se predlaže utemeljeni metodički okvir za primjenu animacije u nastavi općeg obveznog tehničkog odgoja i obrazovanja.

I. UVOD

Tehničko (tehnološko) odgojno-obrazovno područje i tehničko obrazovanje u cjelini svoju pravu odgojno-obrazovnu vrijednost može iskazati danas, u uvjetima izraženog komunikacijskog hendikepa mlade generacije [1] i stoga bitno narušenog razvoja metakognitivnih sposobnosti učenika. Naime, današnji rapidni rast tehničko-tehnoloških spoznaja i dinamika radno-socijalnih odnosa determiniraju kompetencije koje učenici trebaju steći tijekom obrazovanja, zbog čega do izražaja dolazi potreba za razvojem samosvijesti, samoučinkovitosti i samoreguliranog učenja, potrebnih za brzu prilagodbu novom tehničko-tehnološkom okružju i svijetu rada. S obzirom da glavni cilj tehničkog odgoja i obrazovanja podrazumijeva osposobljenost učenika za uspješno snalaženje i djelovanje u kontekstu tehnologije i u kontekstu svakodnevnog životnog okružja, djelovanja i rada u današnjem tehnološkom društvu [2], [3], [4], [5], razvidna je važnost istog za razvoj društva u cjelini.

Konstruktivizam, kao paradigmatska osnovica odgoja i obrazovanja, ali i kao teorija spoznavanja i pristupa obrazovanju, danas čini važnu i prihvatljivu podlogu na kojoj bi trebalo graditi svaki odgoj i obrazovanje pojedinca. Konstruktivistički pristup stvaranje znanja zasniva na aktivnoj ulozi učenika, važnosti osobnog i društvenog iskustva, te na spoznaji da će se kakvoća znanja, pa tako i percepcija stvarnosti, individualno razlikovati [6]. U tom smislu, kontekstualni pristup učenju i poučavanju, kao konkretizacija konstruktivističke paradigme, pridaje veliku važnost nastavnom kontekstu čija uloga, uz ključnu ulogu nastavnika, može bitno doprinijeti kakvoći spoznaje kod pojedinca i razvoju primjerenih kompetencija. U kontekstualnom pristupu nastavi proces učenja i poučavanja zasniva se na autentičnom nastavnom kontekstu na kojem se grade aktivnosti značajne za učenike [7]. Opće obvezno tehničko obrazovanje, kao važan preduvjet profesionalnog obrazovanja nužnog za održivost gospodarstva, pri razradi nastavnih strategija, modela, metoda i pristupa u obzir treba uzeti ograničenja odgojno-obrazovnog procesa na toj razini. Ograničenja se odnose na: iskustvo učenika na koje treba „nadograditi“ nove spoznaje; poznavanje pojmova, činjenica i generalizacija s obzirom na dosadašnje obrazovanje; učenikove mogućnosti apstrahiranja činjenica i generalizacija (mogućnosti za razvoj mentalnih modela) s obzirom na dob učenika; te ograničeno vrijeme i materijalno-tehničku podlogu s kojom na toj razini obrazovanja školski sustav može raspolagati. Navedena ograničenja, kao polazišta, uvjetuju mogućnosti realizacije kontekstualnog pristupa nastavi, te pružaju priliku za razvoj kontekstualnih artefakata, pa tako i računalnih animacija koje će se koristiti u nastavnom procesu.

Uporaba računalne animacije u nastavi danas, unatoč uvriježenom vjerovanju kako može bitno utjecati na kakvoću razumijevanja složenih fenomena i koncepata, ima dvojben učinak na postignuća učenika. Naime, empirijska istraživanja učinka animacije na ishode učenja variraju od pozitivnih do negativnih ili pak rezultata koji pokazuju neutralni učinak animacije [8]. Izvjesno je da učinak same animacije na znanje i razumijevanje ovisi o više različitih čimbenika koji mogu utjecati na način na koji ljudi uče s animacije, te da će uspjeh takvog učenja ovisiti o interakciji tih čimbenika [8]. Dosadašnja iskustva, također, opovrgavaju radikalna shvaćanja kako računalna animacija, ali i uporaba IKT-a u nastavi u cjelini, mogu zamijeniti iskustva učenika zasnovana na autentičnoj stvarnosti. Stoga je ovaj rad usmjeren razradi primjerenog i smislenog metodičkog okvira i mehanizma za razvoj i primjenu računalne animacije u općem obveznom tehničkom odgoju i obrazovanju, u okviru kontekstualnog pristupa nastavi. U tom smislu će se iznijeti okvir i znanstveno utemeljeni način uporabe animacije u nastavi ovog specifičnog nastavnog područja te tako praktičarima olakšati primjenu iste.

II. TRENDOVI PRIMJENE RAČUNALNE ANIMACIJE U NASTAVI

Danas se animacija u nastavi različito shvaća i tumači, zbog čega se često poistovjećuje s video-zapisom ili pak s računalnim simulacijama, što nikako nisu istoznačnice.

Naime, video-sadržaji su izvorno namijenjeni prikazivanju stvarnih objekata, dok animacija prikazuje simulirane sadržaje i objekte [20]. S druge strane, u prirodoslovnom, matematičkom i tehničkom obrazovanju računalne simulacije se uglavnom odnose na relativno složene simulacije procesa, pojava, sustava, modela ili tehnologije. Dok simulacija treba odražavati stvarne ulazno-izlazne veličite bitne za rad sustava, primarno svojstvo animacije je dinamička vizualizacija sadržaja na simboličkoj razini. Uporaba simulacija u nastavi ovog područja često je uvjetovana dostupnošću specijaliziranih softverskih alata, specijaliziranim znanjem nastavnika i visoko razvijenim učenikovim mentalnim modelima potrebnim za praćenje i razumijevanje sadržaja. Dakle, računalna simulacija i nastavna animacija u pozadini mogu dijeliti istu strukturu, ali se bitno razlikuju u primarnoj namjeni, ali i u percepcijskoj i interakcijskoj prilagodbi krajnjeg produkta mentalnim modelima učenika.

Sam pojam animacije može se odrediti kao dinamička reprezentacija prikaza procesa koji se mijenja s vremenom [8], odnosno, kao serija okvira na način da se svaki okvir pojavljuje kao alternacija prethodnome [10]. Suvremene nastavne animacije razvijaju se uglavnom kao računalne animacije, a koriste se onda kad postoji potreba za prikazivanjem onog što nije lako ili moguće vidjeti u stvarnosti, promjena koje nije moguće vidjeti u realnom vremenu, apstraktnih prikaza koji nisu sami po sebi vizualni [8], te kada je stvarna provedba određene aktivnosti nepraktična, opasna za sudionike ili su sredstva za provedbu aktivnosti nedostupna [11]. Za izradu nastavnih animacija nastavnici i stručnjaci danas koriste čitavu lepezu dostupnog softvera, od onog univerzalnog poput Flash-a, Jave i različitih alata za izradu animiranih slika, do specijaliziranih aplikacija kao što su *Moviestorm*, *Autodesk Maya*, *GoAnimate*, *Blender* i sl. Pri tom kakvoća konačnog produkta uvelike ovisi o umješnosti autora, dok je učinak animacije na postignuća učenika upitan, uglavnom zbog nedostatne znanstveno-metodičke podloge i preporuka za razvoj i primjenu animacije u nastavnom procesu. Prema načinu prikazivanja sadržaja u nastavi, animacije se najčešće dijele na ekspozitivne (opisne), čiji sadržaj učenici gledaju i slušaju na zaslonu, i na interaktivne, kojima učenici mogu upravljati na višoj razini. Određeni autori u načine prikazivanja svrstavaju i kvizove [12], koji služe testiranju učenika uz pomoć animiranog sadržaja. Većina animiranih nastavnih sadržaja današnjice razvijena je u 2D tehnologiji, dok se sadržaji razvijeni u 3D tehnologiji najčešće prikazuju na 2D zaslonima. Razlog tome je zahtjevna i skupa uporaba 3D tehnologije u tipičnom školskom okružju. Zbog što jasnijeg uvida u aktualne načine, trendove i probleme prikazivanja animacije u nastavi neophodno je pojasniti glavne tehničke značajke dvije temeljne vrste animacija, ekspozitivnih i interaktivnih 3D animacija.

A. Ekspozitivna animacija (animacija za izlaganje)

Ekspozitivna animacija, koja se danas najčešće koristi u nastavi, uglavnom se primjenjuje u vidu animiranih video sadržaja [12], ali i animiranih slika. Za izradu animiranih video sadržaja češće se koriste specijalizirani softverski alati različitih proizvođača (*GoAnimate*, *Moviestorm*, *Animasher* i sl.), dok se za izradu animiranih slika koristi čitava lepeza softvera, od jednostavnih besplatnih alata, uredskih paketa, do specijaliziranog softvera. Ekspozitivna animacija nastoji kod učenika razviti spoznaju o određenom dinamičkom procesu koji se odvija izvan vizualnog i vremenskog dosega učenika tako da im se taj sadržaj frontalno prezentira i demonstrira u nastavnom procesu. U tehničkom smislu, ključno pitanje je kako animacijom pokazati (vizualno i auditivno) ono što je bitno i kako organizirati sadržaj animacije [12]. Problem pri tom predstavlja spoznaja da učinak animacije na postignuće učenika može biti minoran ili čak neprihvatljiv, ponajprije zbog individualnih razlika u načinu učenja i sposobnostima učenika za učenje složenog sadržaja s animacije. Ovakvi problemi se nastoje zaobići uvođenjem određene interaktivnosti, poput izbora značajki objekata ili osobina subjekata u animaciji, dodavanjem dijaloga, mogućnostima upravljanja tempom animacije i sl. [12]. Ipak, bitan nedostatak autentičnosti, realističnosti i stvarne interaktivnosti ostaje i dalje veliko ograničenje animacije, a učinak iste na postignuća učenika upitan.

B. Interaktivna 3D animacija

Glavna intencija trenda razvoja interaktivnih 3D animacija općenito, pa tako i u nastavi, je pokušaj realističnog i interaktivnog nadomještanja izvorne stvarnosti uporabom tehnologije za 3D vizualizaciju i simulaciju u realnom vremenu [12]. Tehnologija 3D interaktivnih animacija, simulacija i tzv. virtualne stvarnosti permanentno se razvija već više od dva desetljeća, a napredak je osobito uočljiv u svijetu igraćih konzola, profesionalnih simulacija i računalnih igara. No, prava primjena u nastavnom procesu je još daleko od željene i široko prihvaćene. Problemi primjene u nastavi mogu se svesti na tri razine: složena je za korištenje (manipulaciju) od strane učenika – zahtijeva posebno osposobljavanje učenika; izrada animacija je vrlo složeni proces – neophodno je profesionalno osposobljavanje nastavnika; tehnološki je zahtjevna – primjena zahtijeva skupu specijaliziranu opremu, softver i računalnu podlogu. Unatoč tome očekivanja od 3D animacije u nastavi su velika jer se predviđa stjecanje iskustava bliskih iskustvima koja bi učenici stekli u stvarnom svijetu.

U tehničkom smislu, arhitektura 3D interaktivnih animacija, koje se danas primjenjuju u nastavi, zasniva se na podlozi koju čine [12]: *senzori*, čija primarna funkcija je praćenje i reagiranje na korisničke pokrete, te permanentno slanje signala o tim pokretima jezgri sustava. U trenutnoj školskoj praksi ovaj dio se, na žalost, svodi uporabu standardne tipkovnice i miša računala; *Upravljački programi*, koji čine interaktivnu sponu različitog računalnog sklopovlja i sadržaja nastavne animacije; *Fizikalni model (jezgra)* koja ustvari predstavlja matematički model fizičkih procesa kojim se pokušava što realnije simulirati stvarnost, odnosno, autentičnost interakcije korisnika i sadržaja. Modelom se obavljaju potrebni proračuni koji determiniraju ponašanje sustava u skladu s korisnikovim pokretima (interakcijom); *3D zasloni* kojima se nastoji prikazati visoka razina realističnosti sadržaja (3D virtualne stvarnosti), usklađena s razinom interakcije korisnika s 3D sadržajem. U trenutnoj nastavnoj praksi glavna interakcija se ustvari odvija na 2D razini, dok su samo određeni prikazi predstavljeni u 3D tehnologiji. Sam 3D prikaz načelno se zasniva na „varanju“ ljudskog mozga pri interpretaciji

dvije 2D slike u jednu koja ima dimenziju dubine. Danas se u tu svrhu pokušavaju koristiti različite tehnologije, poput stereoskopskih prikaza, polarizacijskih sustava, sustava koji koriste *Eclipse* metodu, sustava koji koriste *Pullfrich efekt*, volumetrijskih zaslona, holografskih zaslona i sl. [13]. Viša razina interaktivne 3D animacije (tzv. virtualne stvarnosti) pretpostavlja stvarnu 3D interakciju korisnika i sadržaja uporabom posebnih senzora, poput haptičkih i sličnih senzora, ali i prethodno spomenutih namjenskih sustava za 3D prikaz. Ovakvi sustavi bi trebali imati i određeni emocionalni i fizički utjecaj na učenika, a moguće ih je, uporabom glasovne, tekstualne i videokomunikacije, iskoristiti za udaljeno suradničko učenje u realnom vremenu, u okviru odgovarajućeg interaktivnog sustava za učenje [12]. Unatoč realno ostvarivoj tehnologiji 3D interaktivne animacije, uporaba takvih sustava u današnjem školskom okružju nije nešto na što se školski sustav može osloniti, osobito s obzirom na upitan i nedostatno istražen učinak ove tehnologije na postignuća učenika. Osim toga, terminološki gledano, sofisticirani sustavi virtualne stvarnosti najčešće se određuju kao simulatori određenog iskustva, što njihovo određenje kao računalne ili bilo koje druge animacije može činiti upitnim.

III. OGRANIČENJA I PREDUVJETI PRIMJENE ANIMACIJE U NASTAVI

Koncept primjene animacije u nastavi može se analizirati na tehničkoj, semiotičkoj i psihološkoj razini [20], koje u nastavnom procesu čine jedinstvenu i neodvojivu cjelinu s ciljem doprinosa spoznajnim procesima i postignućima učenika. Stoga bi glavni cilj uporabe računalne animacije u nastavi trebao uključiti doprinos boljem razumijevanju određenih procesa, pojava, sustava, situacija, tehnologija i sl., koje na drugi način nije moguće predstaviti učenicima. Drugim riječima, nastavna animacija bi trebala doprinijeti razvoju mentalnih modela učenika nužnih za razumijevanje apstraktnih prikaza, ali i utjecati na razvoj učenikove sposobnosti apstraktnog predočavanja izvorne stvarnosti. Zbog toga nastavnu animaciju treba promatrati kroz prizmu doprinosa postignućima učenika, preduvjeta koje treba zadovoljavati i načina na koji je treba implementirati u nastavni proces kako bi polučila željeni učinak. U tom smislu nije primjereno izdvojeno tražiti odgovore na to pomažu li animacije učenicima pri učenju, već treba istražiti i uvjete u kojima je uporaba animacije korisna za učenike [8].

Suvremena istraživanja obilježja nastavne animacije i učinka na postignuća učenika su, unatoč dugotrajnoj primjeni iste, relativno skromna, često kontradiktorna [8], a nisu niti univerzalno primjenjiva na sva nastavna područja. Ipak, dosadašnja istraživanja ukazuju na vrlo egzaktne kriterije koji mogu odrediti uvjete i načine izvedbe nastavne animacije. Razumijevanje animacije u nastavi najčešće je uvjetovano optimizacijom kognitivne opterećenosti učenika [14], [15], [16], [17], [18], [19], razvijenošću učenikovih mentalnih modela za praćenje animacije [21], [22] i individualnim razlikama svakog pojedinca [23], [8]. Individualne razlike učenika su ponajprije odraz dispozicijske dominantne vrste inteligencije kod pojedinca [24], pri čemu se osobito ističe bolje razumijevanje animacije od strane učenika koji imaju razvijenije prostorne sposobnosti [21]. Kod ekspozicijskih animacija, koje se najčešće koriste u nastavi, empirijska istraživanja pokazuju kako sama animacija nema bolji učinak na razumijevanje sadržaja u odnosu na statičke slike istog sadržaja [21], [22], već je razumijevanje sadržaja potaknuto dodatnim glasovnim pojašnjenjima i opisima [8], [21] te pitanjima kojima, na temelju statičkih slika, učenici mogu predvidjeti ponašanje dinamičkog procesa [21]. Ovaj nalaz donekle potvrđuju i istraživanja u kojima bolji učinak na postignuća ima segmentirana grafika u odnosu na kontinuiranu animaciju [25]. Iako su minimalni tekstualni pojmovi, zbog temeljne orijentacije učenika, često dijelom animacije, istraživanja pokazuju kako tekstualni komentari i objašnjenja, suprotno prvotnim istraživanjima, nisu učinkoviti zbog ignoriranja takvih informacija ali i neučinkovitosti s obzirom na razinu predznanja učenika [22]. Dakle, temeljni problem učinka ekspozitivne animacije najčešće predstavlja nedostatno predznanje i nedovoljno razvijen mentalni animacijski model učenika potreban za praćenje i učenje pomoću ili s animacije, čiji nedostatak može kod učenika izazvati nerazumijevanje animiranog sadržaja, tzv. „iluziju“ o razumijevanju [8], [26], ali i razvoj potpuno pogrešnog mentalnog modela [21], [22], [27]. Sličan učinak ustanovljen je i primjenom interaktivne animacije s dodatnom signalizacijom, pri čemu interaktivnost i dodatni elementi pomoći nisu kod učenika izazvali bolje razumijevanje, već su naprotiv, razvijeni mentalni modeli bili potpuno pogrešni [22]. U navedenom istraživanju se nadalje sugerira kako je prethodno dobro poznavanje sadržaja, uz višestruko ponavljanje i pružanje učenicima različitih iskustava sa sličnim sadržajem, neophodno za razvoj ispravnih mentalnih modela učenika. U skladu s tim Bétrancourt [26] navodi kako animaciju u nastavi možemo primijeniti samo kada: a) animirani fenomen uključuje promjene tijekom vremena, te realno možemo pretpostaviti da učenici iz statičkih slika ne bi izveli ispravne zaključke; b) ako su učenici početnici u navedenom području ili su suočeni s visokom razinom kognitivne opterećenosti. U skladu s prethodno navedenim općim razlozima i uvjetima primjene animacije, te ograničenjima proizašlim iz empirijskih istraživanja, Ainsworth [8] navodi moguća obilježja animacije, koja se ovdje iznose u skladu s posebnostima tehničkog nastavnog područja:

Izražajna obilježja – odnose se na svojstvo animacije da izrazi apstrakciju neovisno o prirodi tumača. Može se objasniti i kao količina „izračuna“ u glavi pojedinca potrebnih za izvođenje „ispravnih“ zaključaka. Ova obilježja važna su za prikaz određenih aktivnosti, procesa, tvorevine ili tehnologije koje se odvijaju određenim redoslijedom. Zbog toga je izražajnost animacije iznimno važna kada se prezentira dinamička aktivnost koja treba i biti shvaćena kao vremenski slijed događanja, što može biti problem kod animacija koje nisu takve.

Kognitivna obilježja – govore nam o tome koliko animacija može doprinijeti razvoju spoznajnih procesa kod učenika. U tom smislu, cilj i sadržaj animacije treba biti usmjeren na to da određenu apstrakciju prikaže jasno, nedvosmisleno, usklađeno s kognitivnim predispozicijama učenika, te tako kod učenika potakne i izazove razumijevanje sadržaja koji proizlazi iz animacije.

Motorička obilježja – odnose se na to koliko ishod animacije može potaknuti ili doprinijeti aktivnostima učenika u nastavni (i izvan nastave), odnosno, potaknuti tzv. akciju učenika kojom će realizirati određene aktivnosti u nastavi, koje vode do razumijevanja sadržaja.

Perciptivna obilježja – uključuju tendenciju animacije da umanji (optimizira) količinu kognitivnog napora učenika potrebnog za spoznavanje sadržaja. O kakvoći ovih obilježja animacije ovisi hoće li animacija utjecati na percepciju učenika tako da on izvede pravilne zaključke, ili će izazvati nepremostivu opterećenost učenika te zbunjivanje, nezainteresiranost i nerazumijevanje sadržaja.

Afektivna i motivacijska obilježja – trebaju odgovoriti na pitanje kako animacija može utjecati na stavove, mišljenja i ponašanje učenika, te koliko može motivirati učenike za aktivnosti u nastavi. Znanstvenici često navode kako multimedijalna nastava, pa tako i animacija, pozitivno utječu na motivaciju i interes učenika [28], što može, ali i ne mora značiti da će polučiti razvoj pozitivnih vrijednosti i stavova.

Strategijska obilježja – mogu biti ključna za razumijevanje animacije, a odnose se na one karakteristike animacije koje mogu poboljšati učenikovu strategiju učenja. Učenika bi stoga trebalo osposobiti za učenje s animacije (razviti mentalne animacijske modele), a animaciju prilagoditi različitim strategijama učenja. Ova obilježja usko su povezana s metakognitivnim obilježjima animacije.

Metakognitivna obilježja – uključuju mogući utjecaj animacije na razvoj samoreguliranog učenja, odnosno, na to koliko animacija može utjecati na učenikovo planiranje učenja, samoprocjenu i vlastitu kontrolu na učenjem. U praksi je razvoj metakognicije učenjem iz animacije provediv samo kao nadogradnja znanja učenicima koji već imaju stečeno autentično iskustvo i razvijene pravilne animacijske mentalne modele. Kod učenika bez takvoga iskustva problem često predstavlja pogrešno razumijevanje animacije, odnosno, stvaranje tzv. iluzije o razumijevanju.

Retorička (socijalna) obilježja – odnose se na pitanje u kojoj mjeri animacija može doprinijeti socijalnom učenju kroz verbalizaciju (objašnjenja) sadržaja animacije u socijalnoj sredini. Dakle, animacija bi trebala biti poticajna za rad na zajedničkom i suradničkom razumijevanju sadržaja tako što će potaknuti učenike na verbalno pojašnjavanje sadržaja. Pri tom problem predstavlja osmišljavanje takve animacije koja može poboljšati verbalizaciju i učenje u socijalnoj sredini učenika, pri čemu je nastavnikovo vođenje presudno za uspjeh takvog učenja.

Pojedina obilježja nastavne animacije nisu neovisna jedna od drugih već se nalaze u složenoj međusobnoj interakciji, a naglašenost jednog obilježja može pozitivno ali i negativno utjecati na druga obilježja [8]. Izražajna obilježja bi sadržajem trebala biti vrlo ograničena jer su zasnovana na vremenskom slijedu događaja, što ograničava mogućnost učeničkog zapamćivanja istih. Učeničke strategije učenja ovise o dostupnim kognitivnim resursima učenika (metakognitivnim strategijama), ali i o motivaciji učenika potrebnoj za ulaganje napora kojima će otkriti prikladne strategije učenja. Povećanjem učenikove kontrole nad sadržajem animacije može se utjecati na višu razinu kognitivnih obilježja, ali takvo učenje pretpostavlja metakognitivne vještine i strategije koje učenik nema.

Današnja istraživanja animacije u nastavi uglavnom su usmjerena na istraživanja kognitivnih obilježja u umjetno stvorenim okolnostima, dok se druga navedena obilježja i njihov učinak na postignuća u nastavi sustavno ne proučavaju. Stoga je potrebno više pažnje posvetiti sveobuhvatnom učinku animacije u nastavi, u realnim i tradicijski uvjetovanim okolnostima nastave. Pri tom suvremene spoznaje iz područja kognitivnih znanosti i neuroznanosti mogu odigrati važnu ulogu u istraživanju ovog vrlo složenog područja.

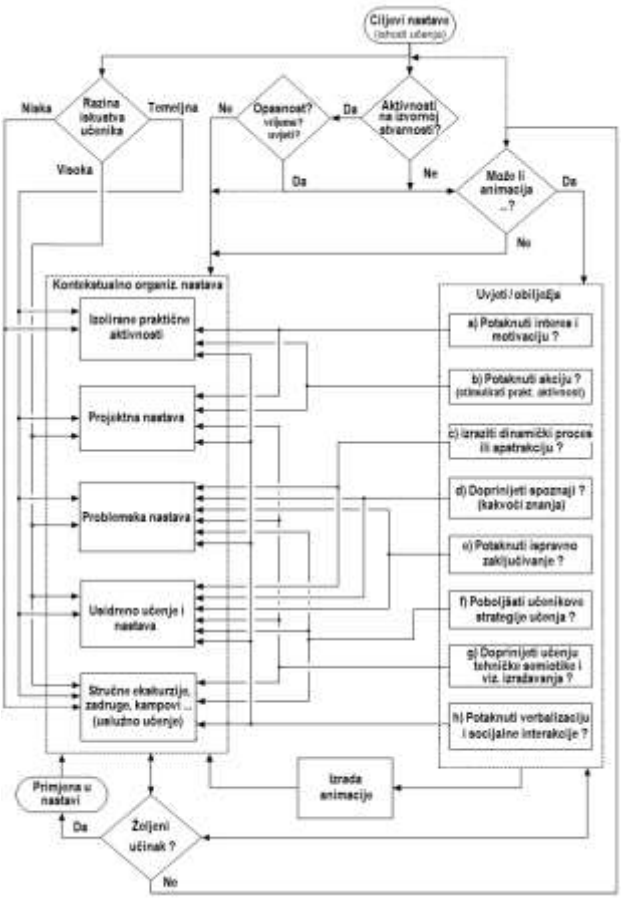
IV. PRIMJENA ANIMACIJE U TEMELJNOJ EDUKACIJI TEHNIKE

S obzirom na aplikativnu i kontekstualno omeđenu prirodu temeljnog tehničkog odgoja i obrazovanja, okvir za primjenu animacije razrađen je kao dio kontekstualnog pristupa nastavi, koji se smatra osobito učinkovitim u tehničkom nastavnom području [29]. Pristup se zasniva na kontekstualnom učenju i poučavanju, kao holističkom sustavu koji ima za cilj pomoći učenicima vidjeti smisao akademskog sadržaja kojeg uče povezivanjem sadržaja s kontekstom svakodnevnog života, odnosno, s kontekstom svojih osobnih, društvenih i kulturalnih okolnosti [7]. U nastavi općeg tehničkog odgojno-obrazovnog područja se, zbog ograničenja navedenih u uvodu, kao kontekstualno organizirana nastava može primijeniti: projektna nastava i učenje, rad u učeničkim zadrugama, kampovima, vrtovima i radionicama, stručne ekskurzije, problemska nastava i učenje, usidreno učenje i nastava te izolirane praktične aktivnosti učenika [30]. Navedeni pristupi ukazuju na to kako se kontekstualno učenje i poučavanje u tehničkom području zasniva na iskustvu učenika stečenom smislenim aktivnostima pri radu s izvornom stvarnošću. No, kako tu stvarnost nije uvijek niti u potpunosti moguće transferirati u školsko okružje, otvara se mogućnost uporabe računalnih animacija, kao kontekstualnih artefakata u nastavi, koji mogu poboljšati motivaciju učenika i razumijevanje tehničko-tehnoloških fenomena.

U sadržajnom smislu, svako tehničko obrazovanje se može zasnivati na procesima koji obilježavaju današnju tehniku i tehnologiju, kao što su: tehnički dizajn, rješavanje tehničkih problema, sistemski pristup, izum (inovacija) u tehnici, te proizvodnja [31]. Najvažnije generičke aktivnosti učenika, kojima se realizira takva nastava, uključuju: procjenu, komunikaciju, modeliranje, generiranje ideja, istraživanje i ispitivanje, proizvodnju i dokumentiranje [31]. Problem koji se pri tom često javlja je realizacija nastave prema unaprijed određenim (diktiranim), sadržajno zasićenim ili pak isključivo proizvodno inspiriranim aktivnostima, pri kojima do izražaja ne dolaze složeni spoznajni procesi kod učenika. Ovi procesi su osobito važni za razvoj metakognitivnih vještina i tehničkih kompetencija učenika. U tom smislu, tehničke kompetencije se mogu definirati kao posebna ponašanja u okviru tehničkog područja djelovanja, a formiraju se integracijom znanja, i/ili sposobnosti, i/ili vještina potrebnih za uspješno ostvarivanje konkretne

aktivnosti [32]. Iz takve određenosti kompetencija razvidno je kako se sposobnosti učenika za tehničko dizajniranje, stvaranje i razvoj inovacije, rješavanje tehničkog problema, ali i razvoj sistemskog pristupa radu, mogu razvijati uglavnom kontekstualnim pristupom nastavi. Zbog toga je važno utvrditi u kojim uvjetima i s kojom svrhom animaciju možemo primijeniti u kontekstualnim pristupima nastavi, vodeći se pri tom dosadašnjim znanstvenim spoznajama i posebnostima nastave općeg tehničkog odgoja i obrazovanja.

Kako bi se nastavnicima i praktičarima olakšalo donošenje odluke o razvoju, te načelno usmjerila izrada i primjena animacije u ovom nastavnom području, razrađen je dijagram uvjeta primjene animacije (slika 1), kao dio metodičkog okvira za primjenu animacije u nastavi. S desne strane dijagrama prikazana je moguća svrhovitost primjene animacije, koja u velikoj mjeri odgovara mogućim obilježjima animacije koja navodi Ainsworth [8], dok su na samom početku prikazani opći uvjeti bitni za svaku, pa tako i kontekstualno organiziranu nastavu tehničkog odgoja i obrazovanja. Među tim uvjetima, donošenje početne odluke o primjeni animacije u nastavi izravno je u svezi s mogućnostima uporabe izvorne stvarnosti na kojoj će se graditi glavne aktivnosti učenika. Iako je nastavnu animaciju moguće, a katkad i poželjno, koristiti i uz aktivnosti na izvornoj stvarnosti, animacija pri tom gubi prioritet, izuzev kao dio kontekstualizacije pri usidrenoj i problemskoj nastavi. Dakle, odluku o izradi i primjeni animacije je potrebno donijeti odmah na početku pripremanja kontekstualno organizirane nastave i u skladu s tom odlukom predvidjeti mogući doprinos animacije za postignuća učenika. Svaki od navedenih uvjeta povezan je s kontekstualnim pristupima nastavi, zbog posebnosti svakog pristupa i očekivano različite funkcije animacije u tom pristupu. Treći početni uvjet, iskustvo učenika, je neizravno povezan s odlukom o uporabi animacije putem kontekstualnih pristupa nastavi, a uvjetovan je očekivanim djelovanjem animacije na učenike. U tom smislu učenici s niskom razinom iskustva primarno trebaju pristupati izoliranim praktičnim aktivnostima radi stjecanja temeljnih znanja i vještina potrebnih za složenije aktivnosti. Za takvu početnu razinu učenja, animacije trebaju poslužiti kao sredstvo za poticanje interesa i aktivnosti učenika, ali i za poticanje verbalizacije, socijalnog pregovaranja i vlastitog, često i neprikladnog ali nužno potrebnog, tumačenja stvarnosti. Motivacijske animacije trebaju biti atraktivne za učenike, olakšavajuće po pitanju kognitivne opterećenosti i dobro vođene od strane nastavnika. Za projektnu nastavu, animacija, također, treba biti motivirajuća, poticajna za praktičnu aktivnost i za socijalne interakcije učenika. No, kako projektna nastava uključuje materijalizaciju tehničke tvorevine koja je produkt samostalnog rada učenika, očekuje se da se ovdje animacija koristi i kao sredstvo učenikovog predstavljanja vlastitog rada. Stoga do izražaja dolazi učenje temeljne tehničko-tehnološke semiotike i načina vizualizacije sadržaja. Za realizaciju problemske nastave te usidrenog učenja i nastave, kao kognitivno najsloženijeg segmenta temeljnog tehničkog odgoja i obrazovanja, potrebna je viša razina iskustva učenika, ali i više različitih animiranih i statičkih prikaza iste izvorne stvarnosti. Stoga, uz obilježja navedena za početno učenje, do izražaja dolaze izražajna, spoznajna, strategijska i obilježja povezana s mogućnošću izvođenja ispravnih zaključaka iz animacije. Ipak, u problemskoj



Slika 1. Dijagram uvjeta primjene animacije u kontekstualno organiziranoj nastavi

nastavi dominira samostalni rad učenika, stoga animacije koje učenici koriste trebaju biti izvedene na različitim spoznajnim razinama (od jednostavnih do zahtjevnih), prilagođene mogućim različitim dispozicijama učenika i strategijama učenja (vizualno, tekstualno i glasovno), te imati temeljne značajke interaktivnosti (upravljivost, oznake i signalizacija, poveznice s vanjskim sadržajima i sl.). Pri usidrenom učenju i nastavi samostalni rad učenika odvija se nakon i uz nastavnikovo kognitivno naukovanje zbog čega animacije trebaju biti kognitivno zahtjevne, temeljno upravljive te raznolike u smislu prikazivanja različitih situacija iste tehničko-tehnološke stvarnosti. Aktivnosti stručnih ekskurzija i različitih inačica radnog i uslužnog učenja, usmjerene su samostalnim aktivnostima učenika isključivo na izvornoj tehničko-tehnološkoj stvarnosti, zbog čega primjena animacija pri glavnoj aktivnosti nema osobitog smisla. Animacija je ovdje primjenjiva kao rezultat učenikovih spoznaja o toj stvarnosti, zbog čega se očekuje temeljno poznavanja alata za izradu iste, ali i upoznavanje i primjena dijela tehničko-tehnološke semiotike, odnosno načina tehničkog vizualnog izražavanja. Dakle, izrada animacije je pri ovakvoj nastavi dio produkta učenikovog rada i predstavljanja toga rada, a manifestacija razvijenog animacijskog mentalnog modela u završnoj etapi takve nastave je predmet evaluacije, učeničkih višestrukih tumačenja, te nastavnikovog moderiranja, usmjeravanja i dopunjavanja toga modela. U konačnici, nastavnik bi

učinak svake primjene animacije u nastavnom procesu, kao refleksivni praktičar, trebao vlastitim instrumentima vrednovati te, sukladno rezultatima, donijeti odluku o budućoj primjeni i/ili redizajnu animacije. Evaluacija učinka treba u prvom redu uzeti u obzir očekivanja od animacije i ciljeve nastave, a rezultat treba biti usmjeren primjeni, poboljšanjima ili pak odbacivanju iste. S obzirom na dominirajuće obilježje i svrhu primjene, animacije u općem tehničkom odgoju i obrazovanju mogu uključivati četiri osnovne skupine:

a) **motivacijske i poticajne ekspozitivne animacije** – predstavljaju kratkotrajni slijed slikovito i atraktivno predočenih događanja, procedura ili tehnološkog procesa, važnog za aktivnosti učenika koje slijede. Animacija treba biti popraćena verbalnim nastavnikovim pojašnjenjima i tek osnovnim tekstualnim sadržajima (elementi, glavni dijelovi, nazivlje, događaji). Dizajn ovakvih animacija treba biti jednostavan, s dominirajućim središnjim subjektom i minimalnim dodatnim elementima (optimalna kognitivna opterećenost učenika). Sadržaj animacije treba uskladiti s razvijenim mentalnim modelima učenika (poznati elementi), uz relativno nisku razinu apstraktnosti koju su učenici u mogućnosti razumjeti.

b) **izražajne i kognitivno značajne ekspozitivne animacije** – predstavljaju nositelje informacija bitnih za učenikovu spoznaju, zaključivanje i proces učenja. Ovakve animacije sežu od jednostavnih dinamičkih prikaza tehničkih tvorevina, sustava, procesa ili tehnologije, koji ustvari predstavljaju apstrahiranu izvornu stvarnost, preko složenih animacija koje prikazuju sadržaj na nižem stupnju apstrakcije (bliže izvornoj stvarnosti) ali s puno više činjenica i detalja, do visoko-apstraktnih dinamičkih shematskih prikaza i dijagrama za čije praćenje i zaključivanje su potrebni razvijeni mentalni modeli učenika. Uporaba sve tri razine ove vrste animacija je poželjna za potpuno razumijevanje određenog sadržaja, učenje tehničke semiotike i razvoj strategija učenja. Dizajn ovakvih animacija nužno treba biti na višoj razini apstrakcije, sa središnjim subjektom, ali i popratnim objektima, koji trebaju izazvati optimalnu kognitivnu opterećenost učenika. Kako bi se izbjegla kognitivna preopterećenost, učenicima se animacija višestruko ponavlja uz nastavnikovo pojašnjenje fokusirano svaki put na drugom skupinom elemenata animiranog prikaza.

c) **izražajne i kognitivno značajne interaktivne animacije** – imaju ista obilježja kao i prethodne, izuzev ugrađene dodatne interaktivnosti, glasovne podrške i poveznica na cjelovita tekstualna pojašnjenja i/ili izvore informacija. Primarno, ova vrsta animacija namijenjena je samostalnom (istraživačkom) radu učenika tijekom problemske nastave ili usidrenog učenja i nastave, ali može služiti i za samoučenje i udaljeno učenje. S obzirom da se radi o podvrsti prethodnih animacija, sve izražajne i kognitivno značajne animacije poželjno je razvijati s navedenim dodatnim elementima, s tim da je sve elemente moguće isključiti tijekom vođene nastave. Ovakve animacije trebale bi biti sastavnim dijelom repozitorija za nastavne materijale i e-učenje, koje bi nastavnici razvijali i razmjenjivali, čime bi se poboljšala dostupnost animiranih materijala, njihova raznolikost, a s vremenom i kakvoća.

d) **učeničke animacije** – predstavljaju jednostavne i kratkotrajne dinamičke prikaze učenikove percepcije tehničko-tehnološke stvarnosti, koja nam daje vrijednu povratnu informaciju o razvijenom mentalnom modelu učenika. Učenička izrada i prezentacija animiranog materijala je sastavni dio autentične evaluacije postignuća u kontekstualno organiziranoj nastavi. Ovakve animacije trebaju nastati kao učenikov izbor, a ne na izričit zahtjev nastavnika, jer se određena aktivnost, projekt, proizvod ili tehnologija mogu predstaviti i drugim načinima.

Sve vrste animacija bi, zbog optimalnog zadržavanja pažnje učenika, trebale trajati najdulje nekoliko minuta, što nikako ne determinira vrijeme i broj ponavljanja potrebnih za obradu sadržaja animacije izravno u nastavi. Dizajn i izradu animacija poželjno je zasnivati na alatima koji su široko dostupni nastavniku i učenicima, a za čije korištenje nastavnik ima dostatne vještine ili se te vještine mogu steći u relativno kratkom vremenu. U tom smislu motivacijske i učeničke animacije se mogu izraditi dostupnim uredskim aplikacijama i besplatnim softverom za izradu animiranih slika, kao na primjeru [33], kojima se relativno jednostavno može dodati govorna podrška. Ovakvi alati najčešće su dostatni i za izradu jednostavnih dinamičkih prikaza rada tehničkih tvorevina, procesa, sustava ili tehnologije, primjer [34]. Za složenije i realnije 3D dinamičke prikaze tehničko-tehnološke stvarnosti, kao i za visoko-apstraktne prikaze s elementima poželjne interaktivnosti, često su potrebni složeniji specijalizirani alati, koji ne moraju nužno biti komercijalni i skupi. Za izradu realnijih dinamičkih prikaza tehničko-tehnološke stvarnosti nastavnici tehničkog područja trebali bi ovladati osnovnim kompetencijama korištenja barem jednog alata za tehnički dizajn, koji ima mogućnost animiranog prikaza tehničke tvorevine, dok su za izradu visoko-apstraktnih i interaktivnih dinamičkih prikaza (i simulacija) poželjne kompetencije i vještine korištenja nekog široko rasprostranjenog skriptnog programskog jezika. Zbog jednostavnije distribucije sadržaja, osobito za potrebe problemskog i suradničkog učenja, poželjno je sve animacije razvijati na način da budu iskoristive u različitim sustavima za udaljeno učenje.

Ipak, vrijedi napomenuti kako animacija u tehničkom obrazovanju nije sredstvo na kojem se zasnivaju glavne aktivnosti, pa time nije ni osnova za spoznaju, već kvalitativni kontekstualni artefakt koji može, ali i ne mora, doprinijeti spoznaji i postignućima učenika.

V. ZAKLJUČAK

Uspješnost učenja i nastave u općem tehničkom odgojno-obrazovnom području često se zasniva na kontekstualno organiziranim učeničkim aktivnostima, pri kojima je autentičnost artefakata neizostavna. Ipak, zbog nedostatnog iskustva učenika i uvjeta školskog okružja, aktivnosti je potrebno prilagođavati učenicima, a izvornu stvarnost često nadopunjavati didaktički prerađenom. Kontekstualizacija je pri tom važan segment nastavnog procesa pri čemu, uz prioritetnu izvornu stvarnost, različiti kontekstualni artefakti, poput računalne animacije, mogu doprinijeti postignućima učenika i motivirati ih za rad.

Dosadašnja istraživanja uporabe računalne animacije u nastavi često osporavaju njen izravni doprinos boljem

razumijevanju sadržaja i postignućima učenika. Unatoč tome, rezultati istraživanja ukazuju kako u primjerenim uvjetima optimalne kognitivne opterećenosti, dobre pripremljenosti učenika, predviđanja dinamike procesa, višestrukog opetovanja i interaktivnosti, animacija može doprinijeti razvoju ispravnih mentalnih modela učenika.

Uzimajući u obzir rezultate empirijskih istraživanja i posebnosti kontekstualno organizirane nastave općeg tehničkog odgoja i obrazovanja, izdvojen je metodički okvir za primjenu animacije u ovom nastavnom području. Primjenu animacije uvjetuje početno iskustvo učenika, kontekstualno organizirana glavna aktivnost, krajnji ciljevi nastave i obilježja kojima animacija može doprinijeti postignućima učenika. U skladu s navedenim uvjetima, za izolirane praktične aktivnosti preporučuje se uporaba motivacijskih animacija prije glavne aktivnosti učenika, za usidreno i problemsko učenje izražajne i kognitivno značajne animacije tijekom aktivnosti, dok se za varijante projektnog i uslužnog učenja preporučuje korištenje učeničke animacije nakon glavne aktivnosti.

Potrebno je istaknuti kako su dosadašnje spoznaje o uporabi animacije u nastavi i učenju zasnovane na teorijskim pretpostavkama i empirijskim istraživanjima u umjetnim, laboratorijskim uvjetima. Stoga daljnja istraživanja treba usmjeriti na uvjete i utjecaj animacije u stvarnim nastavnim uvjetima i situacijama i u različitim nastavnim područjima. Rezultati takvih istraživanja mogu bitno utjecati na dosadašnje spoznaje o utjecaju animacije na učenike, pa tako i na predloženi okvir za primjenu animacije u nastavi općeg tehničkog odgoja i obrazovanja.

LITERATURA

[1] S. Petrina, Advanced Teaching Methods for the Technology Classroom, Hershey-London-Melbourne-Singapore: Information Science Publishing, 2007.

[2] T. Kelley, N. Kellam, "A Theoretical Framework to Guide the Re-Engineering of Technology Education", Journal of Technology Education, Vol. 20 (2), pp. 37-49, 2009.

[3] S. Kovačević, Kurikulumska matrica tehničkih kompetencija u odgoju i općem obrazovanju (Doktorska disertacija), Zagreb: Filozofski fakultet Sveučilišta u Zagrebu, 2012.

[4] J. Milat, "Tehnička kultura bitna je odrednica sustava obrazovanja", Društvena istraživanja, Vol. 5 (1, 21), pp. 109-128, 1996.

[5] A. A. Verbitsky, Aktivnoye obucheniye v vysshey shkole: kontekstnyj podhod [Active Teaching in College: Contextual Approach]. Moscow: Vysshaya shkola, 1991.

[6] D. Purković, "Konstruktivistički pristup operacionalizaciji kurikuluma tehničke kulture", Pedagogijska istraživanja, Vol. 1/2013, pp. 49-64, 2013.

[7] E. B. Johnson, Contextual teaching and learning: what it is and why it's here to stay. Thousand Oaks, CA: Corwin Press, INC, 2002.

[8] S. Ainsworth, "How do animations influence learning?", in Current Perspectives on Cognition, Learning, and Instruction: Recent Innovations in Educational Technology that Facilitate Student Learning, D. Robinson and G. Schraw, Eds., pp. 37-67, 2008.

[9] E. Mayer, R. Moreno, "Animation as an aid to multimedia learning", Educational Psychology Review, Vol. 14(1), pp. 87-99, 2002.

[10] M. Bétrancourt, B. Tversky, "Effect of computer animation on users' performance: a review", Le Travail Humain, Vol. 63(4), pp. 311-330, 2000.

[11] Moviestorm. [Online]. Using animation in schools: a practical handbook for teachers. Moviestorm Ltd, 2011, Available: http://www.moviestorm.net [Accessed:12.05.2014].

[12] L. Xiao, "Animation Trends in Education", International Journal of Information and Education Technology, Vol. 3 (3), pp. 286-289, June, 2013.

[13] E. Luader, 3D Displays. Chichester, UK: John Wiley & Sons (ISBN: 978-1-119-99151-9), 2012.

[14] J. Sweller, "Cognitive load during problem solving: Effects on learning", Cognitive Science, Vol. 12, pp. 257-285, 1988.

[15] R. E. Mayer, "Multimedia learning: Are we asking the right questions?", Educational Psychologist, Vol. 32, pp. 1–19, 1997.

[16] R. E. Mayer, Multimedia learning. New York: Cambridge University Press, 2001.

[17] W. Schnotz, T. Rasch, "Enabling, Facilitating, and Inhibiting Effects of Animations in Multimedia Learning: Why Reduction of Cognitive Load Can Have Negative Results on Learning", Educational Technology Research and Development, Vol. 53 (3), pp. 47–58, 2005.

[18] P. Wouters, "How to optimize cognitive load for learning from animated models", Nieuwegen: The Netherlands Organisation for Scientific Research, Project no. 411-02-163, 2007.

[19] T. van Gog, F. Paas, J. Sweller, "Cognitive Load Theory: Advances in Research on Worked Examples, Animations, and Cognitive Load Measurement", Educational Psychology Review, Vol. 22, pp. 375-378, 2010.

[20] W. Schnotz, R. K. Lowe, "External and internal representations in multimedia learning", Learning and Instruction, Vol. 13, pp. 117–123, 2003.

[21] M. Hegarty, S. Kriz, C. Cate, "The Roles of Mental Animations and External Animations in Understanding Mechanical Systems", Cognition and Instruction, Vol. 21(4), pp. 325-360, 2003.

[22] S. Kriz, M. Hegarty, "Top-down and bottom-up influences on learning from animations", International Journal Human-Computer Studies, Vol. 65, pp. 911–930, 2007.

[23] M. Hegarty, V. Sims, "Individual differences in mental animation during mechanical reasoning", Memory & Cognition, Vol. 22(4), pp. 411-430, 1994.

[24] H. W. Gardner, "Educating for understanding", The American School Board Journal, Vol. 180(7), pp. 20-24, 1993.

[25] S. F. Fong, "Effects of segmented animated graphics among students of different spatial ability levels: a cognitive load perspective", The Turkish Online Journal of Educational Technology, Vol. 12(2), pp. 89-96, 2013.

[26] M. Bétrancourt, "The Animation and Interactivity Principles in Multimedia Learning", in The Cambridge Handbook of Multimedia Learning, R. Mayer, Ed. New York: Cambridge University Press.

[27] R. K. Lowe, "Animation and learning: selective processing of information in dynamic graphics", Learning and Instruction, Vol. 13, pp. 247–262, 2003.

[28] J. Bezjak, "Contemporary engineer pedagogic's project research - using multimedia at technology classes in technical and vocational schools", Procedia Social and Behavioral Sciences, Vol. 2, pp. 407–411, 2010.

[29] D. Parnell, Contextual Teaching Works! Increasing Students' Achievement. Waco, TX: CCI Publishing, 2001.

[30] D. Purković, "Kontekstualni pristup učenju i poučavanju u nastavi temeljnog tehničkog odgoja i obrazovanja", Školski vjesnik, Vol. 1/2015, prihvaćen za tisak.

[31] P. J. Williams, "Design: The Only Methodology of Technology", Journal of Technology Education, Vol. 11(2), 2000.

[32] D. Bjekić, M. Bojović, S. Dragičević, M. Bjekić, "Razvijanje tehničke kompetentnosti", Pedagogija, Vol. LXXIII (1), 2008.

[33] D. Purković, Nastavna animacija "Ohmov zakon", [Online]. Available: http://polifem.ffri.hr/cte/anim/Ohmov_zakon.htm.

[34] D. Purković, Nastavna animacija "Robot EMoRo", [Online]. Available: http://polifem.ffri.hr/cte/anim/robot_naputak.htm.

Žene i informacijsko-komunikacijske tehnologije: zašto su studentice rijetkost na studiju računarstva?

Tihana Babić, Ana Ogrin i Klaudija Šarkanj
Visoko učilište Algebra - visoka škola za primijenjeno računarstvo, Zagreb, Hrvatska
tihana.babic@algebra.hr, ana.ogrin@algebra.hr, klaudija.sarkanj@algebra.hr

Rad se bavi pitanjem vrlo slabog interesa žena za studije računarstva odnosno informacijsko - komunikacijskih tehnologija. Statistički podaci Agencije za znanost i visoko obrazovanje pokazuju da u Hrvatskoj samo 13% studentica izabire studije u području tehničkih, biotehničkih, inženjerskih i prirodnih znanosti (STEM).

Na Visoko učilište Algebra – visoku školu za primijenjeno računarstvo upisan je 491 student na preddiplomske i diplomski studij u školskoj godini 2014./2015., od čega samo 43 studentice, što znači da je zastupljenost studentica manja od 9%. Svrha ovog kvalitativnog istraživanja, u kojem je sudjelovalo 36 studentica, je identificirati motive upisanih studentica te, suprotno, moguće razloge zbog kojih studij računarstva nije popularniji odnosno atraktivniji među djevojkama.

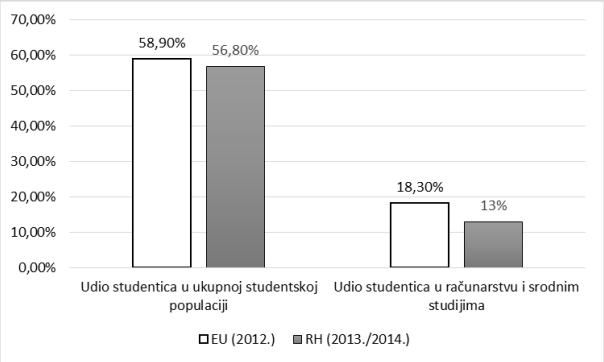
Studentice ističu intrinzične motive za upis, ali i dobre poslovne mogućnosti nakon završenog studija. Smatraju da glavni uzroci slabe zastupljenosti žena u ICT-u leže u tradicionalnom vjerovanju da je to muški posao, slabijem predznanju u području tehnologija, društvenom neprihvaćanju te slaboj informiranosti.

Kao moguća rješenja proizašla iz ovog istraživanja ističu se obvezna rana informatička edukacija, adekvatno informiranje učenika te kreativne promotivne aktivnosti.

I. UVOD

Prema podacima Državnog zavoda za statistiku u zimskom semestru školeske godine 2013./2014. na visoka učilišta Republike Hrvatske, uključivo upis na sve godine studija, upisalo se ukupno 161 911 studenata, od čega 91 997 studentica. Udio studentica je dakle 56,8% [1]. Ako se taj podatak usporedi sa statističkim podacima iz analize interesa mladih za STEM Balena i Drvodelića prema kojoj samo 13% studentica odabire studije u području znanosti, tehnologije, inženjerstva i matematike (STEM) [2], pitanje "zašto?" nameće se samo po sebi.

Statistički podaci iz godine ranije, odnosno 2012. godine, za 27 zemalja Europske unije govore da je preddiplomski i diplomski studij upisalo 4.800,964 studenata od čega 2.826,619 žena [3] (u postotku 58.9). Usporedbom podataka možemo zaključiti da žene i u drugim zemljama Europske unije, kao i u Hrvatskoj, u većem broju upisuju studij od muškaraca (u prosjeku oko 15%). No, udio žena u ukupnom broju studenata upisanih na studij računarstva u 27 zemalja Europske unije (njih 28,246 od ukupno 153,856 studenata) je 18,3% i postotak Hrvatica je ipak niži od europskog prosjeka [3].



Grafikon 1. Usporedba udjela studentica u ukupnoj studentskoj populaciji i studijima u računarstvu u zemljama EU i Republici Hrvatskoj

Na Visoko učilište Algebra – visoku školu za primijenjeno računarstvo upisan je 491 student na preddiplomske studije primijenjenog računarstva i multimedijskog računarstva i diplomski studij primijenjenog računarstva u školskoj godini 2014./2015., od čega samo 43 studentice, što znači da je zastupljenost studentica manja od 9%. Prema tome taj postotak je još manji od gore navedenih podataka.

Slika 1. Promocija 25.04.2014.g.: diplomirali na Visokom učilištu Algebra u šk.god. 2013./2014. (29 prisutnih studenata koji su diplomirali od čega 2 diplomantice)

Zašto su žene rijetkost u informacijsko-komunikacijskim znanostima pitanje je na čiji se odgovor ne smije dugo čekati. Općenito se u hrvatskom obrazovnom sustavu ne ulaže dovoljno u informacijsko–komunikacijske tehnologije (ICT) iako je Strategija obrazovanja, znanosti i tehnologije RH, koju je donio Hrvatski sabor 17. listopada 2014. godine, veoma naglasila značaj ICT-a: „Europski tehnološki prioriteti vezani uz industrijsko vodstvo usmjereni su potpori istraživanjima i inovacijama u informacijskoj i komunikacijskoj tehnologiji (ICT), nanotehnologijama, naprednim materijalima, biotehnologiji i naprednoj proizvodnji. Procjenjuje se da više od 10% svjetskog BDP-a ovisi o informacijskoj i komunikacijskoj tehnologiji, a internet i web ključni su za pokretanje inovacija i kreativnost u društvu i gospodarstvu.“ [4]. Nažalost, vrlo izgledno je da se radi o problemu financijske prirode, jer je istom propisano da će se u Republici Hrvatskoj do 2020. godine ulagati 6% BDP-a u obrazovanje [4], dok se sada ulaže tek do 3,5% [5].

Potaknute navedenim podacima te činjenicom da je udio žena u ICT-u drastično manji od udjela muškaraca, autorice su pokušale do odgovora kako bi se taj broj mogao povećati ne zadirući u državne financije već u mentalitet svih sudionika odgojno-obrazovnog procesa: učenika, nastavnika i roditelja.

II. METODOLOGIJA ISTRAŽIVANJA MOTIVA, ZADOVOLJSTVA IZBOROM I UZROKA SLABE ZASTUPLJENOSTI ŽENA U PODRUČJU RAČUNARSTVA

A. Sudionici

U istraživanju koje je provedeno na Visokom učilištu Algebra - visokoj školi za primijenjeno računarstvo, sudjelovalo je 36 studentica stručnih studija računarstva. U istraživanju je korišten prigodni uzorak.

U Tablici I. prikazana je struktura sudionica s obzirom na studij, razinu i godinu studija.

TABLICA I. STRUKTURA SUDIONICA PREMA STUDIJU

	Primijenjeno računarstvo	**Multimedijsko računarstvo**
Preddiplomski studij	**17**	**10**
1. godina	*8*	*6*
2. godina	*3*	*2*
3. godina i apsolventice	*6*	*2*
Diplomski studij	**5**	**4**
1. godina	*4*	*2*
2. godina	*1*	*2*

B. Instrumentarij

Podaci su prikupljeni pomoću upitnika koji je sadržavao pitanja višestrukog izbora i pitanja otvorenog tipa bez unaprijed ponuđenih odgovora. Nakon objašnjenja svrhe ispitivanja, sve studentice dobile su upitnike te su zamoljene da iskreno odgovore na pitanja. Odgovori ispitanica obrađeni su kvantitativnom i kvalitativnom analizom.

III. REZULTATI

A. Motivi za studij računarstva

Studenticama je postavljeno pitanje kako su se odlučile za studij računarstva. U upitniku su unaprijed ponuđena tri moguća motiva koji se u javnosti često ističu kao prednosti studija u ovom polju te dodatna kategorija „nešto drugo“ odnosno otvoreno polje za unos konkretnog odgovora. Sudionice su mogle birati više odgovora.

U Tablici II. prikazane su frekvencije odgovora studentica na pitanje o motivima izbora studija u računarstvu.

TABLICA II. MOTIVI IZBORA STUDIJA (N=36)

Razlog upisa	**Postotak pojedinog odgovora**
To ih oduvijek zanima	78%
Visoke plaće	8%
Velika potražnja za tim profilom stručnjaka	14%
Nešto drugo *(želja za učenjem, priroda posla)*	11%

Iz odgovora studentica vidljivo je da su glavni motivi za upis studija intrinzične prirode odnosno da proizlaze iz njihovih unutarnjih pobuda jer je to područje njihovog interesa. Njih 78% navelo je kako je to područje koje ih oduvijek zanima. Tzv. ekstrinzični faktori vezani uz tržište rada (kao što su npr. plaće i laka zapošljivost) odnosno izvana potaknuti motivi pojavljuju se, ali u puno manjim frekvencijama, visoke plaće predstavljaju izvor motivacije za 8% ispitanih studentica, a velika potražnja za ovim profilom stručnjaka za 14% studentica. Njih 11% navelo je da ih je motivirala priroda tog posla i želja za učenjem, . Zanimljivo je da se s druge strane prilikom razgovora o potrebi povećanja stručnjaka u ICT-u često posebno naglašavaju upravo pozitivni aspekti tržišta rada kao nešto

što bi trebalo motivirati učenike na odabir ovakvih zanimanja, što nije u skladu s odgovorima studentica.

B. Zadovoljstvo izborom (područja i polja) studija

Na pitanje bi li sada promijenile područje studija da ponovno upisuju fakultet, njih 5, odnosno 14% je odgovorilo da bi mijenjale studij. Od toga je samo jedna studentica navela razlog promjene i to „diskriminaciju i podcjenjivanje od strane muških kolega". Ostale studentice (njih 31, odnosno 86%) ne bi mijenjale studij, a razlozi zašto žele ostati u području ICT-a prikazani su u Tablici III.

TABLICA III. RAZLOZI OSTAJANJA U ICT-u (N=31)

Razlog ostajanja u ICT-u (kategorije odgovora)	**Postotak pojedinog odgovora**
zanimljivo područje	45%
mogućnost kontinuiranog učenja i razvoja vlastitih potencijala	29%
to je zanimanje budućnosti (traženo je i dobro plaćeno)	13%
dinamika razvoja i ljubav prema tehnologijama	13%
vrlo izazovno područje	10%
široko područje rada (raznovrsni poslovi)	10%
logičan nastavak dosadašnjeg obrazovanja	6%
kreativnost područja	3%
internacionalnost područja	3%

Navedeni razlozi ostajanju u području informacijsko-komunikacijskih tehnologija povezani su s inicijalnim motivima upisa studija te su u tom smislu i očekivani. Gotovo polovica studentica (45%) ne bi mijenjale studij zato jer ovo područje smatraju zanimljivim. Ostali češće zastupljeni razlozi ostajanja u području računarstva odnose se na mogućnost kontinuiranog učenja i razvoja potencijala (29%), potražnja na tržištu rada i perspektiva u budućnosti (13%) te dinamika razvoja tehnologija (13%). Opet se pokazuje da unutarnji motivi, interes za to područje i želja za učenjem, predstavljaju glavne faktore ustrajanja u odabranom studiju.

C. Uzroci slabe zastupljenosti žena

Na pitanje koji su, po njihovom mišljenju, uzroci slabe zastupljenosti žena u računarstvu i ICT-u iskristalizirali su se sljedeći odgovori (N=36):

- Predrasude da je to muški posao (36%).
- Karijera žene u ICT-u još uvijek je „društveno neprihvatljiva" (6%).
- Žene zanimaju druga područja (8%).
- Djevojke se od rane dobi usmjeravaju u društveno područje (17%).
- Djevojke se u startu smatraju manje sposobnima za takve poslove pa se trebaju dodatno dokazivati, na što one reagiraju tako da izbjegavaju takve izazove (6%).
- Nisu dovoljno i adekvatno informirane (19%).
- Postoji nedostatak socijalne podrške i diskriminacija prema ženama koje se odluče za takve poslove (8%).

Od svih navedenih, prilikom odabira karijere među studenticama najčešće prevladavaju razlozi koji upućuju na tradicionalnu svijest da je to muški posao (njih 36%), te razlozi koji govore da su djevojkama tehnologije daleke, odnosno da prirodno biraju društvena područja (17%) i nisu dovoljno upoznate s njima, odnosno nisu adekvatno informirane (19%).

D. Koraci za povećanje udjela žena

Jedan dio studentica smatra da se zapravo ne može puno učiniti kako bi se povećao udio žena u računarstvu i općenito u ICT-u jer udio žena u toj branši zapravo je razmjeran udjelu žena koje pokazuju sklonost prema tehničkim znanostima u ukupnoj ženskoj populaciji. Osim toga, smatraju da je teško mijenjati stavove, a i žene idu linijom manjeg otpora. To je navelo 8% studentica, dok 17% ispitanih studentica ne zna što učiniti. Međutim, ostalih 75% studentica je ipak predložilo neke korake kojima bi se situacija u budućnosti mogla poboljšati, a koji su nabrojeni u nastavku:

- Pružati više informacija o IT-u i zanimanjima iz tog sektora te o njihovim opisima od najranije dobi (11%).
- Približiti znanost i demistificirati je odnosno razbiti mitove o podjeli na „muške" i „ženske" karijere (11%).
- Već u osnovnoj školi započeti s obaveznim informatičkim obrazovanjem, koje će pripremiti i motivirati za zanimanja u području ICT-a (8%).
- Osvijestiti društvo i poslodavce (kao i same žene) da se žene u jednakoj mjeri mogu posvetiti ovoj vrsti posla (djelatnosti) kao i muškarci (6%).
- Educirati žene, a i promijeniti stav određenih muškaraca u ICT svijetu odnosno uvesti pozitivnu diskriminaciju (6%).
- Približiti IT ženama kroz određene marketinške kampanje (6%).
- Učiniti način na koji se podučava zanimljivijim (3%).
- Prikazivati računala kao alat pomoću kojeg se mogu napraviti razne kreativne stvari, a ne kao nešto dosadno (3%).
- Uvesti osnove programiranja u srednje škole (3%).
- Organizirati radionice s jednostavnim primjerima koje bi mogle pomoći u porastu interesa, a također ženama pomoći da pronađu granu ICT-a koja im se sviđa i kojom bi se jednog dana mogle baviti (3%).

- Promovirati primjere uspješnih žena u ICT-u te timove i poduzeća s dobrom radnom atmosferom (3%).
- Prezentirati ICT kao industriju otvorenu svima koja ne trpi diskriminiranje, s naglaskom na raznolikost karijera u tom području (3%).
- Uvesti posebne programe za uključivanje žena na projekte na fakultetu i razmjenu studenata (3%).
- Kroz poticajne mjere podržati razvoj ICT poduzetništva čiji su osnivači žene (3%).
- Više naglasiti da su žene dobrodošle i stvoriti atmosferu u kojoj bi se osjećale ugodno i na ovakvim do sada muškim studijima (3%).

IV. ZAKLJUČAK

Iako je zbog prirode uzorka korištenog u ovom istraživanju nemoguće napraviti generalizaciju rezultata na cjelokupnu studentsku populaciju u Hrvatskoj, ovi podaci se ipak mogu smatrati indikativnima te potaknuti daljnja istraživanja i konkretne akcije za povećanje udjela žena u ICT-u.

Osnovni motivi za izbor studija su unutarnje prirode za razliku od vanjskih faktora koji se u radu s učenicima i studentima češće naglašavaju (sigurna radna mjesta i visoke plaće). Stoga je važno od rane dobi učenike, a osobito učenice, „izlagati“ pozitivnim stranama tehnologija i omogućiti im da razviju znatiželju i strast za istraživanjem.

U obrazovanje je nužno uključiti sadržaje iz informatike i računarstva na način da oba spola imaju jednaku priliku upoznati područje, ali i svoje sposobnosti za bavljenje računarstvom. To ne mora nužno značiti informatiku kao obavezan predmet, već se informatički alati trebaju uklopiti u sve ili većinu školskih predmeta.

Isto tako, u postupcima profesionalnog usmjeravanja učenika tijekom školovanja organizirati terensku nastavu kroz npr. posjete tehnološkim tvrtkama. Zatim, kroz ponudu izvannastavnih aktivnosti omogućiti učenicima da se iskustveno bave tehničkim aktivnostima.

Ipak, osnovni problem identificiran u ovom radu jest naš „mentalni sklop“, odnosno tradicionalno utemeljene podjele zanimanja i predrasude prema ženama u ICT-u. To je problem koji treba započeti rješavati u sustavu obrazovanja, ali i izvan njega, u politikama zapošljavanja i poslovnom sektoru. Jedan od mogućih načina promjene (koja nikako neće biti brza) jesu kampanje koje bi trebale imati za cilj upoznati široku javnost (učenike, roditelje, nastavnike) s poslovima u području ICT-a te prikazati primjere dobre prakse, odnosno žene koje su uspjele u tom području.

Temeljni je zaključak kako je potrebno provesti široku raspravu o ovom problemu i ne prepustiti strukturu stručnjaka u ICT-u pukoj stihiji i tradicionalnim društvenim stavovima. To je obrazovno, sociološko, ali i ekonomsko pitanje koje se nikako ne smije zanemariti.

LITERATURA

[1] Priopćenje DZS, GODINA LI. ZAGREB, 13. KOLOVOZA 2014. BROJ: 8.1.7., ISSN 1330-0350

[2] eSkills for Jobs 2014 (2015), http://eskills.hr/wp-content/uploads/2014/11/Predstavljanje-analize-interesa-mladih-za-STEM-studijima.pdf. Pristupljeno 01.02.2015.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Home - Eurostat - European Commission, http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do. Pristupljeno 05.02.2015.P. Perić, “Naslov rada,” naziv časopisa, prihvaćen za tisak

[4] 124 24.10.2014 Strategija obrazovanja, znanosti i tehnologije, http://narodne-novine.nn.hr/clanci/sluzbeni/2014_10_124_2364.html M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] Državni proračun 2014. godina - Ministarstvo financija, http://www.mfin.hr/hr/drzavni-proracun-2014-godina

Prema uspješnijem razvoju algoritamskog načina razmišljanja

K. Blažeka, Tehnička škola Čakovec, Čakovec, Hrvatska
kristinka.blazeka@ck.t-com.hr

Sažetak - Novi strukovni kurikulumi, iako još u eksperimentalnoj provedbi, svojim naglaskom na temeljne ishode učenja uveli su niz pozitivnih pomaka u pristupu i provedbi obrazovanja za zanimanje Tehničar za računalstvo. Očekivano, najvažniji ishodi učenja povezani su sa stjecanjem vještina traženih na tržištu rada jer su od presudne važnosti za potvrđivanje kompetentnosti za obavljanje poslova i radnih zadataka odabrane struke. Među istaknutim vještinama, usvajanje vještine ili umijeća programiranja računala značajno je zbog toga što već duži niz godina omogućava povećanu zapošljivost. Kako u osnovnim školama usvajanje algoritamskog načina razmišljanja i stjecanje osnovnih vještina programiranja još uvijek ne obuhvaća čitavu populaciju, pohvalno je i za učenike korisno što pristup sadržan u novim kurikulima daje veći naglasak na usvajanje ovih temeljnih koncepata. U radu se ističe važnost odvajanja dovoljnog broja sati za te prve, vrlo važne korake za ulazak u svijet ispravno razvijenog algoritamskog načina razmišljanja i usvajanje vještine programiranja, uspoređuju se primjeri i sadržajni obuhvat u postojećim odobrenim udžbenicima, te analiziraju mogućnosti koje za podučavanje razvoja algoritamskog načina razmišljanja nude neka kvalitetna web sjedišta odnosno nacionalni sustavi natjecanja. U zaključku se daju prijedlozi za primjenu i poboljšanja.

I. UVOD

Svakodnevni je život neprekidno obilježen mnogim rutinskim nizovima aktivnosti stvorenim izvođenjem pripadajućih algoritama tijeka akcija i informacija. Takav složen svijet na odgovarajući način postaje shvatljiv samo onima kojima su koncepti takvih aktivnosti bliski, odnosno onima koji posjeduju potrebne vještine za njihovo stvaranje, izvođenje ili praćenje, a obuhvaćaju algoritamsko razmišljanje, logičko zaključivanje i strukturirano rješavanje problema.

S ciljem postizanja važnih kompetencija 21. stoljeća [1] svojim je ishodima poučavanje ICT-a u primarnom i sekundarnom obrazovanju obično usmjereno primarnom stjecanju spomenutih koncepata i vještina, odnosno osiguravanju i ostvarivanju sposobnosti učenika da probleme rješavaju poduzimajući odgovarajuće akcije.

Pri tom je sposobnost algoritamskog načina razmišljanja jedna od najvažnijih kompetencija koja se sve više utemeljuje upravo na tim razinama obrazovanja. Za ilustraciju, ciljevi koje postavlja Model Curriculum for K–12 Computer Science su predstavljanje „osnovnih koncepata računalne znanosti svim učenicima, počevši od razine osnovne škole" [2] pa se tako zahtijeva da "do završetka petog razreda učenici trebaju razviti jednostavno razumijevanje pojma algoritma.", što je ujedno i osnova i preduvjet za usvajanje jednog od najtežih ishoda učenja o ICT-u - programiranja.

Na sličan se način i u kurikulumu obrazovnog sustava Australije očekuje kako će učenici petog i šestog razreda biti sposobni dizajnirati, mijenjati i slijediti jednostavne algoritme (prikazane dijagramima i riječima) koji uključuju nizove koraka, grananja i iteracija (ponavljanja), a u osmom bi trebali implementirati i mijenjati programe koji uključuju grananja, iteracije i funkcije u nekom programskom jeziku opće namjene [3].

Kod nas se prvi koraci u učenju programiranja u novom strukovnom kurikulumu za zanimanje Tehničar za računalstvo odvijaju kroz predmet Algoritmi i programiranje. Nastava ovog predmeta podijeljena je na uobičajeni način na predavanja i prateće laboratorijske vježbe, odnosno usvajanje vještina kroz praktičan rad u laboratoriju, pri čemu je u okviru redovne nastave poželjno predvidjeti i vrijeme namijenjeno učenicima bez dovoljnog predznanja.

II. AKTIVNO UČENJE U NASTAVI ICT-A

Tijekom predavanja nastavnik obično daje naglasak na prezentiranje teoretskih osnova. Stoga je takav način uglavnom usmjeren na nastavnika uz minorizirane učeničke aktivnosti, pa se traže načini uvođenja aktivnog učenja.

Naime, učenici aktivnije uključeni u vlastiti proces učenja motiviraniji su, razvijaju kvalitetnije vještine u povezivanju predznanja, novih ideja i prisjećanja informacija [4]. U okolinama aktivnog učenja učenici su aktivni u procesu učenja: osim što slušaju, oni i reflektiraju vlastiti proces učenja, razmišljaju o rješenjima problema i diskutiraju u grupama [5] [6].

U takvim, okolinama usmjerenim na učenika, učenici su u žarištu procesa učenja, postaju intrinsično motiviraniji, proaktivni su i iziskuju više informacija vezanih uz njihove osobne strategije učenja. Pri tom se uloga nastavnika mijenja u vođu koji im je podrška, pa kad imaju neka pitanja, vođa im daje sugestije, ali ne i konačno rješenje. Po potrebi im dodatno opisuje vježbe i postavlja im kritična pitanja kako bi im pomogao da počnu razmišljati o njihovim problemima.

S obzirom se kod poučavanja informacijskih tehnologija u srednjim školama (i u višim razinama obrazovanja) znanja mijenjaju i šire vrlo brzo, pokazuje se neadekvatnim žarište staviti na učenje i reprodukciju faktografskog znanja. Neizbježno i nastavnici i učenici neprestano trebaju obnavljati svoje predmetne vještine i

znanja. U takvoj se situaciji, stoga, adekvatnijim čini konstruktivistički naglasak na poučavanje vještina rješavanja problema umjesto na inzistiranje na faktografskom znanju.

Za ostvarenje takve nastave preporučene smjernice uključuju pristupe kao što su postavljanje autentičnih zadataka koji potiču radoznalost učenika, korištenje pogrešaka u smislu produktivne komponente procesa učenja, naglasak na konstrukciju, a ne reprodukciju znanja i ohrabrivanje dijaloga i kolaborativnog učenja [7].

Obećavajući pristup razvoju konstruktivističkih strategija unutar bihevioristično sustava je minimalistički pristup [8] koji podupire prirodne strategije učenika [9], a obilježavaju ga navedene odrednice:

• okolina učenja treba biti usmjerena na korisnika;

• učenici trebaju što prije početi na stvarnim zadacima;

• osigurati potporu za istraživanje, zaključivanje i improvizaciju;

• pogreške su neophodan dio procesa učenja;

• potrebno je snažno podržati prepoznavanje i popravljanje pogrešaka;

• na najmanju mjeru treba svesti ometajuće nastavne materijale;

• nastavni materijali trebaju biti strukturirani i modularni;

• korisnici ne bi smjeli biti ograničeni na predodređeni slijed učenja;

• korisnici bi trebali imati slobodan pristup onim nastavnim materijalima koje trebaju i ignorirati ostale.

Dodatno, sugestije nastavnicima koji žele prihvatiti konstruktivističke strategije, odnosno konstruktivistički (minimalistički) način poučavanja ICT-a bi bile:

• hrabriti i očekivati od učenika preuzimanje odgovornost za vlastito učenje;

• njegovati radoznalost i interes učenika;

• upotrebom suvremenih alata postavljati stvarne zadatke u autentičnim kontekstima;

• radije oblikovati doživljaje i okoline učenja nego birati nastavne strategije koje vode specifičnim ishodima učenja;

• postavljati otvorene zadatke očekujući da će ih učenici modificirati;

• u opisu zadataka koristiti različite medije i prikaze (priče, brojeve, pravila, statističke sažetke, fotografije, zvuk i video);

• koristiti stvarne, sirove podatke i primarne izvore i osigurati mnoštvo analitičkih alata;

• dozvoliti učenicima razvoj vlastitih razumijevanja iz različitih perspektiva, odnosno prihvatiti višestruke perspektive;

• zahtijevati od učenika „oluju mozgova“, stvaranje, otkrivanje, propitkivanje, klasificiranje, analiziranje i predviđanje;

• zahtijevati od učenika elaboraciju njihovih inicijalnih reakcija;

• usmjeravati učenike prema zadacima koji im mogu pomoći identificirati pogreške, suočavati se sa zabludama ili izazivati kontradikcije sa njihovim inicijalnim hipotezama;

• uvijek ispitivati razumijevanje učenika, čekati reakciju učenika i prihvatiti njihove odgovore prije dijeljenja opće prihvaćenih koncepata (naglašavati konstrukciju, a ne reprodukciju znanja);

• poticati dijalog i kolaborativno učenje;

• podupirati razvoj kognitivnih strategija i ponašanja kao i memoriranje verbalnih ili činjeničnih informacija, motoričkih i intelektualnih sposobnosti (Gagne, 1992);

• ocjenjivati i procese i ishode;

• više pomagati i voditi nego podučavati;

• zadržati otvoren um i prihvaćati da učenici neće uvijek učiti ono što nastavnik namjerava.

U obzir ipak treba uzeti i to da postoje sugestije kako konstruktivistički pristupi ne funkcioniraju za sve učenike, „malo je vjerojatno da su svi učenici jednako podesni da vrše vlastitu kontrolu, stimulacije i usmjeravanja“ (Davidson, 1995), no potreba za boljim utemeljenjem, istraživanjima i opisivanjem nove pedagogije poučavanja i učenja ICT-a kako bi se pomoglo nastavnicima primijeniti adekvatnije strategije u učionici, više je nego očita.

III. POUČAVANJE ALGORITAMSKOG RAZMIŠLJANJA

U širem smislu, algoritam može biti definiran kao niz “uređenih koraka koje treba poduzeti kako bi se riješilo problem ili dostiglo neki završetak.” (CSTA/ISTE, 2011a, p. 15). Prema toj definiciji bilo koji opis niza koraka za izvođenje bilo kojeg zadatka je algoritam (matematički dokazi, laboratorijski postupci, poslovni procesi i upute za sastavljanje proizvoda, bilo koji skup uputa koji može biti izražen dijagramom tijeka). Sukladno tome, algoritamsko razmišljanje primjenjivo je na širok raspon različitih domena: poslovnu, znanstvenu, matematičku, tehnološku i tehničku i time važno za poučavanje u školama te se nužnim pokazuje oblikovati kurikulum i pedagogiju poučavanja algoritamskog razmišljanja. U tablici u nastavku je primjer podskupa projekata predviđenih za K-12 stupnjeve.

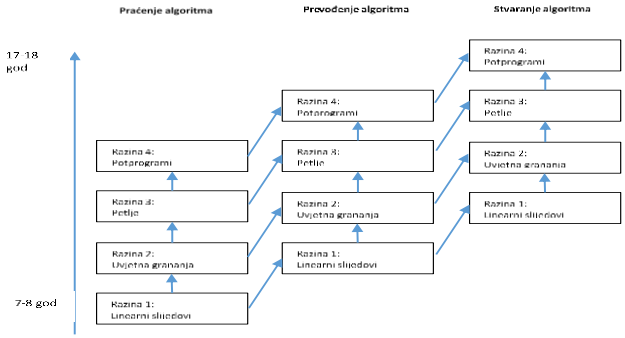
Pri tom je u procesu rješavanja problema potrebno osigurati da se “problemi oblikuju na način koji omogućuje korištenje računala i drugih alata koji će ih pomoći riješiti… automatiziranjem rješenja kroz algoritamsko razmišljanje” [10] što implicira da agentu koji će rješenje izvoditi (računalu ili osobi) ono treba biti jasno prenijeto.

Zbog toga možemo razmišljati o poučavanju algoritamskog razmišljanja kao poučavanju učenika da budu sposobni stvoriti niz uređenih koraka za izvršavanje zadatka i jasno komunicirati te korake drugoj osobi ili stroju koji će ih automatski izvoditi.

TABLICA I. KAKO POUČAVATI ALGORITAMSKO RAZMIŠLJANJE?

Starost	Algoritmi i postupci
7-8	Kreiraj grupu smjerova od škole do glavnih točaka za orijentiranje u susjedstvu.
8-11	Dizajniraj igru na ploči i napiši upute za igranje. Ispitaj upute na kolegama koji će pokušati igrati igru. Razradi upute na temelju povratnih informacija.
11-14	Programiraj robota da pronađe put iz labirinta, tako da iz bilo kojeg zadanog labirinta može u određenom vremenskom periodu uspješno izaći.
14-18	Diskutiraj proces donošenja odluke odabira fakulteta pa kreiraj algoritam koji opisuje taj proces. Algoritam treba biti u mogućnosti rukovati nepoznatim varijablama (npr. dostupnost financijske pomoći, uspjeh na prijemnom i sl.) kako bi doveo do nedvosmislene odluke.

Time se ujedno vrši dekompozicija algoritamskog razmišljanja na tri povezane vještine učenja koje učenici trebaju zadobiti: 1. Razumijevanje i ispravno praćenje algoritma (niza instrukcija rastuće složenosti); 2. Jasno izražavanje poznatog rješenja u obliku algoritma drugoj osobi ili računalu da ga izvede (prevođenje algoritma); 3. Stvaranje algoritma koji izvodi zadatak (vještina dizajniranja niza uređenih koraka za izvođenje zadataka.



Slika 1 Detaljna struktura plana učenja algoritamskog razmišljanja [11]

Na slici se vidi kako se na početno učenje linearnih nizova naredbi obično nadovezuju tri glavne upravljačke strukture koje modificiraju podrazumijevani niz: uvjetna grananja, petlje i potprogrami. Ostale specifične strukture i koncepti naprednih vještina algoritamskog razmišljanja (rekurzija, mjerenje brzine algoritama, dužine koda, različite optimizacije, konkurentni, paralelni i nedeterministički algoritmi) nadogradnja su na spomenute osnovne i mogu biti dodane kao nove razine povrh ovih.

U slučajevima kad poučavanje algoritamskog razmišljanja započinje prije poučavanja vještine programiranja isključena je upotreba programskih jezika za izražavanje algoritama [6], pa se koriste različiti drugi načini kao što su dijagrami tijeka (česti za izražavanje poslovnih procesa i koraka u računalnom programu), pseudokod (neformalni jezik koji programeri često koriste) ili riječi govornog jezika kojima se također mogu opisati nizovi koraka, zatim tehnički opisi karakteristični za priručnike o korištenju ili piktogrami (slični uputama za sastavljanje namještaja) kada nije poznat govorni jezik korisnika.

IV. RAZINE KOMPETENCIJE ALGORITAMSKOG RAZMIŠLJANJA

A. *Prepoznavanje i razumijevanje algoritama*

Najelementarnija razina algoritamskog razmišljanja je prepoznavanje algoritama i problema za čije rješavanje mogu poslužiti. Kako prepoznati što je algoritam? To je postupak koji može biti podijeljen na korake; moguće ga je implementirati (i postoji implementator kojem pripada); ima precizno određen poredak operacija koje izvodi; za vrijeme koraka nešto se s nečim događa; koraci su bilo neke elementarne akcije (koje implementator razumije) ili drugi algoritmi (što podrazumijeva daljnje specifikacije). Pri tom se razumijevanje odvija na dvije razine: razumijevanje onog što se treba učiniti i razumijevanje zašto se mora učiniti to, a ne nešto drugo.

Pitanje "zašto" pripada višim razinama razmišljanja, analizi algoritama, što znači prepoznavanje, poricanje, varijacije i izbor alternativa. Slično, sposobnost razumijevanja poznatog (već objašnjenog) i sposobnost razumijevanja apsolutno novog algoritma također predstavljaju različite razine znanja.

B. *Implementacija algoritama*

Sposobnost implementiranja algoritama je razina viša od razine razumijevanja, jer uključuje nadzor nad međufazama procesa za vrijeme implementacije (pamćenje, registriranje itd.). Pažnja se ne treba usmjeravati na ono što se događa, već na ono što treba učiniti slijedeće ovisno o određenim faktorima.

C. *Analiziranje algoritama*

Analiziranje algoritama dijelom obuhvaća prepoznavanje osnovnih pravila stvaranja algoritama: svaki se osnovni korak izvodi (danim redoslijedom); treba izabrati jedan od osnovnih koraka i izvesti ga; osnovni se koraci trebaju izvesti više puta ponavljajući ih.

Analiziranje algoritama uključuje i razumijevanje uloge njihovih dijelova, razumijevanje cilja prilikom njihove izrade i razumijevanje načina kako je čitav proces rješavanja problema rastavljen na manje podzadatke. Tu također spada i sposobnost čitanja algoritama, što znači sposobnost razumijevanja složene aktivnosti oblikovane od strane nekog drugog nekim algoritamskim jezikom: uviđanje svrhe dijelova, njihovih veza sa drugim dijelovima i odgovarajućih objašnjenja.

D. *Izrada algoritama*

Uključuje razmatranje: što je poznato; što bi trebalo postati poznato; što će se dogoditi; sa kojim podacima treba raditi; kako razbiti zadatke u manje podzadatke.

Pri tom je potrebno nalaziti načine sužavanja beskonačnog skupa algoritama na relativno manji. Kreativni proces ovdje je izbor i spajanje odgovarajućih algoritamskih shema i njihovo prilagođavanje konkretnim aktivnostima. Uspostava sustava shema zahtjeva razmišljanje o analogijama i apstraktno razmišljanje, razlikovanje između esencijalih i irelevantnih obilježja, apstrakciju, prepoznavanje sličnosti, prepoznavanje analogija.

E. *Realizacija algoritama*

Algoritam je potrebno opisati alatom i na način pogodan za izvođenja računalom. Podrazumijeva se potreba učenja upotrebe alata (programskog jezika) za zapisivanje algoritma. Pri tom pojedini alati također imaju svoje specifične načine razmišljanja koje treba usvojiti.

Kako se besprijekorne stvari rijetko izvedu odmah, potrebno je uočiti da li je stvoreni algoritam ispravan i prepoznati i ispraviti pogreške ako ih ima. Programska okolina učinit će ovu aktivnost učinkovitijom, no ispitivanje ipak zahtjeva dosta razmišljanja. Nakon otkrivanja pogreške, potrebne su analitičke sposobnosti za njihovo uspješno otklanjanje.

F. *Modifikacija i promjena algoritama*

Za razumijevanje algoritama drugih autora potrebno je shvatiti implementaciju, razumjeti način razmišljanja i razloge za baš takav način stvaranja, prepoznati mogućnosti izmjena i učinkovitih intervencija itd. kako bi ih se prilagodilo ili učinilo upotrebljivim.

G. *Dizajniranje složenih algoritama*

U ovim postupcima uloženi rad ne raste linearno s rastom veličine algoritma, a važno je uočiti i da cjelokupan zadatak može biti obavljen postizanjem podciljeva i dobrom sintezom [12].

V. RAZINE UČENJA PROGRAMIRANJA

Na usvajanje vještine algoritamskog razmišljanja tijekom usvajanja ICT kompetencija nadovezuje se učenje programiranja. Opsegom ono je najzastupljenije u strukovnim školama u sektoru Elektrotehnika i računalstvo i to za zanimanje Tehničar za računalstvo (programerski izborni modul novog kurikuluma).

Za učenje programiranja može se definirati skup važnih kompetencija koje se mogu klasificirati u razine:

Slika 2. Razine učenja programiranja

Simulacija algoritama znači „igranje“ danim algoritmom (npr. brojevima). To je korak u kojem svi učenici ne moraju u potpunosti razumjeti dani algoritam. Kako bi se utvrdilo razumije li stvarno učenik algoritam može se od njega tražiti da ga objasni nekom drugom.

Dopunjavanje algoritama je kompetencija u kojoj učenici trebaju biti sposobni dopuniti neke nedostajuće informacije u nedovršenim algoritmima.

Pisanje algoritama znači kompetenciju nalaženja i zapisivanja rješenja problema. Pri tom se zapisivanje može vršiti riječima govornog, materinjeg jezika, pseudokodom ili naredbama specifičnog programskog jezika.

Poboljšavanjem se primjeri algoritama popravljaju prepoznavanjem redundatnog koda ili traženjem učinkovitijeg rješenja.

U tablici je naveden primjer ciljeva učenja za temu varijable i pridruživanje [13].

TABLICA II. RAZRADA CILJEVA UČENJA

Koraci	**Cilj učenja: Upotreba varijabli**	**Objašnjenje**
Simuliranje	Simuliranje pridruživanja varijabli	Simuliranje pridruživanja upotrebom jedne ili dvije varijable.
Razumijevanje	Razumijevanje pridruživanja	Učenik razumije jednostavne algoritme u kojima se koristi do pet varijabli.
Upotpunjavanje	Upotpunjavanje pridruživanja	Učenik je sposoban dopuniti nepotpune ili izostavljene naredbe pridruživanja varijabli.
Pisanje	Pisanje pridruživanja	Učenik je sposoban pisati jednostavan algoritam u kojem se na odgovarajući način koristi do tri varijable.
Poboljšanje	Poboljšanje pridruživanja	Učenik je u stanju poboljšati dana pridruživanja u aspektu izbjegavanja redundantnog koda ili da bi pronašao brže rješenje problema.

Pri učenju programiranja učenici trebaju razviti vještinu sustavnog otkrivanja preciznih koraka za rješavanje problema i savladati tri tipa programerskog znanja: sintaktičko, konceptualno i strategijsko [5]. Sintaktičko znanje odnosi se na razumijevanje značenja izraza u nekom programskom jeziku i usporedivo je sa poznavanjem riječi i gramatičkih pravila bilo kojeg govornog jezika. Konceptualno znanje tiče se konstrukata i principa programiranja s ciljem razvoja odgovarajućih mentalnih modela. Strateško znanje odnosi se na vještine razumijevanja i strukturiranja problema i rješavanja problema konstruiranjem upotrebljivog algoritamskog rješenja.

Kroz povratnu vezu važno je učenike motivirati, davati im naputke za poboljšanje njihovog rješenja i upute za druge vrste vježbi čijim bi rješavanjem poboljšali svoje kompetencije u programiranju.

Uz to, sadržaje predavanja poželjno je harmonizirati i produbiti predviđenim pratećim laboratorijskim vježbama namijenjenim prakticiranju odgovarajućih vještina na različitim razinama, najbolje u trajanju 45 do 60 minuta uz individualne načine, procese i strategije učenja i to nakon što se prodiskutiraju otvorena pitanja i nakon objašnjavanja novih zadataka. Posljednjih petnaestak minuta može se predvidjeti za prezentiranje rješenja, odnosno rezultata.

VI. SADRŽAJI IZ UDŽBENIKA KAO POTPORA RAZVOJU ALGORITAMSKOG NAČINA RAZMIŠLJANJA

U svrhu uvida koliko pristupom, koncepcijski i sadržajno udžbenici iz informatike pomažu razvoju algoritamskog načina razmišljanja izvršen je uvid u pet odobrenih udžbenika različitih autora i nakladnika.

- Galešev, Korać, Soldo, Sokol, Kocijan: Informatika i računalstvo, udžbenik za srednje strukovne škole, SysPrint Zagreb, 2014.
- Galešev, Brođanac, Dmitrović, Korać, Sokol, Babić, Soldo, Kovač: Informatika i računalstvo, udžbenik za gimnazije i srednje škole, SysPrint Zagreb, 2014.
- Grundler, Šutalo: Računalstvo, udžbenik za 1. i 2. razred četverogodišnjih strukovnih škola, Školska knjiga, Zagreb, 2008.
- Gvozdanović, Ikica, Kos, Milijaš, Milijaš, Srnec, Šekulić-Štivčević, Zvonarek: Računalstvo/Informatika udžbenik za 1., 2.. i 3. razred strukovnih škola, PRO-MIL Varaždin, 2009.
- Kralj, Linardić, Sudarević: www informatika, udžbenik informatike i računalstva za gimnazije i srednje škole, Profil, Zagreb, 2013.

Pregledana su poglavlja koja obrađuju srodne sadržaje sa nazivima: Rješavanje problema programiranjem, Agoritmi i osnovni algoritamski postupci, Uvod i osnove programiranja, Programski jezici i načela programiranja, Vrste instrukcija i računalni program, Rješavanje problema i programiranje, Programiranje i programski jezici.

Svi udžbenici korektno slijede nastavni plan, odnosno kurikulum i faktografski objašnjavaju ideje, pristupe i znanja o pojmovima i povezanosti algoritama, programskih jezika i načelima programiranja.

Razlike najčešće proizlaze iz opsežnosti stranica namijenjenih određenim temama čemu su ponajviše razlog razlike u trajanju i vrstama škola za koje su namijenjeni (srednje strukovne škole trogodišnjeg i četverogodišnjeg trajanja, gimnazije), tako da primjerice pojedini udžbenici osnove razvoja algoritama ne vežu uz određeni programski jezik nego samo uz odgovarajući tekstualni ili slikovni zapis (pseudokod, dijagrami tijeka), dok drugi uključuju poglavlja sa objašnjenjima naredbi odabranog programskog jezika (C++, C#, Python).

Iako korektni u svom pokrivanju važećih kurikuluma (ne sasvim i za nove eksperimentalne programe), moguće je u njima otkriti iznošenje činjenica koje bi mogle izazvati polemiku kao što je npr. usporedba djeteta sa računalom, činjenica da gotove dijelove programa zovemo objektima, zatim da definirati varijablu znači pridružiti joj vrijednost ili termin „petlja s uvjetom na vrhu“, te da li je strukturno programiranje isto što i top-down programiranje.

Opći je dojam da prezentiranim pristupom sam razvoj algoritamskog razmišljanja u pregledanim udžbenicima nije odgovarajuće zastupljen, da ima dobrih primjera i zadataka no da se o svim relevantnim područjima još uvijek piše uglavnom faktografski bez dovoljno potpore uvođenju suvremenijih nastavnih metoda i nuđenja učenicima aktivnije uloge u nastavnom procesu.

VII. RAZVOJ ALGORITAMSKOG NAČINA RAZMIŠLJANJA BEZ RAČUNALA NA NATJECANJIMA UČENIKA

Natjecanja u programiranju podupiru razvoj vještina algoritamskog razmišljanja i programiranja, no prilagođena su grupi učenika koji znaju i mogu pisati programe i koji su upoznati s bar nekim tehnikama programiranja. Kako bi privukli učenike koji programiranju (još) nisu vješti, trebalo bi im omogućiti rješavanje zadataka koji traže određeno algoritamsko razmišljanje, ali ne i programiranje. Popularizacija učenja programiranja među takvim učenicima može se izvesti dobro izabranim zadacima koji zahtijevaju različite razine algoritamskog razmišljanja bez da im se prezentiraju kao zadaci za programiranje. Ujedno, takvi zadaci najčešće su vrlo pogodni za inicijalnu edukaciju vještine algoritamskog razmišljanja.

Dobro je da su zadaci utemeljeni na jednostavnim konceptima, po mogućnosti popraćeni crtežom kako bi im se povećala atraktivnost i da imaju kratku formulaciju. Učenik bi trebao biti sposoban jedinstveno rješenje dobro formuliranog zadatka s minimalnim brojem proračuna pronaći u 5 minuta. Zaključno, poželjno je da rješenje uključuje neke od ključnih koncepata računalne znanosti i algoritama, no iz same formulacije zadatka treba izuzeti korištenje bilo kojeg koncepta programiranja. Također, prihvatljivi su i zadaci koji ohrabruju učenike u izvođenju verifikacije ispravnosti njihovih rješenja [14].

Posljednjih godina niz država prihvatio je slične pristupe (South Africa od 2003. sa značajnim stupnjem sudjelovanja učenika, Australia od 2005., Litva kroz Bebras (Beaver) u 2004. i otada se širi u mnoge europske države). Bebras kontest jedna je od važnih međunarodnih aktivnosti koje promoviraju računalnu znanost u primarnim i sekundarnim školama (the Beaver competition, Beaver of Informatics [15]). Žarište natjecanja nije na korisničkom pristupu i informacijskim tehnologijama, nego preferira pitanja i teme iz osnova područja računalne znanosti, kao što su npr. algoritmizacija, razumijevanje informacija i njihovog predstavljanja, kao i matematičke i logičke osnove računalne znanosti. Natjecanju pristupaju i učenici u kategoriji osmogodišnjih.

Već spomenuti Australian Informatics Competition (AIC) je natjecanje sa zadacima koji se rješavaju bez računala i koji ne zahtjevaju programersko iskustvo ni poznavanje pseudokoda. Većina pitanja fokusira se na ispitivanje algoritamskih sposobnosti, dok se ostatak odnosi na primjenu pravila i upotrebu logike u rješavanju problema. Algoritamska pitanja uključuju standardne algoritme kao što su pretraživanje u širinu, dinamičko programiranje i igre za dvije osobe, kao i ad-hoc algoritme specifične za određeni scenario. Podatak koji ne uključuje on-line natjecatelje govori o 7000 učenika koji pristupaju natjecanju. Dodatno, AIC je postao relevantniji uvođenjem algoritamskog razmišljanja kao komponente predmeta kurikuluma za digitalne tehnologije.

Sustavan pristup rješavanju problema na natjecanju se promovira zadacima u tri stupnja predviđenih rješavanju malog problema sa tri skupa podataka. Prvi je skup podataka mali ili dovoljno jednostavan kako bi bio osjetljiv na ad-hoc tehnike, uz nadu da će biti dovoljno dobra baza kako bi učenici dobili osjećaj za problem i razvili algoritam koji se može upotrijebiti u preostalim skupovima podataka.

Najveći broj pitanja ulazi u četiri široke kategorije: primjena pravila, logika, analiza i algoritmi. [16].

i) Primjena pravila - zahtjeva od učenika da primjenjuju dobro definirana pravila na skup podataka. Uglavnom su to jednostavni zadaci koji od učenika traže da prolaze kroz zadani postupak ili slijede dobro definirani skup naredbi. Time se ispituje sposobnost učenika da razumiju i slijede zadani algoritam, što je važan preduvjet razvoja vlastitih algoritama.

ii) Logika - uključuje nealgoritamske zagonetke koje mogu zahtjevati prilično rigorozno rasuđivanje i analizu slučaja, esencijalnih vještina za dizajn algoritama.

iii) Pitanja analize - od učenika traže da određuju broj operacija koje određeni algoritam na skupu podataka iziskuje. Učenici pri tom ispituju dobre strane i slabosti danog algoritma ili problema što im omogućuje upoznavanje sa analizom složenosti algoritma. Također, učenici trebaju analizirati ponašanje danog algoritma i/ili problema koji je u pozadini algoritma. Takav zadatak može tražiti i približan broj koraka ili scenarij najgoreg slučaja, ohrabrujući tako razmišljanje o računskoj složenosti i patološkim slučajevima. Zatim, od učenika se može tražiti da nađu pogrešku u nizu naredbi *("debugging")* ili da upotpune iscrpan skup ulaznih scenarija *("testing")* i tako razviju odgovarajuće vještine. Svrha zadataka analize je ohrabrivanje vještina važnih za dizajn algoritama jer programeri trebaju biti sposobni analizirati ispravnost, učinkovitost i robustnost njihovog vlastitog koda [17]. Popularno pitanje analize je brojanje različitih puteva u mreži.

iv) Algoritamska pitanju su centralna i sva su optimizacijska. Učenike se ohrabruje da razviju neformalni, ali učinkovit i ispravan algoritam kako bi riješili zagonetku. Popularna su pretraživanja u širinu, a uobičajeno je i dinamičko programiranje kao i igre pitanja za dvije osobe te dobri ad-hoc problemi koji zahtjevaju vlastite algoritme posebne namjene.

Bugarsko natjecanje Chernorizets Hrabar uključuje zanimljive zadatke analize različitog tipa. Za dani algoritam izražen kodom, pseudokodom ili dijagramom tijeka od učenika se zahtjeva da razumiju algoritam i otkriju koji on zadatak izvodi (tipično neka vrsta numeričkog zadatka, kao što je potenciranje, zbrajanje ili traženje korijena). Žarište je na algoritmima i algoritamskom razmišljanju.

ZAKLJUČAK

Ovaj je rad pokušaj naglašavanja kako i kurikulum i pedagogija poučavanja važnih vještina u području informacijskih tehnologija, osobito algoritmskog razmišljanja, možda središnje komponente računalske znanosti, zahtjevaju više pažnje i bitnije mjesto u obrazovanju učenika za izazove tehnološki naprednog društva.

Upravo je kod poučavanja informacijskih tehnologija posebno očit konflikt između tradicionalnog i suvremenog pristupa obrazovanju učenika. Obrazovanje upotrebom modela poučavanja temeljenih na linearnim i konačnim konceptima znanja i razumijevanja neće im pomoći da uspiju u globalno tehnološki bogatoj budućnosti. Sasvim konkretno, učenje i poučavanje ICT-a inherentno su konstruktivističke aktivnosti. Nova bi pedagogija (u kurikulumu, izvođenju nastave i pratećim nastavnim materijalima) trebala voditi nastavnike u upotrebi konstruktivističkih pristupa i osigurati učinkovitije strategije poučavanja.

Zato ne čudi zabrinutost zbog toga što je usvajanje algoritamskog načina razmišljanja podzastupljeno u školskim kurikulima mnogih država, a i tamo gdje se pojavljuje postoje poteškoće u implementaciji zbog nedostatka potpore ili resursa.

LITERATURA

[1] **Recommendation of the European Parliament and of the Council on Key Competences for Lifelong Learning**, 2006.

[2] Tucker, A**.: A Model Curriculum for K-12 Computer Science: Final Report of the ACM K-12 Task Force Curriculum Committee**, The Association for Computing Machinery, New York, 2003.

[3] **Australian Curriculum** – http://www.australiancurriculum. edu.au/technologies/digitaltechnologies/Curriculum/F-10

[4] M. Cavanagh, **Students' experiences of active engagement through cooperative learning activities in lectures**. Active Learning in Higher Education, 12(1), 2011, pp. 22-33.

[5] Baldwin, L. P; Kuljis, J. **Learning programming using program visualization techniques**. In Proceedings of the 34th Annual Hawaii International Conference on System Sciences, pages 1051-1058, IEEE, Washington, DC, USA, 2001.

[6] Lu, J. J., & Fletcher, G. H. L.: **Thinking about computational thinking**. SIGCSE Bull., 41(1), 2009.

[7] Crawford R. **Teaching and learning IT in English state secondary schools - towards a new pedagogy**, Education and Information Technologies, 1999.

[8] Carroll, J. M., **The Nuremberg funnel: designing Minimalist Instruction for practical computer skill**, MIT Press, 1990, in Sneiderman, B. Designing the User Interface: Strategies for Effective Human-Computer Interaction, Addison Wesley Publishing Company, 1992.

[9] Boyle, T., **Design for Multimedia Learning**, Prentice Hall, 1997.

[10] Computer Science Teachers Association and the International Society for Technology in Education: **Computational Thinking in K-12 Education Teacher Resources**, 2011.

[11] Pat Ko & Cesar Delgado: **A Proposal for a Hypothetical K-12 Learning Progression Set for Algorithmic Thinking**, University of Texas at Austin, San Francisco, California 2013.

[12] Zsakó, L., Szlávi, P. **ICT Competences: Algorithmic Thinking,** Acta Didactica Napocensia, 2012.

[13] Moschitz, J.: **An individual training program for learning programming at university**, 2014.

[14] Radoszewski, J.: **More Algorithms without Programming**, Olympiads in Informatics, 2014, Vol. 8, 157–168 © 2014 Vilnius University, IOI

[15] **Beaver Competition**, www.bebras.org.

[16] Clark D., Clapper M.: **The Australian Informatics Competition** Olympiads in Informatics, 2014, Vol. 8, 179–189 © 2014 Vilnius University, IOI

[17] A. Burton, B. **Encouraging Algorithmic Thinking Without a Computer** Olympiads in Informatics, 2010, Vol. 4, 3–14 3 © 2010 Institute of Mathematics and Informatics, Vilnius

[18] Dagien˙e, V. **Information technology contests – Introduction to computer science in an attractive way**, Informatics in Education, 2006.

Kriteriji vrednovanja rada nastavnika od strane njihovih studenata

Mandi Orlić Bachler* i Marija Marinović **
*Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska
**Odjel za informatiku, Sveučilište u Rijeci, Rijeka, Hrvatska
morlic1@tvz.hr, marinm@inf.uniri.hr

Sažetak - Proteklih nekoliko godina, putem ISVU sustava, na visokim učilištima Repubike Hrvatske provodi se anketiranje studenata u svrhu utvrđivanja stupnja zadovoljstva ili nezadovoljstva radom nastavnika i asistenata koji sudjeluju u realizaciji njihova nastavnog procesa. U ovome radu prikazujemo rezultate ankete, koju smo proveli među studentima sa Odjela za informatiku Sveučilišta u Rijeci i studentima Tehničkog veleučilišta u Zagrebu. Istražili smo koji čimbenici i u kojoj mjeri utječu na valorizaciju rada nastavnika od strane njihovih studenata.

I. UVOD

U Pravilnik o mjerilima i kriterijima za vrednovanje kvalitete i učinkovitosti visokih učilišta i studijskih programa iz 2004. godine, Ministarstvo znanosti, obrazovanja i sporta Republike Hrvatske uvrstilo je anketiranje studenta kao obvezni unutarnji mehanizam za osiguranje kvalitete na visokim učilištima [1]. Međutim, uz tu zakonsku odredbu, anketiranje studenata nužno je provoditi i stoga što je pozitivna ocjena nastavnika od strane studenata, dobivena na osnovi analize rezultata *Ankete za procjenu rada nastavnika*, jedan od uvjeta za izbor u više znanstveno-nastavno zvanje. Posljednjih nekoliko godina visoka učilišta anketiranje studenata provode jednom semestralno putem ISVU sustava. Na taj način studentima je omogućeno da potpuno anonimno iskažu svoje zadovoljstvo ili nezadovoljstvo kvalitetom rada nastavnika i asistenata na svakom pojedinom kolegiju. Međutim, na rezultate ankete djeluju brojni čimbenici, poput zainteresiranosti studenata za određeni kolegij, redovitost pohađanja nastave, uspjeh (ocjena) postignut na kolokvijima ili završnom ispitu, percepcija da se ispunjavanjem anketa ništa ne postiže, zadovoljstvo komunikacijom na relaciji nastavnik student i sl.

Cilj ovoga rada je na osnovi provedene ankete među studentima sa Odjela za informatiku Sveučilišta u Rijeci i studentima Tehničkog veleučilišta u Zagrebu, istražiti koji čimbenici i u kojoj mjeri utječu na valorizaciju rada nastavnika od strane njihovih studenata.

II. METODOLOGIJA ISTRAŽIVANJA

Rad se temelji na empirijskom istraživanju, koje je provedeno tijekom siječnja 2015. godine među studentima Odjela za informatiku Sveučilišta u Rijeci i studentima Tehničkog veleučilištu u Zagrebu. Ispitivanjem je obuhvaćeno ukupno 227 studenata (80% muškaraca i 20% žena), od čega ih je 80 sa Odjela za informatiku, a 147 s Tehničkog veleučilišta u Zagrebu.

Anketom su bili obuhvaćeni samo studenti viših godina jer studenti prve godine u siječnju 2015. godine još nisu imali mogućnost putem ISVU sustava popuniti *Anketu za procjenu rada nastavnika*.

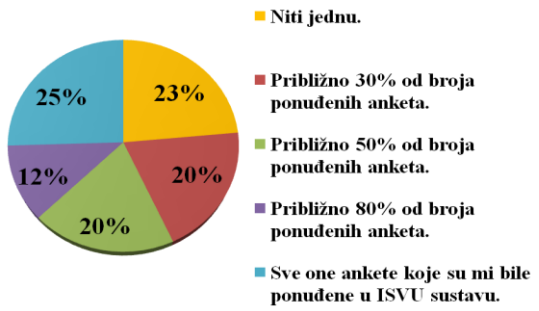
Anketni upitnik koji smo koristili sačinjen je isključivo za potrebe ovog istraživanja, a sastoji se od ukupno 13 pitanja. Od tog broja 10 pitanja je zatvorenog tipa dok su tri pitanja poluzatvorenog tipa, jer pored ponuđenih odgovora studenti imaju mogućnost dopisivanja odgovora po želji. Rezultati su se ovisno o tipu pitanja dobivali zaokruživanjem jednog ili više ponuđenih odgovora ili pak slobodnim upisivanjem odgovora na za to predviđeno mjesto. Popunjavanje anketnog lista provedeno je individualno, a ispitanicama su upute za popunjavanje te objašnjenje o svrsi uporabe dobivenih podataka dane u pismenom obliku u prilogu anketnog lista.

III. REZULTATI ISTRAŽIVANJA

Zbog lakšeg snalaženja i bolje preglednosti, u ovom djelu radu svako je pitanje pojedinačno obrađeno.

U prethodnom poglavlju navedeni su rezultati prva dva pitanja iz ankete, koja se odnose na spol studenta i visoko učilište kojeg pohađa.

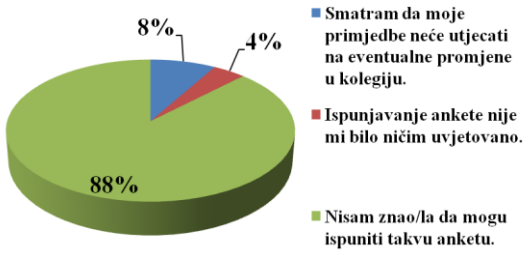
Pitanjem broj tri (**Koliko *Anketa o procjeni rada nastavnika* ste do sada ispunili?**) željeli smo utvrditi koliko često studenti popunjavanju *Anketu za procjenu rada nastavnika*, odnosno koliko ih nikad nije ispunilo takvu anketu. Na grafu sa Slike 1. vidi se da 53 (23%)



Slika 1. Koliko *Anketa o procjeni rada nastavnika* ste do sada ispunili?

ispitanika nikad nije sudjelovalo u vrednovanju rada nastavnika, dok je njih 58 (25%) ispunilo sve Ankete o procjeni rada nastavnika koje su bile ponuđene u ISVU sustavu. Približno 30% od broja ponuđenih anketa ispunilo je 45 (20%) studenta. Isto toliko studenata odgovorilo je da je ispunilo približno pola od ponuđenog broja anketa. Njih 27 (12%) procjenjuje da je odogovorilo na oko 80% od broja ponuđenih anketa.

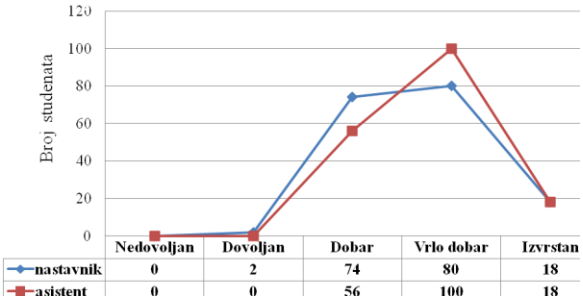
Četvrtim pitanjem (***Anketu o procjeni rada nastavnika* nikada nisam ispunio/la iz sljedećeg razloga: (a) Nisam bio/la raspoložen/a za ispunjavanje ankete. (b) Smatram da moje primjedbe neće utjecati na eventualne promjene u kolegiju. (c) Ispunjavanje ankete nije mi bilo ničim uvjetovano. (d) Nisam znao/la da mogu ispuniti takvu anketu. (e) Držim da anketa nije u potpunosti anonimna, jer se ispunjava putem ISVU sustava. (f) neki drugi razlog**) pokušali smo utvrditi razloge eventualnog neodazivanja studenata popunjavanju ankete. Od 53 studenta, koji nikada nisu popunili *Anketu za procjenu rada nastavnika*, njih čak 88% odgovorilo je da to nisu načinili jer nisu znali da takvu anketu uopće mogu ispuniti. Nadalje, 8% od tog broja studenta to nije napravilo jer smatra da njihove primjedbe neće utjecati na eventualne promjene u kolegiju, dok dva (4%) studenta kao jedini razlog navodi da im popunjavanje ankete nije bilo ničim uvjetovano (Slika 2.). Preostale ponuđene odgovore nije zaokružio niti jedan student.



Slika 2. *Anketu o procjeni rada nastavnika* nikada nisam ispunio/la iz sljedećeg razloga: (a) Nisam bio/la raspoložen/a za ispunjavanje ankete. (b) Smatram da moje primjedbe neće utjecati na eventualne promjene u kolegiju. (c) Ispunjavanje ankete nije mi bilo ničim uvjetovano. (d) Nisam znao/la da mogu ispuniti takvu anketu. (e) Držim da anketa nije u potpunosti anonimna, jer se ispunjava putem ISVU sustava. (f) neki drugi razlog

Na peto pitanje (**Koje od sljedećih tvrdnji najbolje pojašnjavaju razloge zbog kojih ste ispunili *Anketu o procjeni rada nastavnika*? (a) Ankete sam ispunio/la samo zato što mi je to bilo uvjetovano u ISVU sustavu, inače ne bih. (b) Ankete sam ispunio/la uglavnom zato što sam želio/željela pohvaliti rad nastavnika/asistenta. (c) Ankete ispunio/la uglavnom zato što sam želio/željela iskazati nezadovoljstvo radom nastavnika/asistenta. (d) Ankete sam ispunio/la jer to smatram svojom dužnošću. (e) Ankete sam ispunio/la jer ih držim dobarim načinom za iskazivanje svoga mišljenja o radu nastavniku/asistentu te sadržaju kolegija. (f) Neki drugi razlog**) odgovorila su 173 studenta. U ovom pitanju studenti su mogli zaokružiti više ponuđenih odgovora. Odgovor da je *Anketa o procjeni rada nastavnika* dobar način za iskazivanjem mišljenja o radu nastavnika/asistenta te sadržaju kolegija misli 74 studenta, dok je 62 studenata ispunilo ankete samo zato što im je to bilo uvjetovano. Pohvaliti rad nastavnika/asistenta ispunjavanjem ankete željelo je 49, dok je svoje nezadovoljstvo njihovim radom na isti način željelo iskazati 42 studenata. Popunjavanje ankete 74 studenata smatra svojom dužnošću. Nitko od ispitanika nije naveo niti jedan drugi razlog.

Odgovori na pitanja broj šest (**Zaokružite ocjenu kojom ste najčešće ocijenili rad nastavnika?**) i sedam (**Zaokružite ocjenu kojom ste najčešće ocijenili rad asistenta?**) prikazani su grafički na Slici 3. Prosječna ocjena kojom studenti, neovisno o kolegju, ocjenjuju rad asistenta iznosi 3.78, a nastavnika 3.66.



Slika 3. Zaokružite ocjenu kojom ste najčešće ocijenili rad nastavnika/asistenta?

U osmom pitanju (**Na ocjenu rada nastavnika/asistenta utjecalo je sljedeće: (a) Pokušala sam biti objektivn/a te dati ocjenu koja najbolje opisuje njegov rad; (b) Zadovoljstvo konačno postignutim rezultatom; (c) Jednostavnost odnosno složenost nastavnog sadržaja; (d) Jednostavnost odnosno složenost polaganja ispita; (e) Zadovoljstvo odnosno nezadovoljstvo komunikacijom na relaciji nastavnik student; (f) nešto drugo.**) studenti su odgovorili koji čimbenici utječu na njihovu ocjenu o radu nastavnika/asistenta. Na ovo pitanje studenti su mogli odgovoriti zaokruživanjem više ponuđenih odgovora. Od 173 studenta, koja su dala odgovor na ovo pitanje, njih 141 smatra da je bilo objektivno pri ocjenivanju rada nastavnika/asistenta. Na danu ocjenu kod 45 ispitanika utječe kvaliteta komunikacije na relaciji nastavnik/asistent student, dok ih 31 ocjenu temelji na zadovoljstvu svojim konačnim uspjehom na kolegiju. Jednostavnost, odnosno složenost, nastavnog sadržaja kod njih 29 utječe na ocjenu danu nastavniku/asistentu, a kod 18 je to jednostavnost, odnosno složenost, polaganja ispita.

Na pitanje broj devet (**Imate li saznanja o tome kako se vaši kolege dogovaraju da određenog nastavnika/asistenta treba negativno ocijeniti?**) odgovorilo je 174 studenata. Od toga broja 64% dalo je negativan odgoovor. Preostala 62 (36%) studenata odgovaralo je na deseto pitanje (**Jeste li i koliko često sudjelovali u načinu ocjenjivanja rada nastavnika/asistenata na način opisan u pitanju broj 9?**). Odgovor "Nikad" zaokružilo je 18 (29% od 62), "Ponekad" 44 (26% od 62) studenta, dok odgovore "Uvijek" i "Ne želim odgovoriti na ovo pitanje" nije zaokružio nitko.

Na posljednja tri pitanja odgovorilo je svih 227 anketiranih studenata.

Pitanjem broj 11 (**Znate li da je pozitivna ocjena o radu nastavnika, temeljena na rezultatima Ankete za procjenu jedan od uvjeta za njegov izbor ili reizbor u znanstveno-nastavno zvanje?**) željeli smo ispitati koliko su studenti informirani tome na koji način i u kojoj mjeri rezultati *Ankete za procjenu rada nastavnika* utječu na ocjenjene nastavnike, odnosno asistente. Pozitivan odgovor na postavljeno pitanje dalo je 30% od ukupnog broja anketiranih studenata.

Osim putem službenih Anketa za procjenu rada nasatvnika visokih učilišta, studenti imaju mogućnost, potpuno anonimno i javno, iskazati svoje mišljenje o radu nastavnika, odnosno asistenata, putem web stranice www.studokracija.com, a koju uređuju bivši i sadašnji studenti [2].

S posljednja dva pitanja u anketi željeli smo ispitati jesu li studenti upoznati s navedenom web stranicom te kolika je njihova zainteresiranost za takav način iskazivanja mišljenja o svojim nastavnicima/asistentima. Na pitanje broj 12 (**Jeste li čuli za web stranicu studokracija.com?**) pozitivno je odgovorilo samo 20 (9%) anketiranih studenata. Od tog broja 78% pozitivno je odgovorilo na pitanje broj 13 (**Jeste li ikada ocijenili ili komentirali rad nastavnika/asistenta putem web stranice studokracija.com?**).

IV. ZAKLJUČAK

Na temelju rezultata, dobivenih ovim istraživanjem, proizlazi da je relativno mali broj studenata zainteresiran za ispunjavanje Anketa o procjeni rada nastavnika. Iz čega posljedično slijedi da se na temelju studentskih anketa, za samo mali broj nastavnika (odnosno kolegija) mogu dobiti relevantni podaci o kvaliteti njihova rada.

Iako smo utvrdili da na kriterije vrednovanja rada nastavnika od strane njihovih studenata utječu brojni čimbenici, rezultati dobiveni analizom studentskih anketa predstavljaju jedan od bitnih elemenata za postizanje odgovarajuće razine kvalitete visokih učilišta.

LITERATURA

[1] Ministarstvo znanosti, obrazovanja i sporta Republike Hrvatske, Pravilnik o mjerilima i kriterijima za vrednovanje kvalitete i učinkovitosti visokih učilišta i studijskih programa, 2004.

[2] www.studokracija.com

Budućnost eUčenja

Martina Sertić
Industrijsko-obrtnička škola Slavonski Brod, Republika Hrvatska
Klasična gimnazija Slavonski Brod, Hrvatska
sertic_79@yahoo.com

Sažetak - Ovaj rad trebao bi predstaviti neke od sadržaja vezanih uz eUčenje. Navedeni sadržaji mogli bi i trebali bi se u budućnosti sve više razvijati, a potom i koristiti u obrazovne svrhe kod budućih naraštaja učenika, studenata i obrazovanja odraslih. Kroz kraće istraživanje, u radu će se provjeriti stavovi učenika srednjih škola i polaznika obrazovanja odrasli o ovoj temi. U navedenom istraživanje iznijeli bi se stavovi učenika, polaznika obrazovanja odraslih o upotrebi obrazovnih sadržaja u njihovom školovanju i budućem zanimanju. Obrazovni sadržaji obuhvatili bi dobar dio *online* alata koji se danas koriste u nastavi kao dio potrebe globalnog umrežavanja društva i gospodarstva.

I. UVOD

Za Web 2.0 može se reći da je to interaktivni i društveni web čija je funkcija olakšati suradnju među ljudima. On se razlikuje od ranog Weba 1.0 koji je bio statična informacija i mjesto gdje se web stranice samo čitaju, ali rijetko kada ostvaruju bilo kakve interakcije s njima. A što je zapravo Web 3.0? To je sljedeća temeljna promjena na razini interneta i informacija koje će se prenositi i trebale bi se prenositi kao web stranice u interakciji s ljudima.
U radu će se ukratko definirati pojam informacije, informacijske infrastrukture bez koje ne bi bilo ni Weba 1.0, 2.0, 3.0. Ponajviše će se osvrnuti na razvoj Web 2.0 usluga, njihovih osnovnih svojstava, načina primjene i razvoja tehnologije.

Podrobnije će biti navedeni neki Web 2.0 servisi na društvenim mrežama, servisi za obradu fotografija, video zapisa i ostalih grafičkih sadržaja. Tako će se spomenuti i nešto ukratko reći o YouTubu, Vimeu, Slidesixu, Masheru, Stupeflix servisima, web sadržajima za razmjenu fotografija Aviary; sadržajima društvenih mreža Linkedlnu, Deliciousu, Diigou i StumbleUponu.

U zadnjem poglavlju navedeno je nekoliko osnovnih svojstava i razloga moguće primjene Web 3.0 tehnologije. [1]

II. WEB 2.0

Web 2.0 ostvaruje se kroz World Wide Web tehnologiju koja se zasniva kao društvena veza s korisnicima i omogućava sudjelovanje u stvaranju sadržaja weba, te podržava interaktivnu komunikaciju između korisnika, računala, korisnika i drugih korisnika, čime korisnik od pasivnog postaje aktivni sudionik. Društveno umrežavanje osnova je Web 2.0 tehnologije, pri čemu je korisnik u središtu pozornosti kao individua koja kreira sadržaje pomoću raznih mrežnih servisa *blogova,* foruma, fotografija, mrežnih enciklopedija, društvene mreže, video zapisa i dr. [1]

Osnovne karakteristike Web 2.0 su otvorenost, sloboda, cjelovitost i praktična primjena. Korisnici mogu koristiti programe preko web preglednika, a sama arhitektura Weba 2.0 u osnovi potiče korisnika da za vrijeme korištenja aktivno sudjeluje u stvaranju novih sadržaja. [1]

Web 2.0 kao primijenjena *online* tehnologija zahtijeva objavljivanje i aktivan rad. Većinom su to *hosting* usluge s dovoljno prostora za objavu sadržaja. Internet je postao otvoreno tržište na kojem možemo pristupiti bilo čemu kad god želimo. Koncept današnjeg interneta potpuno je drugačiji nego u početku. Današnje Internet usluge neprestano potiču svoga korisnika na interakciju i kreiranje sadržaja kojem pristupa. Najatraktivniji sadržaji nastali primjenom web 2.0 alata su: *blogovi, wiki* sadržaji, *podcastovi*, društvene mreže i slične usluge. [2]

Web 2.0 alati skupina su društvenih programskih alata koji korisnicima omogućavaju interakciju i razmjenu podataka te objavu i izmjenu internetskog i drugog digitalnog sadržaja pri čemu korisnik sam stvara svoje sadržaje. Podržavaju *online* učenje, kupovinu, prodaju, oglašavanje, distribuciju i na taj način potiču veću razinu kreativnosti i interakcije. Mnoštvo besplatnih sadržaja i alata na jednome mjestu, te inovativno korištenje tehnologije, zajednički pristup internetu i informacijama učinili su web 2.0 popularnim kod mnogobrojnih korisnika. [2]

Web 2.0 stvorio je novog čovjeka koji je spreman kvalitetno koristiti sve što mu web 2.0 tehnologija nudi. On postaje konzument onoga što stvara, usavršava se i oblikuje u skladu s novim mogućnostima i zahtjevima

tržišta i korisnika. Stvorena je neka nova internet Google generacija stvaralaštva, komunikacije i kreacije.

III. WEB SADRŽAJI ZA PRIKAZ GRAFIČKIH APLIKACIJA

Web sadržaji koji omogućavaju prikaz grafičkih aplikacija pružaju korisniku interneta velike mogućnosti. Ovi web sadržaji dopuštaju na jednostavan način pregledavanje, postavljanje, kreiranje zvučnog, video i ostalih grafičkih sadržaja. Neki od takvih alata navedeni su dolje u tekstu. Jednostavni su i praktični za korištenje, a nalaze se većinom na navedenim web adresama. Ovi sadržaji pružaju razne mogućnosti npr. *tagiranje* ili označavanje pojedinog sadržaja, kreiranje biblioteka i profila sličnih interesnih mogućnosti. Korisnici isto tako mogu organizirati rezultate pretraživanja po kategorijama i grupama relevantnosti pojedinog sadržaja. [3]

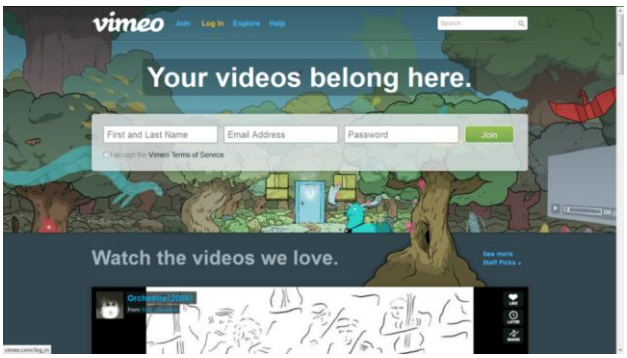
A. Youtube

YouTube su osnovali Chad Hurley, Steve Chen i Jawed Karim, svi rani zaposlenici PayPala. Riječ je o poznatoj web stranici na kojoj korisnici mogu gledati, postavljati i komentirati video sadržaje.
Za komentiranje i postavljanje sadržaja korisnik mora imati otvoren račun, a gledanje video sadržaja je slobodno. Web adresa YouTube stranice je http://www.youtube.com/.
YouTube sadržaji podijeljeni su po kategorijama, a korištenjem YouTube tražilice korisnicima je olakšano pronalaženje video sadržaja na kanalu.

B. Vimeo

Vimeo je *online* video servis, tj. *video sharing* web stranica na kojima korisnici mogu postavljati, dijeliti i gledati video sadržaje raznih profila i kategorija. Osnovali su je Jake Lodwick i Zach Klein u studenom 2004. Web adresa stranice je http://vimeo.com/.

Vimeo je zapravo druga verzija dobro poznatog YouTube kanala, sa malo ljepšim i izraženijim dizajnom samog weba, ali složenijom navigacijom i zahtjevnijom pretragom i postavljanjem sadržaja.



Slika 1. Vimeo web stranica

C. SlideSix

SlideSix je web stranica na kojoj možemo pronaći sav potreban alat za izradu i uređivanje multimedijalnih prezentacija. Ova web stranica daje korisnicima mogućnost prezentacije materijala uživo i to u virtualnom svijetu u stvarnom vremenu širom svijeta. To znači da se korištenjem pojedinih program može predstaviti neki materijal baš kao što bi to učinili koristeći recimo Power Point. Jedina velika razlika je u tome što se postavljeni materijal može uživo prezentirati i uređivati koristeći savjete ostalih sudionika weba.
Za početak je potrebno preuzeti SlideSix Presenter i instalirati ga prateći upute (zahtijeva najnoviju verziju Adobe AIR koji treba biti instaliran na računalu). Nakon preuzimanja programa i pozornog slušanja uputa moguće je krenuti s radom. [3]

D. Masher

Masher je web stranica koja pruža laku izradu video sadržaja, obradu video isječaka, izradu glazbenih zapisa (pjesama i spotova) te obradu i izradu fotografija.
Masher je izrađen 2008. godine i posjeduje poveću bazu raznovrsnih snimki glazbe, koju možemo koristiti u svojim kreacijama besplatno. Besplatni video arhiv - baza ima tisuće kvalitetnih spotova, uključujući i snimke iz CBS News, BBC Motion Gallery, Kine, Japana, Australian Broadcasting Corporation, Rip Curl, Huntley arhiva i još mnogo toga. Web adresa je http://www.masher.com/.

E. Yodio

Yodio je web stranica preko koje se mogu postavljati slike ili koristiti već postojeće slike priložene na webu. Yodio također omogućava prijenos zvuka kojeg možete snimiti putem interneta ili telefona (samo SAD) ili odabrati iz glazbene biblioteke weba koja je podijeljena po kategorijama. Moguće je organizirati vlastite medije u mape ako se namjerava stvoriti više projekata. Web stranica na kojoj se može pronaći ponuđeni sadržaj je http://www.yodio.com/.

Slika 2. Yodio web servis

F. *Aviary*

Aviary je koristan i zabavan web editor za izradu slika koji se koristiti na mobitelima, računalima, tabletima i iPhone uređajima. Praktičan je i efikasan za osnovnu obradu slika i za manipuliranje fotografijama kroz web preglednik. Jednostavan je za korištenje i pristupačan za druge aplikacije koje omogućavaju još veću kreativnost i rad nad fotografijama.
Korištenjem Aviary editora možemo izraditi i čestitke, razglednice, plakate, pozadine za Power Point prezentacije i dr.
Besplatan je za sve korisnike, a pristupamo mu putem web adrese http://www.aviary.com/. [4]

Slika 3. Aviary web servis

IV. WEB SADRŽAJI DRUŠTVENIH MREŽA

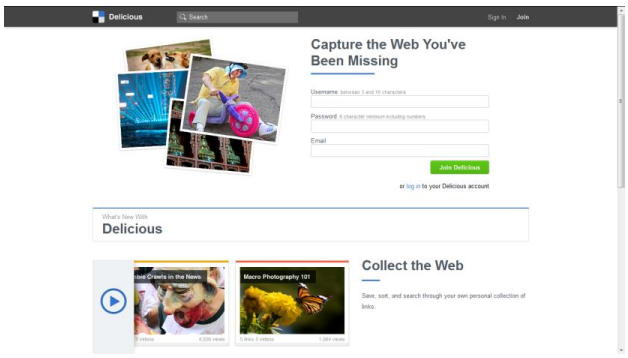
Danas je gotovo nemoguće ne spomenuti alate za društvene mrežne stranice raznih sadržaja. Društvene mrežne stranice postale su neizostavni trenutak života svake mlađe osobe, a i starijih generacija koje sve više počinju koristiti i primjenjivati ovakve sadržaje u svojem životu ili radi zabave ili radi posla. One korisnicima omogućavaju različite poglede na podatke i resurse, ali i pronalaženje pojedinaca s istim područjem interesa što dovodi do društvenog umrežavanja. Dalje u tekstu navesti ću neke od tih mrežnih sadržaja koji se danas koriste. [4]

Društvene mreže omogućavaju korisnicima spremanje i uređivanje svojih podatka u obliku profila, postavljaju *tagove* (oznake) na pojedine sadržaje i tako ih čine aktivnima. Jedan od primjera je i Amazon gdje korisnici mogu označiti proizvode i postaviti ih u košarice, te predložiti ih drugima. [4]

A. *Delicious*

Delicious je društvena mreža za spremanje, dijeljenje i otkrivanje zanimljivih sadržaja. Stranicu je osnovano Joshua Schachter 2003. godine, a 2005. preuzeo ju je Yahoo. Do kraja 2008. stranicu je posjetilo više od 5,3 milijuna korisnika koji su označili oko 180 milijuna raznih URL adresa. Sjedište tvrtke je grad Sunnyvale u Kaliforniji.

Sadržajima na Deliciousu pristupa se preko korisničkog računa kojeg se kreira na stranicama weba. Nakon kreiranja računa možemo pretraživati nove web sadržaje, označavati zanimljive stranice, razmjenjivati mišljenje s drugim korisnicima i dr. Web adresa stranice je: http://www.delicious.com/. [5]

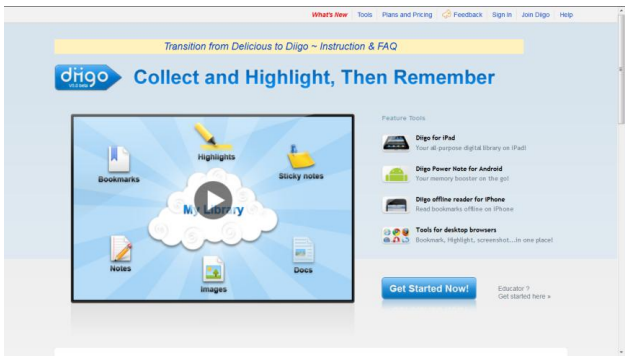


Slika 4. Delicious društvena mreža

B. *Diigo*

Diigo je društvena mreža koja radi na principu oblaka i omogućava svojim korisnicima stvaranje univerzalne vrste sadržaja do kojeg taj isti korisnik na jednostavan način može pristupiti s nekog drugog uređaja. Zasnovana je na principu upravljanja informacijama i alatima koji omogućavaju korisnicima da prikupe, označe, pristupe i dijele razne sadržaje na različitim uređajima.

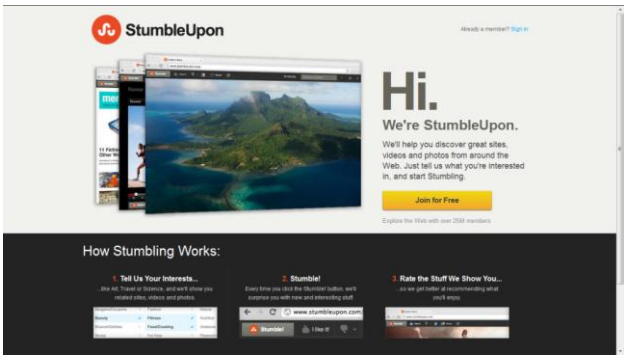
Diigo je zapravo odličan alat koji omogućava svojim korisnicima označavanje *(tagiranje)* traženog sadržaja. Osim toga on omogućava označavanje bilo kojeg dijela web stranice i kopiranje bilješke na određene dijelove neke druge stranice ili kreirane mape. Web adresa Diigo stranice je: http://www.diigo.com/. [3]



Slika 5. Diigo društveni servis

C. *StumbleUpon*

StumbleUpon je prema mnogim statistikama jedna od stranica koja ima najveću posjetu nakon Googla. Stranica je zamišljena tako da se instalira StumbleUpon Toolbar te da se pomoću njega pronalaze nove stranice koje još nitko nije otkrio. Kada te stranice budu pronađene, korisnik će zauvijek biti označen kao korisnik koji ih je pronašao. Web adresa ove stranice je: http://www.stumbleupon.com/. [6]



Slika 6. StumbleUpon društveni servis

D. *Linkedln*

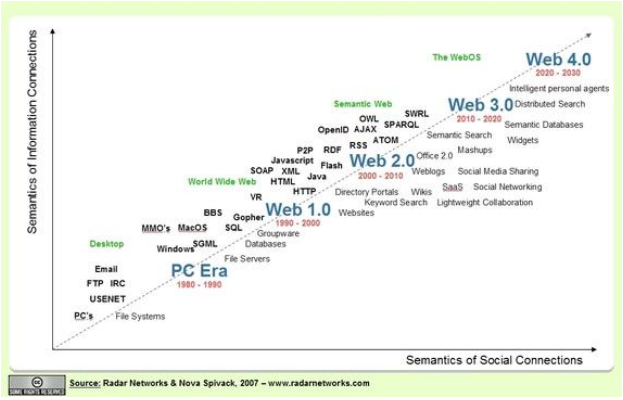
Linkedln je društvena mreža orijentirana više prema poslovnim i profesionalnim korisnicima nego prema mladeži i zabavi. Pristupom na ovu mrežu preko web adrese http://www.linkedin.com/.

Korisnik kreira Linkedln profil koji zapravo predstavlja svojevrsni virtualni poslovni životopis. Oko 175 miliona profesionalnih korisnika koristi ovu društvenu mrežu kako bi efikasnije i efektivnije razmijenilo poslovne informacije, stvorilo kreativne ideje o napredujućim projektima ili tek planiranim projektima i šansama o potencijalnom ostvarivanju ekonomske dobiti.

V. WEB 3.0

Što je to zapravo Web 3.0 i kako bi on trebao izgledati? Je li to inteligentni web i što to znači? Mnogi ljudi razmišljaju o korištenje napredne umjetne inteligencije kao sljedećeg velikog koraka razvoja weba. No, ideja da će cijeli web prerasti u jednu virtualnu zajednicu svijeta sa zgradama, trgovinama i drugim područjima za istraživanje i to sve u interakciji s ljudima nije tako ni nevjerojatna u tehnološkom smislu. [7]

Virtualni web bi trebao biti napravljen od mnoštva web stranica koje bi bile podržane jednim standardom na globalnoj razini i prihvaćene kao glavni medij/servis koji će omogućiti korisnicima dostupnost informacija svugdje i u svakome trenutku. Nešto poput današnjeg računalstva u oblacima. [7]



Slika 7. Prikaz evolucije Weba [7]

Predstavljanjem elektroničkih uređaja i interneta kao izvora za glazbu, filmove, trgovinu, poslovanje, platnu transakciju i još mnogo toga u središtu rada i igre našao bi se Web 3.0.

Unutar jednog desetljeća, pristup internetu na našim mobilnim uređajima (mobiteli, SmartPhone, Pocket PC) bit će (već je i postao) popularan kao SMS poruka. To će učiniti internet uvijek prisutnim u našim životima: na poslu, kod kuće, na putu, na odmoru gdje god idemo, Internet će biti tamo, a samim tim i Web 3.0. [7]

VI. REZULTATI ISTRAŽIVANJA

A. *Anketni upitnik*

Zanimanje:_____
Dob: _____
Spol: _____

1. Što su za Vas elektronički sadržaji/web alati za eUčenje?_____

2. Možemo li elektroničke sadržaje/web alate koristiti u nastavi? Zašto? Na koji način?
_____

3. Koje elektroničke sadržaje/web alate treba koristiti u nastavi? _____

4. Koristite li Vi elektroničke sadržaje/web alate?
da ne

5. Koje elektroničke sadržaje/web alate koristite?
_____

6. Zbog čega i zašto koristite navedene elektroničke sadržaje/web alate u nastavi?
_____

7. Jeste li ikada čuli za web alat *Vimeo* ili za alat *Aviary*?_____

8. Ako ste koristili gore navedene web alate što mislite o njima, njihovom dizajnu, primjeni, praktičnosti?_____

9. Koristite li web sadržaje na *Youtub-u*?
da ne

10. Za što koristite *Youtube*?
_____

11. Znaš li što je web alat *Delicious*, a što *web alat Diigo*? _____

12. Koriste li neki od alata iz 11.pitanja? Ako da, napišite zašto i navedite na koji način koristite navedeni alat.
_____

13. Po Vašem mišljenju koji je od dva navedena web alata kvalitetniji i zašto?
    a) *StumbleUpon* _____
    b) *Linkedln* _____

14. Što mislite o korištenju web alata u nastavi?
    a) neizostavni su dio mojeg obrazovanja
    b) vjerujem da će u budućnosti postati glavni sadržaji u nastavi
    c) mislim da nisu značajne za moje obrazovanje

15. Što mislite o korištenju elektroničkih sadržaja/web alata u nastavi računalstva?
_____

B. Prikaz rezultata istraživanja

Istraživanje je provedeno na uzorku od 50 ispitanika. Od toga je 19 ispitanica ženskog spola i 31 ispitanik muškog spola. Ispitanici su u prosjeku imali od 15 do 25 godina (Tablica I.).

Istraživanje je provedeno u dvije skupine. Prva skupina su bili učenici, a druga skupina odrasli. U prvoj skupini sudjelovalo je 30 ispitanika, a u drugoj skupini sudjelovalo je 20 ispitanika. Obje skupine popunile su anketni upitnik. Kroz anketni upitnik dobiveni su podaci koji su predstavljali stavove polaznika o gore navedenim web alatima za eUčenje. Anketni upitnik sastojao se od 15 pitanja raznih oblika.

Tablica I. Prikaz broja ispitanika po skupinama i spolu

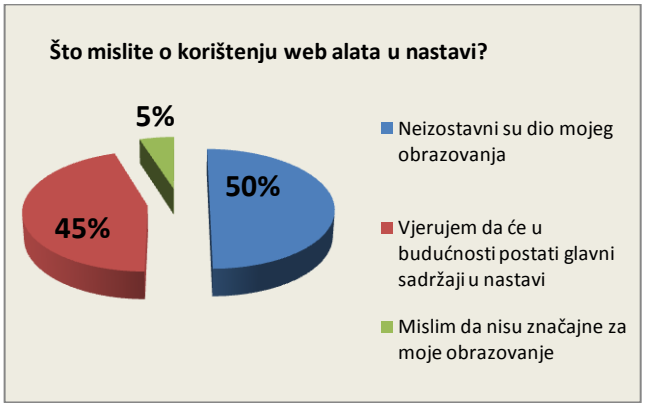
	Žene	**Muškarci**	**Ukupno**
Učenici	**19**	**21**	**30**
Odrasli	**10**	**10**	**20**
Ukupno	**19**	**31**	**50**

Cilj istraživanja je istražiti mišljenje o korištenju i primjeni besplatnih web alata za eUčenje u nastavnom procesu učenika, pa tako i odraslih. Rezultati anketnog upitnika su zanimljivi.

U ostatku teksta ovoga poglavlja navedeno je nekoliko značajnijih primjera odgovora iz anketnog upitnika prikazanih u brojčanim i grafičkim vrijednostima.
Svakoj od skupina postavljena su pitanja vezana uz korištenje web sadržaja kao što su društvene mreže, web alati, web video kanali i dr.

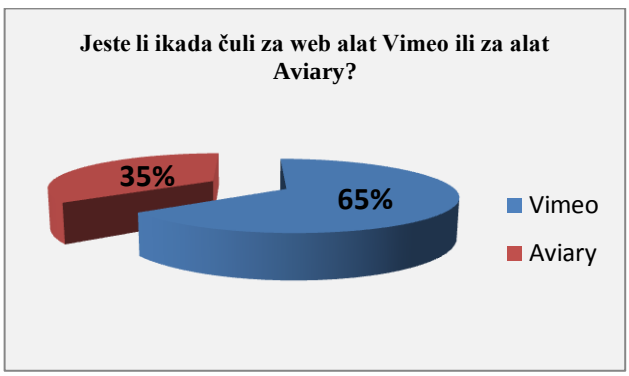
Kroz postavljena pitanja trebali su se provjeriti stavovi obje skupine. Na ovaj način pokušalo se doći do potrebnih informacija vezanih uz mišljenje ispitanika o korištenju, primjeni i funkcionalnosti gore navedenih web sadržaja i alata.

Grafikon I. Prikaz rezultata o korištenju web alata u nastavi



Iz sveukupne analize rezultata anketnih vrijednosti, vidljivo je da veći broj učenika, polaznika prihvaća ovakav rad i pristup nastavnim sadržajima.
Na 14. pitanje: "Što mislite o korištenju web alata u nastavi?" 50% polaznika odgovorilo je da su web alati neizostavni dio njihovog obrazovanja, 45% polaznika odgovorilo je kako vjeruje da će u budućnosti web alati postati glavni sadržaj u nastavi, a svega 5% polaznika da web alati nisu značajni za njihovo obrazovanje. (Grafikon I.).

Grafikon II. Prikaz rezultata o korištenju web alata Vimeo i Aviary



Kod pitanja vezanih uz konkretne primjere korištenja web alata u nastavi odgovori polaznika bili su očekivani.

Na 7. pitanje: Jeste li ikada čuli za web alat Vimeo i Aviary odgovori izraženi u postotnim vrijednostima prikazani su Grafikonom II.. Polaznici nastave, njih 65% odgovorilo je da su čuli za Vimeo, a njih 35% da su čuli za Aviary.
Rezultati odgovora na 11, 12 i 13 pitanje anketnog upitnika bili su slični rezultatima iz 7. pitanja. Dio polaznika napisao je da koristi web alate Delicious, Digo, StumbleUpon i Linkedln. Na 12. pitanje: "Zašto i na koji način koristite navedene alate?" polaznici su odgovorili da navedene alate koriste u svoje slobodno vrijeme za komunikaciju s prijateljima i rodbinom. Dio polaznika je napisao da ih koristi za traženje posla (odrasli polaznici), te za poboljšanje svojeg obrazovanja i efikasnijeg učenja nastavnih materijala.

Rezultati istraživanja su očekivani i zadovoljili su postavljenu hipotezu.
Hipoteza je glasila: Veliki dio web sadržaja danas se koristi u obrazovne i komunikacijske svrhe kao jedan od modela cjeloživotnog obrazovanja budućnosti.

Iz prikazanoga je vidljivo da ispitanici obje skupine koriste razne web sadržaje u svojem obrazovanju i da imaju pozitivne stavove prema korištenju navedenih sadržaja.

VII. ZAKLJUČAK

Web 1.0 definiran je kao standard objavljivanja sadržaja ljudima kako bi ih u što kraćem vremenu obavijestili i informirali o događanjima na globalnoj razini.

Web 2.0 definiran je kao standard objavljivanja sadržaja i informacija na globalnoj razini, ali sa nakanom međusobnog informiranja, novčanih transakcija, rada na projektima, stvaranja novih ideja i druženja jednih s drugima.
Iz navedenih primjera u poglavlju o Web 2.0 tehnologiji vidljivo je da su sadržaji weba postali interaktivni. *Podcasting*, Wikipedia, društvene mreže, razni *online* servisi te *online* prezentacije postali su neizostavni alat koji omogućava jednostavno i brzo stvaranje *online* sadržaja.
Tim sadržajima moguće je pristupiti u bilo koje vrijeme, preuzeti ih na bilo koje računalo ili drugi digitalni uređaj te pomoću njih prezentirati složen ili zanimljiv interaktivni sadržaj na efikasniji način.

Što je zapravo Web 3.0?, Što nam donosi budućnost? Istina je da je predviđanje budućnosti Web 3.0 čisto nagađanje i igra analize postojećih sadržaja. Web 3.0 definiran je kao spoj ljudi i strojeva s velikom razinom interakcije, korištenjem umjetne inteligencije i to ponajprije za objavljivanje i interakciju jednih s drugima, informiranje i povećanje efikasnosti i ljudske efektivnosti u radu na globalnoj razini.
Temeljnih promjena u načinu na koji koristimo web sigurno će biti, ali kako će se one odvijati i u kojem pravcu krenuti može se samo pretpostaviti. Web će se možda temeljiti na virtualnoj stvarnosti i većoj količini interakcije među korisnicima, složenijoj primjeni umjetne inteligencije. U svakom slučaju doći će do još većeg proboja web tehnologije u ljudske živote. Sve to moglo bi se razviti na neki zanimljivi način, prema kojem će umjetna inteligencija kreirati svijet i tržište kapitala.

LITERATURA

[1] Edudemic. The 35 best Web 2.0. classroom tools chosen by you. Dostupno na: http://edudemic.com/2010/07/the-35-best- web-2-0-classroom-tools-chosen-by-you/ (20.1.2015.)

[2] Isabel Peters. Folksonomies: Indexing and Retrieval in Web 2.0. DeGruyter Saur, 2009. (str. 13-107). Dostupno na: http://books.google.hr/books?id=Aeib_wy18gkC&printsec=frontcover&dq=Isabel+petersIndexing+and+Retrieval+in+Web+2.0&source=bl&ots=MItfzFHvyJ&sig=9SNqxvuH5pQPpduQ52KmLJmeass&hl=hr&sa=X&ei=SYJtUIzSJI_asgaKID4DA&ved=0CCwQ6AEwAA#v=onepage&q&f=false. (29.1.2015.).

[3] Radulović, V.; Radulović, Č.; Đurić, Z. Društvene mreže i semantički web. Dostupno na: http://www.infoteh.rs.ba/zbornik/2012/radovi/RSS-8/RSS-8-4.pdf (7.1.2015.)

[4] Slavić, A. Semantički Web, sustavi za organizaciju znanja i mrežni standardi. U: Informacijske znanosti u procesu promjena (ur. J. Lasić-Lazić). Zagreb: Zavod za informacijske studije, 2005.

[5] Wikipedia. Delicious (website); Preuzeto: http://en.wikipedia.org/wiki/Delicious_%28website%29, (24.1.2015).

[6] Wikipedia. StumbleUpon. Preuzeto: http://en.wikipedia.org/wiki/StumbleUpon, (24.1.2015).

[7] The Paisano. Web 2.0 vs Web 3.0. Preuzeto: http://thepaisano.wordpress.com/2008/03/08/web-20-vs-web-30/(11.1.2015)

ANALIZA UPOTREBE PROGRAMA U OBRAZOVANJU NA PRIMJERU SUSTAVA LINEARNIH JEDNADŽBI

mr. Said Čerkezović*, Hajrija Šišić**, mr. Izet Jagodić***
* Općina Živinice, Rudarsko geološko građevinski fakultet, Univerzitet u Tuzli, Živinice, BiH
** MSŠ Banovići, Banovići, BiH
*** MSŠ Živinice, Živinice, BiH
Public office for research and protection
cerkezovic_s@hotmail.com

Sažetak—**Rad prezentira analizu osnovnih metoda i programa koji se koriste u odgojno-obrazovnim ustanovama. Analiza je provedena na primjeru rješavanja sustava linearnih jednadžbi koje su nezaobilazni faktor izučavanja od osnovne škole, srednje škole, fakulteta a kasnije i u samom području djelovanja. Na primjeru sustava linearanih jednadžbi sa tri nepoznanice kroz programe Excel, FORTRAN, PASCAL i Matlab vršena je analiza rješavanja navedenog problema od strane učenika i studenata različitim metodama. Analizira se jedan segment tijesno povezanih i srodnih predmeta matematike i informatike kroz navedene računalske programe, da bi se i u budućem radu pravila paralela između ovih programa.**

I. UVOD

Kada je u pitanju obrazovanje kroz informatiku, računalo se koristi kao sredstvo za dublju analizu razumijevanja predmeta i pronalaženja novih pristupa rješavanju određenih problema, pri čemu važno mjesto zauzima u učenju i rješavanju matematičkih problema. Koliko će određena osoba iskoristiti računalo i dostupne programe zavisi od same osobe i njezine informatičke pismenosti. Pod pojmom informatičke pismenosti (eng. Computer literacy) definiše se sposobnost korištenja računala i programa.

Programi koji se najčešće koriste u osnovnim školama su BASIC, srednjim FORTRAN ili PASCAL, dok izučavanje na fakultetima zavisi od samog nastavnog plana i programa i kreće se od FORTRANA, PASCALA, Matlaba, C, C++ itd. Ono što je karakteristično za svo obrazovanje jeste korištenje uredskog programa Microsoft Excel, koji se izučava u osnovnim i srednjim školama, kao i na pojedinim fakultetima.

Svi programi imaju za cilj da polaznike nauče osnove programiranja i logičkog razmišljanja, kroz algoritme, a nakon toga i kroz kod datog programa. Pronalaženjem metoda za rješavanje određenog problema razvija se kreativnost i apstraktno mišljenje, te produbljuje intuicija osoba. Traži se put kako najefikasnije, praktično i kreativno riješiti određeni problem. Shvatanjem osnove programiranja, razvijanjem logičkog razmišljanja, postavljanja, analiziranja i rješavanja problema, učenici će naučeno moći primijeniti na bilo kojem području djelovanja. Edukacija polaznika mora posjedovati metode za razvoj vještina za rješavanje problema, logičkog zaključivanja, programiranja i ispravljanja pogrešaka [1]. Navedeni problem rješavanja sustava linearnih jednadžbi u nekom od programskih jezika bit će napravljen i analiziran jednom od direktnih metoda sistemskih procedura koje se zasnivaju na principu eliminacije, odnosno mogućnosti pronalaska inverzne matrice A^{-1}. U slučaju neregularnog sustava, odnosno sustava koji nema jedinstveno rješenje, neće postojati inverzna matrica. Cilj analize je pokazati različitost ~~prihvatanja~~ ponuđenih metoda rješavanja i ponuđenih programskih jezika, odnosno programa od strane učenika i studenata.

II. OSNOVE PROGRAMA ZA RJEŠAVANJE SUSTAVA LINEARNIH JEDNADŽBI

Odgovori na rješavanje sustava linearnih jednadžbi nisu jednostavni niti jednoznačni. Oni se mogu tražiti različitim analizama i na različite načine. Ovo je samo jedan od načina analize navedenih programa na primjeru koji se vrlo često spominje kako u obrazovanju tako i u samom praktičnom radu. Programer prilikom programiranja će definirati konačan cilj, koji će se ogledati kroz stvaranje aplikacije koja će služiti određenoj grupi korisnika, pri čemu se nastoji prilagoditi aplikacija domenu krajnjeg korisnika.

Problem koji se javlja kod krajnjeg korisnika programa je da oni ili nemaju vremena ili nemaju interesa naučiti programski jezik. Kada je u pitanju primjena proračunskih tablica ističe se jednostavnost i formula i ugrađenih funkcija. Jednostavnost se ogleda kroz omogućavanje svega što se radi u svakodnevnom obrazovanju:, zbrajanje, oduzimanje... Podaci su složeni u dvodimenzionalnu mrežu koja je korisnicima laka za pregled i izmjenu. Isti se mogu osjećati kao programeri jer direktno rade na određenom problemu, pri čemu vide aritmetičke operatore i funkcije koje pozivaju, kao i manipulirati radom bez uobičajnih aktivnosti programiranja. Može se reći da je lakše shvatiti proračunsku tablicu od općenitosti namjene programskog jezika. Iz istih razloga se koristi u velikom broju različitih aplikacija, od financijskih, administrativnih i sl.

FORTRAN (Formula Translation) je jezik za programiranje prvenstveno namijenjen rješavanju numeričkih problema. Najčešće je sastavljen od:

- glavnog programa;
- podprograma;
- funkcijskog podprograma i
- bloka podataka.

Pascal je programski jezik razvijen za učenje strukturiranog programiranja, kod koga algoritam mora biti dobro razrađen i podijeljen na jednostavnije cjeline, a strukturu programa određuje struktura podataka.

Matlab – matrični laboratorij omogućava brzo i lako programiranje. Način zapisivanja je blizak načinu na koji se inače zapisuju matematičke formule. Ima mnoge prednosti u odnosu na konvencionalne programske jezike (Fortran, C) za rješavanje tehničkih problema [2]. Često jedna naredba u Matlab-u mijenja stotine redova koje su napisane u nekom programskom jeziku opće namjene (Fortran, C-a). Matlab čini skup funkcija za jednostavno, prirodno i efikasno manipuliranje matricama. Određene funkcije su pisane u C-u, pri čemu izvorni kod nije dostupan. Programski jezik koji sa svojim skupovima ugrađenih funkcija većinu problema puno jednostavnije rješava u odnosu na programske jezike Fortran, Pascal, C, ...

Microsoft Excel je program za tablično računanje, uglavnom rješavanje problema matematičkog tipa pomoću tablica i polja koje je moguće povezati različitim formulama. Kada je u pitanju bilo kakvo jednostavnije istraživanje, analiza je pokazala prednosti korištenja Excela koji omogućuje jednostavno integriranje i transformiranje podataka u jednu cjelinu koja je pogodna za analizu računalom. Najčešće se ti podaci prikazuju u obliku tablica. Proračunske tablice se koriste u velikom broju različitih aplikacija. Najčešće su to financijske aplikacije poput proračuna i administrativne aplikacije kao što su upravljanje malim tvrtkama. Ovakve tablice su sadržavale vrlo mali broj proračuna.

Excel ima kompletnu paletu funkcija koje uz određena objašnjenja daje rezultate. Korištenjem odgovarajuće tablice razvijamo kod učenika sposobnost rada sa standardnim alatom današnjeg poslovnog svijeta [3].

Ono što je karakteristično za proračunske tablice jeste da je kvanitativnom analizom na 135 proračunskih tablica prikupljenih od Finskih poslovnih tvrtki i vladinih agencija pokazalo da su tablice sadržavale vrlo malo proračuna. 90% slučajeva su korišteni aritmetički operatori (+, -, *, /), a od 80% upotrebljenih funkcija su korištene funkcije IF i SUM [4]. Snaga trenutnog modela proračunske tablice je direktna manipulacija njegovog interfejsa.

Navedenim programskim jezicima operacije su sagrađene od operacija nižih razina i kontrolnih struktura, pri čemu programeri moraju imati puno veće znanje o osnovama programiranja. Mnogi problemi prilikom projektiranja i analize će proizvesti niz složenih linearnih sustava jednadžbi, kod kojih se algoritmi razvijaju da bi se riješili određeni sustavi.

III. NAČIN RJEŠAVANJA I PROGRAMSKI KOD LINEARNE JEDNADŽBE

Promatramo sustav linearnih jednadžbi sa tri nepoznate

$$a_{11}x_1+a_{12}x_2+a_{13}x_3=b_1$$
$$a_{21}x_1+a_{22}x_2+a_{23}x_3=b_2$$
$$a_{31}x_1+a_{32}x_2+a_{33}x_3=b_3$$

- Ako navedeni sustav jednadžbi rješavamo pomoću FORTRANA, primjenićemo Cramerovo pravilo. Cramerovo pravilo nije efikasno za rješavanje velikih sustava jednadžbi, što u obrazovanju i nije čest slučaj.

$$x_i=\frac{\det A_i}{\det A}$$

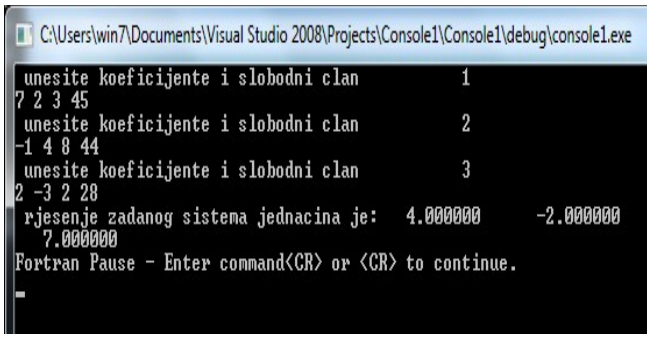
$$\det A=\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$\det A_1=\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}$$

$$\det A_2=\begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}$$

$$\det A_3=\begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}$$

Pomoću podprograma za računanje determinante, i Cramerovog pravila koji mijenja kolone $a(I,1)$ za prvu determinatu, $a(I,2)$ za drugu determinatu i $a(I,3)$ za treću determinatu pomoću DO petlje za $I=1,3$ vrijednostima $B(I)$, dobijamo rješenja X_I [5].



Slika 1. Prozor Fortran-a rješenog navedenog problema

Kod kojim se rješava navedeni problem u fortranu je:

```
program sustav_linearnih_jednadzbi
real A(3,3), X(3), B(3)
```

```
      do I=1,3
      Print*,'unesite    koeficijente    i
slobodni clan', I
      read*,(A(I,J), J=1,3), B(I)
      end do
call CRAMER (A,X,B)
print*,'rjesenje      zadanog      sustava
jednadzbi je:', (X(I), I=1,3)
end

function DET(A)
 real A(3,3)
                  DET=A(1,1)*(A(2,2)*A(3,3)-
A(2,3)*A(3,2))-&
 A(1,2)*(A(2,1)*A(3,3)-A(2,3)*A(3,1))+&
 A(1,3)*(A(2,1)*A(3,2)-A(2,2)*A(3,1))
 return
 end

subroutine CRAMER (A,X,B)
real   A(3,3),   X(3),   B(3),   A1(3,3),
A2(3,3), A3(3,3)
 DENOM=DET(A)
 A1=A
 A2=A
 A3=A
    DO I=1,3
    A1(I,1)=B(I)
    A2(I,2)=B(I)
    A3(I,3)=B(I)
    END DO
 X(1)=DET(A1)/DENOM
 X(2)=DET(A2)/DENOM
 X(3)=DET(A3)/DENOM
 return
 end
```

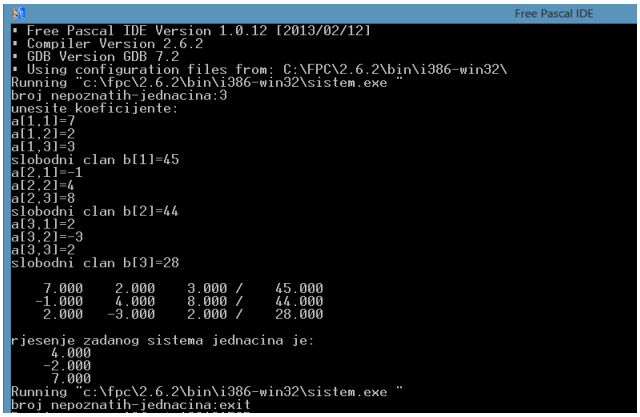
- Navedeni problem u Pascalu ćemo riješiti pomoću Gauss-Jordanove metode eliminacije u kojoj se elementi iznad kao i elementi ispod glavne dijagonale eliminiraju (izjednačavaju sa nulom).

Suština navedenog algoritma je da se učitaju koeficijenti, izvrši proširivanje matrice slobodnim članovima a zatim se nizom transformacija navedena matrica *A* transformira u dijagonalnu matricu, odnosno jediničnu matricu. Vektor *b* tada postaje vektor rješenja *x*. Kod koji je iskorišten za rješavanje problema sustava linearnih jednadžbi i koji je iskorišten u obrazovanju učenika i studenata je:

```
program sustavjednadzbi( input,output );
type matrica = array [1..6,1..7] of real;
var a:matrica;
    n:integer;
 procedure citaj (n:integer;var
a:matrica);
 var i,j:integer;
 begin
    writeln('unesite koeficijente:');
    for i:=1 to n do
        begin
           for j:=1 to n do
                begin
                 write
('a[',i,',',j,']=');
                 readln (a[i,j]);
                end;
           write('slobodni clan
b[',i,']=');
           readln(a[i,n+1])
          end
 end;
 procedure pisi(n:integer;a:matrica);
 var i,j :integer;
 begin
    writeln;
    for i:=1 to n do
        begin
           for j:=1 to n do
write(a[i,j]:9:3);
           writeln(' / ',a[i,n+1]:9:3);
        end
  end;
  procedure ga(n:integer; var a:matrica);
  var i,j,k:integer;
      regularna:Boolean;
      function najveci:integer;
       var naj:real;
       i:integer;
         begin
          naj:= abs(a[k,k]);
          najveci:=k;
          for i:=k+1 to n do
              if (naj<abs(a[i,k])) then
              begin
              naj:=abs(a[i,k]);
              najveci:=i
              end
        end;
    procedure zamjenaredova(k,l:integer);
      var r:real;
      j:integer;
      begin
        for j:=k to n+1 do
            begin
             r:=a[k,j];
             a[k,j]:=a[l,j];
             a[l,j]:=r
            end
      end;

      procedure ispisrjesenja;
      var i:integer;
      begin
      writeln;
      writeln ('rjesenje zadanog sistema
jednacina je:');
```

```
      for i:=1 to n do
writeln(a[i,n+1]/a[i,i]:10:3)
      end;
      begin
        regularna:=true;
        k:=1;
        repeat
         if (k<>najveci) then
zamjenaredova (k,najveci);
         if (a[k,k]<>0) then
         begin
          for i:=1 to n do
            if i<>k then
                 for j:=k+1 to n+1 do
                  a[i,j]:=a[i,j]-
a[i,k]*a[k,j]/a[k,k]
         end
         else regularna:=false;
         k:=k+1
         until (k>n) or not regularna;
         if regularna then ispisrjesenja
                       else   writeln
('neregularna matrica sistema')
      end;
```



Slika 2. Prozor Pascal-a rješenog navedenog problema

- Navedena jednadžba u Excelu bi se predstavila u obliku $Ax=b$ pri čemu bi matrica sustava imala oblik:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Navedeni sustav jednadžbe bi riješili računanjem inverzne matrice A^{-1}, tako da je $x=A^{-1}b$

B9 {=MINVERSE(B1:D3)}

	A	B	C	D
1		7	2	3
2	[A]=	-1	4	8
3		2	-3	2
4				
5		45		
6	[b]=	44		
7		28		
8				
9		0,130612	-0,05306	0,016327
10	Inv[A]=	0,073469	0,032653	-0,24082
11		-0,02041	0,102041	0,122449
12				
13			4	
14	[x]=Inv[A][b]=		-2	
15			7	

Slika 3. Prozor Excel-a rješenog navedenog problema

Inverznu matricu A^{-1} u Excelu bi dobili pomoću funkcije MINVERSE, a množenje matrica pomoću MMULT. Uvrštavanjem određenih promjenjivih dobijamo rješenja X_1. Ubacivanjem određenih parametara sustava linearne jednadžbe u Excel dobija se rješenje navedene jednadžbe (Slika 3.).

- I u Matlabu, rješavanje sustava linearnih jednadžbi može se izvršiti pronalaženjem inverzne matrice *inv(A)*. Jednostavniji način rješavanja je primjenom operacije dijeljenja s lijeve strane, pri čemu se koristi pristup Gaussove eliminacije (*x=A\b*).

```
>> A=[7 2 3;-1 4 8;2 -3 2]
     A =
       7     2     3
      -1     4     8
       2    -3     2
>> b=[45;44;28]
     b =
      45
      44
      28
>> x=inv(A)*b
     x =
      4.0000
     -2.0000
      7.0000
```

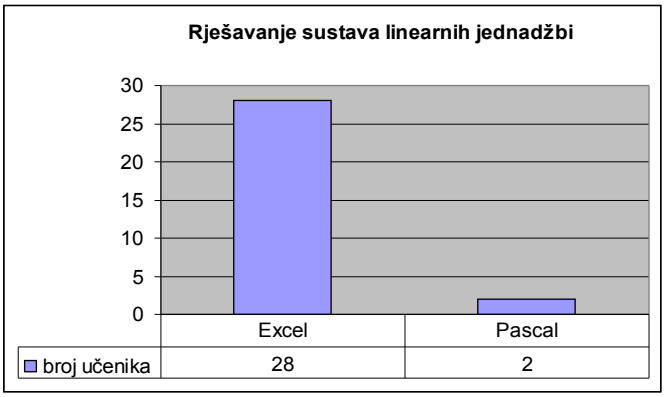
Jednostavniji način je:

```
>> A=[7 2 3;-1 4 8;2 -3 2]
     A =
       7     2     3
      -1     4     8
       2    -3     2
>> b=[45;44;28]
     b =
      45
      44
      28
>> x=A\b
     x =
      4.0000
     -2.0000
      7.0000
```

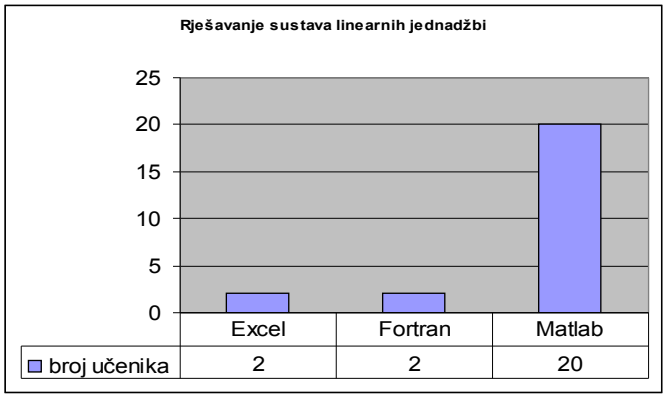
IV. ANALIZA NAVEDENIH PROGRAMA U OBRAZOVANJU

Analizirajući rješavanje navedenih problema linearnih jednadžbi sa učenicima srednje škole koji izučavaju Pascal i Excel, i studentima II godine koji izučavaju i poznaju rad u Fortran-u, Excel-u i Matlab-u,ostavljena im je mogućnost da prilikom rada izaberu metodu i programski jezik (grafikon 1., grafikon 2.). Zahtijevano je da obrazlože zašto su izabrali određeni program i način rješavanja.

Analiza urađenih zadatka sustava linearnih jednadžbi sa tri nepoznanice i dobivenih odgovora je pokazala da su učenici izabrali Excel a studenti Matlab zbog jednostavnosti i bržeg načina rješavanja. I učenici i studenti koji su dati problem rješavali u drugim programima su naveli razlog samo izazova da navedeni problem riješe na drugi način, pri čemu je komentar bio da je lakše riješiti pomoću Excela, odnosno Matlaba. Prednost je bila u tome što su se morale znati samo dvije formule MINVERSE i MMULT u Excel-u i operacija dijeljenja s lijeve strane u Matlab-u, dok kod ostalih programa uvijek postoji mogućnost pogreške prilikom unosa podataka, DO ili FOR petlji, IF naredbi kao i samim podprogramima.



Grafikon 1. Rješavanje sustava linearnih jednadžbi kroz programe Excel i Pascal



Grafikon 2. Rješavanje sustava linearnih jednadžbi kroz programe Excel, Fortran i Matlab

V. ZAKLJUČAK

Proračunske tablice-Excel je program koji je dobro prihvaćen i koji se često koristi. Zbog jednostavnosti većina korisnika iz promatranih grupa će koristiti navedeni program, jer se njime navedeni problemi rješavaju najjednostavnije. I učenici i studenti su pokazali da će rješavati određeni problem korištenjem najjednostavnijeg načina i programa, a suštinu programiranja ne žele i neće da znaju. Međutim, i pored odgovarajuće funkcionalnosti izazov programiranja leži u razvoju alata i metodologija koje mogu pomoći krajnjim korisnicima. Na taj način korisnici će pored programiranja i shvaćanja programskog koda imati mogućnost shvatiti i matematski način rješavanja određenog problema određenim metodama. Koliko će sve to imati pozitivnog efekta na same korisnike zavisi i od njihove zainteresiranosti kao i od nastavnog plana i programa na osnovu kojeg je potrebno predstaviti načine riješavanja navedenih problema, ili objasniti suštinu samog programiranja. Tako se postiže dvostruki cilj: razvija se kreativnost i sposobnost logičkog razmišljanja korisnika u izboru najefikasnijih metoda pronalaska rješenja problema, i istovremeno se ostvaruje kvalitetna veza između matematike i informatike kao srodnih predmeta.

VI. LITERATURA

[1] A.O. Ajayi, E. A. Olajubu, and D. F. Ninan, "Development and Testing of a Graphical FORTRAN Learning Tool for novice Programmers", Interdisciplinary Journal of Information, knowledge, and Management, Volume 5, 2010.

[2] David Houcque, "Introduction to Matlab for engineering students", Northwestern University, august 2005.

[3] Shkelqim KUKA, Brunela KARAMANI, "Using Excel and VBA for Excel to Learn Numerical Methods", 1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011).

[4] Sajaniemi J. Pekkanen J., "An empirical Analysis of Spreadsheet Calculation", Softvare-Practice and Experience 18 (6): 583-596.

[5] Williame E.Nayo, Martin Cwiakala, "Programming with FORTRAN 90, teory and problems", Schaum's outline series.

E-učenje podržano rješenjima u oblaku

Ivan Livaja, Frane Urem, Anita Grubišić, Tanja Radić Lakoš
Veleučilište u Šibeniku, Trg Andrije Hebranga 11, 22000 Šibenik, Hrvatska
Telefon: 00-385 22 311060 Fax: 00-385 22 216716
E-mail:ilivaja@vus.hr, frane.urem@vus.hr, anita@vus.hr, tanja@vus.hr

Igor Žaja
Hrvatski Telekom d.d., Roberta Frangeša Mihanovića 9, 10110 Zagreb, Hrvatska
Telefon: 00-385 1 49 11 919 Fax: 00-385 1 49 12 800 E-mail: igor.zaja@t.ht.hr

Sažetak – Novi tip učenja zove se e-učenje, odnosno kombinacija e-učenja i računarstva u oblaku (cloud). Bazirana je na velikom širenju računarstva koje se oslanja na dijeljenje računalnih resursa, a uključuje isporuku sadržaja, aplikacija, infrastrukture i platforme kao usluge. Padom cijena internet usluga kao i golemim vremensko – troškovnim uštedama e-učenje postaje dio naše svakodnevnice. Mogućnost gotovo trenutne disperzije novih znanja u praktički prostorno neograničenim uvjetima (decentralizirani sustavi, udaljene destinacije, nemogućnosti prisutva nastavi). Ova fleksibilnost je ono što privlači učenike, studente, nastavni kadar da koriste sustave e-učenje. U radu je dan pregled rješenja usluge u oblaku za implementaciju sustava e-učenja. Analizirane su platforme i mogući sustavi koje pružaju Hrvatski Telekom (HT) i Sveučilišni računski centar (srce) za implementaciju sustava E-učenja. Znanje koje dobiveno ovakvim načinom edukacije ekvivalentno je onome u učionici, te ima brojne prednosti.

I. UVOD

E-učenje je učenje uz pomoć računala i, vrlo često, Interneta. E-učenje je prisutno u klasičnom obrazovanju u školama i na fakultetima, tako da se računalo i Internet koriste za vrijeme nastave ili u izradi domaćih zadaća. U drugim slučajevima se obrazovni programi djelomično izvode online, tj. uz pomoć Interneta, a djelomično na standardni način. E-učenje često se primjenjuje u izvođenju stručnih seminara i ostalih obrazovnih programa za koje bi bilo preskupo ili nemoguće okupiti polaznike na dulje vrijeme. Moguće je postojanje obrazovnih programa koji se u potpunosti izvode online, što znači da se od upisa do diplome, sva komunikacija između polaznika i organizatora obrazovanja odvija putem Interneta. Kad se neka ustanova odluči svoje obrazovne programe ponuditi putem e-learninga, treba pretpostaviti da pritom ima kao cilj poboljšati kakvoću obrazovanja ili omogućiti pohađanje onim polaznicima kojima bi to inače bilo onemogućeno ili omogućiti pristup izvorima i stručnjacima koje je nemoguće imati u blizini. Znanje dobiveno ovakvim načinom edukacije ekvivalentno je onome u učionici, te ima brojne prednosti u odnosu na nju, kao npr: nema izostanaka s nastave ili radnog mjesta, nema neadekvatne edukacije, jer možete odabrati odgovarajući sadržaj za svakog korisnika, nema nametnutog tempa učenja, jer ga svaki korisnik bira za sebe, troškovi su znatno niži od klasične obuke, voditelj ima potpuni pregled nad svim aktivnostima i rezultatima edukacije, polaznici imaju mogućnost osvježavanja znanja kroz duži period i nakon kompletiranja seminara. Također, potreba za pristupom svim podacima sa bilo koje lokacije, bez korištenja računala ili hard diska omogućava računarstvo u oblaku ili programska rješenja u oblacima (Cloud computing). Uvođenjem cloud tehnologije studentima i djelatnicima omogućujemo e-učenje, pristup sa bilo koje lokacije, te studenti imaju mogućnost aplikacijama rješavanje zadataka i seminara u bilo kojem trenutku sa bilo koje lokacije korištenjem potrebnih alata. Rješavanjem zahtjevnijih zadataka potrebne su zahtjevnije aplikacije, a samim time i veći resursi samog računala tj. hardvera i licenci za program što utječe na cijenu samog računala. Hrvatski Telekom (HT) je danas jedan među rijetkim telekomunikacijskim operatorima u Hrvatskoj koji je u mogućnosti pružiti kompletno ICT cloud rješenje – od pristupa internetu do razvoja i implementiranja aplikacije i korisničkih zahtjeva.

II. MODELI RAČUNARSTVA U OBLAKU

A. Javni cloud

Davatelj usluga omogućava pristup resursima kao što su aplikacije, skladišta za podatke i drugi resursi dostupni za javnost, nezavisno da li se radi o pojedincima ili organizacijama, putem interneta. Infrastruktura se nalazi u vlasništvu operatora i nije dostupna za uvid ili kontrolu korisnicima. Infrastruktura javnog clouda podrazumijeva dijeljene resurse za korisnike. Dijelovi javnog clouda mogu biti i pod isključivom upotrebom samo jednog korisnika, čineći tako privatni centar podataka [1], [2], [6].

B. Privatni cloud

Privatni cloud je napravljen isključivo za upotrebu jednog klijenta, rezerviran centar podataka za klijenta, koji može biti unutar organizacije ili na serveru od strane cloud davatelja usluga. IT odjeli kompanija ili davatelj cloud usluga grade privatne cloude i upravljaju njima.

Organizacije koje poseduju privatni cloud imaju potpunu kontrolu nad strukturom oblaka.

C. Zajednički cloud

Zajednički cloud je cloud koji se djeli na nekoliko organizacija. Infrastruktura podržava posebne zajednice koje imaju zajedničke potrebe, zahtjeve sigurnosti i slično. Njima mogu upravljati same organizacije ili drugi pružatelj usluga. Ovdje se troškovi dijele između nekoliko klijenta tako da su mogućnosti uštede ograničene.

D. Hibridni cloud

Strukturu clouda čine dva ili više različitih oblaka (privatni, zajednički ili javni) koji ostaju jedinstveni entiteti, ali su međusobno povezani standardiziranim ili prikladnim tehnologijama koje omogućavaju efikasan prijenos podataka ili aplikacija. Hibridni cloud povezuju javne i privatne modele clouda. Mogućnost proširivanja privatnog clouda s resursima javnog clouda može se koristiti za održavanje nivoa kako bi se lakše izdržala velika opterećenja. Hibridni cloud se također može koristiti za upravljanje planiranim velikim opterećenjima. Hibridni cloudi susreću se sa složenosti određivanja kako raspodijeliti aplikacije po javnom i privatnom cloudu.

III. CLOUD USLUGE ZA IMPLEMENTACIJU E-UČENJA

U ovom poglavlju dan je pregled cloud usluga za implementaciju sustava e-učenja. Analizirane su cloud usluge koje pružaju Hrvatski Telekom (HT) i Sveučilišni računski centar (srce). Na temelju analize usluga na hrvatskom tržištu potrebno se opredijeliti za cloud rješenja za koje se smatra da će na najbolji način zadovoljiti potrebe e-učenja tj. održavanja nastave i praktičnog laboratorijskog rada studenata u narednom razdoblju. Sustavi e-učenja pomoću rješenja računarstva u oblaku detaljnije su opisani u poglavlju IV.

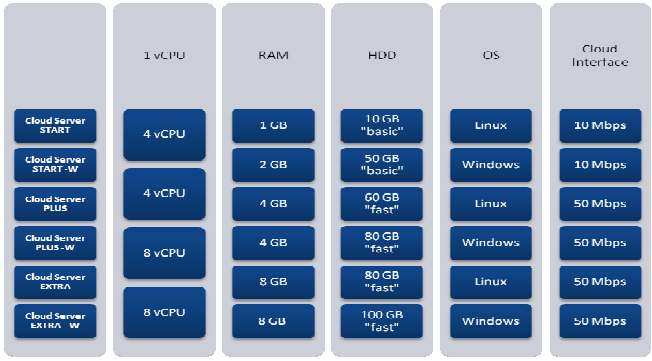
A. Cloud Virtula Private Server

Ustanovama iz sustava znanosti i visokog obrazovanja uslugu korištenja virtualnih poslužitelja (VPS) pruža Sveučilišni računski centar (srce). Usluga Cloud VPS omogućava zadovoljavanje potreba visokoškolskih ustanova za poslužiteljskim kapacitetima uz sljedeću specifikaciju standardnog poslužitelja:

- 2 procesora (vCPU) [Intel Xeon E7-4830 ~2 GHz]
- 8 GB radne memorije (vRAM)
- 100 GB diskovnog prostora
- jedna IPv4 adresa u javnom adresnom prostoru rezerviranom za pružanje usluge.

Ovisno o svojim potrebama ustanove mogu zatražiti dodatne računalne resurse, te tako kreirati optimalnu poslužiteljsku okolinu koja može zadovoljiti sve njihove zahtjeve.

Uslugom Cloud Server HT-a omogućuje se krajnjim korisnicima najam virtualnog poslužitelja smještenog u Data centru. Virtualni poslužitelji su dio virtualne platforme, čija iskoristivost i fleksibilnost čine uslugu prilagodljivom potrebama korisnika. Virtualni poslužitelji adekvatna su zamjena za fizičke poslužitelje, a korisniku omogućuju transparentno upravljanje troškovima bez kapitalnih investicija i potrebe za vlastitom IT infrastrukturom. Prebacujući brigu o IT infrastrukturi na ICT operatora korisniku je omogućeno fokusiranje na svoju primarnu djelatnost. Usluga Cloud Server nudi se u obliku nekoliko paketa određene procesorske snage, radne memorije i diskovnog prostora. Svi paketi uključuju operacijski sustav, pristup prema Internetu te uslugu sigurnosne pohrane podataka. Pristup prema privatnoj mreži korisnika također je moguće ostvariti putem L2 ili L3 veze. Na svakom odabranom virtualnom poslužitelju (Cloud Serveru) moguće su mjesečne promjene u smislu nadogradnje procesorske snage, radne memorije i diskovnog prostora. Sa stanovišta korisnika bitno je da poslužiteljima može pristupiti u bilo koje vrijeme, s bilo kojeg mjesta i preko bilo koje platforme, a cjelokupno korištenje uslugom naplaćuje se mjesečno [3].



Slika 1. Pregled Cloud Server paketa

B. Cloud Virtual Computing Lab

Usluga Virtual Computing Lab (VCL) je usluga koja djelatnicima ustanova iz sustava znanosti i visokog obrazovanja omogućava upotrebu skupova virtualnih poslužitelja. Usluga VCL omogućava korisnicima fleksibilno zadovoljavanje njihovih dinamičkih potreba kao što su virtualne učionice za potrebe nastave ili virtualni laboratoriji u nastavi ili istraživanjima. Usluga osigurava značajne uštede ustanovama, ali i ukupne uštede na razini sustava znanosti i visokog obrazovanja u odnosu na nabavu i održavanje fizičkih poslužitelja. Usluga nije namijenjena zahtjevnom računanju i ostalim oblicima intenzivnih obrada podataka, i ne zamjenjuje usluge sustava visoke učinkovitosti, poput računalnih klastera ili drugih specijaliziranih poslužiteljskih sustava Usluzi mogu pristupiti isključivo djelatnici akademskih i istraživačkih ustanova (odnosno oni koji su tako identificirani u sustavu AAI@EduHR). Korisnik pristupa usluzi korištenjem svog identiteta u sustavu AAI@EduHR [4].

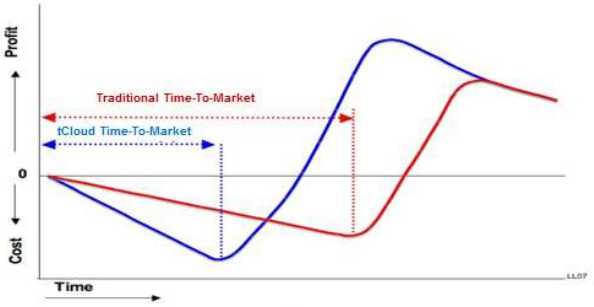
Standardni virtualni poslužitelj u okviru usluge je sljedećih svojstava:

- 1 procesor (vCPU),
- 2 GB radne memorije (vRAM),
- 20 GB diskovnog prostora,
- jedna IPv4 adresa u javnom adresnom prostoru rezerviranom za pružanje usluge.

Usluga cloud računalo (Virtual Desktop) HT-a je usluga koja korisniku osigurava u svakom trenutku, gdje god se nalazio, pristup svim svojim podacima kao da je za računalom, ali bez troškova kupnje računala, instalacije skupe opreme, softvera i održavanja. Ova usluga omogućuje efikasno korištenje računala i računalne infrastrukture bez troškova kupnje računala, instalacije skupe opreme, brige o održavanju IT infrastrukture, kapitalnih ulaganja. Pristup računalu je omogućen sa bilo kojeg uređaja koji je povezan na Internet. Cloud Računalo sastoji se od sljedećih komponenata: Windows OS, Antivirusna zaštita, 50GB HDD prostora, Flash player, Media Player, Internet Explorer, Integrirani backup, SMS autentifikacija.

Slika 2. Pristup virtualnom poslužitelju omogućen je sa bilo koje platforme

Cloud Računalo je DaaS (Desktop as a Service) platforma, proizvod u koju su integrirane Citrix, Cisco i NetApp komponente. Rješenje uključuje Cisco Unified Computing System (USC) komponente, konvergiranu Data Centar platformu sa NetApp sustavom za pohranu podataka te multiprotokolarnu povezivost koja koristi CIFS, NFS i Fibre Channel okruženje na jedinstvenoj platformi za pohranu podataka. Rješenje povezuje NetAppSnapshot i SnapRestore tehnologije. Za povećanja performansi instaliran je NetApp Flash Cache, inteligentni sustav koji ubrzava boot-up proces, ubrzava performanse samih aplikacija i omogućava brži pristup bogatom sadržaju platforme te zbog toga drastično povećava korisničko iskustvo. Sigurnost podataka počiva na NetApp FA6000 sustavu za pohranu podataka, zaštite na mrežnom nivou, layer 7 vatrozidu, integriranoj antivirusnoj i antispam zaštiti, SMS verifikaciji, višestrukim mehanizmima na svim aplikacijskim nivoima te integrirani backup podataka. Za optimizaciju performansi koristi se više tehnologija: NetApp Flash Cache uparen sa "Golden Image-om" koji ubrzavaju boot-up process, blade šasije visoke propusnosti osiguravaju da neće nastati nikakvo "usko grlo" u svakodnevnom radu te je omogućena brza migracija samih servera u slučaju hardverskog zastoja. Integrirani su napredni load balanceri za optimalno korištenje servera i optimizaciju TCP prometa unutar samog proizvoda. Platforma nudi visoko kvalitetno korisničko iskustvo neovisno da li je na krajnjem uređaju Windows, Apple, Android ili Linux operativni sustav [3].



Slika 3. DaaS značajno skraćuje vrijeme implementacije i prelaska na novi sustav

C. Smještaj i upravljanje mrežnim poslužiteljima

Sveučilišni računski centar omogućava fizičko, električko i komunikacijsko udomljavanje mrežnih poslužitelja u svojoj sistem dvorani. Od dodatnih usluga pružaju se usluge vezane za upravljanje mrežnim poslužiteljima kao što je instalacija operacijskih sustava i pomoćnih programa, sistemsko održavanje operacijskih sustava, nadzor računalnih resursa i backup [4].

Ustanove koje žele izbjeći kapitalna ulaganja po pitanju tehničke i sigurnosne infrastrukture, a pri tome žele koristiti svoju opremu, odnosno tvrtke koje imaju postojeću serversku infrastrukturu smještenu u neadekvatnim prostorima, mogu koristiti uslugu kolokacije koja omogućuje smještaj korisnikove opreme u primjerenom tehničkom prostoru ICT pružatelja usluga (Data centru). Kolokacija je posebno zanimljiva financijskim ustanovama koje žele implementirati Disaster Recovery Site, a žele izbjeći kapitalna ulaganja i koristiti svoju opremu.

Sa stanovišta sigurnosti poželjno je da Data centar podržava sljedeće funkcionalnosti i standarde:

- Redundantna klimatizacija
- Objekt otporan na velika opterećenja i utjecaj voda
- Neprekidno napajanje (UPS, generatori)
- Fizička zaštita (kontrola pristupa, upozoravajući sustavi)
- Zaštita od požara (vatrodojavni sustav i automatsko gašenje požara plinom)
- Nadzor 24/7 (napajanja, klimatizacije, dostupnosti)
- Fizički i konzolni pristup kolociranoj opremi
- TIER 3 (raspoloživost 99,982%; 1,6 sati dozvoljenih nenajavljenih prekida godišnje; redundancija (N+1))

Cloud Storage usluga omogućuje korisnicima zakup diskovnih prostora, koji su smješteni u HT Data centru, te koji su dostupni putem Interneta. Uslugom se omogućava jednostavno upravljanje i administracija zakupljenog diskovnog prostora (upravljanje resursima, pravima i brojem korisnika). Prednosti za korisnike su prilagodba troškova trenutnim potrebama, korištenje diskovne infrastrukture bez kapitalnih investicija, potpuna kontrola diskovnog prostora i sadržaja putem korisničkog portala, brže i učinkovitije upravljanje resursima nadogradnje prema stvarnim potrebama [3].

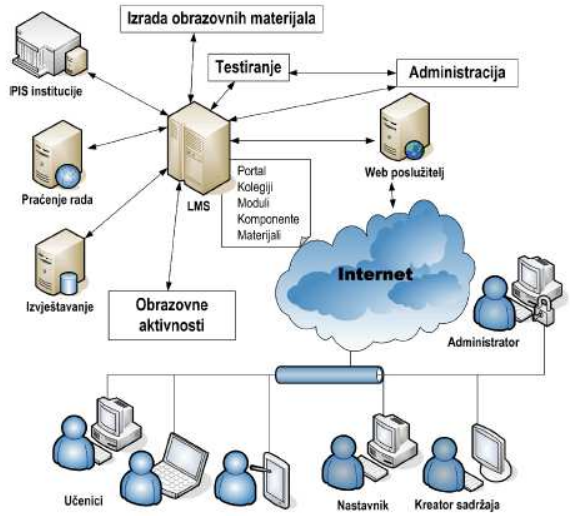
Slika 4. Cloud Storage u konekciji sa ostalim Cloud uslugama

IV. SUSTAVI E-UČENJA

Ne postoji standardna ili prihvaćena definicija što čini jedan e-kolegij. Općenito možemo reći da su to kolegiji koji koriste mogućnosti novih tehnologija za poboljšanje kvalitete izvedbe nastavnog programa i procesa učenja. Razlikujemo više oblika ili razina e-kolegija, od onih jednostavnih koji imaju samo nastavne materijale i informacije o kolegiju dostupne na lnternetu, pa do kompleksnih koji su osmišljeni upravo za internetsko, interaktivno i suradničko okruženje, te spajaju nekoliko obrazovnih sastavnica u jedinstven studijski kolegij ili program. Primjer sustava za e-učenje „Merlin“ (srce) omogućava ustanovama iz visokoga obrazovanja i znanosti, ali i pojedincima iz akademske zajednice da u srcu aktiviraju svoje e-kolegije i koriste zajedničke resurse za potrebe e-učenja. Pri tome srce osigurava i preuzima brigu o tehničkoj pouzdanosti i stalnoj i neprekidnoj dostupnosti toga servisa, a korisnici koriste sustav, brinu se i odgovaraju za sadržaj i podatke, kao i za administriranje podataka o kolegijima i njihovim polaznicima.

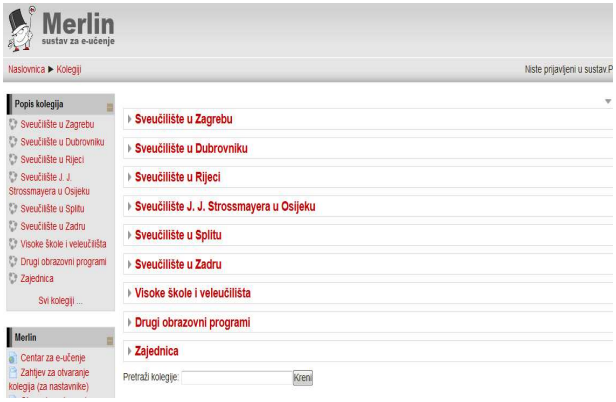
Sustav za e-učenje Merlin bazira se na sustavu otvorenog koda Moodle koji se kontinuirano razvija i prilagođava potrebama korisnika. Sustav održavaju djelatnici centra za e-učenje, a omogućuje ustanovama, nastavnicima i studentima pouzdanu platformu za e-učenje. Korisnicima sustava dostupni su razni standardi kao i dodatni moduli sustava Moodle koji su nadograđeni i prilagođeni za potrebe korisnika. Sustav Merlin je povezan sa e-portfolio sustavom, čineći virtualno okruženje za e-učenje. Također, sustav Merlin je povezan i sa sustavom ISVU čime je omogućeno otvaranje kolegija te upis nastavnika i studenata iz sustava ISVU na sustav Merlin. Pored postavljenih modula u sustavu Merlin, nastavnici mogu zatražiti i uključivanje drugih modula sustava Moodle.

LMS je krovna aplikacija u modelu e-učenja. Najčešće je to Internet aplikacija smještena na određenom Web serveru kojoj mogu pristupiti različiti korisnici preko nekog od mnogobrojnih Web pretraživača. LMS vrši konstantnu evaluaciju percepcije polaznika o efektivnom obrazovanju i učenju različitih procesa te kontrole voditelja tečaja. Glavni cilj LMS-a je upotreba dostupnih ljudskih i tehnoloških resursa kako bi se omogućila najveća kvaliteta, fleksibilnost, adekvatan način učenja te najbolje okruženje za voditelje tečaja, ali i polaznike.



Slika 5. Primjena LMS modela

Moodle - Open Source Course Management System (CMS), također poznat i kao Learning Management System (LMS) ili čak Virtual Learning Environment (VLE). Izuzetno je popularan među brojnim predavačima diljem svijeta kao alat za kreiranje dinamičnog web sadržaja. Vrlo je popularan budući da se može instalirati na bilo koje cloud rješenje ili na jednom od vlastitih računala ili kod neke od web hosting tvrtki. Moodle je vrlo primamljiv za uvođenje u pojedinu poslovnu organizaciju jer ima odlike koje dopuštaju objavu velike količine podataka, a istodobno ga mogu koristiti na stotine ili čak tisuće polaznika. Sustav sadržava forum, omogućuje stvaranje različitih baza podataka, kvizova, zadataka, resursa i sl. Moodle je programski paket koji nam nudi uređivanje i upravljanje tečajevima preko mreže. Alat Moodle nam nudi mogućnost postavljanja virtualne učionice, tj. e-učionice. U virtualnim učionicama učitelji mogu dodavati gradivo, najaviti različite aktivnosti, testove, provjeravati znanje, komunicirati s ostalim sudionicima. Studenti, učenici mogu gledati gradivo, prenositi ga na svoje računalo, rješavati testove, pisati zadaće, komunicirati s nastavnim osobljem i drugim studentima. Ideja Moodlea temelji se na tome da učenici najviše nauče, ako nastavni materijal aktivno koriste. Znači, ne samo učenjem gradiva, već i aktivnim raspravljanjem o njemu te izrađivanjem novog. nastavnog materijala u suradnji sa drugim studentima i nastavnicima [4], [5].



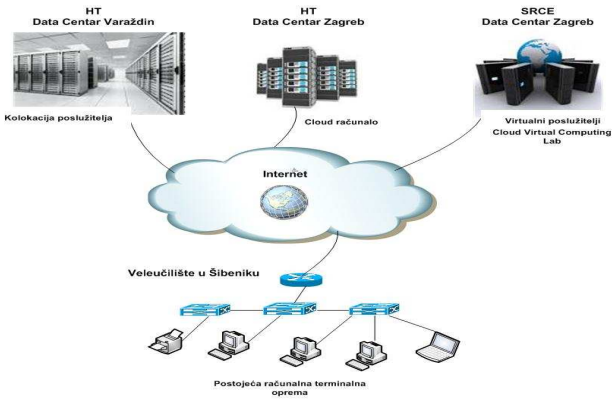
Slika 6. Sustav za e-učenje “Merlin”

Jedna od najvećih koristi e-učenja pomoću cloud rješenja za sve djelatnike i studente su nove mogućnosti u radu. Jednostavna razmjena podataka i zajednički rad na dokumentima povećava fleksibilnost i brže izvršavanje zadataka. Pristup podacima sa bilo koje lokacije olakšavaju rad od pripreme predavanja za stručno osoblje, te odradu vježbi i projekata za studente. Imaju pristup svojim datotekama neovisno o mjestu gdje se nalaze. Rad na projektima je olakšan jer zahtjeva mnogo zaposlenika, odjela i lokacija. Projekti zahtijevaju puno individualnih zadataka i obaveza koji su u međusobno zavisnoj vezi. Sudionici projekta se mogu uključiti sa bilo koje lokacije kako bi upravljali rasporedom, zadacima i informacijama. Promjene se vide u realnom vremenu, što ubrzava proces i lakše se držati rokova na projektu. Svakom djelatniku i studentu omogućen je pristup edukativnim alatima bilo gdje i bilo kada, te će na vježbama moći koristiti i vlastita prijenosna računala. Studentima dostupnost osigurava i moguća podrška sustava za tablete, te druge prijenosne uređaje. Tehnologija virtualnog računala će nam omogućiti veliko poboljšanje nastavnog procesa kroz održavanje interaktivne nastave korištenjem prijenosnika i tableta, te mogućnost izvođenja nastave na daljinu.

Također, mogućom primjenom virtualnih računala pristupamo svom desktopu preko Interneta, pritom sve izgleda kao da radimo na svom računalu, no operativni sustav, aplikacije i podaci nisu smješteni lokalno “na disku”, već na udaljenom serveru gdje su sigurni. Upravljanje virtualnim računalima i aplikacijama s jednog mjesta, produljavamo životni vijek računala, jer usluga cloud računalo ne ovisi o sklopovskoj snazi vašeg računala. Ukoliko posjedujete određenu vlastitu poslužiteljsku infrastrukturu koja je za sada smještena u prihvatljivoj okolini, ali uskoro će trebati raditi povećavanje kapaciteta poslužiteljske infrastrukture. Time bi se smanjila raspoloživost trenutnih prostora i

poslovanje i izvođenje nastave bi moglo biti ugroženo. Korištenjem kolokacije opreme u Data centru Hrvatskog Telekoma bilo bi nam omogućeno da umjesto ulaganja u uređenje vlastitog Data centra za koji trebamo osigurati prostor, napajanje, hlađenje, u potpunosti osiguramo svoje poslovanje kolokacijom u HT Data centar. Na ovaj način elimirali bi sve moguće probleme uzrokovane varijacijama napona zbog grmljavine te potrebe za hlađenjem poslužitelja što je izraženo naročito u ljetnim mjesecima. Tada bi postojeća poslužiteljska infrastruktura bila premještena u neki Data centar(npr.HT data centar u Varaždinu). Time bi se oslobodio prostor ustanove, koji bi se preuredio za dodatne učionice/pomoćne prostorije. Istovremeno korištenjem virtualnih poslužitelja u SRCU imali bi infrastrukturu smještenu na dvije nezavisne lokacije i na različitim tektonskim pločama (Varaždin i Zagreb) čime bi dobili maksimalnu sigurnost podataka, naročito u slučaju prirodne nepogode/katastrofe kao što je npr. potres.

Korištenjem kolokacije vlastitog poslužitelja i Cloud virtualnog poslužitelja od SRCA zadovoljene su trajne potrebe za poslužiteljskim kapacitetima i to s razinom podrške standardnom za moderni podatkovni centar (visoka raspoloživost, pouzdanost, skalabilnost, fleksibilnost upravljanja) uz eliminiranje potrebe investiranja u vlastiti poslužitelj. Također primjenom Cloud Virtual Computing Lab povećavamo kvalitetu i efikasnost izvedbe same nastave [7].



Slika 7. Primjena Cloud rješenja u Veleučilištu Šibenik

Korištenjem navedenih cloud rješenja SRCA i Hrvatskog Telekoma, ustanove ostvaruje velike uštede i sigurnost podataka, te maksimalnu iskoristivost vlastite postojeće infrastrukture na zadovoljstvo studenata i stručnog osoblja.

V. ZAKLJUČAK

Razvoj i napredak društva nameću nužnu potrebu da se obrazovni proces usmjeri ne samo na prijenos znanja, nego i na istraživanja. ICT daje svoj doprinos u svakodnevnoj organizaciji obrazovnog procesa, u vremenskoj fleksibilnosti te neograničenoj dostupnosti obrazovnih sadržaja. Primjenom cloud tehnologije, omogućava se kvalitetna i jednostavna komunikacija nastavnika i studenata u obrazovnom procesu. Najrazličitiji nastavni materijali polaznicima mogu biti dostupni uvijek, kao i pristup s bilo kojeg dislociranog mjesta. Organizacija nastave zasniva se na aktivnoj suradnji između nastavnika i polaznika, te je prilagođena potrebama polaznika s jasno postavljnim obrazovnim ciljevima. Naime, budući da postavljanje web-sustava za e-obrazovanje zahtijeva znatno uloženo vrijeme te računalne i druge resurse, potrebno je u fazi pripreme napraviti procjenu ciljeva i željenih učinaka on-line obrazovnog sustava. Ukoliko se većina nastave izvodi na klasičan način u predavaonici ili laboratoriju, a korištenje materijala na webu je samo dopuna takvoj nastavi, u pravilu nije potrebno kreirati kompleksan sustav za on-line edukaciju jer se većina obrazovnih sadržaja dobiva tijekom klasične nastave i iz tiskane literature, a savjetovanje s predavačem moguće je u osobnom kontaktu, tj. licem u lice. Međutim, ako polaznici najvećim dijelom pristupaju obrazovnim sadržajima on-line te su vrlo rijetko u neposrednom kontaktu licem u lice s nastavnikom, obično je potrebno uložiti vrlo mnogo truda i sredstava u oblikovanje sustava koji će biti jednostavan za korištenje i barem donekle ispuniti potrebu korisnika za socijalnim kontaktom s predavačem i drugim polaznicima. Primjena određenih metoda u e-obrazovanju u vrlo velikoj mjeri ovisi o raspoloživoj tehnologiji. Polaznici mogu koristiti brze širokopojasne internetske veze, a predavač može snimiti predavanja te ima na raspolaganju poslužitelj odgovarajućeg memorijskog kapaciteta i brzine pristupa Internetu. Obrazovni je sadržaje moguće pohraniti i u obliku videozapisa predavanja. Ako korisnici pristupaju obrazovnim sadržajima putem sporijih veza ili ukoliko server kojim raspolaže predavač ima slab kapacitet veze prema Internetu, tada je vjerojatno najbolje sadržaj predavanj pripremiti u obliku pisanog teksta i statičnih vizualnih ilustracija koje ne zauzimaju mnogo memorijskog prostora. Primjenom rješenja javnog clouda i e-učenja omogućava se ustanovama fokusiranje na njihovu osnovnu djelatnost (stručno i znanstveno obrazovanje studenata). Cloud rješenja se svakodnevno sve više i više koriste u osobne, edukacijske i poslovne svrhe. Primjenom cloud rješenja korisnicima se omogućava pristup putem niza web usluga koje upravljaju računalnim resursima kao što su na primjer: računala, mreža, sustav za pohranu podataka, operativni sustav, okružje za razvoj aplikacija, aplikacijski programi. Implementacijom e-učenja pomoću cloud rješenja bez većih ulaganja, te uz postizanje maksimalne sigurnosti podataka na adekvatan način povezati ćemo postojeću poslužiteljsku infrastrukturu (postojeće

poslužitelje koji se planiraju premjestiti u Data centar) i nove cloud servise koji će omogućiti napredna rješenja e-učenja, te fleksibilnije održavanje nastave i vježbi (virtualna računala, virtualni poslužitelji i virtualni laboratorij). Fleksibilnost održavanja nastave i vježbi ogleda se u tome što je pristup podacima omogućen s bilo koje platforme i u bilo koje vrijeme.

LITERATURA

[1] R.Buyya, J. Broberg, A. Goscinski, Cloud Computing: Principles and Paradigms, John Willey and Sons Inc, 2011

[2] B. Furht, A. Escalante, Handbook of Cloud Computing, Springer Science + Business Media, 2010

[3] Interni materijali Hrvatski Telekom d.d.

[4] Web stranice www.srce.unizg.hr

[5] Web stranice www.carnet.hr

[6] George Reese, Cloud Aplication Architectures, O'Reilly, 2009

[7] SoftCOM 2014: Primjena cloud rješenja u Veleučilištu Šibenik

Oblikovanje trodimenzionalnih objekata u sustavima virtualne stvarnosti prilagođenih učenju

Viktorija Filipašić, Mihael Kukec i Nikola Poljanec
Međimursko Veleučilište u Čakovcu,
Bana Josipa Jelačića 22a, 40 000 Čakovec, Hrvatska
viktorija.filipasic@mev.hr, mihael.kukec@mev.hr, nikola.poljanec@mev.hr

Primjena virtualne (VR) i proširene stvarnosti (AR) postaje sve popularnija u mnogim granama znanosti i obrazovanja. Štoviše primjena virtualne stvarnosti u edukacijske svrhe koristi se već dulje vrijeme. Mnoga istraživanja pokazuju kako unatoč evidentnim nedostacima primjena virtualne stvarnosti povoljno utječe na rezultate učenja. Tome pridonosi velika sloboda bez fizičkih ograničenja stvarnog svijeta što djeluje motivirajuće u odnosu na klasične metode učenja. Potrebno je istaknuti važnost i složenost izrade sadržaja za sustave virtualne stvarnosti prilagođene učenju. Među oblicima sadržaja koji čine virtualnu stvarnost suvremeni trend su trodimenzionalni modeli koji stvaraju doživljaj tzv. uranjanja u virtualni svijet. U radu se navode i komentiraju rezultati postojećih istraživanja sustava učenja pomoću virtualne stvarnosti s osvrtom na njihove nedostatke, te će se nakon toga prikazati postupak stvaranja virtualnog trodimenzionalnog sadržaja od početka nastajanja pa do njegove realizacije. Tijek izrade virtualnog objekta započinje procesom modeliranja, koje je moguće provesti s nekoliko različitih tehnika ili njihovom kombinacijom. Ukoliko se radi o statičnom modelu, poput objekta ili stvari, tada je završna faza izrade modela teksturiranje ili bojanje objekta te priprema za animiranje koja se izvršava izradom skeleta.

I. UVOD

Učenje je sposobnost bez koje nema napretka. Ono počinje od vrlo ranog djetinjstva, rani stadiji učenja počinju kroz igru. Takvo je učenje najbrže, najzanimljivije i najefikasnije [1].

Nove metode i alati za učenje evoluiraju i olakšavaju učenje. Time se ne dovodi u pitanje efikasnost tradicionalnih metoda učenja, međutim često su nove metode učenja zanimljivije od dobro poznatih [1,2]. Novi pristup učenju kroz virtualnu stvarnost potiče zainteresiranost, motiviranost i mnoge druge pozitivne učinke.

Tehnologija računalne grafike napreduje velikom brzinom, te omogućuje stvaranje sadržaja gotovo bezgraničnih mogućnosti. Mnogo takvog sadržaja može poslužiti za unapređenje procesa učenja [3]. Mnogo je praktičnih primjera korištenja tehnologija VR i AR za učenje u širem smislu. Primjerice, primjena u vojne svrhe gdje se pomoću vrlo preciznih simulacija leta trenira pilote u bezopasnom virtualnom okruženju i olakšava upravljanje letjelicom pomoću kacige sa proširenom stvarnošću [7]. Veliki uspjeh i pozitivne reakcije virtualna stvarnost postiže u umjetnosti i prikazu kulturološkog naslijeđa u smislu virtualne restauracije i rekonstrukcije umjetničkih djela bez fizičkog utjecaja na original. Nadalje, virtualna stvarnost omogućuje osobama s poteškoćama da nadomjeste onemogućena osjetila. „The Museum of Pure Form“ je projekt koji daje mogućnost slijepima da opipom dožive izgled određene statue ili umjetničkog djela [6]. Također, AR se koristi u medicini za vizualizaciju i treniranje operativnih postupaka, te za treniranje manufakturne proizvodnje i popravak uređaja i strojeva. Najčešće je korištenje AR i VR namijenjeno zabavi [7] a postoje razni projekti koji ujedinjuju AR ili VR sa učenjem. Efikasnost učenja astronomije u osnovnoj školi pomoću virtualnog modela naprama fizičkog modela testirali su S. Fleck i G. Sinom [8]. Isto tako je tim u Španjolskoj istražio kako se VR može upotrijebiti u stvaranju trodimenzionalne umjetnosti [9]. Osim samostalnog učenja unutar virtualnog okoliša, poprilično interesantna okolina za učenje ostvarila bi se mogućnošću suradnje unutar virtualne stvarnosti. Takav aspekt učenju istražili su R. L. Jackson i E. Fagan u svom radu [10]. Virtualna stvarnost otvara nove mogućnosti te novi pogled na svijet i njegovu percepciju.

Integracija virtualne stvarnosti u edukacijske primjene ima pozitivnih i negativnih konotacija. Prema pretpostavkama, nova metoda učenja djeluje pozitivno na rezultate učenja, tako prednost primjene tehnologije VR može poboljšati performanse učenja i konceptualnog razumijevanja pri određenim zadacima. Mnoge činjenice koje učimo teško je smisleno razumjeti, virtualna stvarnost omogućuje vizualizaciju koja olakšava poimanje i shvaćanje. Međutim, razumijevanje načina i modaliteta primjene tehnologije VR u smislu unaprjeđivanja ishoda učenja je za sada razmjerno ograničeno [4].

Upotrebom kacige (engl. Head gear) s audio-vizualnim zaslonom, senzorima za pozicioniranje i uređajem s taktilnim sučeljem moguće je potpuno uranjanje (engl. immersion) u virtualni svijet u kojem možemo vidjeti, čuti, pa čak i stvarati, izmijeniti i manipulirati virtualnim objektom te tako učiti o njemu bez zapreka fizičkog svijeta [3].

Problematika VR-a u edukaciji je odvraćanje pozornosti. Bogati sadržaj i mogućnost manipuliranja potiču na ponekad pretjerano istraživanje unutar sustava, te tako sudionik bude preplavljen mogućnostima, informacijama i sadržajem te lako izgubi primarni cilj [5].

Također, da bi se izbjegle negative konotacije potrebno je razriješiti barem tri primarna problema prije integracije sustava virtualne stvarnosti u edukacijske svrhe: (i) cijena, (ii) funkcionalnost, (iii) strah od tehnologije [3].

Troškovi za nabavku opreme potrebne za komercijalni virtualni sustav dovoljno sofisticirane za složene modele i zahtjevnije operacije relativno su viši od troškova nabavke prosječnog računala za osobnu upotrebu. Međutim, razvojem tehnologije računalna snaga postaje te sklopovska oprema visoke kvalitete i performansi postaje sve dostupnija te se ujedno se razvijaju pristupačniji i efikasniji sustav virtualne stvarnosti.

Slijedeća prepreka je funkcionalnost. U stvarnom svijetu zakone fizike ubrajamo u svojstvene elemente prirode. Gravitaciju, svjetlosti, Newtonove aksiome podrazumijevamo kao normalnu pojavu, no u virtualnom svijetu je potrebno sve te pojave uzeti u obzir i simulirati kako bi se ostvario što bolji dojam uranjanja u virtualni svijet. Dobro oblikovanje je temelj funkcionalnosti takvog sustava, oblikovanje mora slijediti mnoge zahtjeve. Primjerice, u učionici gdje je mnogo učenika potreban je veliki broj senzora i ostale opreme, podešavanje velikog broja parametara, te značajna ulaganja i promišljanje o načinu primjene kako bi se takav sustav polučio pozitivne rezultate.

Posljednji izazov za prihvaćanje VR i AR tehnologije za širu upotrebu, pa tako i u školstvu, je strah od računalne tehnologije. Brenda Laurel istražila je strah od računalne tehnologije i identificirala njegove komponente [3]:

- tabu imitacije boga,
- strah od pogrešivosti: inteligentni softverski entiteti će možda postati bezosjećajni i beživotni stvorovi koji posjeduju računalnu tehnologiju, započinju rat i uništavaju okoliš,
- strah od gubljenja identiteta: nastanak ovisnosti o tehnologiji, robovi kibernetske simbioze,
- strah gubljenja kontrole nad stranim entitetom: softverski entiteti će možda razviti osjećajnost te prevladati.

Nadalje, kao produkt straha računalne tehnologije pojavili su se strahovi od VR kao što je [3]:

- strah od gubljenja kontrole nad drugima: kako možemo znati tko stoji iza virtualne osobe, kako će se afirmirati prava pojedinca, kako će se rješavati konflikti, te tko će vladati,
- strah od odbijanja pristupa: kada se kreira VR bez obzira na njezinu izvedbu i oblik, tko će imati pravo biti u njoj i kako će ući u nju,
- strah od konfuzije: što ako izgubimo osjećaj razlike između virtualnog i stvarnog svijeta,
- strah od napuštenosti: što ako se većini društva svidi virtualna svijet i ono zanemari stvarnost. Hoće li pojedinci koji će ostati u stvarnosti biti zanemareni i odbačeni.

U konačnici realna zabrinutost s kojom se sučeljavaju nastavnici je nedostatak istraživanja koja jasno ukazuju na prednosti primjene ove tehnologije te jasne pedagoške paradigme i načine njene primjene u procesu podučavanja.

II. OBLIKOVANJE TRODIMENZIONALNOG OBJEKTA

Ukoliko bi se razriješio barem jedan dio problematike implementacije sustava virtualne stvarnosti u primjenama pogodnim za učenje, povećala bi se učinkovitost podučavanja. Jedna od tri primarne prepreke je strah od tehnologije, strah koji proizlazi iz nedostatka znanja i informacija te posljedičnog nerazumijevanja tehnologije. S ciljem povećavanja razine znanja o procesu stvaranja sadržaja za sustave virtualne stvarnosti, u ovom radu je prikazan proces nastanka sadržaja za virtualni sustav, točnije proces oblikovanja trodimenzionalnog objekta. Programski alat koji je pri tome korišten je Blender, verzije 2.73a.

A. Modeliranje

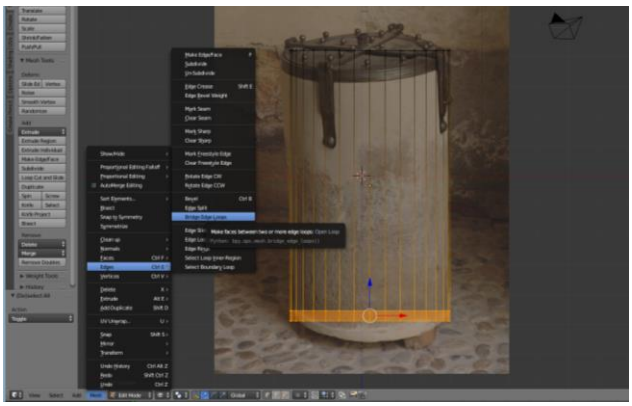
Stvaranje objekta započinje korakom modeliranja. U glavnom prozoru u kojem će se modelirati pogled definiran se tip pogleda na „Front Ortho“, odnosno na prednji, ortogonalni pogled na objekt.

Radi lakšeg modeliranja potrebno je umetnuti sliku objekta koji se modelira u pozadinu tako da se kursor namjesti na glavni prozor za modeliranje i pritisne tipka „N“ koja otvora prozor s detaljima i opcijama. U tom prozoru potrebno je omogućiti pozadinsku sliku pritiskom na ikonu pokraj „Background Images“ kako bi se aktivirala slika te u istom prozoru odabrati datoteku slike pomoću „Add Image“. Otvorit će se dodatni prozor u kojem se odabire slika pomoću naredbe „Open“. Na pozadini prozora s pogledom pojavit će se odabrana slika. Ukoliko se promijeni pogled iz primjerice „Front Ortho“ u „Top Ortho“ ili „User Persp“ ili bilo koji drugi, slika se neće prikazivati. Da bi se pozadinska slika prikazivala u svim pogledima potrebno je otvoriti prozor postavki (prečac: tipka N), te u postavkama pozadinske slike za opciju „Axis“ odabrati „All Views“.

Objekt koji je odabran za izradu ovog primjera je kamena kasa iz razdoblja stolovanja Nikole Zrinskog Sigetskog (16. stoljeće), stajala je na mostu te je služila za uplatu takse prilikom ulaska u grad Čakovec (slika 8).

Izrada kamene kase počinje dodavanjem jednostavnog valjka pomoću „Add-Mesh-Cylinder“ (prečac: Shift + A) koji se preoblikuje naredbama „transformation manipulators“ koje uključuju translaciju (g), rotaciju (r) i skaliranje (s). Kako bi ga se detaljnije preoblikovalo, potrebno je promijeniti način rada u uređivanje, odnosno „Edit mode“ (prečac: Tab). Potrebno je skalirati valjak po z osi tipkama „S“ pa „Z“ te pomicanjem pokazivača miša.

Način rada za uređivanje, tzv. „Edit mode“ pruža mogućnost oblikovanja objekta sa točkama „vertices“, bridovima „edges“ i plohama „faces“. Nakon što valjak prebacimo u „Edit mode“ u kojem možemo manipulirati točkama, bridovima i plohama, potrebno je dodati još bridova kako bi napravili prsten na dnu objekta. Kako bi dodali bridove koji se protežu oko cijelog objekta, potrebno je odabrati funkciju „Edge Loop“ kao što je



Slika 1. Dijeljenje objekta funkcijom “Edge loop”

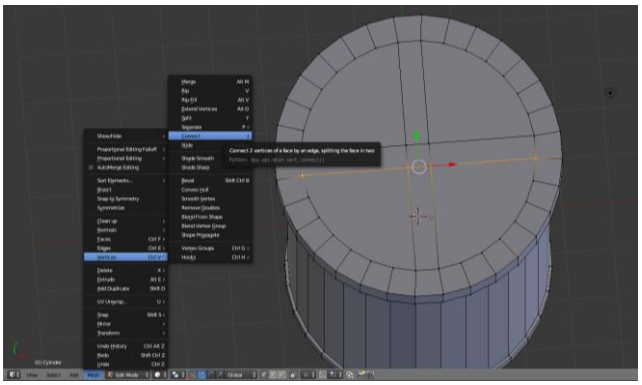
prikazano na slici 1.

Osim funkcija translacije, rotacije i skaliranja postoji još jedna osnovna funkcija naziva „Extrude“. Ta funkcija omogućava izvlačenje dijelova mreže površine.

Nakon što se izradi „Edge loop“ i pomoću funkcije „Edge Slide“ izrađeni element se postavlja na željeno mjesto, odabiru se plohe te se povlače opcijom „Extrude“ a veličina im se prilagođava skaliraju tako da se izradi izbočeni prsten na donjem dijelu objekta. Na jednak način. kombinacijom izvlačenja, skaliranja i translatiranja ploha objekt se udubljuje. Ovim postupcima izrađeno je tijelo kase.

Sljedeći korak je izrada poklopca za kasicu pomoću novog valjka. Skaliranjem je potrebno spljoštiti valjak te kombinacijom funkcija skaliranja, translacije i dodavanja materijala izvlačenjem („Extrude“) tako da valjak prianja na tijelo kase. Kako bi se izradio metalni križ tj. okov na poklopcu kase, najprije je potrebno označiti brid na gornjem dijelu poklopca, te funkcijom „Select – Edge Loops“ (prečac: Alt + Shift + desna tipka miša) označiti sve bridove u krug funkcijom „Extrude“. Potrebno je odabrati nasuprotne točke koje se spajaju funkcijom „Mesh – Vertecies-Connect“ tako da na tvore križ. Kada je na taj način gornja ploha poklopca definirana, odabiru se plohe koje tvore oblik križa te im se treća dimenzija dodaje funkcijom „Extrude“. Na isti način stvaraju se bočne plohe na poklopcu koje se nastavljaju na križ s gornje strane, kako je prikazano na slici 2.

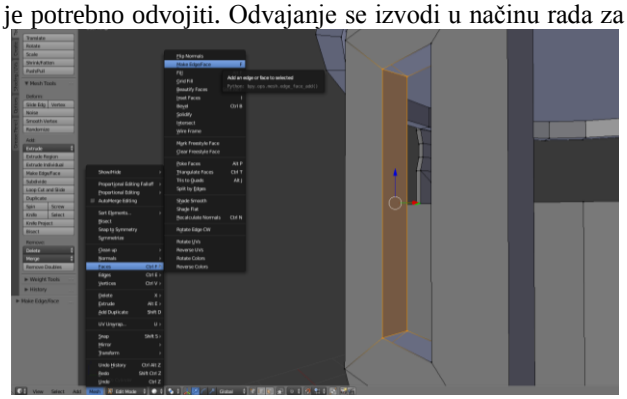
Kako bi se izradile cilindrične spojnice na okovima, dodaju se bridovi funkcijom „Edge Loop“, označavaju se određene plohe koje se tada mogu izbočiti funkcijom „Extrude“. Kako bi se bolje oblikovale spojnice, moguće je ručno namjestiti točke za manipulaciju.



Slika 2. Spajanje točaka funkcijom “Connect”

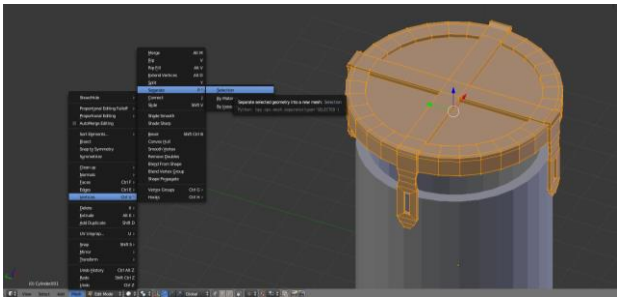
Okovi u obliku križa nastavljaju se niz tijelo kase te čine četiri zatvarača. Oni se izrađuju kao i prethodno izrađeni dio okova, kombinacijom funkcija i alata „Extrude“, translacijom, rotacijom i skaliranjem. Kako bi se izradili otvori, koristi se funkcija „Knife tool“, koja se nalazi u prozoru „Tools“ čiji prečac je „T“. Zbog simetričnosti otvora, prvo je potrebno funkcijom „Extrude“ i skaliranjem unutar plohe izraditi manju plohu te tu manju plohu obrisati. Odabirom najmanje triju točaka i funkcijom „Mesh – Faces – Make Edge/face“ spajaju se točke te nastaje nova ploha. Unutar nastale praznine uslijed brisanja plohe potrebno ih je popuniti tako da se ne vidi šupljina unutar objekta, a ostane otvor kao na slici 3.

Poklopac kase i tijelo kase čine jedan objekt, što onemogućava manipuliranjem svakim posebno. Stoga ih je potrebno odvojiti. Odvajanje se izvodi u načinu rada za



Slika 3. Spajanje točaka funkcijom “Connect”

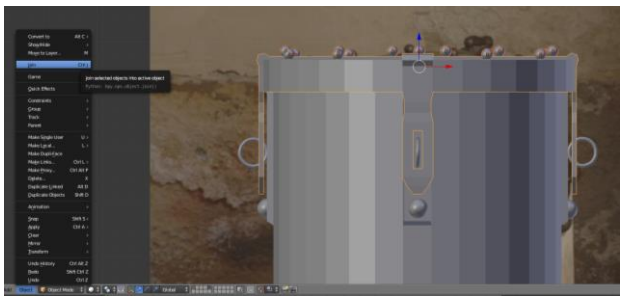
uređivanje, tzv. „Edit mode“ gdje se označe točke poklopca te upotrijebi funkcija „Mesh - Vertices – Separete - Selection“ te na taj način poklopac učini zasebnim objektom. Kako bi se jasno moglo razaznati koji cilindar čini poklopac, a koji tijelo kasice u prozoru „Outlineru“ dvostrukim pritiskom na ime objekta otvaramo opciju preimenovanja. Poklopac se preimenuje iz „Cylinder“ u „Poklopac“, a tijelo kasice u „Tijelo“. Poklopac je potrebno namjestiti na tijelo kasice translacijom i skaliranjem. Potrebno je dodati „Edge Loop“ na tijelo kasice kako bi izradili metalne podloge na koje prianjaju okovi poklopca kao što je prikazano na slici 4.



Slika 4. Odvajanje objekata funkcijom "Separate"

Zakovice i otvor za novčiće na kasici moguće je izraditi na jednostavan način. Potrebno je dodati sferu „Add – Mesh – UV Sphere" i namjestiti je na željeno mjesto. Zakovice na tijelu kasice izrađuju se tako da se u načinu rada uređivanja označi pola sfere i obrišu odabrane točke pomoću tipke „Delete" i „Delete Vertices".

Prsten za učvršćivanje okova izrađuje se dodavanjem jednostavnog oblika, primitiva torusa „Add – Mesh-Torus", te se u prozoru „Tools" (prečac: T) mijenja polumjer „Major Radius" i „Minor radius" tako da prsten bude tanji. Mali dijelove poput zakovica i prstena spajaju se sa poklopcem ili tijelom funkcijom "Object – Join" (prečac: Ctrl + J) kao na slici 5.



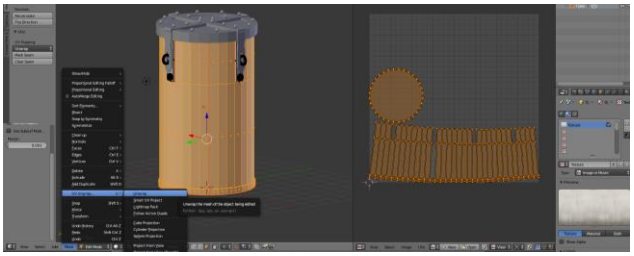
Slika 5. Spajanje detalja na poklopac funkcijom "Join"

Slika 5 pokazuje oštre bridove na kasa što izgled čini razmjerno umjetnim. Kako bi se izgled učinio što realističnijim, potrebno je zaobliti cjelokupan model, jednostavnom funkcijom "Smooth" koja se nalazi u prozoru "Tools" (prečac: T) te je time modeliranje kasice završeno.

B. Teksturiranje

Teksturiranje se u prvom koraku provodi dodjeljivanjem materijala određenoj skupini ploha. U prozoru „Properties – Material" potrebno je odabrati „New material", te podesiti parametre sjaja, disperzije svjetla, boje i slično. Međutim, umjesto boje biti će dodana tekstura, odabirom teksture prljavog, starog kamena ili betona te teksture starog metala. Potrebno je označiti plohe na koju se želi postaviti tekstura betona. U prozoru „Properties - Texture" kod parametra „Type" potrebno je odabrati „Image or Movie" te niže, u prozoru „Image" dodati sliku teksture korištenjem naredbe „Open". Otvara se prozor u kojem je potrebno pronaći putanju mjesta gdje se nalazi željena tekstura.

Kako bi tekstura bila primijenjena na objekt potrebno je izraditi radni prozor povlačenjem crta u gornjem desnom kutu. Tada se u tom prozoru u lijevom donjem kutu mijenja tip iz „3D View" u „UV/Image Editor". Na tijelu kase je potrebno označiti bridove po kojima će se izrezati tj. stvoriti kreirana UV mapa označavanjem bridova koje je potrebno izdvojiti te odabirom opcije „Mesh – Edges - Mark Seam". Bridovi će postati crveni te je nakon toga potrebno „odmotati", odnosno projicirati sve površine trodimenzionalnog objekta na dvodimenzionalnu površinu, odnosno UV mapu što je moguće izvršiti naredbom „Mesh – UV Unwrap -

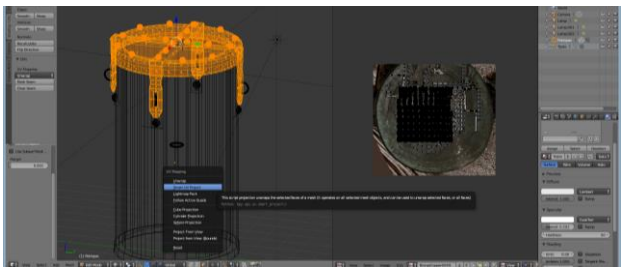


Slika 6. Projeciranje UV mape tijela kasice

Unwrap" kao na slici 6.

U prozoru „UV/Image Editor" odabire se tekstura betona te se namještaju točke objekta koje se približno poklapa s teksturom u UV mapi.

Isti postupak primjenjuje se za poklopac i metalne dijelove kase. Dakle, u „Edit mode-u" označavaju se bridovi po kojima se želi razrezati objekt za UV mapu te ga je potrebno projicirati pomoću „Mesh – UV Unwrap –



Slika 7. Projeciranje UV mape poklopca kasice funkcijom "Smart UV Project" za konju nije potrebno označiti bridove rezanja

Smart UV project" opcije kako je prikazano na slici 7.

U prozoru svojstava potrebno je dodati novi materijal pritiskom na tipku označenu simbolom „plus". Na taj materijal dodaje se nova teksturu i odabire se tekstura metala. Nadalje, u prozoru „UV/Image Editor" otvara se tekstura metala i podešavaju se točke tako da pokrivaju željeni dio teksture na način da na objektu ne bude iskrivljena na teksturi. Time model kase poprima izgled koji je vrlo sličan početnom modelu temeljem kojeg je nastao.

Kako bi se kasa dobro vidjela sa svih strana, potrebno ju je odgovarajuće osvijetliti. Osvjetljenje se podešava prema izboru s ciljem što boljeg isticanja pojedinih detalja, pri čemu je najbitnije uzeti u obzir da jedan izvor svijetlosti nije dovoljan. Razmislimo li o fizici svjetlosti u stvarnosti, ono se raspršuje i reflektira od predmeta tako da sjena nikada nije stopostotan nedostatak svjetla što znači da je s one strane modela koji nije osvjetljen potrebno postaviti dodatan izvor svjetlosti manjeg intenziteta.

Dodavanje izvora svjetlosti vrši se postavljanjem opcije „Object mode" i odabirom „Add – Lamp - Sun".

Postavke izvora svjetlosti podešavaju se u prozoru "Properties". U ovisnosti o izvoru svjetla potrebno je postaviti odgovarajući intenzitet i boju. Izvori svjetla u sjeni postavljaju se na blago plavo svjetlo s malim intenzitetom, a glavni izvor svjetla blago žutu boju i veći intenzitet.

Nadalje, osim svjetlosti potrebno je namjestiti kameru, tj. točku pogleda na objekt. Uz zadovoljavajuće osvjetljen 3D model kase i namještenu kameru, model je spreman za postupak renderiranja tj. postupak izrade slike modela. Ikona alata za renderiranje predstavljena je slikom fotoaparata i nalazi se u prozoru „Properties". Prozor za renderiranje posjeduje razne parametre za renderiranje od kojih su najbitniji „Output", dakle mjesto gdje želimo pohraniti renderiranu datoteku i „Dimensions" gdje postavljamo rezoluciju datoteke. Kada su parametri

Slika 8. Konačna slika modela kasice

namješteni, potrebno je pritisnuti tipku „Render" u istoimenom prozoru čime se stvara slika slična slici 8.

III. ZAKLJUČAK

Virtualna i proširena stvarnost su nove tehnologije smještene u centar mnogih istraživanja njihove primjene. Radi iznimnih mogućnosti koje će te tehnologije otvoriti u skoroj budućnosti, uz konstantno usavršavanje i napredak, nije upitno njihovo integriranje u gotovo sve aspekte društva.

Ideju implementacije tehnologija virtualne i proširene stvarnosti u edukacijske sustave je potrebno temeljito istražiti, štoviše ona se već provodi u obrazovanju u određenim granama znanosti. Mnogo je konkretnih primjera praktične primjene tehnologija virtualne i proširene stvarnosti.

Pitanje na koje je potrebno odgovoriti je koliko je takva tehnologija poželjna, odnosno jeli je moguće i koliko je isplativo integrirati je u široj primjeni u komercijalnije edukacijske sustave. Podijeljena su mišljenja između dobrih i loših konotacija koje nosi takva tehnologija. Kako bi društvo prihvatilo takvu tehnologiju u edukacijske sustave potrebno je razriješiti negativne strane tehnologije kao što su: cijena, korištenje, strah od tehnologije i poticati pozitivne strane kojih je mnogo.

Cijena tehnologije pada kada ona napreduje do određenog stupnja usavršavanja za primjenu u komercijalnije svrhe. Primjena u komercijalne svrhe te široka dostupnost tehnologije VR i AR povećava i širi područja njezine primjene. Širokom primjenom tehnologije umanjuje se njezino nepoznavanje te ona postupno postane prihvaćena što rezultira smanjenjem ili potpunim nestankom straha od njena korištenja.

Ulaganje u tehnologiju virtualne i proširene stvarnosti rezultirat će njenim eksponencijalnom rastom u razvitku i korištenju što jedino može rezultirati napredovanjem edukacijskih sustava te izravno utjecati na napredak i evoluciju ljudskog društva.

Ovim radom je prikazan postupak stvaranja jednostavnog trodimenzionalnog modela kase iz 16. stoljeća. Unatoč razmjerno jednostavnom modelu, potrebno je istaknuti kako je praćenje ovdje iznijetog postupka njegovog stvaranja razmjerno složeno. Mnogo bolji način prikaza ovog postupka je izrada video materijala kojim će postupak biti prikazan a što autori rada uključuju u buduća nastojanja prema stvaranju novih spoznaja o primjeni tehnologija virtualne i proširene stvarnosti u procesu stjecanja znanja.

LITERATURA

[1] M. C. Juan, D. Furió, I. Seguí, N. R. Aiju, and J. Cano, "Lessons Learnt from an Experience with an Augmented Reality iPhone Learning Game," in Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology, 2011, pp. 52:1–52:8.

[2] A. Shirazi and A. H. Behzadan, "Technology-enhanced learning in construction education using mobile context-aware augmented reality visual simulation," in Simulation Conference (WSC), 2013 Winter, 2013, pp. 3074–3085

[3] M. Bricken, "Virtual Reality Learning Environments: Potentials and Challenges," SIGGRAPH Comput. Graph., vol. 25, no. 3, pp. 178–184, 1991.

[4] E. A.-L. Lee and K. W. Wong, "Transactions on Edutainment I," A. El Rhabili, Ed. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 231–241

[5] O. De Troyer, F. Kleinermann, and A. Ewais, "Enhancing Virtual Reality Learning Environments with Adaptivity: Lessons Learned," in Proceedings of the 6th International Conference on HCI in Work and Learning, Life and Leisure: Workgroup Human-computer Interaction and Usability Engineering, 2010, pp. 244–265.

[6] M. Carrozzino, C. Evangelista, A. Scucces, F. Tecchia, G. Tennirelli, and M. Bergamasco, "The Virtual Museum of Sculpture," in Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts, 2008, pp. 100–106.

[7] B. H. Thomas, "A Survey of Visual, Mixed, and Augmented Reality Gaming," Comput. Entertain., vol. 10, no. 1, pp. 3:1–3:33, 201

Izborna ili redovna nastava u Osnovnim školama

Davor Šokac
Osnovna škola "Matije Petra Katančića"
I. L. Ribara 3, 31550 Valpovo, Hrvatska
tel. 031-651-576, fax: 031-652-641 e-mail: davor.sokac@skole.hr

Marijan Antolović
Srednja škola Valpovo
Dr. Franje Tuđmana 2, 31550 Valpovo, Hrvatska
tel. 031-651-577, fax: 031-651-497 e-mail: marijan.antolovic@os.htnet.hr

Sažetak

Rezultati ovog istraživanja na malom broju ispitanika mogu predstavljati indikativne pokazatelje stanja i odnosa između nastave informatike u osnovnim i srednjim školama

Iako je nastava informatike u osnovnim školama prisutna preko 20 godina još uvijek se provodi kao izborna nastava. Dok se u svim gimnazijama (osim malog broja PMG) i većini trogodišnjih i četverogodišnjih strukovnih škola informatika provodi jednu ili dvije godina kao redovni predmet. Nameće se pitanje iz kojih razloga učenici osnovnih škola odabiru informatiku kao izborni predmet i koliko im to pomaže ili ne u daljnjem školovanju kroz srednju školu. Isto tako je pitanje kakvi su izgledi ovako obrazovanog učenika sutra u životu, privredi, Europi i svijetu?

Ako usporedimo izbornu nastavu informatike u osnovnim školama sa redovnom nastavom informatike u srednjim školama dolazimo do saznanja da je se osnovnoškolskom planu i programu izborne nastave informatike svladavaju osnovna znanja i vještine korištenja informacijsko komunikacijske tehnologije, kako za svakodnevne potrebe tako i za potrebe budućih zanimanja. Ali isto tako se i u srednjoškolskom planu i programu redovne nastave informatike svladavaju **osnovna znanja i vještine korištenja informacijsko komunikacijske tehnologije**. Na žalost radi se po planovima (plan i program) starim dvadesetak godina na području gdje se ideje i potrebe mijenjaju i razvijaju iz dana u dan.

Možemo uočiti sličnost između programa ovom grubom usporedbom svrhovitosti i provedbe nastave informatike u osnovnim i srednjim školama

1. Uvod

Kako je danas lako ispitati i anketirati određenu skupinu ispitanika pojavila se želja i potreba za saznavanjem odgovora na prethodno navedeno pitanje: Iz kojih razloga učenici osnovnih i srednjih škola odabiru informatiku kao izborni predmet i koliko im to pomaže ili ne u daljnjem školovanju?

Svjedoci smo da je u osnovnim školama osjetni pad zainteresiranosti učenika za izbornu nastavu općenito, pa tako i za izbornu nastavu informatike. Do toga je dovelo nekoliko činjenica. Svjesni samo da današnji osnovnoškolci, skoro svi, imaju „svoja“ računala. Sami svladavaju dijelove osnova rada s računalima, upotrebe Interneta i internetskih servisa, ali je to znanje, kod velikog djela učenika, na niskoj razini znanja. Nezainteresiranost za svladavanjem veće i svrhovitije količine znanja i vještina predstavlja jedan od problema manje zainteresiranosti za izbor informatike kao izbornog predmeta. Učenici koji ipak odabiru izborni predmet informatike u osnovnoj školi u srednjoškolskom dobu susreću se s drugim problemom: uče većinu znanja i vještina koje su već savladali za vrijeme osnovnog školovanja u izbornoj nastavi informatike – te im stoga informatika postaje dosadna.

Ova saznanja su nas potakla da napravimo istraživanje, bolje rečeno da se približimo mislima i razmišljanjima učenika. Iako je istraživanje rađeno na malom broju učenika i samo u jednom dijelu Lijepe naše može se smatrati indikativnim za probleme koji se mogu uočiti u analizi ankete.

Razlozi odabira ili ne odabira izborne nastave u osnovnoj školi mogu se povezati s nekima od slijedećih razlog.

2. Provedba istraživanja

Istraživanje je zamišljeno tako da se kroz jednu grupu pitanja ispitaju trenutni učenici osnovnih škola, a kroz slična pitanja s pogledom „u natrag“ ispitaju učenici srednjih škola. Usporedbom razmišljanja i stavova o istim problemima iz dva različita životna perioda dobit će se jasniji rezultati o izbornoj nastavi informatike u osnovnoj školi.

Danas se ankete mogu vrlo lako napraviti uz uzimanje vrlo malo slobodnog vremena ispitanika, a također i uz minimalne troškove – online. Najjednostavniji način provedbe ankete je izrada anketnog ispitnog obrasca u digitalnoj formi i postaviti ga na web poslužitelj te u odgovarajuće vrijeme i poveznicu do dokumenta napraviti dostupnu ispitanicima. Nakon anketiranja odabrane populacije učenika sumirati podatke, a onda izraditi zaključke.

3. Izrada anketnih pitanja

Za izradu anketnih pitanja oformili smo tim sastavljen od učitelja informatike osnovnih i srednjih škola, psihologa i pedagoga osnovnih i srednjih škola. Zadatak članova tima je bio da kroz neformalne razgovore sa malim brojem učenika otkrijemo neke od razloga odabira ili ne odabira nastave informatike u osnovnoj školi (npr. Slika 1). Članovi tima pedagoške službe imali su zadatak kontrolirati izradu pitanja na osnovi prikupljenih informacija.

Neki od razloga za odabir izborne nastave informatike u osnovnoj školi:

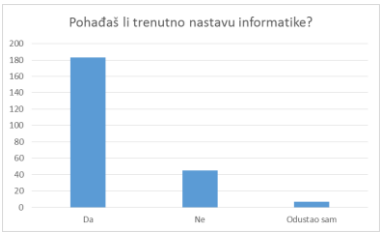
- Izbor izborne nastave informatike zbog velikog interesa rada s računalom,
- Izbor izborne nastave informatike zbog svladavanja određenih znanja i vještina koje ne mogu svladati

sam ili ih je lakše svladati uz vodstvo nastavnika informatike,
- Izbor izborne nastave informatike zbog toga što i druga djeca iz ulice/naselja idu na informatiku,
- Izbor izborne nastave informatike zbog toga što su to htjeli ili preporučili roditelji,
- Izbor izborne nastave informatike zbog povećanja ukupne prosječne ocjene (blaži kriteriji u izbornoj nastavi) te
- Izbor izborne nastave informatike zbog lakšeg svladavanja obaveznog predmeta informatike u srednjoj školi.

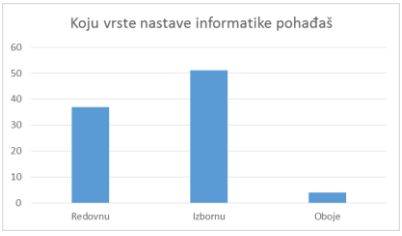
Neki od razloga za ne odabir izborne nastave informatike u osnovnoj ili srednjoj školi:

- Imam svoje računalo i znam se služiti s njime,
- Imam previše obveza i drugih nastavnih i izvan nastavnih obveza te bi to bilo dodatno opterećenje slobodnog vremena,
- Ne zanimaju me druga znanja i vještine, računalo koristim kao komunikacijski uređaj (facebook, mail, youtube, …),
- Ne želim učiti znanja i vještine koje djelomično znam, a koje se ponovno uče u srednjoj školi kroz redovni predmet informatike,
- Ne želim se dovesti u mogućnost smanjenja ukupne prosječne ocjene zbog lošijeg znanja i zalaganja i
- Ne želim učiti znanja i vještine za koje ne znam hoće li će mi u daljnjem školovanju trebati.

Ovo su samo neka od anketnih pitanja s kojima su ispitani osnovnoškolski i srednjoškolski učenici.



Slika 1. Informatika u osnovnoj školi



Slika 2. Informatika u srednjoj školi

Zadatak nastavnika informatike osnovne i srednje škole bio je digitalizirati kreiranu anketu postaviti je na web poslužitelj, omogućiti njeno korištenje odgovarajućim skupinama učenika te prikupljati rezultate ankete. Svi članovi tima zajedno su pregledali rezultate ankete te nakon analize rezultata izvukli zaključke ankete.

4. Rezultati ankete - ?

Trenutni rezultati ankete pokazuju da Informatika kao izborni predmet dodatno ne opterećuje učenike i da im većinom popravlja ukupnu prosječnu ocjenu. Iz srednjoškolske ankete je vidljivo da je gradivo informatike prelagano, ponavlja se, dok jedan dio učenika uviđa da im je informatika u osnovnoj školi trebala kao temelj. Iako je anketa još uvijek u aktivna, i bit će tijekom mjeseca ožujka i travnja 2015. godine, nameće se zaključak da bi ujednačavanju kvalitete nastave informatike u srednjoj školi doprinijelo uvođenje redovne nastave informatike u osnovnim školama.

Anketa će se provoditi tijekom ožujka i travnja 2015. godine u školama Osječko-baranjske županije, te će se tada vidjeti rezultati ankete na većem broju škola i ispitanika.

5. Zaključak

Iako je ova anketa napravljena na malom broju ispitanika i na malom geografskom području može biti indikativna za generaliziranje stavova o statusu nastave informatike u osnovnoj školi.

Iz ankete je vidljivo da je vrijeme za promjenom statusa nastave informatike u osnovnim školama iz izbornog u redovni predmet već odavno došlo. Učenici se iz raznih, opravdanih i neopravdanih razloga ne uključuju u izbornu nastavu informatike. Takav odnos prema informatičkim znanjima i vještinama za današnje i buduće shvaćanje načina života i rada predstavlja dodatnu kočnicu.

Nastava informatike u osnovnim školama trebala bi postati redovnim predmetom, što bi automatski značilo i promjene u planu svladavanja znanja i vještina u nastavi informatike u srednjim školama. Na ovaj bi se način izbjeglo da učenici završavaju osnovu školu na totalno različitim razinama znanja i vještina i informatike. Izbjegla bi se, također i mogućnost da nekim učenicima u nastavi informatike u srednjoj školi dijelovi gradiva budu preteški, a nekima prelagani ili dosadni.

Vrijeme je da nastava informatike u osnovnim školama iz statusa izborne nastave dobije status redovne nastave.

LITERATURA

[1] The Microsoft Office Team, Windows XP Professional Resource Kit, Microsoft Press,A Division of Microsoft Corporation One Microsoft Way, Redmond, Washington 98052-639

E-učenje i europsko vojno obrazovanje

Vlasta Zekulić i Darko Možnik
Ministarstvo obrane Republike Hrvatske
Hrvatsko vojno učilište „Petar Zrinski“
Ilica 256B, Zagreb, Republika Hrvatska
Telefon: 01-3786 840 Fax: 01-3786 466 E-mail: vzekulic@gmail.com
Telefon: 01-3784 613 Fax: 01-3784 210 E-mail: dmoznik@xnet.hr

Sažetak - Rad opisuje i objašnjava teorijski koncept i praktične primjene odnosno primjere e-učenja u europskom vojnom obrazovanju. Navode se procesi i tehnologije e-učenja kao i primjeri primjene e-učenja u vojnom obrazovanju nekih zemalja članica EU.

Opisuje se, inspiriran programom ERASMUS (engl. *European Community Action Scheme for the Mobility of Univer-sity Students*) inicijativa i program *ERASMUS militaire* po-znat po kratici EMILYO (engl. *Exchange of Military Young Officers*) - razmjena kadeta i mladih časnika među vojnim akademijama zemalja članica EU.

Osnovan je u skladu s Europskom sigurnosnom strategi-jom (engl. *European Security Strategy* - ESS) i djeluje kao jedan od modela implementacije europske Zajedničke sigur-nosne i obrambene politike (engl. *Common Security and Defence Policy* - CSDP) s primarnim ciljem promicanja eu-ropske kulture sigurnosti i obrane tijekom inicijalnog vojnog obrazovanja i obuke budućih nacionalnih vojnih elita. Podređen je Europskom sigurnosnom i obrambenom fakultetu (engl. *European Security and Defence College* - ESDC).

Objašnjena su četiri oblika implementacije inicijative i programa EMILYO kao i sadašnje stanje odnosa i integracije MORH-a i OS RH u EMILYO kroz razmjenu kadeta i mladih časnika među vojnim akademijama nekih zemalja članica EU.

Objavljen je i intervju sa zapovjednikom ESDC-a gsp. Hans Bernhard Weisserth-om.

Kroz pogled u budućnost prikazano je mjesto, uloga i značenje e-učenja u transformaciji Hrvatskog vojnog učilišta (HVU) „Petar Zrinski“ u visoko obrazovnu i znanstveno-istraživačku ustanovu u OS RH i MORH-u , tj. HVU kao sastavnicu akademske zajednice u RH i punopravnu aktivnu članicu europskog vojnog obrazovanja.

Na kraju su navedena zaključna razmatranja.

I. UVOD

Zbog višestrukih prednosti koje pruža e-učenje sve se više primjenjuje i u europskom vojnom obrazovanju.

E-učenje je vrsta komunikacijskog kanala, kanala kroz koji se odvija učenje.

ERASMUS *militaire*, poznat po kratici EMILYO (engl. *Exchange of Military Young Officers*) - razmjena kadeta i mladih časnika među vojnim akademijama zemalja člani-ca EU. To je jedan od modela implementacije europske Zajedničke sigurnosne i obrambene politike (CSDP).

Objašnjena su četiri oblika implementacije inicijative i programa EMILYO.

I MORH i HV su se integrirali u navedeni program razmjenom kadeta i mladih časnika.

Za provedbu e-učenja na HVU koriste se LMS sustavi: ILIAS i MOODLE.

Dio programa EMILYO i e-učenje provode se na HVU koji je u fazi transformacije u visoko obrazovnu i znanstveno-istraživačku ustanovu u MORH-u. HVU će biti sastavnica akademske zajednice RH. HVU je punopravna aktivna članica europskog vojnog obrazovanja.

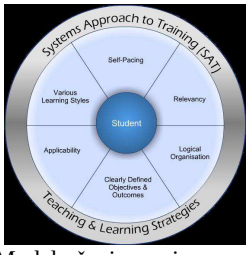
II. KONCEPT, UPRAVLJANJE PROCESIMA I TEHNOLOGIJE E-UČENJA

a) KONCEPT E-UČENJA

Danas je sveprisutna potreba ali i želja da se unaprijedi kvaliteta i dostupnost različitih obrazovnih programa. Pri tome je obrazovanje postalo *cjeloživotnim procesom* i potrebom, pa se briga o dostupnosti i kvaliteti odnosi na sveukupni ciklus *cjeloživotnog obrazovanja*. E-učenje je proces obrazovanja (proces učenja i pouča-vanja) uz uporabu informacijske i komunikacijske tehno-logije, koja pri tome doprinosi unaprjeđenju kvalitete toga procesa i kvalitete ishoda obrazovanja.

E-učenje omogućava: učenje prilagođeno individualnim potrebama studenta (bilo kada, bilo gdje), aktivno sudjelovanje i aktivnu ulogu studenata, aktualnije i ažurnije obrazovne materijale, globalni pristup različitim obrazovnim sadržajima, smanjenje „pritiska“ na fizičke resurse (omogućavajući virtualne učionice, virtualne laboratorije), više komunikacije nastavnik-student i student-student, timski rad, projektni rad nastavnika i studenta, bolju iskorištenost vremena (za kvalitetnije sadržaje) pri kontaktima nastavnik-student, samostalnu provjeru znanja i kontinuirano mjerenje i praćenje napretka studenta [5].

To je učenje i poučavanje koje je usmjereno na studente.



Slika 1. Model učenja usmjeren na studenta

b) UPRAVLJANJE PROCESIMA E-UČENJA

Model upravljanja procesom e-učenja se sastoji od 3 faze: *planiranje, organiziranje i kontroliranje* procesa e-učenja. Proces upravljanja e-učenjem napreduje kroz slijedeće faze: *planiranje* procesa e-učenja (izrada operativnog plana odnosno scenarija e-učenja), *organiziranje* odnosno implementacija procesa e-učenja (uz provedbu scenarija e-učenja u praksi), *kontroliranje* procesa e-učenja (tj. vrednovanje različitih aspekata procesa) i *poboljšanje procesa i platforme*. Osnovni princip ovog procesa je

ponavljanje - jednom kada se potvrde ili odbace početne pretpostavke kroz fazu vrednovanja, ciklus se ponavlja uz primjenu poboljšanja [6].

PLANIRANJE procesa e-učenja je iznimno bitan za cjelokupni uspjeh e-učenja, a najčešće obuhvaća aktivnosti od analize suvremenih trendova u obrazovanju i identificiranje karakteristika polaznika tečajeva za e-učenje; zatim osmišljavanje okoline za učenje koja odgovara sadržaju, ciljanoj publici i tehničkoj platformi te odabira odgovarajuće platforme za e-učenje; pa sve do formuliranja očekivanih ishoda e-učenja i izrade operativnog plana odnosno scenarija e-učenja. Cilj faze planiranja procesa e-učenja je osmišljavanje ishoda i scenarija e-učenja prema određenom kontekstu. Ishodi učenja predstavljaju instrument razvoja plana i programa, te provjere njegove konzistentnosti. Na razini programa preporučeno je definirati 15 do 30 ishoda učenja, na razini predmeta 5 do 10, a na razini nastavne cjeline 5 do 15 ishoda učenja [7].

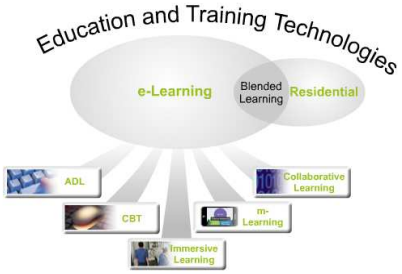
ORGANIZIRANJE osnovnog sadržaja e-učenja
LMS sustavi omogućuju polaznicima da kolegije istražuju nelinearno, a isto tako i nastavnicima omogućuju izradu tečajeva na fleksibilan način, nelinearno. Redoslijed dodavanja materijala i funkcionalnosti u sam kolegij ovisi isključivo o nastavniku. Ako nastavnik i ne zna gdje započe-ti, npr. nije siguran treba li organizirati kolegij oko tjed-nog ili tematskog oblika, treba ipak odabrati jedan oblik i početi s dodavanjem materijala. U obrazovnom okruženju gdje se koriste LMS-ovi nastava se može odvijati i u potpunosti *online*.

KONTROLIRANJE napretka i ponašanja polaznika
Pri kontroliranju je nužno postaviti standarde mjerenja, izmjeriti performanse, usporediti performanse sa standardima te poduzeti potrebne korektivne akcije u svrhu poboljšanja cjelokupnog procesa e-učenja. Svrha faze kontroliranja e-učenja može se najbolje izraziti kao promatranje i procjena cjelokupnog procesa e-učenja i očekivanih rezultata u odnosu na postavljene ciljeve.

c) TEHNOLOGIJE E-UČENJA

Postoji više tehnologija obrazovanja i obuke primjenom e-učenja, a prema NATO-voj podjeli, to su:

1. združeno napredno distribuirano učenje - JADL (engl. *Joint Advanced Distributed Learning*) i napredno distribuirano učenje - ADL (engl. *Advanced Distributed Learning*),
2. učenje pomoću ICT-a odnosno CBT (engl. *Computer-Based Training*),
3. „uronjeno“ učenje (engl. *Immersive learning*),
4. mobilno učenje (engl. *M-learning*),
5. suradničko ili kolaborativno učenje (engl. *Collaborative learning*) i
6. kombinirano ili hibridno učenje (engl. *Blended learning*) [5].



Slika 2. Tehnologije e-učenja prema NATO-u

1. JADL i ADL

a) JADL

Slika 3. Oznaka za JADL

b) ADL

Slika 4. Oznaka za ADL

ADL (engl. *Advanced Distributed Learning*) proizvod je *ADL inicijative* uspostavljene 1997. godine u Ministarstvu obrane SAD-a za standardizaciju i modernizaciju upravljanja „treninzima“ i „edukacijama“ te načinima poučavanja. ADL koristi strukturne i suradničke (engl. *Collaboration*) metode iz industrije, akademskih zajednica i vlada koji razvijaju standarde, alate i sadržaje učenja.

Zemlje članice NATO saveza i EU mogu koristiti izrađene obrazovne on-line tečajeve za različita područja u obrambenim i vojnim sustavima.

Više o JADL i ADL nalazi se na: https://jadl.act.nato.int

2. CBT (engl. *Computer Based Training*)

CBT je obrazovni ili obučni tečaj dostavljen direktno na studentsko računalo putem CD-ROM-a ili nekog drugog uređaja za pohranu. On se razlikuje od ADL-a po tome što se ne isporučuje preko weba ili putem sustava za upravljanje učenjem.

Slika 5. Oznaka ili logo za CBT

3. „URONJENO“ UČENJE (engl. *Immersive learning*)

„Uronjeno“ (engl. *Immersive*) učenje koristi računalno simulirano okruženje koje se događa u stvarnom životu ili hipotetsku situaciju u grafički bogatom i dinamičnom okruženju. Studenti su stavljeni u situaciju i uključeni su u proces učenja kroz interaktivne simulacije i igre temeljene

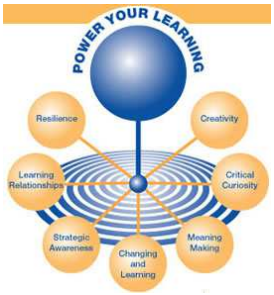
na *game-based* aplikacijama. *Immersive* učenje podupire jednog učenika ili više malih timova da rade zajedno kako bi riješili problem, uvježbali tehnike ili poboljšali svoje vještine.

4. MOBILNO UČENJE (engl. *m-learning*)

Slika 6. Mobilno učenje (engl. *m-learning*)

Mobilno učenje (engl. *m-learning*) je uporaba prijenosnih računalnih uređaja, kao što su različiti iPad-i i drugi tableti, prijenosna računala, osobni digitalni pomoć-nici (PDA) i pametni telefoni povezani s bežičnim mreža-ma. To omogućuje da se podučavanje i učenje proširi i na prostore izvan tradicionalne učionice te daje nastavnicima i polaznicima povećanu fleksibilnost i nove mogućnosti za interakcije koje su prikladne i dostupne virtualno s bilo kojeg mjesta.

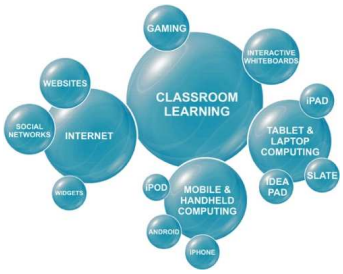
5. SURADNIČKO ILI KOLABORATIVNO UČENJE



Slika 7. Suradničko ili kolaborativno učenje

Suradničko ili kolaborativno učenje (engl. *Collaborative learning*) podupire nastavu kroz interaktivne metode ili online komunikacije koristeći društvene medijske alate kao što su chat sobe, tribine i video alate kao što su webi-nari i web-emitiranja. Ovi alati se mogu implementirani u učionici i online okruženju kako bi podržali tijek nastave. Oni daju mogućnost velikom broju polaznika da rade zajedno s nastavnikom-instruktorom.

6. KOMBINIRANO ILI HIBRIDNO UČENJE (engl. *blended learning*)



Slika 8. Kombinirano ili hibridno učenje

U zemljama članicama NATO saveza i EU koriste se LMS-ovi: ILIAS i MOODLE.

III. EUROPSKO VOJNO OBRAZOVANJE

EMILYO - RAZMJENA KADETA I MLADIH ČASNIKA MEĐU VOJNIM AKADEMIJAMA EU

10. studenog 2008. godine, tijekom francuskog presjedanja EU-om, 27 ministara obrane EU jednoglasno su podržali osnivanje „Europske inicijative za razmjenu mladih časnika“ inspirirana programom ERASMUS (engl. *European Community Action Scheme for the Mobility of University Students). ERASMUS militaire* je ubrzo postao poznat po kratici EMILYO (engl. *Exchange of Military Young Officers*). Od rujna 2009. do rujna 2013. godine EMILYO je omogućio razmjenu 630 kadeta i mladih časnika te znatno ojačao europeizaciju visokog vojnog obrazovanja.

U suvremenom sigurnosnom okružju biti časnik podrazumijeva iznimnu međunarodnu aktivnost, gotovo od prvog dana aktivne službe. Ovaj poziv zahtijeva ne samo razumijevanje kompleksnosti područja operacija i razumijevanje opcija međunarodnog odgovora, nego također poznavanje rada, običaja i standarda drugih nacija i partnera u misijama i operacijama. Dvoje renomiranih europskih vojnih stručnjaka, Hans Wiedner i Herald Gell, navode kako “časnici trebaju kreativnost, prilagodljivost i ostale instrumente vođenja kako bi mogli na pravi način odraditi zadaću… Ako ih moramo učiti vođenju u ekstremnim uvjetima, koja je razlika da li ih to učimo u pustinji, džungli ili na Arktiku? Ako ih moramo učiti radu u međunarodnom okruženju, moramo ih tijekom obuke poslati u inozemstvo. Budući časnici moraju gledati preko horizonta, a profesori i obučavatelji na prvoj razini vojnog obrazovanja moraju ih ohrabriti i učiti tome. Samo tada možemo opravdano reći da smo pružili najbolje moguće obrazovanje našim budući časnicima.“

Ovakve opće težnje razvoja vojne obuke, dodatno su usložene intenziviranjem zahtjeva za stvaranjem prepoznatljivog europskog vojnog identiteta i veće vidljivosti europskog odgovora na međunarodne krize. EMILYO je stoga osnovan u skladu s Europskom sigurnosnom strategijom (*ESS)* i djeluje kao jedan od modela implementacije europske Zajedničke sigurnosne i obrambene politike (*CSDP*) s primarnim ciljem promicanja europske kulture sigurnosti i obrane tijekom inicijalnog vojnog obrazovanja i obuke budućih nacionalnih vojnih elita.

Danas se smatra neophodnim, već na prvoj razini časničkog obrazovanja, upoznati buduće časnike s konceptima interoperabilnosti i to:

- Interoperabilnost njihove nacionalne vojske s europskim načinom sigurnosnog promišljanja i djelovanja,
- Interoperabilnost unutar savezničkih grana, rodova i struka jer kompleksne krize i misije, traže kompleksne odgovore i
- Interoperabilnost s civilnim instrumentima potpore upravljanja krizama u skladu sa suvremenim sveobuhvatnim odgovorom na krize.

Kako bi se ovi višeslojni ciljevi mogli ostvariti, a mladi časnici bili spremni odgovoriti na njih neposredno po stu-

panju na prvu časničku dužnost, potrebno ih je već tijekom njihovog školovanja izložiti radu, uvježbavanju i zapovijedanju u međunarodnom okružju sukladno europskim načelima i standardima. S tim ciljem razvijeni su, akreditirani i provedeni zajednički vojni obučni moduli diljem europskih vojnih akademija.

IV. EMILYO PROGRAMI

EMILYO je načelno podređen Europskom sigurnosnom i obrambenom fakultetu (*ESDC*) te Implementacijskoj grupi (engl. *Implementation Group - IG*) za EMILYO. Ove institucije razvile su način europeizacije visokog vojnog obrazovanja s posebnim naglaskom na poticanje zajedničke kulture sigurnosti i obrane, kroz dvije glavne smjernice razvoja:

1. Obveza obrazovanja i obuke mladih časnika o svim aspektima CSDP-a i
2. Nužnost stavljanja mladih časnika u europsko okruže-nje tijekom raznih faza inicijalne časničke obuke.

Naime, cijela inicijativa je izgrađena na dvije glavne pretpostavke: prva je da mladi časnici sve češće rade u međunarodnom okruženju već u inicijalnim fazama svoje karijere. Ovo se vidi na terenu u mirovnim i humanitar-nim misijama i operacijama, posebno u vojskama zapad-nih zemalja. Druga pretpostavka je da će mladi časnici pu-no bolje raditi na terenu u međunarodnom okružju, ako su već tijekom školovanja bili izloženi radu, obuci i koopera-ciji s pripadnicima drugih nacija. Na taj način se već u najranijoj fazi kulturološki senzibiliziraju, prilagođavaju svoja znanja i vještine primjeni na stranom jeziku, te ubr-zano uče kako ono što je „normalno“ u jednoj državi čla-nici, nije neophodno „normalno“ i u drugoj.

Temeljem ovih smjernica u proteklih pet godina razvijena su četiri oblika implementacije EMILYO inicijative:

1. Zajednički i međunarodno akreditirani akademski moduli u trajanju od tjedan dana. Moduli su razvijeni kao krucijalno važni za sve mlade europske časnike, a to su:

a) „Common module on CSDP“ - jednotjedni ključni zajednički modul koji upoznaje polaznike s konceptom, mehanizmima i izazovima CSDP-a i služi europeizaciji inicijalne časničke obuke. Prvi put je proveden 2009. godine u Vojnoj akademiji Portugala, a danas se provodi u 8 zemalja članica EU za sve ostale države Unije. Iz njega je proizašla i CSDP Olimpijada na kojoj se kadeti i mladi časnici natječu u poznavanju tematike Europske sigurnosne i obrambene politike. Održava se jednom godišnje na Cipru.

b) „Common module on media training“ - jednotjedni, akreditirani, zajednički modul koji upoznaje polaznike kako raditi i ponašati se s medijima u području operacija. Modul je razvila Vojna akademija Belgije.

c) *„Common module on Crisis Management and Peace Support Operations “* - najzahtjevniji zajednički modul koji se provodi u četiri dijela, svaki u trajanju od tjedan dana: Pod-modul A - Temelj upravljanja krizama i međunarodne vojne operacije; Pod-modul B: Vježba na zemljovidu na temu iz PSO; Pod-modul C: Taktička vježba bez postrojbi na temu iz PSO; Pod-modul D: Tečaj za zapovjednike vodova u PSO, a neophodno je važan za uvježbavanje mladih časnika za sudjelovanje u međunarodnim vojnim operacijama. Razvila ga i je i provodi Terezijanska vojna akademija svake godine tijekom veljače.

d) „Common module on Law of Armed Conflict“ - jednotjedni zajednički modul o ratnom pravu koji je razvila Vojna akademija Francuske. Iznimno važna tema obuke i daje neophodna znanja za sve mirovne misije i operacije.

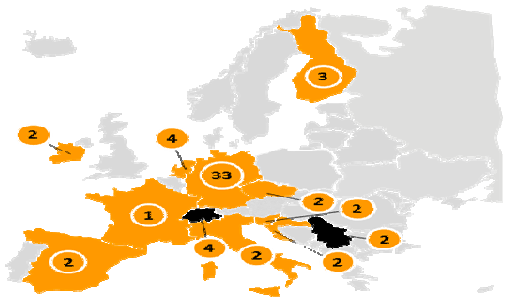
e) „Common module on Maritime Leadership“ - jednotjedni zajednički modul osmišljen za buduće časnike ratnih mornarica EU. Razvila ga je Mornarička akademija Francuske i spada u domenu obuke za „zapovijedanje u ekstremnim uvjetima.“

f) „Common module on Maritime Navigation“ - jednotjedni modul razvijen za upoznavanje sa specijaliziranim mornaričkim procedurama donošenja odluka. Razvila ga je također Mornarička akademija Francuske. Svi ovi moduli se sastoje od obveze polaganja ADL tečaja prije jednotjednog boravka i praktične obuke na akademiji koja je nositelj obuke, te se vrednuje s 2-5 ECTS bodova.

2. Akademski moduli u trajanju 3-6 mjeseci

Svaka od akademija razvija određene semestralne programe koji su otvoreni za međunarodne studente. Npr. Austrija ima razvijen cijeli zimski semestar iz vojnog vođenja *„Bachelor Program Military Leadership“*, Španjolska nudi ukrcaj stranih kadeta na svoje školske jedrilice i brodove u trajanju do 6 mjeseci s ciljem uvježbavanja procedura mornaričkog vođenja i zapovijedanja, Kopnena vojna akademija Poljske također nudi semestar za strane studente iz područja sigurnosnih strategija i međunaro-dnih odnosa, a Francuska vojna akademija u Rennesu za EU vojne kadete nudi *„International political science semester“* u trajanju od 3 mjeseca i nosi 30 ECTS bodova.

Dio zemalja je razvio i vlastite nacionalne programe razmjene, poput austrijskog *„International Training for Job - ITJ“*. Ovaj program obvezuje SVE austrijske kadete da moraju 6 tjedana stručne prakse provesti u stranoj zemlji kao zapovjednici vodova. Na slici 12. prikazane su države koje sudjeluju i broj kadeta koji su upućeni u njih.



Slika 9. Broj austrijskih kadeta na praksi (ITJ) u državama EU

3. Završetak cjelokupnog vojnog obrazovanja u jednoj državi je moguć u Austriji, Poljskoj, Češkoj, Slovačkoj i Francuskoj, ali se rjeđe događa zbog usklađivanja sa standardima civilnog obrazovanja i dijela vojne obuke.

4. Razmjena kadeta i mladih časnika u vojnoj obuci.

Ove se razmjene najčešće događaju tijekom ljetnih i zimskih kampova, vojnih obuka ili stažiranja. Razmjene su intenzivne s preko 100 obučnih aktivnosti godišnje.

Samo u ovom trenutku ponuđeno je više od 36 oblika vojnog obrazovanja i obuke u 10 europskih vojnih akademija u trajanju od tjedan do 6 mjeseci. Zemlja primateljica najčešće snosi sve troškove školovanja, smještaja i prehrane, dok matična zemlja snosi troškove putovanja i dnevnica.

Pet godina programa EMILYO omogućio je razmjenu 630 kadeta i mladih časnika diljem europskih vojnih akademija, te se iz godinu u godinu, broj dana razmjene intenzivno povećavao.

Iako je EMILYO primarno orijentiran na kadete i mlade časnike, on je značajno utjecao i na same vojne obrazovne institucije koje su se morale prilagoditi akreditiranim programima, sadržajima i standardima obučavanja, provođenju obučnih sadržaja na stranom jeziku (najčešće engleskom ili francuskom), apsorbirati strane polaznike i vrjednovati vlastite programe i kroz njihova iskustva, viđenja i stavove.

V. RAZGOVOR SA ZAPOVJEDNIKOM EUROPSKOG SIGURNOSNOG I OBRAMBENOG FAKULTETA (ESDC) gsp. Hans Bernhard Weisserth-om

1. *Možete li pojasniti razloge i prednosti osnivanja "European Security and Defence College" (ESDC)*

ESDC je osnovan 2005. godine s ciljem pružanja zemljama članicama, EU institucijama i agencijama školovano osoblje sposobno učinkovito raditi u području CSDP-a. Od tada smo obučili gotovo 6000 diplomata, policajaca, civilnih administratora i vojnih osoba. Više od 500 polaznika je došlo iz partnerskih zemalja i međunarodnih organizacija, posebno onih koje s EU rade na upravljanju krizama. Ovo je moguće zbog toga što su se 74 nacionalne obučne institucije, uključujući diplomatske akademije, civilni obučni instituti, policijske akademije i nacionalne vojne akademije kao i drugi obrambeni studijski centri i nevladine organizacije, aktivno predale izradi i provedbi tečajeva pod okriljem ESDC-a, s usmjerenjem na CSDP i vezane teme.

Pod ESDC-om, zajednički se razvijaju rasporedi tečajeva u intenzivnoj suradnji sa stručnjacima i relevantnim službama u EU institucijama, koji su certificirani da na najbolji način održavaju EU politike i koncepte. Ovo je i razlog da certifikate svih naših tečajeva potpisuje Visoki predstavnik EU za vanjsku i sigurnosnu politiku.

Gotovo svi ESDC tečajevi spajaju diplomate, policajce, pravne stručnjake, civilne administratore i vojne stručnjake, kako bi se poboljšalo njihovo međusobno razumijevanje i omogućilo im bolji i učinkovitiji zajednički rad prema istim političkim i operativnim ciljevima. Na taj način ESDC aktivno i značajno utječe na razvoj zajedničke europske sigurnosne kulture.

2.*Kako vidite daljnji razvoj ESDC*

U travnju 2013., Vijeće je odobrilo novi pravni dokument za ESDC ("Council Decision 2013/189/CFSP") koji daje sveučilištu pravni okvir djelovanja, proračun i dopunsko osoblje. Kada bude u cijelosti implementiran, ojačati će sveukupne kapacitete sveučilišta i omogućiti daljnju sinergiju obuke u EU kroz modele „*pooling and sharing*". Moguće je još više se posvetiti obuci civila u upravljanju krizama, a u potpori realizacije CSDP-a.

3. *Što to čini europski pristup vojnom obrazovanju toliko drugačijim i jedinstvenim*

Jedna važna činjenica je da vojna obuka i obrazovanje mora biti sagledano u kontekstu zajedničke Europske sigurnosne kulture. Ona je obilježena shvaćanjem da niti jednu krizu danas ili u neposrednoj budućnosti ne može riješiti samo jedna zemlja i to samo samo svojim oružanim snagama. Rješenje leži u uporabi različitih instrumenata koji uključuju političke, diplomatske, razvojne, ekonomske i druge civilne i vojne mjere. Stoga je civilno-vojna suradnja ključna odrednica europskog pristupa i posljedično iznimno utječe na vojno obrazovanje. Drugi važan aspekt su dugoročni ciljevi izloženi u Ugovoru iz EU vezanim za razvoj CSDP-a koji tijekom vremena mogu dovesti do stvaranja zajedničke europske obrane - cilj koji u ovom trenutnu ne dijele sve zemlje članice, ali većina njih ozbiljno promišlja o konceptu. Stoga je iznimno važno da se unutar zemalja članica uspostavi bliska suradnja u području obuke i izobrazbe, ne samo stožernog osoblja, već na samom početku časničkog obrazovanja, tijekom njihove inicijalne integracije u vojni sustav. Europska inicijativa za razmjenu mladih časnika inspirirana Erasmusom, EMILYO je stoga iznimno važan korak u pravom smjeru.

VI. ORUŽANE SNAGE RH I EMILYO inicijativa i program

Osnivanjem programa „Kadet" 2003. godine, RH je dobila mogućnost sudjelovanja u razmjenama i obukama kadeta i mladih časnika diljem Europe. Aktivnosti su se intenzivirale 2008. godine, a Terezijanska vojna akademija iz Bečkog Novog Mjesta formalno je započela ovu suradnju pozivajući hrvatske kadete da po prvi put sudjeluju na austrijskom časničkom balu. Hrvatima u čast, dozvoljeno je da naši kadeti i kadetkinje otvore bal uz zvukove Jelačićevog marša. Bal je označio i simbolički ulazak RH i OS RH u europske sustave vojnog obrazovanja jer od tada kadeti započinju sa sudjelovanjem u vojnim obučnim modulima (C i D) austrijske vojne akademije. S druge strane, Austrija je izabrala OS RH kao mjesto za provođenje stručne prakse svojih kandidata za časnike (programa ITJ), te se svake godine 2 kadeta upućuju u postrojbe hrvatske kopnene vojske na 6 tjedna, gdje uz nadzor našeg mentora zapovijedaju i rade s vodom OS RH.

Uz aktivnosti u Austriji, hrvatski kadeti sudjeluju i u EMILYO inicijativi Njemačke časničke škole u Dresdenu, koja organizira jednotjedne module upoznavanja s njemačkom kopnenom vojskom, postrojbama, opremom i sustavom obuke.

Od 2013. godine OS RH su se i uključile po prvi put u Konferenciju mladih vojnih znanstvenika CEFME, koja se održala u Budimpešti. Na konferenciji se prezentiraju vojno-stručni društveni i tehnički radovi. Navedena konferencija održana je u svibnju 2014. u Slovačkoj i 2015. u Poljskoj. Konferencija uvezuje i rad iMAF-a (engl. *International Military Academic Forum*) i njenog znanstvenog

odbora čiji članovi ocjenjuju dostavljene i prezentirane radove.

Tijekom 2015. godine očekuje se i prvo upućivanje hrvatskih kadeta na jednotjedni EU CSDP modul u Vojnu akademiju Italije u Turin, koji bi im omogućio sudjelovanje i na sljedećoj CSDP olimpijadi koja će se održati 2016. godine na Cipru.

Slika 10. Hrvatski kadeti na modulu D vojne obuke u Austriji

Slika 11. Hrvatski i njemački kadeti na međunarodnom tjednu u Dresdenu

VII. POGLED U BUDUĆNOST

Slika 12. Hrvatsko vojno učilište „Petar Zrinski“

Slika 13. Primjena elemenata e-učenja na terenu

HVU je u procesu transformacije u visoko obrazovnu i znanstveno-istraživačku vojnu ustanovu odnosno instituciju tj. sastavnicu sveučilišne zajednice. Od 2014. godine provodi dva nova preddiplomska sveučilišna studijska programa za potrebe OS RH. Vizija je da se uvođenjem e-učenja želi u procesu obrazovanja ostvariti aktivan partnerski odnos nastavnika i polaznika, prilagođen osobinama i potrebama svakog polaznika, s jasno postavljenim obrazovnim ciljevima i jasno dodijeljenim odgovornostima za postizanje tih ciljeva. Uvođenjem i aktivnom primjenom e-učenja HVU namjerava ostvariti četiri strateška cilja: unaprijediti kvalitetu obrazovanja, omogućiti nastavnicima i studentima ostvarivanje novih uloga u obrazovnom procesu, povećati konkurentnost i dostupnost učilišta i učilišnih programa i osposobiti studente za uporabu tehnologija cjeloživotnog učenja.

I u europskom vojnom obrazovanju ponuđen je određeni broj tečajeva i zajedničkih modula u području e-učenja. Svaki od njih obuhvaća IDL i f2f fazu. Sadašnje sposobnosti su na zavidnoj razini, a njihovim daljnjim razvojem, koje su precizirane Smjernicama za daljnje unapređenje sustava e-učenja, osnažit će se novim sposobnostima, a polaznici će sistematizirati postojeća i usvojiti nova znanja, vještine i sposobnosti kojima će biti osposobljeni za još učinkovitiju njihovu primjenu u svakodnevnim zadaćama.

VIII. ZAKLJUČAK

U informacijskom društvu i vremenu *e-učenje* dobiva sve više na značenju poglavito u europskom vojnom obrazovanju. Može se zaključiti kako je i u europskom vojnom obrazovanju napravljen značajan pomak u razvoju i uporabi sustava e-učenja, ali zbog važnosti stalnog izlaganja međunarodnom okruženju tijekom obuke i obrazovanja, većina ih ipak uvezuje ADL i f2f module u vojnom obrazovanju svojih pripadnika. Razvijen je i ponuđen veliki broj „gotovih“ *online* tečajeva za potrebe obrambenog i vojnog sustava. Stvoreni su potrebni preduvjeti za implementaciju *e-učenja* i u obrambenom i vojnom sustavu RH.

„Znanjem u budućnost“ moto je sadašnjih naraštaja i ostat će moto i naraštajima koji će doći na vojno obrazovanje u škole Hrvatskog vojnog učilišta „Petar Zrinski“ u Zagrebu.

LITERATURA

[1] A. W. (Tony) Bates, „Upravljanje tehnološkim promjenama - Strategije za voditelje visokih učilišta“, *CARNet* - Zagreb i *Benja* - Lokve, 2004. (prijevod), Izvornik: „Managing Techological Change: Strategies for College and University Leaders“, *Jossey-Bass Publishers*, San Francisco, USA, 2000.

[2] B. Shackelford, „Project Managing E-Learning“, *American Society for Training & Development Press*, Alexandria, USA, 2002.

[3] M. M. Lynch & J. Roecker, „Project Managing E-Learning“, A handbook for successful design, delivery and management, *MPG Books Ltd*, Bodmin, Great Britain, 2007.

[4] J. Greene & A. Stellman,“Head First PMP“ A Learner's Companion to Passing the Project Management Professional Exam, A Brain-Friendly Guide, *Oreilly Media, Inc*., CA, USA, 2007.

[5] Grupa autora, „E-Lerning Concept“, NATO unclassified, *NATO, Bruxelles, Belgium*, 8.2.2012.

[6] M. Ćukušić & M. Jadrić, „E-učenje: koncept i primjena“, *Školska knjiga d.d.*, Zagreb, 2012.

[7] B. Divjak, „Ishodi učenja u visokom školstvu“, TIVA, FOI, Varaždin, 2008

Integriranje eTwinning projekata u nastavu informatike i engleskog jezika

Branka Lamza*, Iva Naranđa**
* II. osnovna škola Čakovec, Čakovec, Hrvatska
** II. osnovna škola Čakovec i OŠ Vladimira Nazora Pribislavec, Čakovec, Hrvatska
branka.lamza@skole.hr
iva.naranda@skole.hr

Sažetak - Rad predstavlja pregled iskustava i ideja stečenih provođenjem dvanaest eTwinning projekata u protekle tri školske godine u II. osnovnoj školi Čakovec: jedanaest projekata nagrađeno je Nacionalnom oznakom kvalitete, pet projekata nagrađeno je Europskom oznakom kvalitete i jedan projekt nacionalnom nagradom Comet 2013 u kategoriji viših razreda osnovne škole. eTwinning je portal koji služi za sklapanje kontakata između učitelja širom Europe i pokretanje zajedničkih projekata. Sve se odvija uz podršku Agencije za mobilnost i programe EU, odnosno odgovarajućih institucija u svakoj zemlji. U članku ćemo se baviti integriranjem eTwinning projekata u nastavu informatike i engleskog jezika. Kvalitetan eTwinning projekt treba zadovoljiti kriterije pedagoške inovativnosti i kreativnosti, integracije u kurikulum, razmjene i suradnje među partnerskim školama i kreativne uporabe informacijsko-komunikacijske tehnologije u izradi sadržaja sukladno s dobi učenika. Ostvareni rezultat i učinak vrlo je važno dokumentirati. Kvalitetni projekti uključuju aktivnosti izvan učionice koje su sastavni dio školskih aktivnosti i obrazovnih ciljeva škole te suradnju s roditeljima i lokalnom zajednicom. Projektne aktivnosti osmišljavaju se tako da svim učenicima omoguće aktivno sudjelovanje na način koji im najbolje odgovara te im pruže mogućnost izražavanja kreativnosti, osjećaj uspjeha i učenje uz zabavu. Ukratko, ovi projekti imaju zadaću razvijati i jačati kod učenika vještine 21. stoljeća.

I. UVOD

Integracija eTwinning projekta u kurikulum jedan je od šest kriterija prema kojima Nacionalna služba za podršku iz Agencije za mobilnost i programe EU vrši evaluaciju projekata. Uz integraciju projekta u kurikulum, bitna je i kreativna uporaba informacijsko-komunikacijske tehnologije, također jedan od kriterija vrednovanja projekata.

Integriranje eTwinning projekata u nastavu informatike i engleskog jezika uključuje nov način učenja i podučavanja, poticanje kreativnosti kod učenika te ostvarivanje ciljeva kako samog projekta tako i cjelokupne nastave informatike i engleskog jezika. Pri osmišljavanju projektnih aktivnosti posebnu pažnju posvećujemo odabiru alata primjerenih aktivnostima i ciljevima projekta u odnosu na dob učenika. Radom na eTwinning projektima učenici usvajaju pravila uljudnog ponašanja na internetu, važnost poštivanja autorskih prava i kritičko odabiranje sadržaja pronađenih na internetu, razvijaju vještine timskog rada i samostalnog učenja. Za očekivati je da će im sve navedeno biti od koristi u daljnjem životu.

Uz sve navedeno, aktivnosti na eTwinning projektima pružaju odličnu priliku za izlazak iz informatičke učionice i povezivanje nastavnih sadržaja predmeta informatika s raznim aktivnostima organiziranim na razini škole i u suradnji s lokalnom zajednicom kao što su obilježavanja važnih datuma i obljetnica, ekološke akcije, nacionalne kampanje i slično.

II. INTEGRIRANJE ETWINNING PROJEKATA U NASTAVU

A. Integriranje eTwinning projekata u izbornu nastavu i izvannastavnu aktivnost iz informatike

U provedenim projektima integriranje eTwinning projekata u izbornu nastavu informatike realizirano je u sklopu nastavnih cjelina Osnove crtanja, Multimedija, Izrada prezentacija, Programiranje i Internet te u okviru projektnih zadataka. U petom razredu u cjelini Osnove crtanja učenici su se uvodili u eTwinning projekte izradom crteža na razne teme projekata te prigodnih čestitaka za eTwinning partnere. U šestom razredu projektne aktivnosti izvodile su se u sklopu ponavljanja gradiva iz cjelina Multimedija i Izrada prezentacija. U sedmom i osmom razredu eTwinning projekti integrirani su u nastavni plan informatike u sklopu cjeline Internet i projektnih zadataka koji su planirani kako bi se učenicima omogućilo proširivanje i povezivanje vještina uporabe informacijske tehnologije i upoznavanje s korištenjem raznih Web 2.0 alata. Učenicima su se zanimljivim načinom rada pokazali satovi u obliku radionice na kojima bi jedan od učenika svojim razrednim kolegama pokazao i objasnio rad u nekom Web 2.0 alatu po svom izboru.

eTwinning projekt „E-waste - What Can Citizens Do?“ skrenuo je pozornost na sveobuhvatnost društvenog, ekološkog i ekonomskog aspekta e-otpada, globalnog izazova i podcijenjenog ekološkog problema. Ciljevi projekta bili su dijeljenje informacija o e-otpadu, objasniti i približiti osnove društvenih, ekoloških i ekonomskih aspekata e-otpada i razumjeti mogućnosti za smanjenje, ponovno korištenje i recikliranje e-otpada te podići svijest učenika, njihovih roditelja i lokalne zajednice o problemu

e-otpada i potaknuti ih na odgovorno ponašanje s njime. Jedan od rezultata projekta je zajednička online brošura pod nazivom "E-otpad i ja".

Na satovima informatike učenici sedmog i osmog razreda napravili su animacije alatom Gickr, multimedijske online postere alatom Glogster na temu smanjivanja, ponovnog korištenja i recikliranja e-otpada, prezentacije alatom Prezi na teme vezane uz Svjetski dan zaštite okoliša, održivi razvoj, smanjivanje ugljičnog otiska i e-otpada, upoznati su sa socijalnim aspektima e-otpada te su vodili diskusiju o toj temi koristeći alat TodaysMeet i nakon toga napravili su PowerPoint prezentacije na teme vezane uz socijalni aspekt e-otpada.

Članovi školskog web-tima aktivno su sudjelovali u svim fazama projekta, izradili su mentalne mape alatom bubbl.us te redovito pratili i objavljivali članke i materijale o projektnim aktivnostima na Twinspaceu projekta i školskom webu.

Na početku projekta proveden je upitnik među učenicima četvrtog razreda polaznicima izvannastavne aktivnosti iz informatike II. osnovne škole Čakovec kako bi se ustanovilo inicijalno stanje i dobio uvid u to jesu li učenici već upoznati s pojmom e-otpada. Rezultati su pokazali opravdanost provođenja projekta o ovoj temi jer je od 32 učenika samo njih 7 odgovorilo da je već čulo za e-otpad (Slika 1). Učenici četvrtog razreda zajedničkom prezentacijom predstavili su se partnerima, napravili mentalne mape alatom bubbl.us, oblake riječi alatom Wordle, crtali simbol za odvojeno sakupljanje otpada u programu Bojanje, rješavali online igre vješala, osmosjerke i slagalice.

Tri mjeseca nakon završetka projekta, u prosincu 2014. anketirani su učenici II. osnovne škole Čakovec kako bi se vidjelo je li projekt postigao promjenu ponašanja učenika prema postupanju s e-otpadom. Obuhvaćeno je 54 učenika koji su prethodne školske godine sudjelovali u projektu i koji su još uvijek u osnovnoj školi, dakle učenike petog i osmog razreda školske godine 2014./2015. Iz rezultata može se zaključiti da je provođenje projekta na temu e-otpada bilo opravdano te da je projekt ostvario pozitivan učinak na odnos učenika prema e-otpadu (Slika 2).

Slika 1. Inicijalno stanje u 4. razredu II. OŠ Čakovec

Mogućnost suradnje sa europskim vršnjacima pokazala se vrlo poticajnom za učenike te su aktivnosti na eTwinning projektima korištene i kao pomoć u motiviranju učenika za programiranjem, dijelom gradiva koje učenici često doživljavaju najtežim. Ove aktivnosti su povezane i s profesionalnom orijentacijom učenika te bi trebale potaknuti interes za odabirom daljnjeg školovanja u nekom od informatičkih područja. Primjeri projektnih aktivnosti u kojima smo se u II. osnovnoj školi Čakovec bavili i programiranjem su projekt „2014 Winter Olympics, Sochi, Russia“ i projekt „Safe Internet – A Bridge Between Us“. Osnovni cilj projekta „2014 Winter Olympics, Sochi, Russia“, promicanje ideje olimpizma, povezan je s još jednim događanjem, proslavom Europskog tjedna programiranja 2013. organiziranjem radionica za učenike sedmog i osmog razreda „Create logo using Logo“ i Skype susreta s eTwinning partnerima. Ideja je izuzetno jednostavna, izrada programa u Terrapin Logu koji crtaju logo olimpijskih igara i razmjena iskustava s partnerima. U kurikulumu izvannastavne aktivnosti iz informatike za učenike viših razreda planiran je rad u Scratchu što se realiziralo aktivnošću na projektu „Safe Internet – A Bridge Between Us“ u kojoj su učenici imali mogućnost kreativno izraziti svoje ideje izradom animiranih priča na temu sigurnosti na internetu te razmijeniti iskustva i dojmove s partnerima iz Grčke.

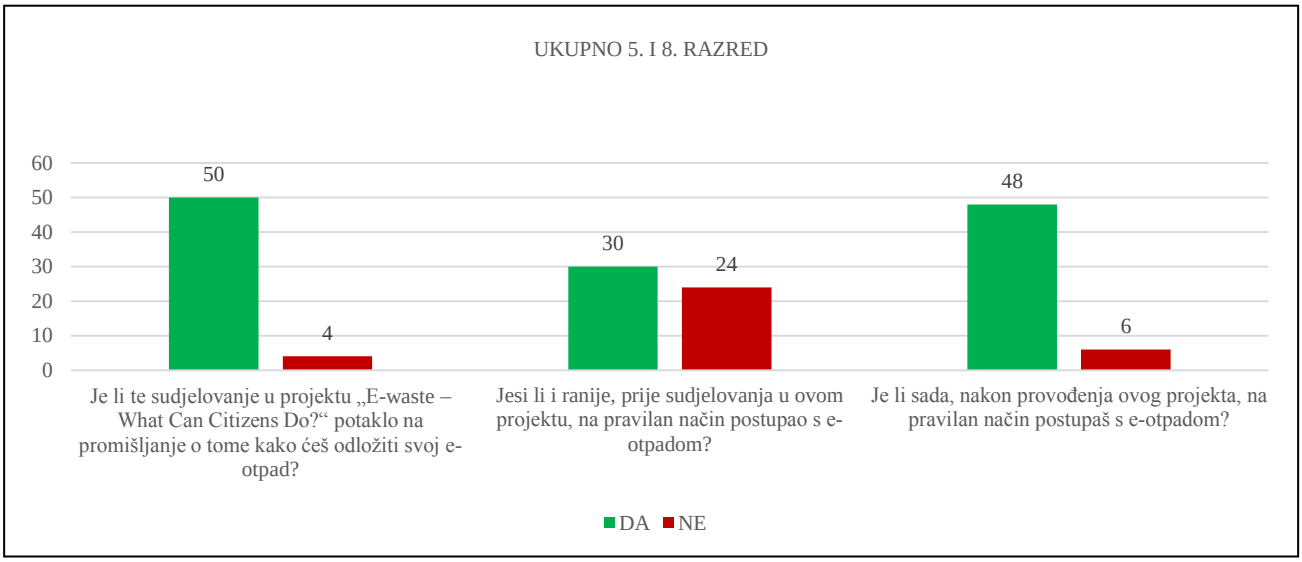
B. Integriranje eTwinning projekata u nastavu engleskog jezika u osnovnoj školi

Iako je portal eTwinning većim dijelom prilagođen za korištenje na većini jezika zemalja Europe činjenica je da je kao i mnogim drugim područjima života engleski jezik postao nezaobilazni jezik sporazumijevanja među većinom korisnika. Samim time on je i radni jezik većine projekata.

Uzevši u obzir navedeno, integriranje eTwinning projekata u nastavu među prvima su prigrlili učitelji engleskog jezika. Naime, integriranje projekata na engleskom jeziku na teme koje svatko može sam izabrati i prilagoditi nacionalnom obrazovnom standardu i vlastitom planu nameće se kao nešto prirodno u nastavi engleskog jezika.

S druge strane, suvremena nastava engleskog jezika ne može se ostvariti bez informacijsko-komunikacijske tehnologije. eTwinning omogućuje učiteljima da usavršavaju svoje informatičke vještine i uče o novim internetskim alatima jedni od drugih.

Možemo reći da se kroz eTwinning pojavila naša ideja o suradnji učiteljica engleskog jezika i informatike. Neodvojivost engleskog jezika i informacijsko-komunikacijske tehnologije, odnosno sadržaja engleskog jezika i informatike kao nastavnih predmeta pružaju jaku podlogu za njihovo uspješno integriranje u praksi.



Slika 2. Rezultati ankete provedene među učenicima petog i osmog razreda tri mjeseca nakon završetka projekta E-waste – What Can Citizens Do?.

C. Integrirana nastava informatike i engleskog jezika u eTwinning projektima

„Safe Internet: A Bridge Between Us“ projekt je koji smo proveli u suradnji s partnerskom školom iz Grčke. Ciljevi su bili potaknuti svijest učenika o važnosti što sigurnijeg korištenja interneta, zaštite osobnih podataka kao i zaštite i poštivanja autorskih prava. Naš učenički tim predvodile su učiteljice engleskog jezika i informatike. Iako je voditeljica grčkog tima bila učiteljica informatike, i ona je tijekom cijelog projekta surađivala s učiteljicama engleskog jezika, tako da možemo reći da je u oba tima ostvarena uska korelacija između ova dva nastavna predmeta. U projektu su sudjelovali učenici starosti od 13 do 14 godina. U našem timu radilo se o učenicima 7. i 8. razreda.

U početnoj fazi svakog eTwinning projekta uobičajeno je predstavljanje i međusobno upoznavanje škola sudionica. Mi smo dogovorili upoznavanje putem videokonferencije koristeći program Skype. Osim nekoliko rečenica koje su obje ekipe trebale pripremiti o svojoj školi, gradu i zavičaju dogovoreno je da se održi i kratki kviz na temu našeg projekta. Učenici su istražili općenita pravila sigurnosti na internetu i najvažnije činjenice o zaštiti autorskih prava i pripremili „točno/netočno“ pitanja za partnersku ekipu. Primjer pitanja bio je: „Preuzimanje glazbe s iTunes portala je nezakonito.“ (netočno).

Kviz je protekao u vrlo ugodnom ozračju na zadovoljstvo obje ekipe. Svaka ekipa postavljala je pitanja drugoj ekipi a rezultat je bio izjednačen. Osim provjere znanja kviz je ispunio sljedeće ciljeve: poticanje učenika na razmišljanje o temi te razvijanje motivacije za daljnji rad na projektu.

Mogućnost korištenja Skypea kao sredstva komunikacije pokazala se kao prava blagodat moderne tehnologije u nastavi. Ne iziskuje nikakve dodatne troškove u odnosu na uobičajene troškove korištenja interneta, a motivacijski učinak ovih susreta je velik.

Videokonferencija putem Skypea bila je uvod u obilježavanje Dana sigurnijeg interneta 2014. koji se u II. osnovnoj školi Čakovec unazad nekoliko godina obilježava uređivanjem prigodnih panoa, izradom prezentacija i njihovim prikazivanjem u holu škole, predavanjima i radionicama u sklopu sata razrednika i na nastavi informatike te sličnim aktivnostima. U veljači 2014. središnja aktivnost bila je integrirana nastava engleskog jezika i informatike za učenike 7. i 8. razreda.

Integrirana nastava provedena je u obliku blok-satova a bila je podijeljena na pretežno teoretski i pretežno praktični dio. Iako se niti u jednom dijelu ne može potpuno razdvojiti nastavne predmete (a to nam i nije bila namjera), možemo zaključiti da u prvom dijelu prevladava engleski jezik, a predvodi ga učiteljica engleskog jezika, dok u drugom dijelu prevladava informatika. Poučavanje o korištenju računalnih programa i/ili Web 2.0 alata odvija se pod vodstvom učiteljice informatike, dok u praktičnom dijelu podjednako sudjeluju obje učiteljice. Budući da se u ovom dijelu nastoji postići što veća samostalnost učenika u radu, učiteljice se uključuju i pomažu samo po potrebi.

Stoga, možemo zaključiti da se ovaj oblik nastave može uspješno provoditi samo ukoliko su oba učitelja upućena minimalno u osnove struke svoga kolege. Ukoliko se zajedničkim radom kolege zainteresiraju za predmet onog drugog i sami prošire svoja znanja, uspjeh budućih zajedničkih projekata zasigurno neće izostati.

U sklopu integrirane nastave povodom Dana sigurnijeg interneta 2014. u prvom dijelu čitalo se na tekstove na engleskom jeziku (u elektronskom i tiskanom obliku), nakon čega su slijedile vježbe razumijevanja i proširivanje vokabulara. U drugom dijelu učenici su se upoznavali s novim Web 2.0 alatom te uz pomoć njega izrađivali radove na engleskom jeziku.

Učenici 7. razreda uz pomoć učiteljice engleskog jezika istražili su savjete kako biti što sigurniji pri korištenju interneta. Dobili su zadatak izdvojiti i napisati savjete koje smatraju najkorisnijima. Nakon toga učiteljica informatike upoznala ih je s mogućnostima Web 2.0 alata Glogster pomoću kojeg mogu izraditi online postere. Nakon toga prionuli su izradi vlastitih radova s prethodno izdvojenim savjetima.

Učenici 8. razreda kroz prezentaciju na engleskom jeziku upoznali su se s najvažnijim odredbama hrvatskog zakonodavstva o zaštiti autorskih prava koje je većim dijelom usklađeno s europskim zakonodavstvom. Posebno su istaknute sljedeće odredbe:

1) Autorsko pravo pripada autoru djela samim stvaranjem.

2) U autorska djela spadaju: jezična djela (pisana djela, govorna djela, računalni programi), glazbena djela, djela likovne umjetnosti, fotografska djela, audiovizualna djela i dr.

3) Za obrazovanje je vrlo važan članak 85 Zakona o autorskom pravu i srodnim pravima: Dopušteno je reproduciranje na papir ili sličan medij i distribuiranje pojedinih odlomaka zakonito objavljenih autorskih djela koja su po svojem sadržaju i sistematizaciji isključivo namijenjena nastavi ili znanstvenom istraživanju uz navođenje izvora, osim ako to autor izričito ne zabrani.

4) Fizička osoba može reproducirati autorsko djelo ako to čini za privatno korištenje, kao i reproducirati autorsko djelo u obliku fotokopije i za drugo vlastito korištenje, koje nema komercijalnu svrhu i nije namijenjeno ili pristupačno javnosti.

Nakon provedenih vježbi razumijevanja i proširivanja znanja neophodnim leksikom engleskog jezika učenici su dobili zadatak izdvojiti po njihovom mišljenju najvažnije zakonske odredbe o zaštiti autorskih prava. U drugom dijelu nastave upoznali su se s Web 2.0 alatom Emaze. Ovaj alat omogućuje izradu prezentacija s 3D efektom. Pomoću njega učenici su prethodno izdvojene odredbe o zaštiti autorskih prava pretvorili u vrlo atraktivne prezentacije.

Praktični dio nastave izgleda tako da učenici samostalno ili u parovima rade na računalima a obje učiteljice nenametljivo obilaze i po potrebi pomažu – svaka vezano uz svoj predmet. Na ovaj način postiže se neformalni oblik nastave kojim se postiže maksimalna moguća samostalnost i kreativnost učenika.

Na sličan način proveli smo aktivnosti u sklopu projekta o održivim izvorima energije „Sustainable Energy – Clean Energy for the Future“. Glavni cilj ovog projekta bio je upoznati se s mogućnostima korištenja energije sunca, vode, vjetra, biomase i geotermalne energije. U projektu su sudjelovali učenici 6. i 8. razreda (12 i 14 godina).

Nakon što su učenici dobili zadatak istražiti ove izvore i upoznati se s najvažnijim činjenicama održana je integrirana nastava engleskog jezika i informatike. U teoretskom dijelu učenici su gledali tematski animirani film na engleskom jeziku. Nakon toga uslijedili su vježbe razumijevanja i proširivanje vokabulara. Učenici su zatim dobili zadatak pismeno izraditi zadatke uparivanja i kvizove.

Učiteljica informatike upoznala je učenike s programom HotPotatoes. U praktičnom dijelu nastave prethodno izrađene zadatke pretvorili su u elektronski oblik.

U narednim satovima informatike učenici su upoznati sa sve popularnijim Web 2.0 alatom Zondle te su i njega iskoristili za izradu kvizova o održivim izvorima energije u elektronskom obliku.

U ovom projektu (kao i u kvizu o korištenju programa Terrapin Logo) parterska škola bila je tehnička škola iz Španjolske. Zbog različite strukture obrazovanja učenici su bili slične dobi (14). Naši partneri posvetili su se temi unutar programa strukovnih predmeta svoje škole, a upravo od nas dobili su ideju za korištenjem alata Zondle.

Kao i u projektu „Safe Internet: A Bridge Between Us“ održali smo tematski kviz putem videokonferencije a pitanja koja su ekipe pripremile objavili smo na školskim web stranicama u obliku Zondle kviza kako bi ga svi zainteresirani mogli rješavati.

Sljedeći primjer uspješne integracije nastave engleskog jezika i informatike je prethodno spomenut projekt „Winter Olympics, Sochi, Russia“. U ovaj projekt bili su uključeni učenici četvrtog i sedmog razreda. Tema sporta sastavni je dio nacionalnog plana i programa nastave engleskog jezika za 7. razred, te je ovaj važni događaj poslužio kao korisna poveznica s temom. Učenici su u sklopu nastave engleskog jezika proširili vokabular vezan uz zimske sportove. U sklopu komunikacijskih aktivnosti dobili su zadatak sastaviti dijaloge na temu omiljenog zimskog sporta i praćenja Zimskih olimpijskih igara.

Na nastavi informatike učenici su se upoznali s Web 2.0 alatom Go!Animate koji služi za izradu kratkih animiranih filmova. Već u 6. razredu učenici su se upoznali s osnovnim mogućnostima ovog alata. Naučili su izraditi animirani film odabiranjem likova i pozadine i unosom željenog teksta. Na ovaj način vježbali su uporabu pravopisa. Naime, neispravno unesene riječi program interpretira jednako neispravnim izgovorom, što učenici u pravilu prepoznaju i ispravljaju pravopisne greške.

Ovaj put učenici su proširili svoja znanja o alatu Go!Animate i upoznali se s dodatnim opcijama: unosom teksta u oblačićima, reguliranjem pokreta likova, izborom izraza lica i pozadinske glazbe. Svoje prethodno sastavljene dijaloge uspješno su pretvorili u zanimljive kratke animirane filmove koji su se izuzetno svidjeli našim projektnim partnerima.

D. eTwinning projekti i aktivnosti izvan učionice

Kao primjere školskih aktivnosti izvan učionice navodimo aktivnosti u sklopu projekta „E-waste - What Can Citizens Do?“: organiziranje predavanja za učenike povodom obilježavanja Dana planeta Zemlje u travnju 2014. u suradnji s Gradskim komunalnim poduzećem Čakom, sudjelovanje u nacionalnoj akciji Zelena čistka

kada je u travnju 2014. organizirano prikupljanje električnog i elektroničkog otpada u dvorištu škole u suradnji s GKP Čakom te obilježavanje Svjetskog dana zaštite okoliša 2014 (World Environment Day – WED).

III. ZAKLJUČAK

Prednosti eTwinning projekata u nastavi 21. stoljeća neupitne su. Rad na ovim projektima nedvojbeno razvija informacijsko-komunikacijske vještine učenika i njihovu svijest kao europskih građana te služe razvijanju kreativnosti i značajno doprinose izrastanju učenika u samostalne i odgovorne osobe sklone timskom radu.

Ono što čini razliku između tradicionalne nastave i nastave u koju su integrirani eTwinning projekti jest mogućnost da učenici rezultate svoga rada prezentiraju ne samo izvan učionice nego izvan svoje zemlje bez ikakvih dodatnih troškova i dobiju povratne informacije. Učenici i učitelji su u mogućnosti iz prve ruke doznati kako se uči širom Europe i koliko ima sličnosti i razlika među obrazovnim sustavima. Provođenje ovih projekata moguće je u okviru bilo kojeg nastavnog predmeta i ne zahtjeva osmišljavanje dodatnih tema pored onih već predviđenih nastavnim planom i programom. U svaku temu unosi interkulturalnost, interaktivnost i toliko potrebnu svijest o činjenici da razlike nisu nešto negativno, naprotiv, one obogaćuju Europu budućnosti.

LITERATURA

[1] Vuorikari, Riina. December 2012. *eTwinning School Teams – Case studies on teacher collaboration through eTwinning*. Central Support for eTwinning (CSS) - European Schoolnet (EUN Partnership AISBL). Brussels.

[2] Zakon o autorskom pravu i srodnim pravima, NN 167/03, 79/07, 80/11, 125/11, 141/13, 127/14

[3] Bluck, Emma; Doris, Aine. November 2012. Best practices in energy efficiency education in schools across Europe – chosen from the U4energy school competition. European Schoolnet (EUN Partnership AISBL). Brussels.

[4] The Web We Want. 2013. European Schoolnet. Brussels.

Stručno usavršavanje nastavnika putem e-učenja

Marina Mirković, dipl. inženjer elektrotehnike
Tehnička škola, Požega, Republika Hrvatska
marina.mirkovic@po.t-com.hr

Sažetak

Društvo u kojem živimo je informatički orijentirano s neprestanim rastom potrebe za cjeloživotnim učenjem. Online učenje je izazov i alat za poboljšanje i unapređenje obrazovnih procesa. Mogućnosti e-učenja povećavaju dostupnost stručnog usavršavanja svim odgojno obrazovnim djelatnicima koji su ujedno potaknuti na uporabu informacijsko komunikacijske tehnologije.

U radu se opisuju mogućnosti i iskustva primjene online tečajeva za stručno usavršavanje nastavnika u Republici Hrvatskoj. Agencija za odgoj i obrazovanje razvila je šest programa e-učenja za stručno usavršavanje koji se odnose se na teme vrednovanja, individualizacije rada, strategije učenja i poučavanja, mentorstva, profesionalnog razvoja i programa Pestalozzi.

CARNetov masovni online tečaj - CARNet Moodle MOOC namijenjen je svima koji žele naučiti kako samostalno osmisliti i izraditi online tečaj u Moodle-u. Slovenska akademska istraživačka mreža - Arnes organizira besplatni masovni online tečaj namijenjen svima koji žele naučiti više o sigurnom korištenju Interneta i suvremenih informacijskih tehnologija.

Stručno usavršavanje za online nastavu omogućuje nastavnicima da postanu svjesni kako učenici doživljavaju e-učenje, a to najbolje mogu postići uključivanjem nastavnika u ulogu polaznika različith online tečajeva. Ovaj oblik iskustvenog učenja izaziva refleksiju na dubokoj razini i poboljšava praksu. Nastavnici pri tome rade na napredovanju svoje stručnosti, poboljšavaju svoje obrazovne i tehnološke vještine.

I. UVOD

Intenzivno uvođenje informacijskih tehnologija u obrazovne procese postao je prioritet modernih visokoobrazovnih institucija diljem svijeta. E-učenje utječe na različita područja društva, a time i na ekonomiju što se odnosi na usavršavanje zaposlenih. Omogućuje uštede vremena i novca. Sredstva koja škole mogu trošiti na stručna usavršavanja nastavnika, nažalost su sve manja, a programi e-učenja ne zahtijevaju dodatna plaćanja. Time se čuva novac koji bi se izgubio na troškove smještaja i prijevoza koji se javljaju prilikom stručnog usavršavanja zaposlenih. Ušteda radnog vremena je značajan čimbenik. Nisu potrebne zamjene zaposlenika jer nastavnici mogu pristupiti online programima poslije njihove nastave. Vrijeme stručnog usavršavanja prilagođava se nastavnikom željama, potrebama i mogućnostima.

Praktičan rad nastavnika sa različitim tehnologijama uz stjecanje informacija o temi e-programa, pruža i dodatna znanja i vještine o uporabi različitih tehnologija, omogućuje nastavnicima da postignu i zadrže razinu infomatičke pismenosti, samostalno učenje i interakciju.

II. PROGRAMI E-UČENJA ZA STRUČNO USAVRŠAVANJE NASTAVNIKA

Agencija za odgoj i obrazovanje u sklopu projekta „Poboljšanje sustava stručnog usavršavanja odgojno-obrazovnih djelatnika“ razvila je pet programa e-učenja za stručno usavršavanje koji su od proljeća 2014. godine dostupni putem interneta. Interdisciplinarni programi e-učenja odnose se na teme vrednovanja, individualizacije rada, strategije učenja i poučavanja, mentorstva, profesionalnog razvoja i Pestalozzi vijednosti i načela. Kreirani programi su:

- Vrednovanje učeničkih postignuća
- Individualizacija u odgojno-obrazovnom procesu
- Strategije učenja i poučavanja
- Izazovi mentorstva
- Putem profesionalnog razvoja
- Učenje i življenje Pestalozzi vrijednosti i načela u odgoju i obrazovanju.

Programi e-učenja Agencije za odgoj i obrazovanje smješteni su na CARNetovom poslužitelju, a verzija Moodle-a koja se koristi u CARNetu nazvana je Loomen [1]. Nalaze se u kategoriji Programi e-učenja AZOO-a. Prijava za sudjelovanje izvodi se putem sustava www.ettaedu.eu. Veliki je broj zainteresiranih za pohađanje programa, te se vrlo brzo popuni broj prijava (50 ili 100 polaznika). Za pohađanje programa e-učenja dovoljne su osnovne vještine korištenja računala i interneta kao što su: korištenje elektroničke pošte, pregledavanje web stranica, upravljanje datotekama na računalu te korištenje programa za obradu teksta i/ili tabličnih kalkulatora.

Odabranom programu može se pristupiti i nesmetano raditi u bilo koje vrijeme, odnosno kada polaznicima najviše odgovara. Ne moraju se odjednom proći svi njegovi dijelovi. Sustav bilježi vrijeme provedeno na

tečaju i mjesto na kojem je polaznik zastao i treba nastaviti kad se ponovno uključi.

Predviđeni oblik rada je individualni. Korisnici se koriste sljedećim strategijama: vizualizacija, sumiranje, stvaranje analogija, popunjavanja ponuđenih obrazaca; postavljanja vlastitih ciljeva učenja i rokova izvršavanja aktivnosti zadanih programom; odabir i primjena odgovarajućih strategija s obzirom na materijale i alate programa te samovrednovanje i samorefleksija. U sadržajima se mogu naći različiti materijali za čitanje, upitnici i kratki filmovi za samorefleksiju i samoprocjenu, alati i aktivnosti za usvajanje, ponavljanje i uvježbavanje. Svaki tematski dio sadrži poveznice na različite alate i obrasce.

Program „Vrednovanje učeničkih postignuća“ omogućuje polaznicima da usvoje osnovne koncepte vrednovanja. Posebna pažnja dana je zakonima i pravilnicima vezanim za ovo područje. Polaznici će upoznati moderne koncepcije vrednovanja u kojima učenik postaje aktivni sudionik vlastitog vrednovanja pri čemu je ono sastavni dio učenja i poučavanja. Prikazano je zanimljivo predavanje dr. sc. Borisa Jokića o kvaliteti vrednovanja. Kroz program, sudionici će osvijestiti važnost dobro definiranih odgojno-obrazovnih ishoda te moći oblikovati različite vrste zadataka otvorenog i zatvorenog tipa.

Program e-učenja „Individualizacija u odgojno-obrazovnom procesu“ pruža polaznicima temeljna znanja o individualiziranim odgojno obrazovnima programima te potrebi i obvezi individualizacije u odgojno-obrazovnom radu. Nužno je razvijati sposobnost prepoznavanja potreba učenika, kako onih s posebnim obrazovnim potrebama, odnosno s teškoćama u učenju, tako i darovite djece.

Svim odgojno obrazovnim djelatnicima preporučujem program „Strategije učenja i poučavanja“ koji će ih kroz razne sadržaje uvesti u svijet suradničkog učenja (slika 1). Polaznici će upoznati nužnost promjene u pristupu učenju i poučavanju te usvojiti osnovna načela poučavanja

Slika 1. Naslovnica programa e-učenja „Srategije učenja i poučavanja“

usmjerenog na učenika. Razrađene su faze suradničkog učenja, posebno opisan rad u paru. Polaznici će upoznati nove uloge učitelja, analizirati primjere aktivnosti učitelja i učenika, proučiti poticanje samostalnosti učenika u učenju i njegove motivacije za učenje. Za lakši rad u aktivnostima, polaznici koriste dnevnike refleksije i radne bilježnice za učitelje.

Program „Izazovi mentorstva“ namijenjen je odgajateljima, učiteljima, nastavnicima i stručnim suradnicima koji se nalaze u ulozi mentora ili to mogu postati. Polaznici tečaja upoznati će i razumjeti uloge i karakteristike mentora, faze mentorskog rada i zadatke te izazove s kojima se susreću u pojedinoj fazi da bi se prikladno pripremili za njih. Osvijestiti će i razumjeti komunikacijske vještine važne za uspješnu interakciju i komunikaciju između mentora i pripravnika te će moći primijeniti različite komunikacijske tehnike za uspješno rješavanje mogućih problema u komunikaciji u mentorskom odnosu.

Zanimljivi su i korisni prijedlozi obrazaca dnevnika stažiranja, plana i programa pripravničkog staža, obrazaca za analizu potreba pripravnika i mentora, refleksiju i samorefleksiju. Zasebni dio je „Kutak za čitanje“ koji sadrži primjere iz prakse i dodatne materijale.

U programu „Putem profesionalnog razvoja“ polaznici se informiraju o konceptu profesionalnog razvoja, cjeloživotnom učenju i ciklusima profesionlane karijere. Kroz različite zadatke, aktivnost promišljanja vodi ih kroz analizu profesionalne uloge i omogućuje razvoj kompetencija samorefleksije profesionalnog razvoja. Predstavljen je kotač ravnoteže (životni i profesionalni) kao alat za planiranje, evaluaciju, provjeru i refleksiju o zadovoljstvu našim radom. Polaznik reflektira vlastito iskustvo na osnovi različitih alata i metoda, istražuje svoje potencijale i područja vlastitog razvoja. Kroz intervju u paru, polaznici razmijenjuju iskustva s kolegama. Svoj profesionalni razvoj polaznici prikazuju putem e-portfolia.

Cilj programa „Učenje i življenje Pestalozzi vrijednosti i načela u odgoju i obrazovanju“ je da polaznici saznaju koje su vrijednosti i načela Programa Pestalozzi Vijeća Europe te da uoče važnost usvajanja i primjene tih vrijednosti i načela u odgoju i obrazovanju. Polaznici će saznati osnovne informacije o organizaciji Vijeća Europe, osnovne podatke o Europskoj kulturnoj konvenciji, teme i aktivnosti stručnih usavršavanja u Programu Pestalozzi. Naglasak je na razvoju interkulturalne kompetencije u nastavi kao i transverzalnih stavova, vještina i znanja. Njih nastavnici razvijaju u nastavi kroz pet područja djelovanja: raznolikost i empatija; suradnja i sudjelovanje; ljudska prava i jednakost; stvaranje znanja i epistemologija; osobnost i interakcija. U programu polaznici kroz radionicu izrađuju i objavljuju svoje primjere primjene transverzalnih stavova, vještina i znanja te ih kroz forum komentiraju u raspravi.

Polaznici na kraju pohađanja pojedinog programa ispunjavaju evaluacijski upitnik te njime autorima daju povratnu informaciju o programu. To je ujedno i preduvjet za dobivanje potvrde o sudjelovanju u programu e-učenja. Programi će se kontinuirano usavršavati i nadopunjavati. Nedostaje kontakt polaznika s trenerima i drugim

sudionicima te se predlaže da budući programi e-učenja budu kombinirani a da se komunikacija ostvaruje putem foruma [2].

III. MASOVNI ONLINE TEČAJEVI

MOOC (Massive Open Online Course) je online tečaj ili predavanje s neograničenim sudjelovanjem i slobodnim pristupom preko interneta. Ovaj oblik edukacije je prilagodljiv rasporedu, potrebama i interesima polaznika. Omogućena je interakcija i razmjena znanja i mišljenja s kolegama diljem svijeta koji pohađaju isti tečaj. Uz internetsku vezu, za učenje je važna i osobna motivacija kao i odgovornost sudionika. Neispunjavanje obveza ne nosi sankcije slične onima u formalnom sustavu obrazovanja.

Prvi CARNetov masovni otvoreni online tečaj - Moodle MOOC do sada je organiziran i održan u dvije grupe polaznika (oko 900 sudionika) koji su željeli naučiti kako samostalno osmisliti i izraditi online tečaj u Moodle-u. Tečaj je podijeljen na sedam tematskih cjelina:

- Upoznajte se sa sučeljem i ulogama u Moodle-u
- Prvi koraci u izradi tečaja
- Izradite sadržaj svog tečaja
- Uključite aktivnosti za vaše polaznike
- Vrednujte znanje svojih polaznika
- Upravljajte polaznicima i tečajem
- Vaš osvrt na tečaj i što nakon tečaja?

Online tečaj traje šest tjedana, tijekom kojih polaznici mogu pristupiti tečaju i nesmetano raditi na njemu u vrijeme kada to njima najviše odgovara. Za svladavanje gradiva, izradu zadataka te sudjelovanje u ostalim nastavnim aktivnostima pretpostavljen je angažman polaznika od šest do osam sati tjedno, ovisno o predznanju i afinitetima polaznika. Polaznici pri tome izrađuju vlastiti online tečaj čiju sigurnosnu kopiju mogu koristiti za kasniju uporabu.

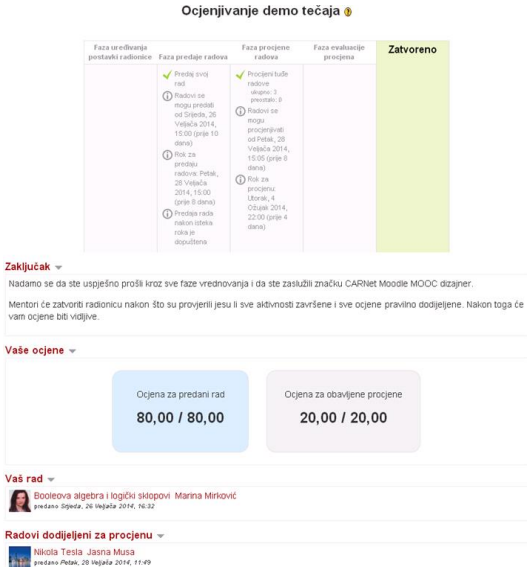
Tijekom trajanja cijelog tečaja polaznici individualno savladavaju gradivo, izrađuju individualne zadatke, ali sudjeluju i u zajedničkim aktivnostima s ostalim polaznicima: diskusijama u forumima i međusobnom vrednovanju postignuća, odnosno izrađenim tečajevima.

Komunikacija s drugim polaznicima omogućena je putem foruma u tečaju s ciljem upoznavanja i razmjene ideja u online okruženju. Mentori polaznicima tijekom pohađanja pružaju minimalnu podršku u savladavanju nastavnih sadržaja i ne sudjeluju u evaluaciji znanja. Ako je potrebno, komunikacija s mentorima može se odvijati putem foruma tečaja.

U tečaju je implementiran novi sustav evaluacije znanja putem Moodle digitalnih znački. Značka je dokaz da je polaznik online tečaja završio zadane aktivnosti i ostvario planirane ishode učenja [3].

Ukoliko uspješno završi neke od jednostavnijih aktivnosti prema definiranim kriterijima, polaznik dobiva značku CARNet Moodle MOOC - Polaznik. Polaznici koji se uključe u aktivnosti dijeljenja znanja i iskustava prema definiranim kriterijima dobit će značku CARNet Moodle MOOC - Istaknuti polaznik. Ukoliko izradi osnovni tečaj pomoću Moodle-a prema definiranim kriterijima, polaznik dobiva značku CARNet Moodle MOOC – Dizajner. Polaznici međusobno vrednuju svoje tečajeve prema unaprijed zadanim kriterijima i uz pomoć odgovarajućeg alata (slika 2).

Kada CARNet kao organizator tečaja dodijeli polazniku značku, ona se pojavljuje u profilu polaznika unutar sustava. Polaznik značku može koristiti kao dokaz određenih znanja i vještina, prenijeti je u svoj profil na sustavu Loomen ili objaviti na drugim web stranicama.



Slika 2. Radionica ocjenjivanja tečajeva polaznika

Slovenska akademska istraživačka mreža - Arnes tijekom studenog je organizirala besplatni masovni online tečaj namijenjen svima koji žele naučiti više o sigurnom korištenju interneta i suvremenih informacijskih tehnologija [4]. Tečaj je trajao tri tjedna a teme proučavanja bile su:

- Zaštita uređaja
- Zaštita podataka i digitalni identitet
- Zaštita zdravlja i okoliša.

Za uspješan završetak tečaja trebalo je postići 75% bodova na kvizovima i izvršiti sve aktivnosti u tečaju, posebno sudjelovati u forumima. U tečaju je sudjelovalo više od 750 sudionika iz Slovenije i inozemstva, odnosno susjednih zemalja. Oko 400 sudionika je uspješno završilo tečaj s odradom svih propisanih aktivnosti. Tečaj je održan na slovenskom jeziku.

U tečaju su sudionici upoznali kako se zaštititi od virusa te što su dobre lozinke zaštite. Analizirani su primjeri slučajeva prijevara i prijetnji s kojima se susrećemo online. Velika pozornost posvećena je privatnosti na internetu. Prikazani su slučajevi objavljivanja osobnih podataka na internetu kao i moguće posljedice takvog ponašanja. U tečaju je također prikazana ovisnost o internetu, igrama i mobitelima te načini zlostavljanja putem interneta i mobilnih uređaja, kao i budući izazovi u korištenju suvremenih tehnologija.

U tečaju su sadržaji proučavanja bili dokumenti, videozapisi, ankete, kvizovi. Forumi su bili vrlo aktivni. Sudionici su raspravljali o brojnim zanimljivim temama u kojima su podijeliti svoje iskustvo o sigurnoj upotrebi interneta.

Zanimljivi su bili brojni kvizovi, među njima i kviz o Web detektivu, s primjerima prepoznavanja problema u uporabi interneta. Kviz je dostupan na mrežnoj stranici: http://www.spletni-detektiv.si/.

Za uspješno sudjelovanje u tečaju polaznici su dobili Mozillinu značku "Pametan korisnik interneta" (slika 3).

Online tečaj koji se temelji na Europskom okviru za razvoj i razumijevanje digitalnih kompetencija (DIGCOMP), završio je sastankom uživo sudionika koji su to željeli. Pri tome su se problemi koji se javljaju "on-line", riješavali "off-line". Nastavnici su analizirali složene probleme sigurnijeg interneta i modernih tehnologija, s kojima se danas vrlo često susreću učitelji, ravnatelji i stručni suradnici u slovenskim, ali i školama ostalih zemalja. Koncept tečaja predstavljen je na konferenciji Arnes 2014. Opisao ga je gospodin Domen Božeglav, voditelj/mentor tečaja. Videozapis prezentiranja možete pogledati na mrežnoj stranici: https://video.arnes.si/portal/asset.zul?id=c1XBVpcAklCaLWTKjSzHkXMg.

Uz CARNetov Moodle MOOC i Arnesov „Sigurno korištenje interneta i modernih tehnologija", vjerujem da će se njihova uporaba i popularnost širiti i na druge organizacije a time promijeniti struktura obrazovanja.

Slika 3. Mozillina značka "Pametan korisnik interneta"

IV. TEČAJEVI UDRUGE „ZVONO"

Učitelji razredne nastave okupljeni na Forumu razredne nastave i u Udruzi hrvatskih učitelja razredne nastave „Zvono" u svom radu i razmjenom iskustva pokušavaju unaprijediti svoj rad pa tako i informatička znanja i vještine. U svrhu edukacija uže vezane za programe i alate koji bi bili korisni učiteljima za izradu nastavnih materijala, pokrenuli su besplatne informatičke online tečajeve za učitelje razredne nastave kao oblik neformalnog učenja [5].

Pri udruzi „Zvono" oformljena je radna skupina Informatički tečajevi čiji članovi sudjeluju u izradi i vođenju online tečajeve u LMS sustavu Moodle. Tako su nastali slijedeći tečajevi:

- Upotreba MS PowerPointa
- Izrada razredne web stranice
- Izrada nastavnih materijala
- Web 2.0 u nastavi
- Programiranje u nastavi likovne kulture
- Izrada kviza Hot Potatoes
- Zondle- igre kao podrška učenju

Tečajevi se odvijaju online što znači da se savladavanje sadržaja odvija u vrijeme kada polaznicima odgovara. Mišljenja polaznika o tečaju su puna pohvala i preporuka za uključivanje i ostalih nastavnika u ovaj oblik stručnog usavršavanja koji pridonosi stjecanju informatičkih kompetencija učitelja. Radionice/tečajevi su dobro organizirani, zanimljivi, primjenjivi u nastavi. Preglednu, jasnu i zanimljivu građu polaznici postupno usvajaju uz konkretne vježbe s preciznim uputama. Vježbe ocjenjuju i komentiraju mentorice tečaja.

V. TEČAJEVI BRITISH COUNCILA

British Council Hrvatska kao međunarodna organizacija vodi brigu o razvijanju vještina koje mogu doprinijeti ekonomskom rastu i pomoći u razvoju otvorenog i kreativnog društva. U svom radu organiziraju online tečajeve koji pomažu nastavnicima razviti razumijevanje konteksta unutar kojeg učenici uče engleski jezik te upoznati načela i praksu učinkovitog podučavanja [6].

Online tečajevi British Councila dostupni su kao moderirani tečajevi i kao tečajevi za samostalno učenje. Prikladni su za nastavnike početnike koji trebaju steći znanje iz osnovnih pojmova u poduci engleskog jezika, kao i za iskusne nastavnike koji žele osuvremeniti svoju metodiku. Polaznicima se nude slijedeći tečajevi:

- Tečaj o posebnim obrazovnim potrebama
- Tehnologija u učionici
- Komunikacijski pristup ocjenjivanju

- Osnove integriranog učenja jezika i nastavnog sadržaja
- Osnove stručnog testa za nastavnike
- Osnove poduke engleskog u nižim razredima osnovne škole

Cijena pojedinog tečaja je od 300,00 do 900,00 kuna, a trajanje najduže do 12 tjedana. Diplomu tečaja polaznici primaju nakon uspješnog ispunjavanja svih zadataka i aktivnosti u tečaju.

Povremeno su dostupni i besplatni online programi British Councila koji omogućuju nastavnicima da poboljšaju engleski jezik i saznaju više o britanskoj kulturi. Zainteresirani polaznici imaju mogućnost istražiti i prijaviti se prema različitim područjima i kategorijama na besplatne online tečajeve u organizaciji vodećih sveučilišta i kulturnih institucija.

VI. POUČAVANJE O E-TEČAJEVIMA

Nastavnici osnovnih i srednjih škola koji u svoju nastavu žele primjeniti e-učenje putem online tečajeva ili online komunikaciju s učenicima moraju steći vještine i znanja koja treba imati online mentor-nastavnik. U tome im pomaže CARNetov E-tečaj za e-mentore.

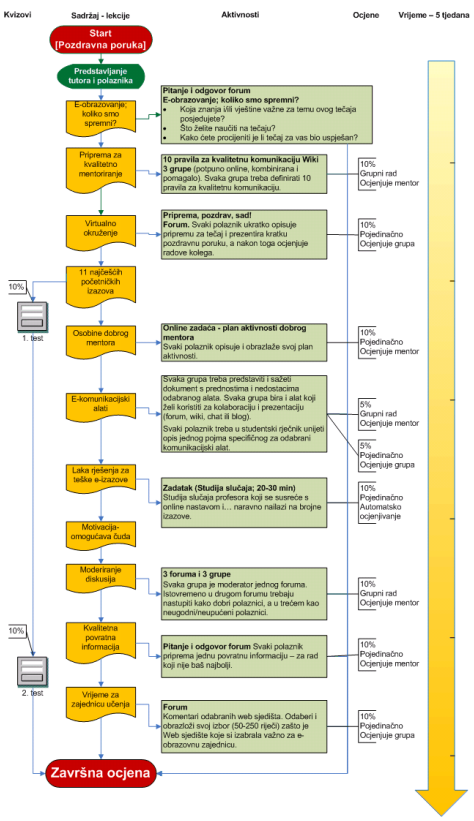
Sadržaj tečaja obuhvaća teme:

- priprema mentora i polaznika za tečaj
- uvođenje polaznika u virtualno okruženje
- 11 najvećih početničkih izazova
- osobine dobrog mentora
- odabir i prednosti pojedinog komunikacijskog alata
- motivacija polaznika
- moderiranje diskusija
- povratne informacije
- mogućnosti i potrebe osobnog razvoja nakon tečaja.

Tijekom održavanja tečaja, polaznici sudjeluju u brojnim aktivnostima (slika 4) sličnim onima koje će kasnije sami provoditi kao mentori: diskusije na relevantne teme, grupni radovi (pravila za kvalitetnu online komunikaciju, prednosti i mane pojedinog komunikacijskog alata i dr.), izrada pripreme za tečaj i pozdravne poruke, studija slučaja, moderiranje foruma, ocjenjivanje kolega i drugo .

Studija slučaja opisuje primjer uvođenja e-obrazovanja u nastavu. Polaznik prolazi kroz određene tipične situacije, te nakon svake bira koji sljedeći korak će napraviti kako bi što kvalitetnije uveo e-learning u nastavu [7].

Tečaj je implementiran u open source alat Moodle pri čemu su korišteni alati koje Moodle nudi, kao što su forum, wiki, blog, chat, lekcije, testovi i dr.



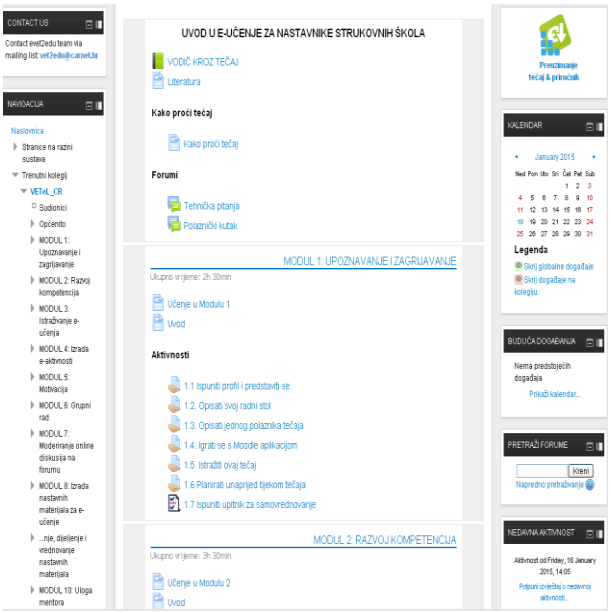
Slika 4. Raspored aktivnosti i zadataka u E-tečaju za e-mentore

Tečaj "Uvod u e-učenje za nastavnike strukovnih škola" koji je nastao u okviru projekta eVET2EDU, namijenjen je nastavnicima strukovnih škola i organizacija koji se žele upoznati s osnovama e-učenja. Tečaju je moguće pristupiti preko naslovnice CARNetovog sustava Loomen.

Sadržaj tečaja je prilagođen nastavnicima strukovnih predmeta koji žele svoju nastavu obogatiti elementima e-učenja pomoću kojih bi se nastava tih predmeta učinila tehnološki unaprijeđenom pomoću alata i metodike pogodne strukovnim područjima. Time nastava postaje usmjerena na učenika, interaktivnija i zanimljivijima današnjim generacijama učenika i polaznika u sustavu obrazovanja odraslih [8].

Tečaj se sastoji od 10 modula (slika 5) koji ukupno traju oko 60 sati učenja:

- Modul 1: Predstavljanje i zagrijavanje
- Modul 2: Razvoj kompetencija
- Modul 3: Istraživanje e-učenja
- Modul 4: Izrada e-aktivnosti
- Modul 5: Izrada nastavnih materijala za e-učenje
- Modul 6: Nastavni materijali na internetu – preoblikovanje, dijeljenje i vrednovanje
- Modul 7: Motivacija



Slika 5. Prikaz sadržaja iz tečaja "Uvod u e-učenje za nastavnike strukovnih škola"

- Modul 8: Grupni rad
- Modul 9: Moderiranje online forumskih diskusija
- Modul 10: Uloga tutora.

Jedna od osnovnih komponenti sadržaja tečaja je tablica za procjenu usvojenosti kompetencija, koja predstavlja svojevrstan alat za procjenu znanja u strukovnim predmetima. Nastavnicima bi takav alat omogućio lakše praćenje realizacije ishoda učenja kroz evidentiranje usvajanja svake pojedine kompetencije potrebne osobama koje se bave određenom strukom.

Svrha ovog tečaja je steći osnovna znanja o e-učenju i uvođenju tehnologije u nastavu strukovnih predmeta. Polaznici će prepoznati vlastite nastavničke kompetencije i planirati načine postizanja strukovnih kompetencija svojih učenika, razlikovati oblike e-učenja u kontekstu strukovnog obrazovanja, osmisliti i izraditi e-aktivnosti, organizirati grupni rad u online okruženju, planirati i voditi online diskusije te organizirati svoj rad kao tutora e-tečaja s temom iz strukovnog obrazovanja te vrednovati vlastite vještine.

VII. ZAKLJUČAK

Važna je promocija i upoznavanje nastavnika s mogućnostima e-učenja. Teme o tehnologijama i pedagoškim/metodološkim aspektima e-učenja potrebno je uvrstiti u programe usavršavanja nastavnika. Stručno i metodički pripremljeni programi pružaju mnogo korisnih sustavnih i temeljitih materijala i informacija u digitalnom obliku na jednom mjestu. Sadržaji se mogu pregledati više puta. Polaznici interdisciplinarnim programima obogaćuju svoje stručno usavršavanje i osposobljavaju se za kvalitetnije provođenje nacionalnog kurikuluma. Potrebna je velika samosvijest i motivacija polaznika u primjeni online programa koji su nadogradnja dosadašnjim načinima stručnog usavršavanja, stručnim predavanjima i radionicama.

Stručno usavršavanje za online nastavu omogućuje nastavnicima da postanu svjesni kako učenici doživljavaju e-učenje, a to najbolje mogu postići uključivanjem nastavnika u ulogu polaznika različith online tečajeva. Ovaj oblik iskustvenog učenja izaziva refleksiju na dubokoj razini i poboljšava praksu. Nastavnici pri tome rade na napredovanju svoje stručnosti, poboljšavaju svoje obrazovne i tehnološke vještine.

Programi e-učenja ispunjavaju potrebe polaznika za stručnim usavršavanjem. Omogućuju nadogradnju znanja, iskustava i vještina. Aktivnosti su usklađene s ciljevima programa i olakšavaju učenje. Produbljuje se razumijevanje teme i potiče razvoj kritičkog promišljanja o vlastitoj praksi. Polaznici su motivirani koristiti u praksi znanja i vještine stečene u programu, te imaju potrebu dijeliti ih sa svojim kolegama. Korisni su kontakti s drugim polaznicima i voditeljem/mentorom tijekom programa učenja. E-učenje omogućuje učinkovit pristup osobnom profesionalnom razvoju te kod polaznika pobuđuje interes da iskušaju program e-učenja u vlastitoj praksi.

Nedostaci koji bi sudionike mogli odvratiti od pohađanja online tečajeva su nedostatak vremena i motivacije, nedovoljno poznavanje tehnologije kao i nemogućnost kvalitetne komunikacije u tečajevima gdje je veliki broja polaznika. Potreban je visok stupanj aktivnosti i discipliniranosti polaznika kako ne bi odustali od pohađanja programa e-učenja. Važno je da imaju sposobnost brzog učenja kako se koristi nova tehnologija i prilagođavanja kod javljanja tehničkih problema.

LITERATURA

[1] Programi e-učenja AZOO-a, dostupni na https://loomen.carnet.hr/course/index.php?categoryid=537 putem prijava na portalu www.ettaedu.eu

[2] Evaluacija pilot provedbe pet novih programa e-učenja za stručno usavršavanje odgojno-obrazovnih radnika, Dr. Justina Erčulj i prof. dr. Paul Roeders, Agencija za odgoj i obrazovanje, ožujak 2014.

[3] Moodle MOOC, dostupno na http://www.carnet.hr/loomen/moodle_mooc

[4] Brezplačni Arnesov spletni tečaj (MOOC) o varni rabi interneta in sodobnih tehnologij, dostupno na http://www.arnes.si/obvestila/obvestilo/article/brezplacni-arnesov-spletni-tecaj-mooc-o-varni-rabi-interneta-in-sodobnih-tehnologij-pricetek.html

[5] Udruga hrvatskih učitelja razredne nastave Zvono, dostupno na http://udruga-zvono.weebly.com/

[6] Online tečajevi za nastavnike, dostupno na http://www.britishcouncil.hr/poduka-engleskog/tecajevi-online

[7] E-tečaj za e-mentore, dostupno na http://www.carnet.hr/onlinetecajevi/e-tecaj

[8] Uvod u e-učenje za nastavnike strukovnih škola, dostupno na http://evet2edu.eu/moodle/course/view.php?id=29

Impress.js - open source prezentacijski okvir u nastavi uvoda u web tehnologije

E. Ciriković *, D. Suk *
* Visoka škola Virovitica, Virovitica, Republika Hrvatska
enes.cirikovic@vsmti.hr, dominik.suk@vsmti.hr

Sažetak - U nastavi uvoda u web tehnologije često nije problem privući pažnju studenata atraktivnošću prezentacije samog sadržaja, međutim u pozadini svega stoji u izvjesnoj mjeri standardizirana jezična struktura koja podsjeća na programske jezike te kao takva studentima instinktivno djeluje upozoravajuće. Zato je u traženju motivacije potrebno pronaći ravnotežu između ponekad suhoparne suštine određene tematike i atraktivnosti te korisnosti konačnog ishoda učenja. U radu se prikazuje pregled java script biblioteke otvorenog koda, impress.js, orijentirane izradi web stranica te interaktivnih web prezentacija, kao motivacijskog alata pri svladavanju kolegija koji studente uvode u web tehnologije. Dobra ravnoteža između zahtjevnosti koda te mogućnosti potpuno personalizirane izrade web prezentacija te samih stranica, praktično se pokazalo kao povoljno u prikazu potencijala integracije zasebnih web tehnologija poput HTML-a, CSS-a i JavaScript-a.

I. UVOD

U nastavi uvoda u web tehnologije, vrlo važnu ulogu imaju HTML i CSS kao okosnice dizajna svih web stranica. Iako su navedene tehnologije izrazito vizualno orijentirane, često su i zahtjevne po pitanju količine pravila koje je potrebno usvojiti prije konkretnih rezultata. Tako svaka HTML web stranica prati standardnu početnu strukturu, na koju se potom nadopunjuje sav ostali kod iz kojega proizlazi ukupna i konačna realizacija svih želja dizajnera. Srećom, motivacija za usvajanje makar osnovnih principa i sintakse navedenih jezika pronalazi se u brojnim potpornim alatima za izradu i oblikovanje web stranica koji svaki na svoj način korisnicima pružaju kreativnu slobodu, u smislu izravne interakcije sa kodom, uz određene pogodnosti raspoloživosti automatiziranih funkcija u slučajevima neophodnih, često ponavljajućih linija koda**.** U skupu potpornih alata, osim komercijalnih rješenja, nalaze se i brojna korisna i zanimljiva rješenja otvorenog koda. Zastupljenost alata otvorenog koda svih vrsta i namjena bilježi znakovit rast te privlači pozornost sve šireg kruga korisnika, među kojima su i studenti, koji se s takvim rješenjima osim privatno susreću i u nastavi. Time se pruža vrlo dobra i prihvatljiva alternativa ustaljenim komercijalnim rješenjima, najčešće zatvorenih prema mogućnosti izravne korisničke interakcije s jezgrenim komponentama sustava te su upravo u tom smislu rješenja otvorenog koda bliža korisnicima od kojih su ustvari i potekla te zbog kojih opstaju i razvijaju se. Upravo takav alat je impress.js, razvijen kao JavaScript biblioteka, koji je kao suvremeni prezentacijski okvir nastao na temeljima CSS3 svojstava transformacija i tranzicija. Za razliku od ustaljenih prezentacijskih desktop alata, poput Microsoft Power Point-a ili Apache OpenOffice Impress-a, impress.js okvir pruža besplatnu izradu web prezentacija s mnoštvom nestandardnih i vizualno atraktivnih mogućnosti. Tako se koraci prezentacije, poznati i kao slajdovi u desktop prezentacijskom okruženju, mogu istovremeno pojavljivati s prikazom u trećoj dimenziji te s mnoštvom primijenjenih efekata nastalih na pravilima dizajna web stranica. Primjenom naprednih svojstava na postojeće CSS transformacije postiže se većina željenih efekata koji se do pojave ovoga okvira nisu mogle često vidjeti među prezentacijama. Tako se pomoću svojstava pozicioniranja prezentacijski elementi mogu smještati u željene dijelove okvira web preglednika, čime se ostvaruje dobra platforma za navigaciju između slajdova, odnosno prezentacijskih koraka. Obraćanje pozornosti na detalje unutar elemenata i općenito upravljanje veličinom prikaza objekata, tzv. skaliranjem, otvaraju se dodatne mogućnosti za pregled i prikaz slajdova, a rotacijom oko zadanih osi uvodi se nova dimenzija i interaktivnost prikaza. Na to se nadovezuje i mogućnost izrade web stranica koje iskorištavaju navedena prezentacijska svojstva i tako postižu dodanu vrijednost u smislu povoljnog odnosa uloženog truda i konačnog rezultata.

Najveća i najvažnija prednost ovog prezentacijskog okvira je njegova licenca koja podliježe MIT i GPLv2+ licenčnim uvjetima, a koja omogućuju besplatno i neograničeno korištenje cijelog okvira kao i slobodnu promjenu izvornog koda. Osim toga, impress.js se u odnosu na većinu današnjih komercijalnih prezentacijskih rješenja također ističe uporabljivošću te međuplatformskom kompatibilnošću. Tako je znatno jednostavnija mogućnost prezentacije i portabilnost materijala u odnosu na klasična desktop rješenja jer se prilikom pregleda prezentacijskog sadržaja ne zahtijevaju posebne prezentacijske aplikacije već je za tu svrhu dovoljan bilo koji web preglednik. Budući su web preglednici dio uobičajene aplikacijske podrške gotovo svih operacijskih sustava jasno je zašto impress.js odskače u smislu portabilnosti i dostupnosti aplikacijske podrške. S druge strane, mogućnost izbora tehnika prezentacije sadržaja znatno premašuje ponudu

mogućnosti desktop aplikacija, a za promjene sadržaja nije potrebno imati instalirane specijalizirane uređivače prezentacijskog sadržaja, već je potreban najosnovniji uređivač teksta kakav je obično prisutan u svim operacijskim sustavima.

II. OPIS IMPRESS.JS PREZENTACIJSKOG OKVIRA

Prije same uporabe potrebno je dohvatiti kopiju impress.js biblioteke smještene na lokaciji *https://github.com/bartaz/impress.js*. Na istoj lokaciji moguće je pronaći prateću dokumentaciju i podršku zajednice okupljene oko projekta te dodatna funkcionalna proširenja.

Da bi se impress.js prezenentacijski okvir mogao koristiti potrebno je iz prethodno raspakirane .zip datoteke izdvojiti datoteku impress.js koja upravo predstavlja biblioteku prezentacijskog okvira. Potom se raspakirana datoteka u stukturi HTML dokumenta navodi kroz referencu na lokaciji gdje je smještena i tako se uključuje u sam sadržaj učitanog web dokumenta. Preporuka je referencu ubaciti na kraj <body> sekcije HTML dokumenta kako bi se datoteka impress.js učitala nakon svih drugih elemenata strukture i dizajna.

```
<!DOCTYPE html>
<html>
<head>
      <title>impress</title>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
      </head>
<body>
<script src="impress.js"></script>
<script>impress().init();</script>
</body>
</html>
```

Slika 1. Inicijalizacija impress.js biblioteke u kontekstu osnovne strukture HTML dokumenta

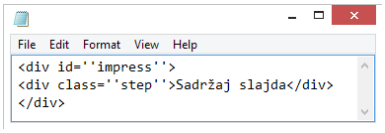
Nakon toga slijedi inicijalizacija impress.js biblioteke unosom sljedećeg koda:

<script>impress().init();**</script>**

Time se pokreće prikaz svih impress.js prezentacija prisutnih u HTML dokumentu s kojega se funkcija poziva. Navedena skripta smješta se nakon skripte za referenciranje na datoteku impress.js, kao u slici 1.

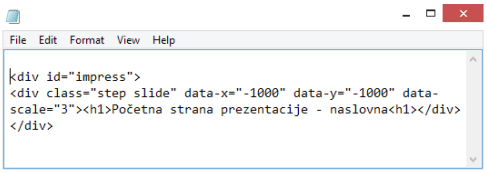
A. *Izrada prezentacija*

Za razliku od standardnih desktop prezentacijskih alata, impress.js zahtijeva prethodno planiranje razmještaja slajdova i općenito tijeka njihova prikaza jer se ovdje radi o neograničenom prostoru unutar kojega je smješten cjelokupan scenarij: slajd po slajd. Upravo taj moguće složen proces planiranja međuodnosa slajdova omogućuje veliku kreativnu slobodu te potiče studente, i općenito dizajnere, na razmišljanje o strukturi rada i planskom pristupu problemu. Za identifikaciju pojedinačnih komponenata prezentacije koriste se predefinirane ID oznake i klase, a sve korake prezentacije potrebno je smjestiti u roditeljsku sadržajnu sekciju (container) označenu ID oznakom ***impress***. Sadržaj pojedine prezentacije ili koraka smješta se unutar sadržajne sekcije s oznakom klase ***step***.



Slika 3. Definicija slajda unutar sadržajne sekcije ***#impress***

Na upravo stvoren slajd moguće je dodati osim samog sadržaja i efekte animacije, kroz HTML svojstva pozicioniranja (data-x, data-y i data-z), skaliranja (data-scale) te rotacije (data-rotate). Prikaz prezentacije ujedno je i prvi korak koji je prema nacrtu **sa slike X** pozicioniran u središtu zaslona te je nešto većih dimenzija u odnosu na ostale slajdove koji ga okružuju. Primjenom sljedećeg koda stvorit će se početni korak prezentacije iz primjera:

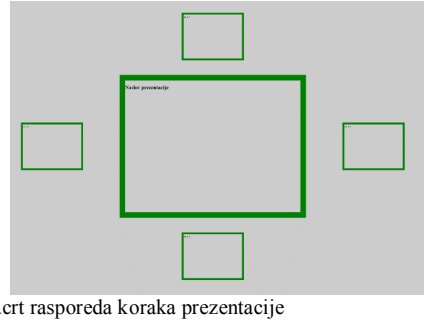


Slika 4. Definicija početnog slajda prezentacije iz primjera sa slike 2

U primjenenama gdje nisu eksplicitno navedene jedinstvene ID oznake sadržajne sekcije za pojedini slajd, za poziv slajdova kroz URL traku koriste se oznake #/step-N, gdje je N broj slajda koji se trenutno poziva na prikaz u prozoru web preglednika. Vodeći se istim principom, moguće je kreirati proizvoljan broj slajdova prezentacije uz potrebu obraćanja pažnje na redoslijed definiranja u kodu jer će se raspored kao takav pratiti i tijekom konačnog prikaza prezentacije.

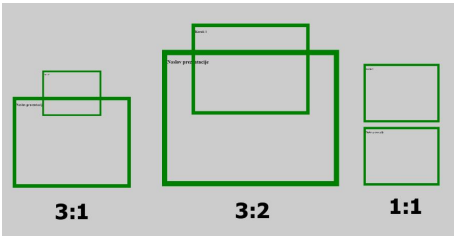
Sljedeći korak izrade prezentacije odnosi se na definiciju otiska ili razmještaja koraka na prezentacijskom prostoru zaslona. To se obavlja unosom *data* atributa *step* elemenata. Za potrebe prvog slajda korišteni su atributi *data-x*, *data-y* te *data-scale*, pa je tako dotični slajd pozicioniran na -1000 piksela na x-osi, te za isti iznos na y osi, što se uzima kao referentna točka ili ishodište za ostale korake prezentacije. Isti učinak na poziciju početnog slajda imala bi sljedeća definicija:

<div class="step slide" data-x="0" data-y="0" data-scale="3" **>**…



Slika 2. Nacrt rasporeda koraka prezentacije

Pretpostavljena vrijednost atributa *data-scale* za skaliranje prikaza slajdova je 1, a u primjeru početnog slajda sa slike 2, korišteno je trostruko skaliranje u odnosu na prethodno definirane dimenzije površine slajda. Time se također postiže povećanje u odnosu na ostale slajdove u prezentaciji i to relativno s obzirom na njihove vrijednosti skaliranja. Tako je moguće postići relativne odnose povećanja 3:2, 2:3 i bilo koji drugi racionalni odnos međusobnog prikaza u skaliranom prikazu koraka (slika 5).

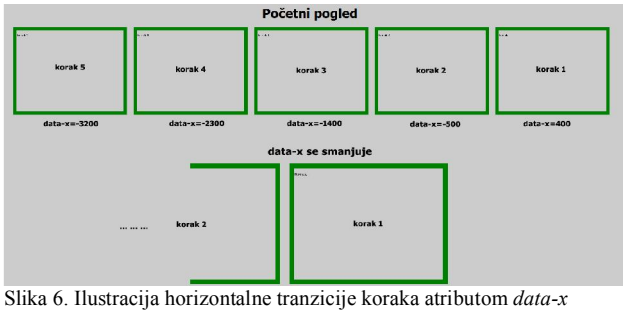


Slika 5. Prikaz različitih skaliranja i međusobnih odnosa dvaju slajdova

B. *Vizualni efekti*

Vizualni efekti ključne su komponente u izradi impress.js prezentacija, a zasnivaju se na temeljima CSS3 transformacija. Prostor efekata za izradu animacija u CSS3 vrlo je velik i u praksi se razlikuje ovisno o web preglednicima u kojima se interpretiraju. Zajednička osobina svih web preglednika praktično je neograničen prostor za izradu dizajna web stranica, no obično se od toga uzimaju dijelovi koji odgovaraju standardima prikaza ne različitim zaslonima. Uobičajena je upotreba grid-a širine 960 piksela neograničene visine, čime se automatski ograničava iskoristivost prostora iznad, lijevo i desno od pozicije grid-a**.** Impress.js, srećom, pruža mogućnost pozicioniranja elemenata izvan granica grid-a čime se ostvaruje bolja iskoristivost inherentnih mogućnosti prikaza web preglednika. Dodatna vrijednost nalazi se i u mogućnosti iskorištenja treće dimenzije što do sada nije bilo ostvarivo u klasičnim prikazima web stranica niti prezentacija.

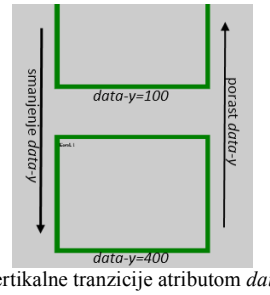
Horizontalno pozicioniranje elemenata prezentacijskog okvira postiže se primjenom atributa *data-x*, a učinak mu se ogleda efektom horizontalnog pomicanja slajdova u tranziciji i to u smjeru lijevo-prema-desno ili desno-prema-lijevo.



Slika 6. Ilustracija horizontalne tranzicije koraka atributom *data-x*

Na primjeru sa slike 6 prikazano je pet koraka svaki sa svojom pozicijom određenom atributom *data-x*. Budući su viši koraci smješteni lijevo od početnog, znači da im je vrijednost atributa *data-x* manja od vrijednosti koraka 1, što znači da se vrijednost atributa *data-x* smanjuje s desna na lijevo, odnosno povećava s lijeva na desno. U primjeru su istovremeno prikazani i ostali koraci za vrijeme prikaza trenutno aktivnog, no to je primjenom odgovarajućih CSS stilova moguće urediti prema želji. Tranzicijom iz koraka 1 u korak 2 dolazi do smanjenja vrijednosti parametra x te će se korak 2, s lijeve strane pomicati prema sredini zaslona dok će se korak 1 pomicati u desnu stranu kao sada već bivši korak. Pomicanje u suprotnom smjeru moguće je postići pridruživanjem većih vrijednosti parametra *data-x* koraku 2 i sljedećim višim koracima. Iz navedenog je moguće zaključiti kako bi se u slučaju istih vrijednosti parametra *data-x* susjedni koraci preklapali u svojem prikazu, što je svakako neprihvatljivo pa je potrebno voditi računa o pravilnom pozicioniranju koraka.

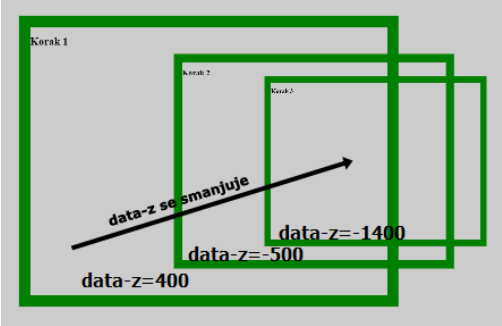
Koraci prezentacije pozicionirani po y-osi poravnavaju se vertikalno, jedan iznad, odnosno ispod drugoga. Primjenom atributa *data-y* određuju se položaji pojedinih koraka u vertikalnom smjeru, dok će se u postupku prijelaza između koraka pojaviti efekti pomicanja koraka odozdo prema gore ili odozgo prema dolje, ovisno o vrijednostima atributa pridruženih koracima.



Slika 7. Ilustracija vertikalne tranzicije atributom *data-y*

Proces vertikalne tranzicije koraka prema slučaju sa slike 7 provodit će se pomicanjem koraka 1, skupa s korakom 2 odozdo prema gore, a suprotan efekt bi se postigao primjenom manjih vrijednosti atributa *data-y* na korake iza koraka 1.

Svi dosadašnji efekti odvijali su se u ravnini x i y osi, a prelazak na trodimenzionalnu perspektivu započinje pozicioniranjem elemenata po z-osi, primjenom atributa *data-z*. Pogledom na ilustraciju pomicanja koraka po z-osi može se uočiti i promjena dimenzija prikaza pojedinih koraka prezentacije s obzirom na veličinu atributa *data-z*, što ujedno izravno utječe na iluziju treće dimenzije, tj. prostorne dubine.



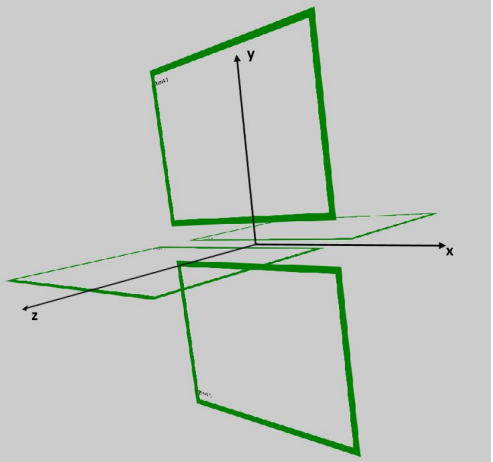
Slika 8. Ilustracija prostorne tranzicije atributom *data-z*

Iz slike 8 očigledno je da su koraci s većim vrijednostima atributa *data-z* prikazom istureniji, odnosno bliži gledatelju, dok su oni s manjim vrijednostima u odnosu na ostale korake udaljeniji i prikazom manji.

Pravu puninu u prostoru trodimenzionalnog dojma pružaju efekti rotacije i skaliranja koraka. Rotacija oko osi x moguća je primjenom atributa *data-rotate-x*.

Primjena efekata pozicioniranja po osi y i osi z, te rotacije oko osi x na četiri uzastopna slajda poredanih na koraku pregleda u kvadraturnom odnosu radi bolje preglednosti prikazana je slikom 9.

```
<div id="impress">
<div class="step slide" data-x="0" data-y="-400" data-z="0" data-rotate-x="0"><h2>Korak 1</h2></div>
<div class="step slide" data-x="0" data-y="0" data-z="-400" data-rotate-x="-90"><h2>Korak 2</h2></div>
<div class="step slide" data-x="0" data-y="400" data-z="0" data-rotate-x="-180"><h2>Korak 3</h2></div>
<div class="step slide" data-x="0" data-y="0" data-z="400" data-rotate-x="-270"><h2>Korak 4</h2></div>
<div id="overview" class="step" data-x="0" data-y="0" data-rotate-x="10" data-rotate-y="45" data-scale="2"></div>
</div>
```



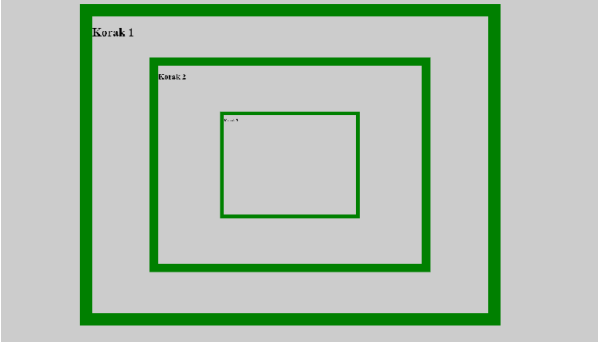
Slika 9. Ilustracija primjera rotacije koraka primjenom atributa *data-rotate-x*

Korištenjem negativnih vrijednosti kutova u primjeru sa slike 9 postignuta je rotacija koraka od dna prema vrhu zaslona, a za suprotan efekt potrebno je promijeniti predznake kutovima atributa *data-rotate-x*. U prikazu je predstavljena bočna perspektiva radi lakše predodžbe međusobnih odnosa između koraka prezentacije koji su, prema definiciji u kodu iznad ilustracije, u kvadraturnom odnosu, što znači da je svaki slajd okomito postavljen na susjedni. Prikazana perspektiva definirana je u posljednjem koraku prezentacije, tzv. koraku pregleda, s postavljenim nagibima ravnina *xy* za 45^0 te ravnine *xz* za 10^0.

Na sličan način postiže se rotacija oko osi y, čime je praktično moguće postići gotovo suprotne efekte onima pri rotaciji po osi x, dok kombinacija navedenih rotacija unosi dodatne prezentacijske mogućnosti uz nešto složenije početne postavke aritmetičkih odnosa.

Posljednja mogućnost rotacije elemenata ostvariva je atributom *data-rotate-z* ili samo *data-rotate* i kao takva ne unosi trodimenzionalne efekte u same prezentacije.
Za razliku od prethodnih efekata, skaliranje se svodi samo na dvije jednostavne funkcije: povećanje (tzv. *zoom-out*) i smanjenje (tzv. *zoom-in*). Skaliranje se definira atributom *data-scale*. Normalan prikaz elemenata prezentacije podrazumijeva skaliranje vrijednošću 1 te ju nije potrebno navoditi prilikom definicije takvih koraka. U primjeru sa slike 10 zadano je skaliranje na tri koraka prezentacije, i to s vrijednostima 1, 2 i 4, redom za korake 1, 2 i 3. Svakom koraku prezentacije pridijeljena je ista vrijednost atributa *data-x*, čime je postignuto poravnanje po središnjoj vertikalnoj osi. Kako bi se izbjeglo preklapanje sadržaja slajdova tijekom prikaza prezentacije, na svaki korak je primijenjena drugačija vrijednost atributa *data-y*.

```
<div id="impress">
<div class="step slide" data-x="400" data-y="300" data-scale="3"><h2>Korak 1</h2></div>
<div class="step slide" data-x="400" data-y="300" data-scale="2"><h2>Korak 2</h2></div>
<div class="step slide" data-x="400" data-y="300" data-scale="1"><h2>Korak 3</h2></div>
<div id="overview" class="step" data-x="400" data-y="300" data-scale="3"></div>
</div>
```

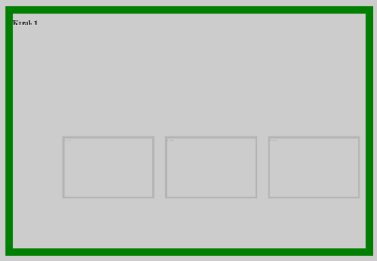


Slika 10. Prikaz efekta skaliranja na primjeru tri slajda

Na početnom koraku prezentacije prikazat će se sadržaj koraka 1 u punoj veličini, dok se ostali koraci neće vidjeti zbog većih iznosa skaliranja. U tranziciji između koraka 1 i 2 dolazi do približavanja koraka 1 prema gledatelju, čime se stvara efekt „*izlaženja iz zaslona*“, odnosno *zoom-out* efekt, a na kraju tranzicije ostaje prikaz sadržaja koraka 1 u koraku 2. Na sličan način odvijaju se i ostale tranzicije i prikazi koraka iz primjera sa slike 10. Bitno je naglasiti kako skaliranje ne podrazumijeva izravne odnose u dimenzijama susjednih elemenata, odnosno koraka, već isključivo na odnose u veličini elementa za koji se atribut *data-scale* određuje. Iako se efekt pozicioniranje može koristiti samostalno, bez primjene ostalih efekata, pravi smisao hijerarhijske organizacije koraka prezentacija postiže se tek primjenom odgovarajućih pozicioniranja koraka preostalim atributima za pozicioniranje. Time se prilikom prikaza određenog koraka postiže dojam podređenog odnosa trenutno neaktivnih koraka u odnosu na onaj koji se prikazuje. U posljednjem primjeru sa slike 11 to je

postignuto promjenom vrijednosti atributa *data-x* i *data*-y za svaki korak prezentacije, pa je tako dobiven efekt ugnježđivanja manjih (starijih) slajdova u veće (trenutno aktivne).

```
<div id="impress">
<div class="step slide" data-x="400" data-y="300" data-scale="4"><h2>Korak 1</h2></div>
<div class="step slide" data-x="-700" data-y="800" data-scale="1"><h2>Korak 2</h2></div>
<div class="step slide" data-x="700" data-y="800" data-scale="1"><h2>Korak 3</h2></div>
<div class="step slide" data-x="2100" data-y="800" data-scale="1"><h2>Korak 4</h2></div>
<div id="overview" class="step" data-x="400" data-y="300" data-scale="5"></div>
</div>
```

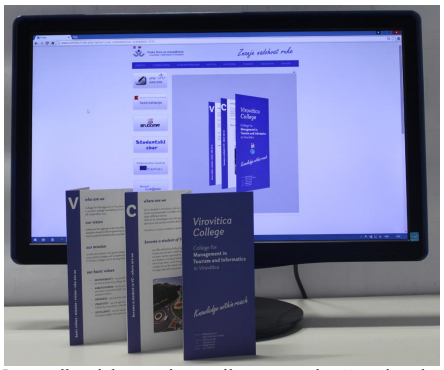


Slika 12. Ugnježđivanje koraka prezentacije primjenom skaliranja i pozicioniranja

Dodatan utjecaj na trodimenzionalne efekte u impress.js prezentacijskom okviru moguće je ostvariti atributom *data-perspective*. Ovim atributom određuje se udaljenost površine zaslona od ishodišta z-osi prezentacije, a utjecaj je kao takav primjenjiv samo na trodimenzionalne kompozicije koraka prezentacije. Osnovna vrijednost atributa *data-perspective* je 1000, dok se manjim, odnosno većim vrijednostima postavlja manja (elementi su bliže zaslonu), odnosno veća udaljenost zaslona od koraka prezentacije. Kako bi se isti efekt udaljenosti zaslona od z-osi istovremeno mogao primijeniti na sve korake prezentacije potrebno je atribut *data-perspective* primijeniti na element s ID oznakom #impress.

III. PRIMJENA U NASTAVI – OD PREZENTACIJE DO WEB STRANICE

Izrada promo brušure Visoke škole Virovitica bila je motivacija za studente da se tiskani proizvod upotpuni i web inačicom u sklopu vježbi iz kolegija E-Business. Tako je na temelju tiskanog predloška nastalo i web izdanje upotpunjeno, osim osnovnim HTML i CSS elementima, i efektima impress.js bilioteke. Usporedba tiskanog izvornika i web ekvivalenta prikazana je sljedećom slikom:



Slika 13. Usporedba tiskanog izvornika promo brošure i web ekvivalenta

Svaka stranica brošure predstavljena je jednim slajdom koji sadržajno i strukturno odgovara web stranici samoj za sebe. Web brošura sastoji se od ukupno jedanaest slajdova – web stranica, međusobno povezanih koracima naprijed-nazad, što je ostvareno navigacijom, posredstvom predefiniranih tipki „gore-dolje“, „lijevo-desno“ te tipke „space“. Redoslijed prikazivanja pojedinih stranica brošure određen je redoslijedom definiranja slajdova u sljedećem kodu:

```
<body>
<div id="impress">
<div id="overview" class="step" data-x="-1325" data-y="0" data-scale="0.9" data-z="600" data-rotate-y="-60" > </div>
<div  style="background-image: url(img/str0.png);" class="step slide" data-x="-940" data-y="0" data-z="-750" data-rotate-y="-150"> </div>
<div  style="background-image: url(img/str1.png);" class="step slide" data-x="-900" data-y="0" data-z="-600" data-rotate-y="0"> </div>
<div  style="background-image: url(img/str2.png);" class="step slide" data-x="-860" data-y="0" data-z="-450" data-rotate-y="-150"> </div>
<div  style="background-image: url(img/str3.png);" class="step slide" data-x="-820" data-y="0" data-z="-299" data-rotate-y="0"> </div>
<div  style="background-image: url(img/str4.png);" class="step slide" data-x="-780" data-y="0" data-z="-150" data-rotate-y="-150"> </div>
<div  style="background-image: url(img/str5.png);" class="step slide" data-x="-740" data-y="0" data-z="0" data-rotate-y="0"> </div>
<div  style="background-image: url(img/str6.png);" class="step slide" data-x="-740" data-y="0" data-z="-1" data-rotate-y="180"> </div>
<div  style="background-image: url(img/str7.png);" class="step slide"  data-x="-780" data-y="0" data-z="-150" data-rotate-y="30"> </div>
<div  style="background-image: url(img/str8.png);" class="step slide" data-x="-820" data-y="0" data-z="-300" data-rotate-y="180"> </div>
<div  style="background-image: url(img/str9.png);" class="step slide" data-x="-860" data-y="0" data-z="-450" data-rotate-y="30"> </div>
<div  style="background-image: url(img/str10.png);" class="step slide" data-x="-900" data-y="0" data-z="-601" data-rotate-y="180"> </div>
<div  style="background-image: url(img/str11.png);" class="step slide" data-x="-940" data-y="0" data-z="-750" data-rotate-y="30"> </div>
</div>
<script src="js/impress.js"> </script>
<script>impress().init(); </script>
</body>
```

Efekti prijelaza između slajdova ostvareni su primjenom atributa *data-x*, *data-y*, *data-z* te *data-rotate*. Radi preglednosti, izostavljene su definicije stilova oblikovanja slajdova, te je prikazan samo sadržaj HTML elementa <body>.

Za uspješnu primjenu navedenih svojstava bilo je potrebno prethodno odrediti apsolutne pozicije pojedinih slajdova te relativne odnose između njih, što je zahtijevalo osnovne trigonometrijske vještine i snalaženje u 3D prostoru. Cjelokupan dizajn brošure, odnosno, pojedinih stranica, određen je na temelju tiskanog izvornika, primjenom svojstava CSS-a. Po završetku projekta, brošura je objavljena na službenim stranicama Visoke škole Virovitica kao dio standardnog Joomla! članka.

IV. ZAKLJUČAK

Prikazom pregleda osnovnih svojstava jedne male biblioteke moćnog jezika JavaScript te pristupačnosti praktičnih mogućnosti koje početnicima entuzijastima pruža impress.js, otvorena je prilika za polivalentnu primjenu od desktop prezentacija do web stranica, bez ograničenih uvjeta korištenja.

Isto se pokazalo i tijekom studentskih vježbi iz uvoda u web dizajn, te je upravo jednostavnost integracije s

neophodnim elementima HTML-a i CSS-a bio poticaj za savladavanje učenja osnova navedenih jezika i motivacija za daljnje istraživanje područja brzo rastućih trendova i potpornih tehnologija. To svakako nije sve što se u ovome trenutku nalazi na tržištu web jezika i pratećih alata otvorenog koda, ali pruža uvid u snagu i potencijale modernih jezika orijentiranih na web i potiče studente i ostale korisnike na istraživanje i primjenu do sada neistraženog ili manje poznatog.

LITERATURA

[1] R. N. Ratnayake, "Building Impressive Presentations with impress.js, Packt Publishing, Birmingham", UK, 2013.

[2] M. Jurčić, M. Horvat, "Nastavna sredstva i nastavna pomagala u predmetnom kurikulu informatike", Školski vjesnik, Vol. 63, No. 1-2, str. 71-89, 2014.

[3] Andi Wahju Rahardjo Emanuel, "Statistical Analysis of Popular Open Source Software Projects and Their Communities", Proceedings of 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE 2014), Yogyakarta, 7-8 October 2014

[4] D. R. Smith, "An Exploration of HTML5 and Javascript - Building a Presentation Engine", A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science (Honours), Rhodes University, 2012.

[5] J. A. Holm, "3 open source tools to make your presentations pop", July 2014, http://opensource.com/life/14/7/3-open-source-tools-make-your-presentations-pop (14.1.2015.)

[6] A. Morrison, "Technology Tools For The Classroom: Presentation Resources", April 2011, http://www.pearsonschoolsystems.com/blog/?p=295#sthash.hGhfSDnq.XsxcLvz3.dpbs (14.1.2015.)

[7] https://github.com/bartaz/impress.js/ (14.1.2015.)

CTS

International Conference on

COMPUTERS IN TECHNICAL SYSTEMS

Steering Committee

Chairs: Vlado Sruk, University of Zagreb, Croatia
Leonardo Jelenković, University of Zagreb, Croatia
Dejan Škvorc, University of Zagreb, Croatia
Jadranko F. Novak, Nomen, Rijeka, Croatia
Nikola Mišković, University of Zagreb, Croatia

Members: Matjaž Colnarič, University of Maribor, Slovenia
Željko Hocenski, University of Osijek, Croatia
Niko Guid, University of Maribor, Slovenia

PAPERS

Design of a domain specific language and IDE for Internet of things applications

A.Salihbegovic*, T. Eterovic * , E. Kaljic* and S.Ribic*
* Faculty of Electrical engineering, University of Sarajevo, Bosnia Herzegovina
asalihbegovic@etf.unsa.ba , teo.eterovic@gmail.com, ekaljic@etf.unsa.ba, sribic@etf.unsa.ba

Abstract - With the goal of relieving the Internet of things (IoT)application designers of the complexities and heterogeneity of wireless sensory networks (WSNs), devices, communication media, protocols and operating systems, the development of higher level domain specific language has been undertaken. The result is DSL-4-IoT Editor- Designer, which is based on high level visual programming language, established on the class of visual domain specific modeling languages (VDSMLs). DSL-4-IoT is using formal presentations and abstract syntax in a metamodel. The visual front-end of the Editor has been developed in JavaScript language. The runtime execution of generated IoT application configuration files is done by the open source project "OpenHAB" runtime engine.
In order to demonstrate the viability and usability of the developed DSL-4-IoT visual model based language, an experimental IoT testbed including 15 heterogeneous wireless sensory devices spanning two application domain (smart home and remote patient monitoring), has been designed and deployed.

I. INTRODUCTION

Application development challenges, diversity of devices (physical variables sensors, biometrics sensors), different software implementation (Android, iOS, Linux), different interaction modes (pub/sub, request/response, command) and different engineering units (deg. C, deg. F, bar, N, etc.) are the ecosystem in which the designers of The Internet of Things (IoT) applications are creating their projects[1].

Designers of IoT applications should be relieved of most of these heterogeneity and specifics (wide range of hardware and software entities running on specific platforms, middleware specific features). They need to use integrated development environment (IDE) based on domain specific high level language which in its entities would abstract most of these intricacies and specifics in hardware, software, communication media and protocols. Furthermore, the language should be capable to support large scale design of these systems by combining several design blocks and saving them in application libraries, importing them, reconfiguring and parameter setting with ease for new tags and locations, enabling reusability and scalability [2].

II. DOMAIN SPECIFIC LANGUAGES IN IOT APPLICATIONS DEVELOPMENT

In this area of very intensive research and design, a short overview of the state of the art and approaches to empower IoT designers with higher level languages for their design and deployment is presented.

A. *PervML Language*

PervML is a Domain Specific Language (DSL) designed with the aim of providing the pervasive IoT system developers with a set of constructs that allow them to precisely describe the pervasive system in a technology independent way [3]. PervML promotes the separation of roles where IoT application developers can be categorized as system analysts and system architects.

B. *DiaSuite: a Tool Suite to Develop/Compute/Control Applications Language*

DiaSuite focuses on a specific domain, namely Sense/Compute/Control (SCC) applications [4]. It comprises of:

- Domain-specific design language,
- Compiler producing a Java programming framework,
- 2D-renderer to simulate an application,
- Deployment framework.

DiaSuite provides a design language, named DiaSpec that is dedicated to the SCC paradigm. The design language consists of two layers: taxonomy layer and application design layer.

The taxonomy layer allows a class of entities to be described. An entity is defined as a set of data sources and actuating capabilities, abstracting over devices, whether hardware or software.

Application Design Layer of DiaSpec allows the application logic to be decomposed into context and control operators. Sensors and actuators are the two facets of an entity described in the taxonomy.

C. *Node-RED*

Node-RED open source IBM Emerging Technology software is visual tool for wiring the Internet of things application configurations. It enables the designer to create IoT applications wiring together hardware devices, APIs and online services. Node-RED provides a browser based flow editor that supports wiring the flows using the wide range of sensory nodes in the palette. Created flows can then be deployed and executed in runtime with a single mouse click.

The paper has been prepared and written on the results achieved within the project that was financially supported by the Federal ministry for education and science, Federation of Bosnia Herzegovina.

Runtime engine is built on Node.js, making use of its event-driven, non-blocking model. Since this runtime module is of small footprint, it can run on low-cost hardware such as Arduino or Raspberry Pi as well as in cloud [5].

D. OpenHAB

OpenHAB represents general-purpose framework for smart home gateways and is fully based on the Eclipse SmartHome project, which allows building smart home solutions as the most common application area of IoT. The framework consists of a set of Java OSGi bundles that can be deployed on an OSGi runtime and which defines OSGi services as extension points, acting as middleware on the sever side of WSN (wireless sensory) network [6].

The project focuses on services and APIs for the following topics [7]:

- *Data Handling*: This includes a basic but extensible type system for smart home data and commands that provides a common ground for an abstracted data and device access, as well as event mechanisms to send this information around. Integration with other systems is done through so called bindings, which are a special type of extension.
- *Declarative User Interfaces*: A framework with extensions for describing user interface content in a declarative way. This includes widgets, icons, charts etc.
- *Rule Engines*: A flexible rule engine that allows changing rules during runtime and which defines extension types that allow breaking down rules into smaller pieces like triggers, actions, logic modules and templates.
- *Persistence Management*: Infrastructure that allows automatic data processing based on a simple and unified configuration persistence services are pluggable extensions, which can be anything from a log writer to an IoT cloud service.

Beside the runtime framework and implementation, Eclipse SmartHome has also its text Editors for editing configuration models and rules.

III. STRUCTURE AND FUNCTIONALITY OF DSL-4-IOT LANGUAGE

The vast versatility of the types of sensors and actuators as well as heterogeneity of communication media and protocols that are immanent when the standards are lagging behind, are facing design engineer within the domains of Internet of things applications. Understandably, all this is posing burden on designer to acquire knowledge and learn many specifics that are far beyond the functionality and application requirements that he/she wants to build into the system.

Therefore, there are many efforts going on among researchers and developers to develop high-level domain specific language that will abstract many of these specifics and peculiarities into building blocks and library modules. Then, he/she is designing its application, specifying the devices and hardware needed for designed system acquisition and deployment on the application site [8].

In this research project, we have opted for the development of one visual domain specific language that shall empower IoT application designer with a visual high level language and based on it, a design Editor named DSL-4-IoT. The Editor shall enable application designer to configure the system structure, select devices, sensors and actuators either from built-in library modules available in Designer itself, or enter the ones that he/she selects, parametrizing the metadata available from manufacturer's design literature. These device modules can then be saved and included in built-in library modules for future use, although for these later actions, more specific knowledge is required, and shall be accomplished by someone that shall be named DSL-4-IoT developer.

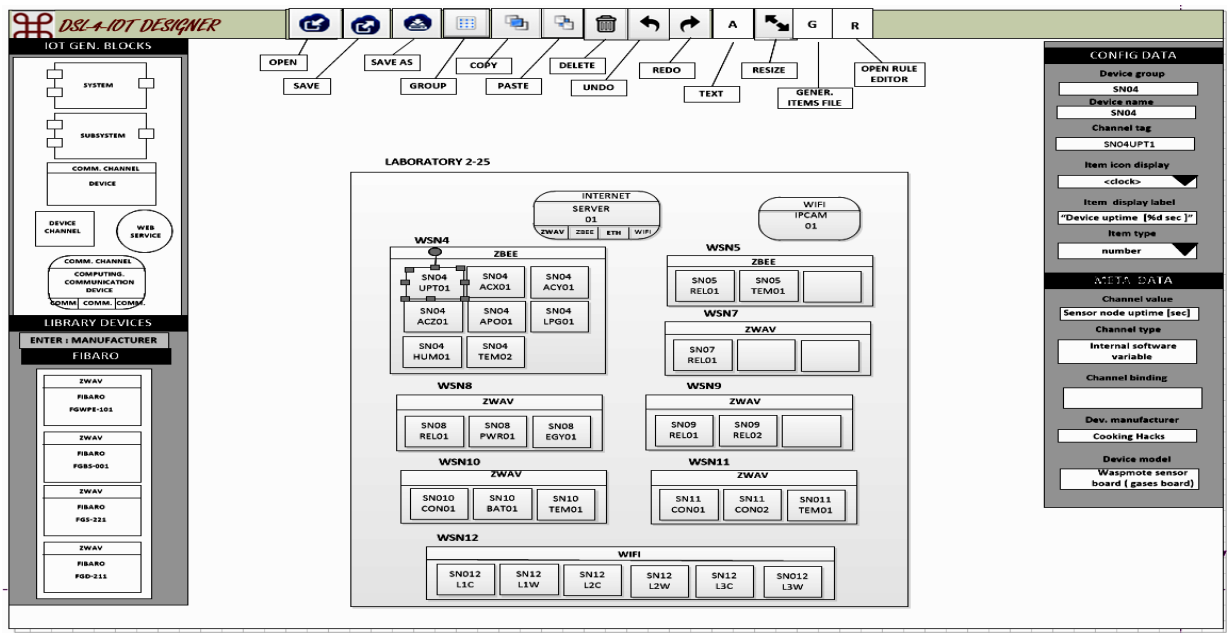


Fig.1 The layout of DSL-4-IoT Editor

The development of the DSL-4-IoT language is based on the class of visual domain-specific modeling languages (VDSMLs) that are using formal presentations and abstract syntax in a metamodel. The front end of the Editor has been developed in JavaScript. The layout of DSL-4-IoT Editor-Designer is shown on Fig. 1.

In the command bar, the Editor is providing the necessary tools for opening and saving the applications configuration, together with additional commands for manipulation of individual or groups of block structures (group, copy, paste, delete, undo, redo, resize, etc.).

Labeling and commenting of the block structures on the hierarchical level of generic blocks (system, subsystem, devices, device channels, etc.) is enabled using the "Text" button.

Two buttons further right, are intended for generation of textual configuration files with extension "items", and "sitemap" that are used directly in the runtime execution.

The last button is used to call an external program named "Rule editor", that enables the Designer to configure any combinatorial, sequential or temporal relationships between "items" created in previous step of entering configuration program. These „items" are the logical representation of physical channels of the configured devices (sensor inputs or actuator or command outputs), or variables associated with messages obtained via web services, SMS messages or emails. In broader terms, items are readable or write-able objects, in order to interact with them. Items can be bound to bindings, i.e. for reading the value or status from device channel input or for updating output channel going to some actuator device. More on this can be found in documentation on OpenHAB project [7].

Bellow the higher abstract layer of DSL-4-IoT domain specific language, additional glue code that is running in Openhab runtime is executing:

- Interfacing software components and middleware

- Interfacing with disparate hardware and software components.

- Mapping code for device and software components

This glue code running in runtime and executing DSL-4-IoT tools developed configuration files, has been selected as Java based OpenHAB 1.x open source project, widely popular in the IoT community.

In order to abstract mentioned heterogeneity of various IoT devices, sensing and actuating physical devices are described as entities in high level manner. These resources are instantiated as many times as necessary with the distinct tag names, to represent all physical and/virtual devices included in the particular project application.

To address scalable operations within the IoT system application logic that connects "items" in interconnected wireless sensory networks of devices and users, hierarchical clustering has to be specified consisting of:

- System
- Subsystem
- Device (autonomous and powered)
- Physical or virtual channel

Use of this hierarchical structure enables: project partitioning at physical level; project partitioning at functional level; defining scope from which software components shall produce or consume data. This multi-stage model driven approach for design of IoT applications supports all stages of the life cycle of these systems. Selected VDSML, model driven approach applied in development of DSL-4-IoT visual design environment, contributes to ongoing search for design tools that shall allow IoT application designers to model their systems at a higher abstraction level, providing them with automatic model transformations to incrementally refine abstract models into more concrete ones.

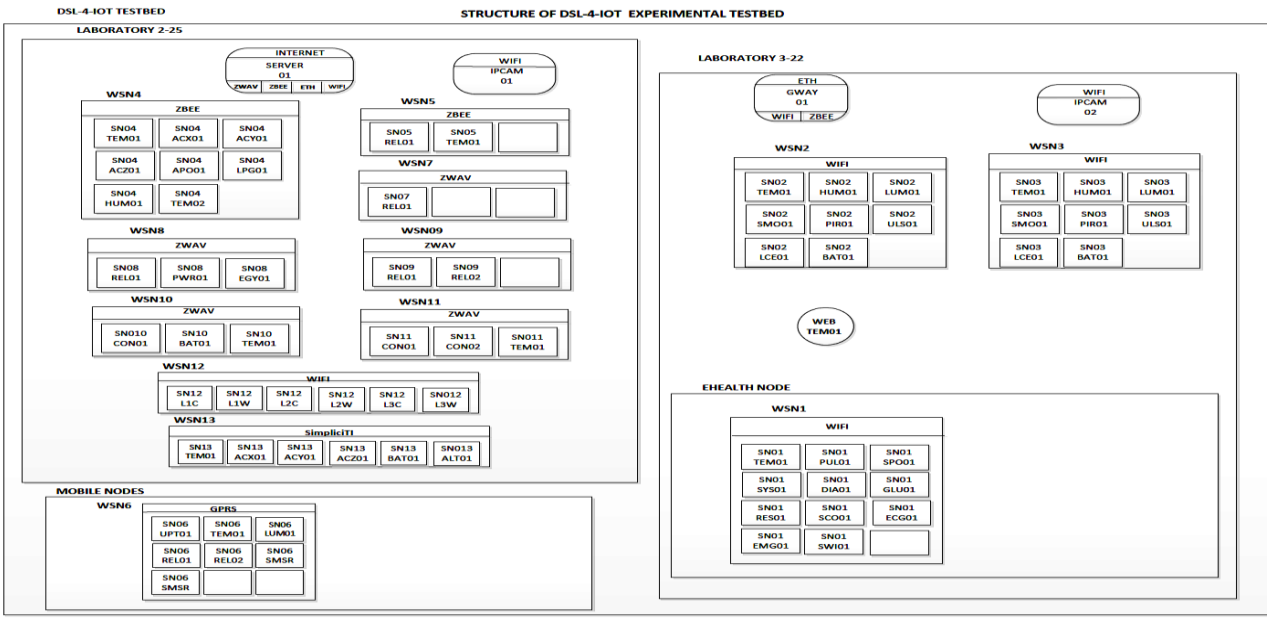


Fig. 2 The structure of experimental testbed generated with DSL-4-IoT Editor

In order to demonstrate the viability and usability of above described DSL-4-IoT model based language and tools, an experimental IoT testbed spanning two domain applications (smart home and remote patient monitoring), has been designed and deployed using the DSL-4-IoT language. The output configuration generated as the result of entering the structure of the experimental testbed is displayed on the Fig. 2.

DSL-4-IoT compiler within Designer is mainly composed of a parser and configuration files generator.

Configuration files generator (items, rules, sitemap) generates formatted text outputs that can be directly downloaded and executed within OpenHAB runtime engine.

Both tools rely on declarative domain-specific language used to graphically represent IoT system topology and "items" embedded in this topology.

Runtime libraries are JavaScript coded based on OpenHAB runtime engine used to execute application with generated configuration files.

DSL-4-IoT for configured IoT system structure, exports the data into one JSON array. This array keeps information about position of all the items within configuration, relationships between items and groups, value of all configured fields associated with items and of data types. JSON configuration array includes OpenHAB specific fields and associated rules. This array is typically huge in size, and is useful as container from which it is possible to restore the state for further editing and setting up of sensory network configuration.

From the configuration stored in JSON array format, it is possible to export data to wider scope of integration platforms for home automation, one of them being OpenHAB. OpenHAB runtime engine requires five kinds of configuration files: Items, Rules, Persistence, Scripts and Sitemaps. These text files differ by file extension and internal structure.

At the time of writing this paper, DSL-4-IoT Designer with visual Editor, parser and configuration files generator for two configuration files: items and sitemaps has been completed and tested, while development of visual Editor for rules configuration is underway.

Items in OpenHAB 1.x runtime can be defined using configuration files located in subfolder configurations/items, relative to OpenHAB installation directory. Item definition files have the file extension *.items. Groups are also defined in the *.items file.

A special conversion program scans the JSON document generated from the visual DSL-4-IoT configuration Editor. Each cell in this document, which is of type iot.device, becomes item, and cells which are recognized as embedding cells, became a group. The items are stored inside groups.

Sitemaps are used to create elements of a user interface. They are stored in OpenHAB runtime directory configurations/sitemaps. Each file has *.sitemap extension. The structure is similar to above described items configuration file. The very first line has an element called Sitemap, and it is followed in next lines by some of elements: Colorpicker, Chart, Frame, Group, Image, List, Selection, Setpoint, Slider, Switch, Text, Video, and Webview [7]. All of them, except Frame, describe user interface elements in single line.

The file which defines rules for IoT application configuration has extension *.rules. A rule file can contain multiple rules, and consists of three sections:

Imports, Variable Declarations and Rules.

The Imports section contains import statement, used to make the imported types available without need of fully qualified name for them. The Variable declaration section declares global variables (with or without initial values), accessible to all rules in this file. The Rules section contains a list of rules. After configuration files are generated, they can be transferred to the respective OpenHAB runtime directory manually or automatically downloaded, using simple web service. Manual transfers are carried out using Blob interface and download links [9].

Configuration files generator within DSL-4-IoT Designer creates these files in broader a more generic way, hence these configuration files are not confined only to OpenHAB runtime for deployment and execution. Since for development and coding of DSL-4-IoT Designer functionality, the interpreted language of JavaScript is used, it is possible to simply replace corresponding JavaScript file to generate and export configuration files for other IoT integration platforms and their runtime engines.

IV. EXPERIMENTAL TESTBED FOR PROOF OF DEVELOPED IDE

For proof of concept and versatility of developed visual domain specific language and development environment based on DSL-4-IoT Designer, the experimental testbed has been built and implemented composed of the variety of wireless static and mobile sensory nodes with multiplicity of sensors, actuators, communication channels and protocols from various manufacturers. From the point of view of IoT application domains, two most frequent domains were chosen to be presented, by selection of the types of sensors and their functionality: smart home and remote patient monitoring.

Sensors and actuators used in build-up of experimental testbed have been either autonomous with integral communication transceiver channel, or hooked up to large number of wireless devices (nodes), based on different communication technologies (Wi-Fi, ZigBee, Z-wave, Bluetooth, NFC) and protocols (HTTP, TCP/IP, UDP, MQTT, Modbus, SimpliciTI, serial protocols, etc.).

The topology of experimental testbed wireless sensor network is depicted on the following Fig. 3:

Experimental testbed is physically divided in several domains:

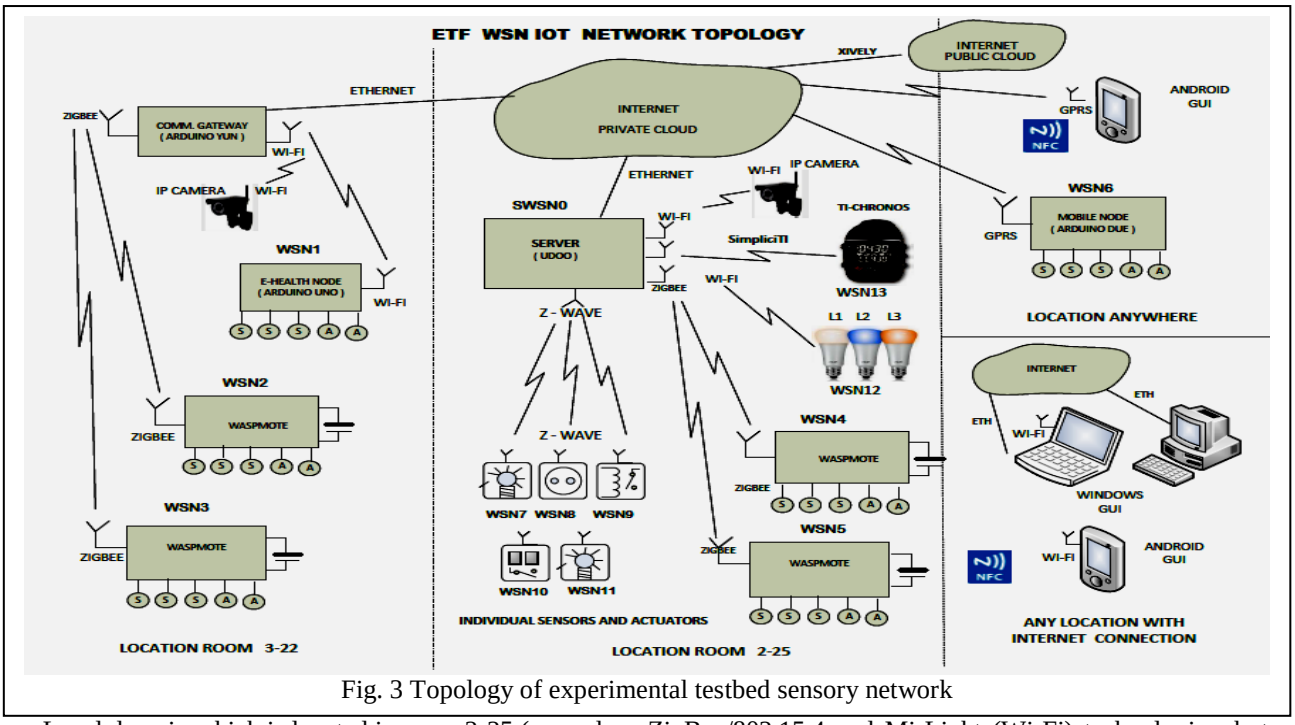


Fig. 3 Topology of experimental testbed sensory network

- Local domain which is located in room 2-25 (second floor within faculty building),

- Remote domain which is located in room 3-22 (third floor within faculty building)

- Mobile domain that can be located anywhere within the reach of mobile network signal.

Local domain (room 2-25) consists of several wireless sensor networks implemented using a variety of wireless accessing technologies (e.g. Wi-Fi, ZigBee, Z-Wave, etc.) and core server (SWSN0).

Z-Wave sensor network consists of a light dimmer (WSN7), window sensors (WSN10), relays (WSN9), outlet with electrical consumption metering (WSN8), and multi-purpose binary sensor (WSN11) connected to the door (open-close) and knock-knock sensor. ZigBee/ 802.15.4 wireless sensor network is implemented using Libelium Waspmote nodes (WSN4 & WSN5) [10]. On the input of the Waspmote nodes are connected various analogue and digital sensors from smart home variety. The sensors perform sensing and measurement of a large number of environmental parameters (e.g. temperature, humidity, luminosity, presence, position, proximity, air pollution with various pollutants, movement, weight, opened, closed, flood, etc.). They are capable, as well, for actuating and controlling functions (switching ON and OFF various electrical devices like lights, fans, heaters, solenoid valves, opening and closing or positioning in any intermittent position blinds, windows, doors, etc.). For control of LED lights including the RGB spectrum and intensity, the Mi-Light sensor network that consists of three RGB light bulbs is used in (WSN12) [11].

Core server (SWSN0) is built on Arduino compatible UDOO module based on embedded computer with Freescale i.MX 6 ARM Cortex-A9 1 GHz quad-core processor and 1 GB of RAM. UDOO server is connected to the communication modules for Z-Wave, ZigBee/802.15.4 and Mi-Light (Wi-Fi) technologies, but also includes the communication modules that make it also a gateway for non-IP devices. UDOO based server executes open source OpenHAB runtime software which is based on the OSGi container. To support the connected communication module protocols, OSGi plug-ins called bindings are used. For Z-Wave and Mi-Light there were already available bindings which have to be included as add-ons. However, for ZigBee/802.15.4 there was no binding readily available, so the support was provided within the Project implementation, developing a dedicated Java application that uses OpenHAB REST API on one side (server), and a serial connection to the ZigBee/802.15.4 module on the other side.

To test and experiment with wearable devices as IoT nodes, one Texas Instrument wristwatch model EZ430-Chronos was integrated with the UDOO server using proprietary SimpliciTI protocol over wireless RF link [12]. This wearable wireless sensor node offers the data on temperature, altimeter (absolute pressure), and three-axis accelerometer.

Remote domain, located in room 3-22, consists of ZigBee wireless sensor nodes (WSN2 & WSN3), of the same type as in the room 2-25 (Libelium Waspmote), Libelium e-Health wireless node (WSN1) and communication gateway (implemented on Arduino Yun embedded processor module). E-Health wireless node is based on the Arduino Uno board, e-Health sensor shield and Wi-Fi module [13]. The e-Health sensor shield empowers biometric and medical applications where human patient local and remote monitoring is needed, using 10 different biomedical sensors. They are: pulse (heartbeat rate), oxygen in blood (SPO2), airflow (breathing), body temperature, electrocardiogram (ECG), glucose meter, galvanic skin response (GSR-sweating), blood pressure (sphygmomanometer), patient position (accelerometer) and muscle/electromyography sensor (EMG). To enable connection of wireless nodes from a

remote domain in room 3-22 to the server in the local domain (room 2-25), the communication gateway is used. Communication gateway is implemented using Arduino Yún with Wi-Fi and ZigBee/802.15.4 communication modules [14].

Mobile domain is based on the GPRS mobile node, and Android based mobile phones and tablets. GPRS mobile node is implemented using Arduino Due board and GPRS shield. For communication between GRPS mobile node and server, OpenHAB REST API was used, utilizing my.openHAB cloud service [15].

Local and remote domains are covered by video surveillance serving as visual feedback of the activities performed in these two rooms, implemented with IR Led night vision IP cameras. Video streams from IP cameras are taken in MJPEG format and presented to the user inside openHAB web interface on desktop and laptop computers, or HABDroid application on Android based mobile phones [16].

User interface that provides monitoring as well as full control of all devices is provided as web application in any web browser that has JavaScript runtime engine included, connecting to IP address of the server (SWSN0). This interface can be accessed either on desktop or laptop computer or any smart phone or tablet with Android or Apple iOS Operating system, downloading the utility program.

V. CONCLUSION

The development of IoT application design environment based on the high level visual programming language in form of Editor within the class of visual domain specific modeling languages (VDSML) using formal presentations and abstract syntax, has been presented in this paper. The front end of the Editor is based on JavaScript language.

The runtime execution of IoT application configuration files, generated with this Editor is carried out by OpenHAB runtime engine developed within the framework of "OpenHAB" open source project.

Work in progress is development of visual rule editor based on Google Blockly blocks for building web based Editors [17]. The entered rules structure shall be converted to text file, compatible with *.rules file within OpenHAB runtime engine. Chris Jackson also announces similar work for OpenHAB project [18], and there is specific version for Vesternet smart home applications as Zipato rule creator [19].

Further extension in the development of IDE is planned on adding new IoT devices into DSL-4-IoT library, which is activity done by IoT developer rather than by IoT application designer. This shall remove the burden from IoT application designers to add new devices not available in the library, and together with devices dynamic discovery feature [20], shall fully automate application library extension in future applications.

REFERENCES

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, vol. 29, pp. 1645-1660, 2013.

[2] J. Clerk J. Kiljander, J. Takalo-Mattila, M. Etelapera, J.-P. Soininen, and K. Keinanen. Enabling end-users to configure smart environments. In Applications and the Internet SAINT), 2011 IEEE/IPSJ 11th International Symposium on, pages 303–308, 2011.

[3] T. S. Lopez, D. C. Ranasinghe, M. Harrison, D. McFarlane, "Adding sense to the Internet of Things An architecture framework for smart object systems," Personal and Ubiquitous Computing, vol. 16, no. 3, pp. 291-308, 2012

[4] Benjamin Bertran, Julien Bruneau, Damien Cassou, Nicolas Loriant, Emilie Balland, et al.. ,"DiaSuite: a Tool Suite To Develop Sense/Compute/Control Applications. Science of Computer Programming", Fourth special issue on Experimental Software and Toolkits, Elsevier, 2012, <10.1016/j.scico.2012.04.001>

[5] Node- RED: Visual tool for writing the Internet of Things, http://nodered.org/ [Accessed on: 10.1.2015]

[6] Eclipse smarthome: A Flexible Framework for the smart home, https://www.eclipse.org/smarthome/ / [Accessed on: 10.1.2015]

[7] OpenHAB : open source project web site, http://www.openhab.org/ / [Accessed on: 10.1.2015]

[8] C. Perera, P. Jayaraman, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Dynamic configuration of sensors using mobile sensor hub in internet of things paradigm. In IEEE 8th International Conference on Intelligent Sensors, Sensor Networks, and Information Processing.

[9] Blob JavaScript API , http://www.javascripture.com/Blob [Accessed on: 10.1.2015]

[10] Waspmote: Libelium wireless sensory network , http://www.libelium.com/products/waspmote/ [Accessed on: 10.1.2015]

[11] Houselogix, Mi-Light Control4 https://www.houselogix.com/shop/limitlessled-milight-easybulb [Accessed on: 10.1.2015]

[12] I. S. SimpliciTI™ - RF software protocol - Texas Instruments http://www.ti.com/corp/docs/landing/simpliciTI/index.htm?DCMP=hpa_rf_general&HQS=NotApplicable+OT+simpliciTI, [Accessed on: 10.1.2015]

[13] e-Health Sensor Platform for Biometric and Medical applications http://www.libelium.com/130220224710/ , [Accessed on: 10.1.2015]

[14] Guide to Arduino Yun , http://arduino.cc/en/Guide/ArduinoYun [Accessed on: 10.1.2015]

[15] My.openHAB , https://my.openhab.org/ , [Accessed on: 10.1.2015]

[16] Y. HABDroid – Android application for openHAB , https://github.com/openhab/openhab/wiki/HABDroid, [Accessed on: 10.1.2015]

[17] Google Developers: Blockly : Library for building visual programming editors, https://developers.google.com/blockly/ [Accessed on: 10.1.2015]

[18] Chris Jackson : Rule designer : Overview, : https://github.com/cdjackson/HABmin/wiki/Rule-Designer%3A-Overview , [Accessed on: 10.1.2015]

[19] Vesternet : basic Guide to the Zipato rule creator, : http://www.vesternet.com/resources/application-notes/apnt-9#.VLzt-0fF_A0 , [Accessed on: 10.1.2015]

[20]] Charith Perera, Prem Prakash Jayaraman, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos," Context-aware Dynamic Discovery and Configuration of 'Things' in Smart environments, , Springer Berlin Heidelberg, Volume 546, Pages 215-241 (2014)

Basic PHP implementations, opcodes, and internal work

dr. sc. Alen Šimec[1], Davor Lozić[2]
[1] Polytechnic of Zagreb, Informatics, Zagreb, Croatia,
[2] Polytechnic of Zagreb, Informatics, Zagreb, Croatia,
alen@tvz.hr, dlozic@tvz.hr

Abstract - The main purpose of this paper is to investigate how PHP works internally and how variables, arrays, objects, and other types are stored in memory. The paper will cover generated series of operation codes, commonly known as "opcodes", and why advanced PHP programmers should know the basics of PHP "under-the-hood". Additionally, it aims to show how function calls and classes work internally, basic concepts such as "pre-increment or post-increment speed comparison", and minor optimization recommendations.

I. INTRODUCTION

PHP is a server side, scripting languge created mostly for web development. It can be integrated with HTML code or it can be used with templating engines or web frameworks.

PHP source code is processed by an interpreter so having a deeper understanding about internals could be crucial for scripts which are executes by many users. When parsing PHP files, the engine generates a series of operation codes, commonly known as “opcodes” so it’s valuable to understand the meaning of basic opcodes.

PHP is written in the C programming language so it’s a mandatory for a reader to understand pointers, structures, unions and other basic constructs of the C language. As this is advanced topic about PHP, the reader must also have some experience with PHP.

II. VARIABLES

By definition, PHP is a dynamic, loosely typed language that uses copy-on-write and reference counting. Dynamic programming language is a term in computer science that describes a language that allows the definition of a program to be updated while the program is running. An example of loose types is defining a string variable and then multiplying the variable by 2. Now, the variable contains an integer instead of a string.

```
$x = "2"; // string
$x = $x * 2; // int (4)
```

In C, similar behavior exists via **unions**. Unions are values that may store different data in the same memory location. The current type of a union must be stored separately.

```
union car {
  char name[20];
  int price;
};
```

When created, the size of the variable will always be 20 bytes (since 1 character in this example is 1 byte long). The value of the variable in PHP is stored in a union like this:

```
typedef union _zvalue_value {
        long lval;
        double dval;
        struct {
                char *val;
                int len;
        } str;
        HashTable *ht;
        zend_object_value obj;
} zvalue_value;
```

Now, this gives a lot of options to PHP. In this union, the value could be stored as a long, double, string, hash table, or object. Notice that strings are represented as a structure where the actual value and the length of the string are stored. Hash tables are used for many things, not only for implementation of PHP arrays but for storing object properties and methods. “zend_object_value” stores an object store handle and handlers. This union

only saves the value of the variable; i.e. this union is actually a part of something bigger. Here is the full implementation of a PHP variable in C:

```
typedef struct _zval_struct {
    zvalue_value value;
    zend_uint refcount__gc;
    zend_uchar type;
    zend_uchar is_ref__gc;
} zval;
```

The code above shows the actual internal presentation of the variable. It contains the following: the value inside a union; the number of current references to that variable; the current type of the variable; and a reference flag. As mentioned previously, PHP uses reference counting, meaning that the engine itself is responsible for references between variables.

III. REFERENCES AND COPY-ON-WRITE

To avoid copying values, PHP utilizes reference counting. When one variable is assigned to another, it always just references the first until the values are changed. For example, a variable is declared and the value 5 is assigned to it. This variable's reference count will be equal to 1.

```
$x = 5;
// zval_1(value=5, ref=1)
```

Now, when two new variables are created with this variable as the value, the struct is not copied but instead the reference count will increment.

```
$y = $z = $x;
// zval_1(value=5, ref=3)
```

When the first variable changes its value, the PHP engine is smart enough to copy just the first variable instead of coping the last two variables.

```
$x = 3;
// zval_2(value=3, ref=1)
// zval_1(value=5, ref=2)
```

The PHP engine simply decremented reference count in zval_1 [($y, $z) => 2] and created a new zval variable for $x with reference count 1.

In the next scenario, variables will be destroyed with the *unset* function. The variable is destroyed when the reference count is equal to 0.

```
unset($y);
// zval_1(value=5, ref=1)
unset($z)
// zval_1(value=5, ref=0)
```

This is a simple example of how the variables are implemented internally. As the examples show, the PHP engine is efficient when it comes to saving memory and variable management.

IV. ARRAYS

Arrays in PHP are stored as hash tables. The C implementation of a hash table looks like this:

```
typedef struct _hashtable {
    uint nTableSize;
    uint nTableMask;
    uint nNumOfElements;
    ulong nNextFreeElement;
    Bucket *pInternalPointer;
    Bucket *pListHead;
    Bucket *pListTail;
    Bucket **arBuckets;
    dtor_func_t pDestructor;
    zend_bool persistent;
    unsigned char nApplyCount;
    zend_bool bApplyProtection;
#if ZEND_DEBUG
    int inconsistent;
#endif
} HashTable;
```

A simple explanation of variables inside a hash table:

- nTableSize specifies the actual size of the C array. Its size is always a power of 2. If there are 128 elements, array is 128 elements long. If there are 129 elements, array is 256 elements long;
- nTableMask is the table size minus 1. It is used for shrinking the hash value;
- nNumOfElements contains the number of elements inside the array;
- nNextFreeElement is the next free integer key for when the next value is added;
- pInternalPointer is a value which stores the current position in the array for iteration;
- pListHead stores the start of the array;
- pListTail stores the end of the array;
- arBuckets is defined as an array of bucket pointers and the bucket is a container that stores

information about each element in the array including: its key; where that key map to; and some other internal metadata such as shared hashes;

- pDestructor is a function called when a value is removed from the hash table. The function will reduce the reference count of the zval;
- persistent determines persistence between multiple requests;
- inconsistent is used in debug mode to catch incorrect uses of hash tables;

To really understand arrays, the bucket type must be explored. As stated above, bucket is a container that stores information about each element:

```
typedef struct bucket {
    ulong h;
    uint nKeyLength;
    void *pData;
    void *pDataPtr;
    struct bucket *pListNext;
    struct bucket *pListLast;
    struct bucket *pNext;
    struct bucket *pLast;
    const char *arKey;
} Bucket;
```

A simple explanation of variables inside a bucket:

- h stores the value of the hash without the mask. It's used when engine does a full search for key;
- nKeyLength is used to save the length of string keys but not used when keys are integers;
- pData and pDataPtr are used to store the actual value (zval);
- pListNext and pListLast specify the order of the array elements;
- arKey is an array that stores a pointer to the first bucket;

The main problem with this array implementation is there is a possibility for collisions on the hash of the key. In that case, buckets will be stored as a linked list, so keys in PHP arrays are actually pointers to a list of buckets [3]. The major drawback with linked list is search efficiency. If the engine sees more buckets in a linked list, it will start a full text search for the key inside that linked list. When classifying this algorithm, instead of O(1) for searching keys in a C array, there is O(n) for searching keys inside a linked list. In terms of security, this can be an exploit. If an attacker only inserts keys which collide, then the whole array would be one big linked list which is much slower and can exhaust CPU power. Developers should check that the *max_input_vars* inside *php.ini* is set to some reasonable number that will limit the number of allowed keys in one PHP array.

V. OPCODES

When parsing PHP files, the engine generates opcodes. These are not the same as bytecode in Java or CIL in .NET but instead they are a stepping stone between source code and execution. Here is a simple script:

```
<?php
$i = 0;
$j = 2;
$k = $i + $j;
```

Now, with the *vld* extension, it's possible to look at the generated opcodes. Instead of variables (*$i*, *$j*, *$k*) compiled variables are used (*!0*, *!1*, *!2*). Variables starting with tilde (~) are temporary variables. It's really simple to get opcodes in Linux with the *vld* extension inside PHP:

```
php -d vld.active=1 script.php
ASSIGN      !0, 0
ASSIGN      !1, 2
ADD     ~2  !0, !1
ASSIGN      !2, ~2
RETURN      1
```

As stated above, opcodes show compiled variables:

TABLE 1 – Variable in source and compiled variable

Variable in source	Compiled variable
$i	!0
$j	!1
$k	!2

The first two lines assign values to variables !0 and !1 ($i and $j). Then, a temporary variable (~2) stores the sum of the first two variables. Next, the value from the temporary variable (~2) is copied to the !2 variable ($k). Finally, as in many other languages, there is a return statement.

VI. CONCLUSION

Knowing about PHP internals can provide many advantages. Firstly, a person who knows about internals could program more securely in their applications, as shown in the array implementations. Secondly, they could create faster scripts, especially with the knowledge of opcodes. Furthermore, if a programmer has a deep understanding of PHP internals, they could more easily give back to the community with patches and improvements to the language that are appreciated and respected in many circles. Finally, it's a really great feeling having your scripts and PHP under control and feeling that your scripts are secure, fast and reliable.

VII. LITERATURE

[1.] Sara Golemon, Extending and Embedding PHP, Sams Publishing, 1st edition, 2006;

[2.] PHP community, php.net, 2014, http://php.net/internals2

[3.] Nikita Popov, Supercolliding a PHP array, 28. December 2011, http://nikic.github.io/2011/12/28/Supercolliding-a-PHP-array.html,

[4.] David Shafik, PHP Performance I: Everything You Need to Know About OpCode Caches, 30. September 2013, https://support.cloud.engineyard.com/entries/26902267-PHP-Performance-I-Everything-You-Need-to-Know-About-OpCode-Caches

[5.] Anthony Ferrara, PHP's Source Code For PHP Developers - Part 1 - The Structure, 12. March 2012. http://blog.ircmaxell.com/2012/03/phps-source-code-for-php-developers.html

[6.] Nikita Popov, Understanding PHP's internal function definitions (PHP's Source Code for PHP Developers - Part 2), 16. March 2012. http://nikic.github.io/2012/03/16/Understanding-PHPs-internal-function-definitions.html

[7.] Anthony Ferrara, PHP's Source Code For PHP Developers - Part 3 – Variables, 21. March 2012. http://blog.ircmaxell.com/2012/03/phps-source-code-for-php-developers_21.html

WebRTC technology overview and signaling solution design and implementation

Branislav Sredojev*, Dragan Samardzija ** and Dragan Posarac*
* Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Republic of Serbia
** RT-RK Institute for Computer Based Systems, 21000 Novi Sad, Republic of Serbia
sredojev.branislav@gmail.com

Abstract - This paper describes the WebRTC technology and implementation of WebRTC client, server and signaling. Main parts of the WebRTC API are described and explained. Signaling methods and protocols are not specified by the WebRTC standards, therefore in this study we design and implement a novel signaling mechanism. The corresponding message sequence chart of the WebRTC communication behavior describes a communication flow between peers and the server. The server application is implemented as a WebSocket server. The client application demonstrates the use of the WebRTC API for achieving real-time communication. Benefits and future development of the WebRTC technology are mentioned.

I. INTRODUCTION

WebRTC (Web Real-Time Communication) is a new web technology which allows browser and mobile applications with functionalities such as audio/video calling, chat, P2P (peer-to-peer) file sharing and all that without any additional third-party software or plugins.

It was published as an open source technology by Google in May 2011 and includes fundamental components for real-time communication on the web. With those components, that can be accessed via the JavaScript API (Application Programming Interface), developers have opportunity to create impressive web applications. Google, Mozilla and Opera support WebRTC and they will participate in the further development of this technology. At the moment, there are still platforms and browsers that either partially or do not support WebRTC [1]. Major components of the WebRTC API are:

- MediaStream - allows a web browser to access the camera and microphone;
- RTCPeerConnection - sets up audio or video calls;
- RTCDataChannel - allows browsers to send data through peer-to-peer connections.

It is important to know that WebRTC is not just a single API, but it is a collection of APIs and protocols defined by the various working groups such as W3C (World Wide Web Consortium) and IETF (Internet Engineering Task Force). Support for each of them develops in different browsers and operating systems [2].

A. *MediaStream*

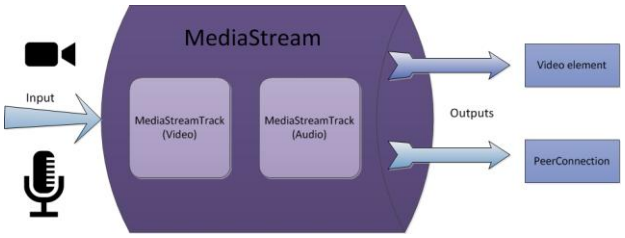


Figure 1. MediaStream

The MediaStream API has a control over the synchronized streams of media [3]. Each MediaStream has an input and output. The input source might be a physical device, such as the camera or microphone. The output might be forwarded to one or more destinations, such as video element or PeerConnection (Figure 1.).

The two main components in the MediaStream API are the MediaStream interface and MediaStreamTrack. The MediaStream interface represents a stream of media content. That media stream can have several parts. Those parts are called tracks and they have explicit type, such as audio and video. MediaStreamTrack represents a type of media that have been obtained from the input source. For example, if it is a microphone, we will get MediaStreamTrack (Audio).

B. *RTCPeerConnection*

The RTCPeerConnection API allows two peers to communicate directly, browser to browser. The WebRTC architecture diagram (Figure 2.) represents a very simple hierarchy. WebRTC provides a framework to make real-time communications possible.

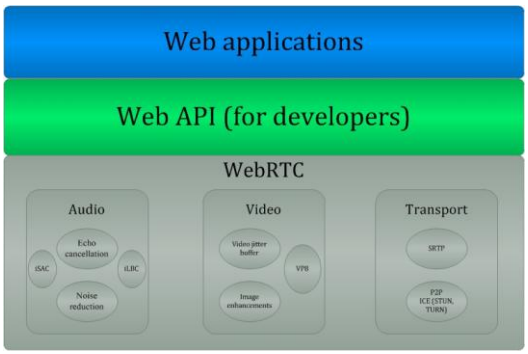


Figure 2. WebRTC architecture

The WebRTC architecture has three basic sections. The first section, WebRTC, consists of an audio, video and transport mechanism.

The main task of the audio mechanism is to handle echo cancellation and noise reduction. Echo cancellation is a method that improves the sound quality, removing the resulting echo or preventing its occurrence. Noise reduction is the process of removing noise from the signal. The audio mechanism contains iSAC and iLBC codecs. iSAC codec represents a wideband audio codec developed by Global IP Solutions. This codec is suitable for audio streams and in June 2011 became a part of the WebRTC technology. Global IP Solutions is also the creator of the iLBC codec. This narrowband audio codec is suitable for voice communication over IP.

The video mechanism is responsible for image enhancement and video jitter buffer. Image enhancement is the process of improving the quality of digital image by manipulation with a certain software. The video mechanism includes VP8 codec. VP8 codec is a video compression format. It was developed by On2 Technologies in September 2008 and is currently supported by Chrome, Mozilla and Opera.

The transport mechanism is responsible for SRTP (Secure Real-time Transport Protocol), P2P and ICE (Interactive Connectivity Establishment). P2P and ICE with STUN and TURN servers will be explained in the following chapter.

The Web API is on the top of the WebRTC section which shields developers from complexities and facilitates their work. The third and last section are Web applications, such as audio, video and chat applications with possibility of file sharing.

C. *RTCDataChannel*

The RTCDataChannel API represents a bi-directional data channel between two peers and enables exchange of arbitrary data between them. Each RTCDataChannel can be configured to provide:

- reliable or unreliable delivery of messages;
- in-order or out-of-order delivery of messages.

The reliable and in-order delivery is equivalent to TCP (Transmission Control Protocol). On the other hand, the unreliable and out-of-order delivery is equivalent to UDP (User Datagram Protocol). There are many potential use cases for the API, including:

- gaming;
- remote desktop applications;
- real-time text chat;
- file transfer.

D. *ICE*

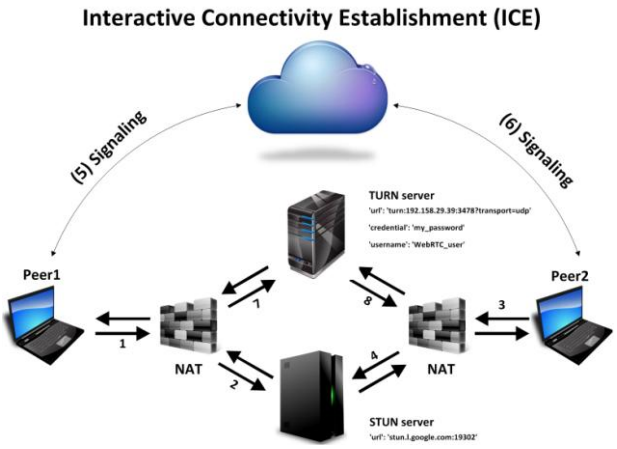


Figure 3. Interactive Connectivity Establishment (ICE)

The Internet addressing system is still using IPv4 (Internet Protocol version 4) and because of that, most of our devices are behind one or more layers of NAT (Network Address Translation). NAT is a mechanism of mapping private addresses to the public address changing address information within IP packets while it is in transit through the device for routing. NAT can be implemented with a home Wi-Fi router.

WebRTC technology developers can use ICE which will simplify complexity of the Internet addressing system. Developers need to pass the ICE server URLs to the peer connection. ICE will try to find the best path to connect two peers (Figure 3.). ICE is using two types of servers, STUN (Session Traversal Utilities for NAT) and TURN (Traversal Using Relays around NAT). ICE is using a STUN server to find out what external address is assigned to a specific peer. If that fails, the traffic is routed via a TURN server. Every TURN server supports STUN, a TURN server is a STUN server with added relaying functionality built in [4].

An analysis by webrtcstats.com in 2014, which lasted 6 months covering over 6 million minutes of one-to-one video calls, shows that 92% of WebRTC video calls are P2P without a relay and just 8% are over a TURN server (Figure 4.) [5].

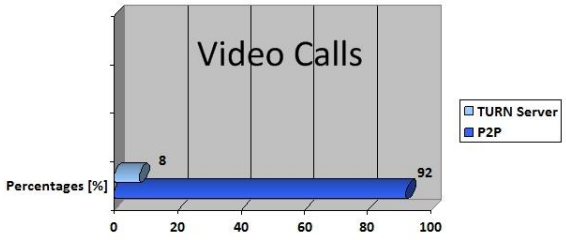


Figure 4. Video calls statistics

II. DESCRIPTION OF REALIZATION

In this section, we describe our implementation of a WebRTC client, server and newly-proposed signaling solution between them. Signaling is a process of exchanging system control information like session

control messages, media metadata, error messages and network data - information important to set up a call. This signaling process needs a two-way (bi-directional) messaging and therefore we have chosen to implement signaling using WebSocket. The server application is implemented as a WebSocket server. The client application demonstrates the use of the WebRTC API for achieving real-time communication.

A. Signaling

Signaling methods are not specified by WebRTC in order to avoid redundancy and to maximize compatibility with established technologies [6].

The signaling process corresponds to protocols such as SIP (Session Initiation Protocol), XMPP (Extensible Messaging and Presence Protocol), XHR (Xml Http Request), and SIP over WebSocket. In this study, the signaling implementation uses the WebSocket mechanism [7]. WebSocket is a relatively new addition to browsers, and it was standardized in 2011. The most important thing about WebSocket is that it opens a session from a peer to the server, leaving it open until the session is closed. Messages can be textual or binary. The message sequence chart of the WebRTC communication describes communication flow between the peers and server. The diagram is depicted in Figure 6.

In our signaling solution we made four types of the control messages: "initialize", "initiator", "got user media" and "peerChannel". Those control messages are textual and they are used to set up flags for certain mechanisms. At start, Peer1 will send "initialize" to the server, because it needs to find out whether it is the session initiator. If so, it will configure the user media after getting the control message "got user media" from the server. Now Peer1 is in a waiting phase until Peer2 appears. Upon its activation, Peer2 sends the equivalent signaling message, and when it finds out that it isn't the initiator, it will first get "peerChannel" and then "got user media". After that, both sides can start establishing a peer-to-peer connection.

The offer and answer messages are transmitted in the SDP (Session Description Protocol) format [8]. SDP plays a main role in the setup of WebRTC sessions. SDP is the protocol that has been chosen by IETF and W3C to exchange media information between peers in WebRTC. SDP carries information about media type, format, codecs, and all associated properties that will be used in the media session. The diagram shows that Peer1 will send the SDP offer and ICE candidates to the server and the server will pass that to Peer2. Peer2 will send the SDP answer and ICE candidates to the server and the server will pass that to Peer1.

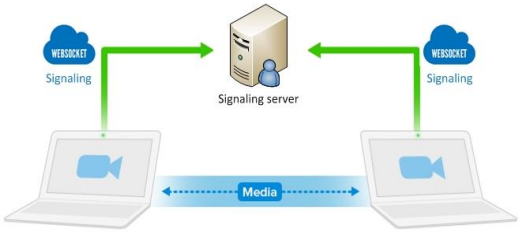


Figure 5. Signaling process

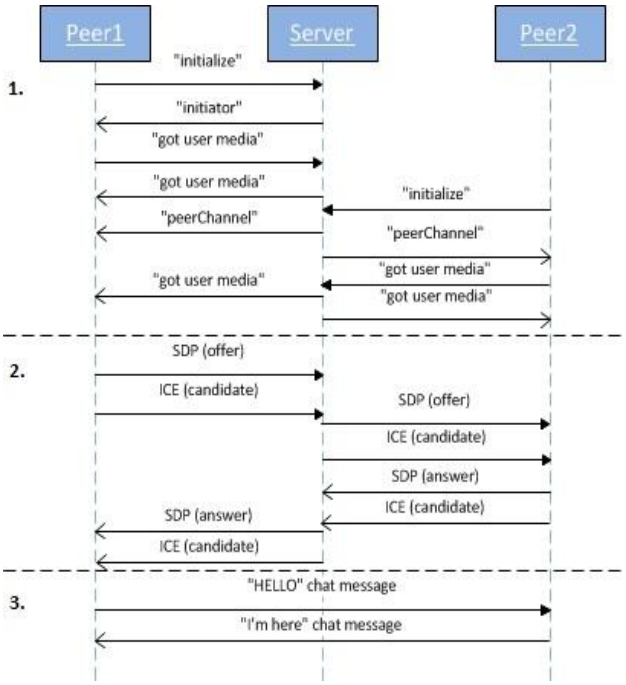


Figure 6. WebRTC message sequence chart

After this, the peer-to-peer connection is established. In addition, we have implemented a real-time text chat and now the peers can directly exchange messages without the server presence.

All signaling SDP and control messages, in our WebRTC application are transmitted as JSON (JavaScript Object Notation) objects (Figure 7.). JSON is a syntactic subset of JavaScript and has a great advantage of being natively interpreted by web browsers. JSON enables information exchange, as well as JSON objects generation and parsing.

In summary, there are three phases in Figure 6. The first phase – the peers are exchanging the control messages with the server, about the initiator, peer channel and user media. The second phase – creation of peer connection based on the SDP exchange. The third phase – the peer-to-peer connection is created and the peers can now exchange messages directly.

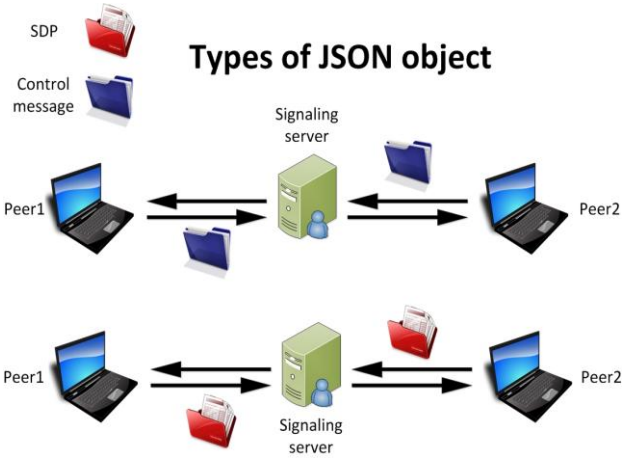


Figure 7. Types of JSON object

B. *Server application*

The server application is implemented as a WebSocket server written in Node and it can be used by any client or a server application. Node.js is an open source, cross-platform runtime environment for the server-side and networking applications. Node.js applications are written in JavaScript. It provides an event-driven architecture and a non-blocking I/O API that optimizes an application's throughput and scalability.

Our WebSocket server is using the WebSocket library. It is running on a localhost at port 2222 and it is available to anyone within the enterprise. The server implements the accepting, processing and forwarding message mechanism.

The WebSocket server is dealing with four types of control messages ("initialize", "initiator", "got user media" and "peerChannel"), two types of media information messages (SDP offer and SDP answer) and one type of network information messages (ICE candidate). Also it manages the number of clients, deciding whether a client is the initiator or should it be removed when the session is over.

C. *Client application*

The client application is implemented using the WebRTC API. The main task is to establish a peer-to-peer connection. Before a peer-to-peer connection gets established, the exchange of local and remote descriptions is performed (the audio and video media information).

In the following we will describe how the signaling offer and answer are exchange using the SDP. First, Peer1 runs the RTCPeerConnection createOffer() method. After that, it sets the local description using the setLocalDescription() method and then sends this session description to Peer2 via the signaling server. Then, Peer2 sets the description that Peer1 sent as a remote description using the setRemoteDescription() method. Peer2 runs the RTCPeerConnection createAnswer() method. In the createAnswer(), it sets the local description and sends it to Peer1. When Peer1 gets the Peer2 session description, it sets that as the remote description with the setRemoteDescription() method.

Figure 9. represents one snap shot of our application screen. Also there is the chat application extension, so peers can send text messages. For example, once installed within an enterprise network, our server (with URL http://localhost:2222/) will enable WebRTC peer-to-peer communication sessions.

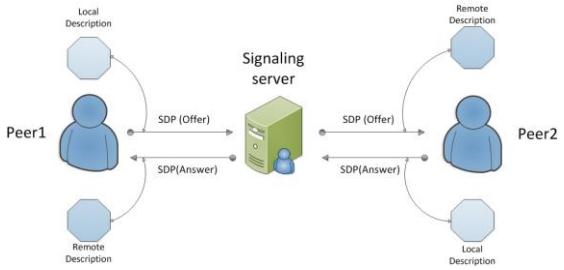


Figure 8. Offer/answer exchange between peers

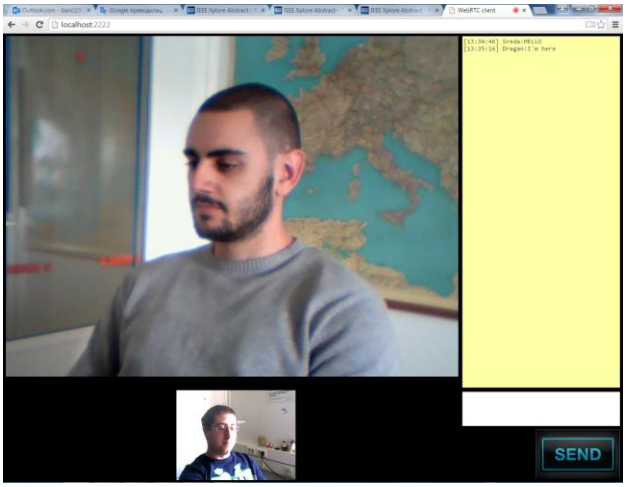


Figure 9. Our application

III. CONLUSION

WebRTC is a powerful technology that benefits all areas of communication accessibility and feature rich tool. This technology has a potential to improve enterprise communications. It can provide a lot of benefits for enterprises. WebRTC technology reduces the costs of using VPN (Virtual Private Network) to connect remote branches and real-time communication is supported without the need for additional third-party software or plugins. Secure communication with employees in within an enterprise and its remote branches is enabled by state of the art encryption standards.

In the future we will develop an additional file sharing option and we will use encryption for both the media and signaling.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education and Science of the Republic of Serbia under the project TR 36029.

REFERENCES

[1] A. Sergiienko, "WebRTC Blueprints, Develop your very own media applications and services using WebRTC," in Birmingham, Packt Publishing, ISBN 978-1-78398-310-0, May 2014.

[2] A. Zeidan, A. Lehmann, and U. Trick, "WebRTC enabled multimedia conferencing and collaboration solution," WTC 2014, in Berlin, Germany, Print ISBN 978-3-8007-3602-7, June 2014.

[3] W3C Editor's Draft, http://w3c.github.io/mediacapture-main/getusermedia

[4] K. Singh, A. Johnston, and J. Yoakum, "Taking on webRTC in an enterprise," Communications Magazine, IEEE, ISSN 0163-6804, April 2013.

[5] WebRTCstats, http://webrtcstats.com/webrtc-revolution-in-progress/

[6] R. Manson, "Getting Started with WebRTC, Explore WebRTC for real-time peer-to-peer communication," in Birmingham, Packt Publishing, ISBN 978-1-78216-630-6, September 2013.

[7] IETF - WebSocket, https://tools.ietf.org/html/rfc6455

[8] IETF - SDP, https://tools.ietf.org/html/rfc4566

Methods for Geotargeting Redirection in Corporate Wide Area Networks

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko miljkovic@hep.hr

Abstract - In business organizations, particularly large ones, it is sometimes necessary to manage web users according to their geographic location. Final destination can be either business related web site or web application server. Such redirection can be accomplished in numerous ways. The choice depends on a goal one want to achieve: distribution of user load to optimal site, achieving high availability or both. Several interesting methods suitable for geographically dispersed clusters of web (and application) servers are presented and analyzed. This includes client side scripts, server side scripts, load balancers, multiple DNS records as well as combinations of these techniques. Reviewed methods are compared and their benefits and shortcomings are discussed.

I. INTRODUCTION

A corporate Wide Area Network (WAN) connects geographically dispersed corporate users across metropolitan, regional, national or international boundaries. It links Local Area Networks (LAN) in multiple locations using private or leased lines. Such network enables collaboration from all regional centers and branch offices. Geotargeting is a technique that determines the geographic location of the client and provides content to the user based on that location. Such content may be targeted advertisements, automatic language selection, content delivery based on user region, regulatory compliance, load balancing and resource allocation between available hosts and network resources, [1,2]. Geographic location of user is most often done using the Internet Protocol (IP) addresses associated with the user's Personal Computer (PC), although other means for accomplishing task exist (e.g. ping delay and traceroute), and some advanced [1-4]. In the corporate network there is a known network topology and full control of the allocated address space. Editing and maintaining of corporate Domain Name System (DNS) is also easy to accomplish on a DNS in corporate ownership. Simultaneous knowledge of IP addresses, locations of PCs and flexibility of maintaining own DNS servers makes geotargeting redirection within the corporate WAN much easier to achieve than geotargeting of general Internet users.

II. USER ADDRESS SPACE

When a corporate user accesses a corporate web page, request contains the IP address of his computer. If a corporate server is accessed from the user located outside the corporation, request will contain the IP address of the proxy server instead and this makes it much more difficult to properly locate such user. However, access to most corporate information services are restricted to users coming from within the corporate WAN. For general users coming from the outside untrusted network (usually the Internet), dedicated web servers are provided located in the demilitarized zone (DMZ). In case of corporate users connecting to the corporate network using the Virtual Private Network (VPN) similar problem appears, however this generally represents a very small share of total number of users. The IP address of these users doesn't correspond to their geographic location. Such users may connect to a particular web server directly, entering the web address of appropriate web server and bypassing redirection mechanisms. Usually a defined range(s) of user IP addresses are allocated to network switches within common corporate locations. In following example four corporate regional centers of Croatian Electricity Company (HEP) are illustrated in Fig. 1 with the network transmission backbone shown in Fig. 2. Each region (represented by the regional center) has defined its own address space as show in Table I.

Figure 1 Four corporate regional centers of HEP (Croatian Electricity Company)

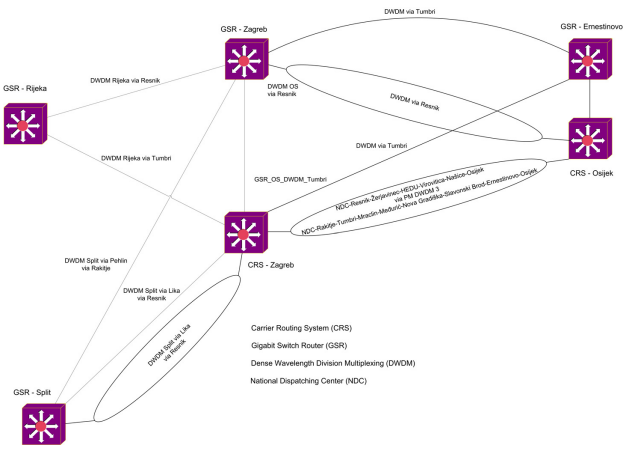


Figure 2 Network transmission backbone

TABLE I EXAMPLE OF CORPORATE ADDRESS SPACE (HEP)

Regions represented by regional center	IP address range
Osijek	10.0.0.0 – 10.63.255.255
Rijeka	10.64.0.0 – 10.127.255.255
Split	10.128.0.0 – 10.191.255.255
Zagreb	10.192.0.0 – 10.255.255.255

Address space shown in Table I provides a clear division of addresses across different geographic regions. Sometimes this division is not straightforward because instead of a single address range for each region, there may be two or more address ranges, e.g. one range including new IP addresses and one range for old addresses, e.g. starting in form of 22x.x.x.x. If this is the case simple DNS based geotargeting method must be augmented by some other technique. Exception must also be provided for users accessing corporate network using VPN which has its own address pool containing group of IP addresses, often from one region.

For most corporate uses there is no need to know the exact geographical location of each user, but only the user membership to a particular geographic region represented by one of corporate regional centers. This is enough in almost all uses. Address space is further divided and can provide necessary granularity as is illustrated in Fig. 3. The regional level that suffices for most uses is depicted by elliptic curve.

Sometimes users of business web applications in premises located near the border of two regions prefer to connect to the web application server in another region. In such rare cases when IP addresses from Table I don't provide enough information for separation each regional address space can be further divided in smaller subranges that correspond to subregions (consisting of cities and their surroundings) as illustrated in Table II. Similar subranges exist in all remaining regions. This additional information can be used for defining necessary exceptions. It is rarely needed to go beyond this level.

TABLE II ADDRESS SPACE OF ZAGREB REGION

Geographic subregion (city and surroundings)	IP address range
Zagreb	10.10.x.x - 10.10.255.255
Zabok	10.12.x.x - 10.12.255.255
Varaždin	10.13.x.x - 10.13 255.255
Čakovec	10.14.x.x - 10.14 255.255
Koprivnica	10.15.x.x - 10.15.255.255
Virovitica	10.16.x.x - 10.16.255.255
Bjelovar	10.17.x.x - 10.17 255.255
Križ	10.18.x.x - 10.18.255.255
Karlovac	10.19.x.x - 10.19.255.255
Sisak	10.20.x.x - 10.20.255.255

III. COMMON METHODS FOR GEOTARGETING REDIRECTION

Users can be geographically directed to the suitable web server in several ways. Here are briefly reviewed common methods for geotargeting redirection. In most cases redirection during the same user session must be tied to the same web (or application) server. First two methods use the application level geotargeting that depends on the application code installed as a script on the client side or server side. Last two methods use transport level geotargeting. Needless to say, common content of web (application) servers must be synchronizing among all servers providing the service to the user.

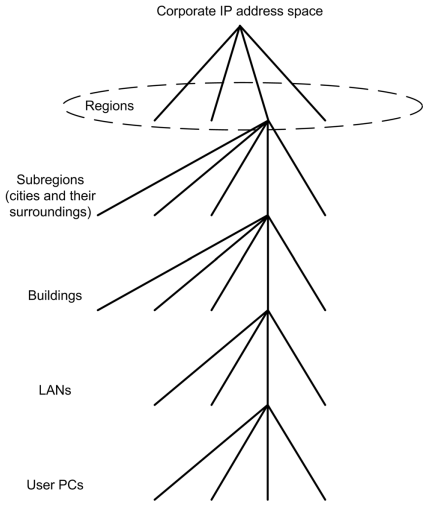


Figure 3 Address subranges down to individual user PCs

A. Client Side Script

In this approach a web server provides a client with the code (e.g. Java or Flash SWF) that contains geotargeting algorithm and a list of web addresses for all regional servers, [5]. Based on the IP address of the client PC the user is directed to the appropriate server, Fig. 4.

```
Fetch list of servers and address ranges
IF User_IP ∈ [Region_1] then redirect to Regional_Server_1
IF User_IP ∈ [Region_2] then redirect to Regional_Server_2
                        …
IF User_IP ∈ [Region_N] then redirect to Regional_Server_N
```

Script can be programmed to access web server data on load and availability for all web (application) servers (destination servers must make these data available) and make a smart choice in case of the overload or unavailability (when a server doesn't respond in a preset time period with the correct answer to a HTTP query).

```
Fetch list of servers and address ranges
IF User_IP ∈ [Region_1] AND IF Server_1 is available then
                        redirect to Regional_Server_1
IF User_IP ∈ [Region_2] AND IF Server_2 is available then
                        redirect to Regional_Server_2
                        …
IF User_IP ∈ [Region_N] AND IF Server_N is available then
                        redirect to Regional_Server_N
ELSE randomly chose one of remaining available servers
```

Disadvantage of the method is its dependence on the code located in a client PC where various browser types exist with the various security settings (that change from time to time via changes in group policy) and may interfere with the necessary (Java) privileges. Also, the end user may intentionally or unintentionally tweak these settings, hence there is a lack of full control in this method. It also poses some security risk because the client

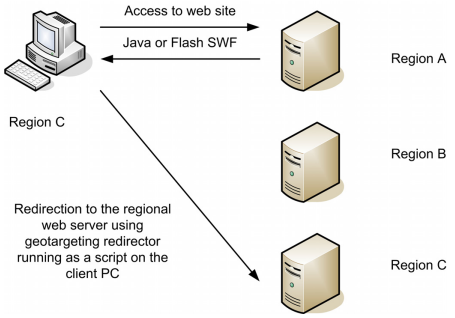


Figure 4 Client side script geotargeting redirection

script knows what lies beneath and infrastructure architecture of web servers (IP addresses of all servers) is revealed. Previous two disadvantages make the client based method an interesting but bad idea that is rarely (if ever) used in a corporate environment.

B. Server Side Script

This approach requires a geotargeting server side script, [6], running on, at least, one web server. User accesses this web server first. The script on the server considering the IP address of the user provides necessary redirection (using HTTP redirect command) to the same or another appropriate web (application) server, Fig. 5.

```
Fetch User_IP address
IF User_IP ∈ [Region_1] then redirect to Regional_Server_1
IF User_IP ∈ [Region_2] then redirect to Regional_Server_2
…
IF User_IP ∈ [Region_N] then redirect to Regional_Server_N
```

When a web browser attempts to open a web address that has been redirected, a page at a new, geotargeted web address is opened instead. Example of HTTP redirect command that redirects to http://destination.com/ immediately is illustrated bellow.

```
<meta http-equiv="refresh" content="0; url=http://destination.com/">
```

Just as in a client side script approach server side script can also have access to data on load and availability and make a smart redirection choice. Server side technique is the preferred approach because it adds negligible server overhead, simultaneously giving server administrators full control of redirection rules (that can be easily edited and changed with the immediate effect).

Common approach is the use of CGI/Perl, PHP or ASP.NET redirector. The IP address of the user is compared to the predefined address ranges for each

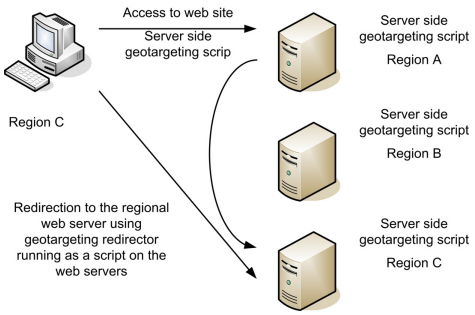


Figure 5 Server side script geotargeting redirection

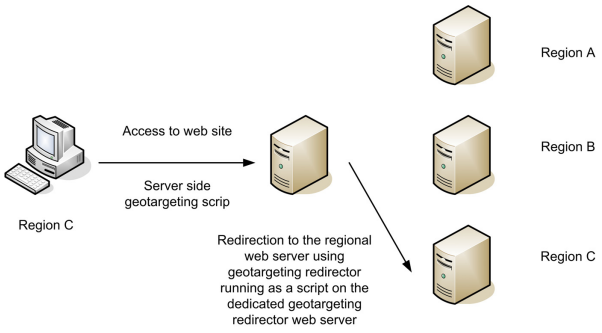


Figure 6 Server side script on a dedicated server

geographical region and the HTTP redirect command is issued toward the target web server (this can be original server too). On a web server one address is dedicated to the redirector script and another to web content. It is possible to dedicate a separate server for server side redirection script, as shown in Fig. 6. Use of additional addresses for redirector script on web servers is avoided.

Following two methods operate at the transport level. It includes TCP/IP level load balancing with the geotargeting redirection and DNS based approach.

C. Load Balancers with Geotargeting Capability

Load balancing is a technique to spread the work between two or more computers in order to get optimal resource utilization, [7-9]. Load balancers are generally hardware/software device combination that balances incoming traffic according to some predefined rules (most often defined by some load metrics). Some load balancers can direct traffic according to the user IP address (different destinations for different IP address ranges corresponding to different geographic locations), Fig. 7. They can handle hundreds of thousands of requests per second and are capable even for the most demanding global operations (generally only small part of their capacity is needed within a corporate WAN).

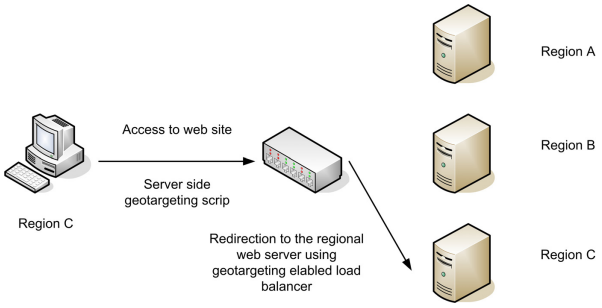


Figure 7 Network load balancer with geotargeting redirection capability

D. DNS Based Method

This method makes use of multiple A-records set for the same web address, [9-12]. It makes use of Subnet Prioritization and client's PC handling of multiple IP addresses defined for the same web address. With enabled Subnet Prioritization, the DNS server can check the IP address of a client requesting name resolution and sort the results it returns to the client according to the proximity of the host address to the querying client IP address. Subnet Prioritization recognizes when a host is

on the same subnet as the client and returns the local host's address first. Because subnets are restricted to particular locations, re-sorted results correspond to the geographically closest server, Fig. 8. By default, the DNS server uses Round Robin to rotate the order of A-records returned in answers to a client queries when multiple A-records exist for the same DNS web address, [7-9,13]. If Round Robin and Subnet Prioritization are disabled on the DNS server, the order in the answer list is based on a static ordering of A-records as they are stored in the DNS server zone file. On the DNS server Subnet Prioritization should be enabled and Round Robin disabled. If both Round Robin and Subnet Prioritization are enabled, Round Robin will supersede. Subnet Prioritization should preferably also be enabled on clients. Use of multiple A-records in the DNS for the geographically dispersed cluster of web application servers located at four corporate regional centers, [14,15], is shown in Table III, [9,14]. Under the web address *hepweb* there are four A-records.

TABLE III MULTIPLE A-RECORDS FOR THE SAME WEB ADDRESS

Web address (same for all users)	Multiple A-records in DNS	
hepweb	10.193.64.116	*heprediros*
	10.65.64.206	*hepredirri*
	10.129.64.116	*hepredirst*
	10.1.200.206	*hepredirzg*

Each A-record corresponds to the web address of one regional cluster server side script redirector (*heprediros*, *hepredirri*, *hepredirst* and *hepredirzg*), but could be server or cluster as well. If Round Robin is enabled, the DNS server will in its answers rotate IP addresses of four A-records in each subsequent query as shown in Table IV.

TABLE IV ROUND ROBIN ROTATION OF MULTIPLE A-RECORDS IN SUBSEQUENT CLIENT QUERIES

Query	DNS answer rotated by round robin
1	10 193 64 116, 10 129 64 116, 10 65 64 206, 10 1 200 206
2	10 129 64 116, 10 65 64 206, 10 1 200 206, 10 193 64 116
3	10 65 64 206, 10 1 200 206, 10 193 64 116, 10 129 64 116
4	10 1 200 206, 10 193 64 116, 10 129 64 11, 10 65 64 206
5	10 193 64 116, 10 129 64 116, 10 65 64 206, 10 1 200 206

However, if Subnet Prioritization is enabled and Round Robin is disabled (not to supersede Subnet Prioritization) DNS answer will depend on the user location. Based on IP proximity metrics it will return the IP address of the closest regional server first, as shown in Table V. In some implementations the IP address of the host closest to a client will appear at the first place while remaining entries will rotate in Round Robin fashion in subsequent answers to DNS queries. DNS based method works well if all regional servers are available and there are not many exceptions in regional IP address ranges.

TABLE V SUBNET PRIORITIZATION AND MULTIPLE A-RECORDS

User region	DNS answer sorted by subnet prioritization (IP metrics)
Osijek	10 193 64 116, 10 129 64 116, 10 65 64 206, 10 1 200 206
Rijeka	10 65 64 206, 10 1 200 206, 10 193 64 116, 10 129 64 116
Split	10 129 64 116, 10 193 64 116, 10 1 200 206, 10 65 64 206
Zagreb	10 1 200 206, 10 65 64 206, 10 129 64 116, 10 193 64 116

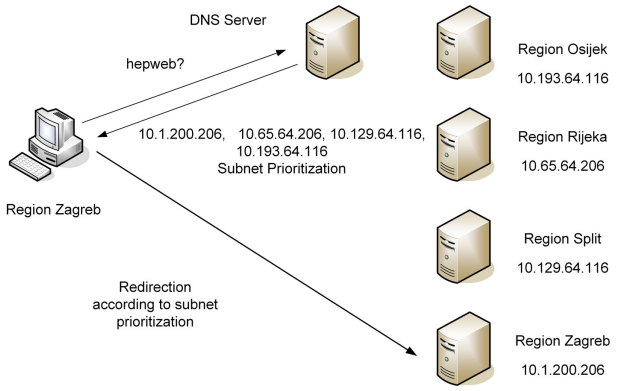


Figure 8 Geotargeting redirection with the DNS based method: multiple A-records and use of subnet prioritization

IV. SOLUTIONS FOR HIGH AVAILABILITY

The goal is to choose a solution that doesn't have single point of failure. It can be achieved by use of multiple servers/load balancers where the redirection takes place or DNS based method that inherently supports failover. This also presumes that DNS servers are redundant.

A. Dual Servers with Server Side Scripts

Server side geotargeting redirection script installed on a separate server presents a single point of failure. To achieve a High Availability two servers are used together, in the active-passive (failover) or active-active configuration (e.g. NLB), Fig. 9. Scripts put negligible overhead and servers can be used for other purpose as well.

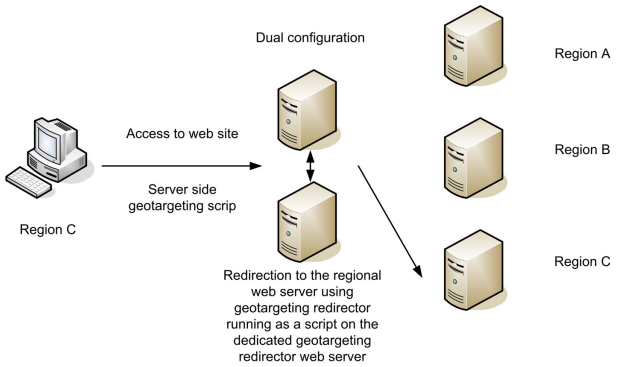


Figure 9 Geotargeting using dual servers with the server side script for the high availability

B. Dual Load Balancers with Geotargeting Capability

Load balancers present a single point of failure. If high availability is required two load balancers must be used

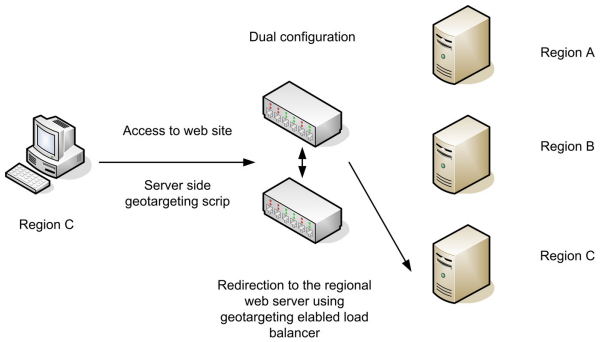


Figure 10 Two load balancers in the active-passive or active-active configuration

together, as illustrated in Fig. 10. In a past two load balancers were often configured in active-passive configuration, but today modern load balancers support load distribution in active-active configuration, [7,8].

C. *DNS Based Method*

Multiple A-records can be used not only for geotargeting redirection but also for automatic failover. This is achieved by the default behavior of the internet browser in the presence of multiple A-records. If multiple IP addresses are supplied for the same web address and browser can't access web site using the first IP address, it will after 20s automatically try the next IP address from the list returned by the DNS, Fig. 11. This works for following browsers: Opera, Safari, IE, and Firefox. Unfortunately it doesn't work with Google Chrome that failed to try the next IP address from the list, [16].

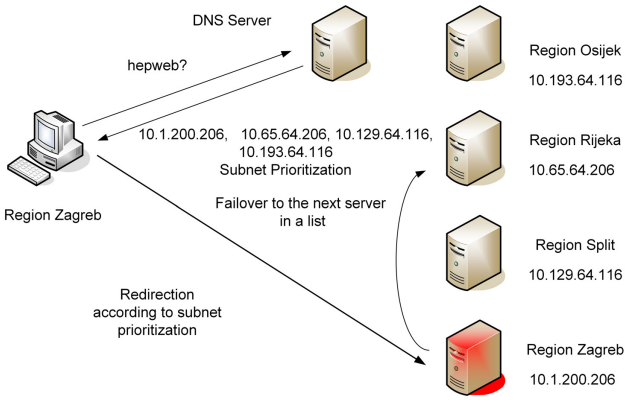


Figure 11 Failover in geotargeting redirection with DNS based method

To address the shortcomings of the DNS based method when High Availability is needed two additions to the method may be implemented. First addition gives the full control of the redirection in case of overloaded or unavailable server(s) and second address exceptions in the default browser behavior when confronted with multiple A-records.

1) *Use of Additional Server Side Script Redirector*

Two levels of user redirection can be implemented by the combination of multiple A-records and the server side script redirector. When the web server proposed by subnet prioritization is not available, the browser will automatically try the next server from the list returned by the DNS. In this way user from one region will be always directed to the second server in a list generating imbalance in the web server load. Better way of handling such situation is use of the additional server side script redirector installed on each web server (making two level redirection) that checks if the user IP address belongs to the intended range of that server. If so, user is served by that server, if not user is depending on the current load or randomly directed to one of remaining active web servers, Fig. 12, avoiding possible overload of the second server in a list. This to be accomplished, web servers must periodically check health and load of other web servers (by some kind of sentry software) and based on that info define a list of remaining choices to a server side script.

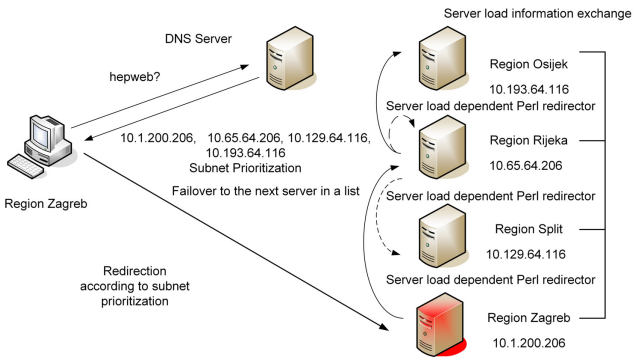


Figure 12 Failover in geotargeting redirection with DNS based method and additional server side script for equal load on remaining servers

2) *Dynamic DNS Entries*

In corporate environment it is relatively easy to dynamically change DNS addresses that can, within several minutes, propagate to all DNS servers in a corporate WAN. When using multiple A-records, user can encounter the situation when the web server that is listed first in an answer returned from a DNS server is not available, Fig. 13. In such case user must rely on the browser behavior where browser after 20s attempts the next IP address in a returned list (as already mentioned Google Chrome unfortunately doesn't work in such usual way). Also, small share of users are impatient to wait even 20s necessary for browser's second try. To address this problem A-record corresponding to the unavailable web server can be (temporarily) removed from the DNS. It can be achieved by a script or specialized small program, [17], installed on some server within a WAN. This script or program needs to periodically check the health (and hence availability) of web servers at all locations. When the particular server is not up and running, (remote) command to the DNS server is issued and A-record removed from the DNS, Fig. 14. Because clients store fetched IP addresses in a local DNS cache it is important that A-records are defined in the DNS with low Time To Live (TTL) values, e.g. few minutes. After the A-record corresponding to a failed server is removed from a DNS and the TTL period for cached entries expires, when users contact the DNS server again, they will get list of A-records corresponding to healthy web servers. When the web server becomes available again similar command can add corresponding A-record to the DNS.

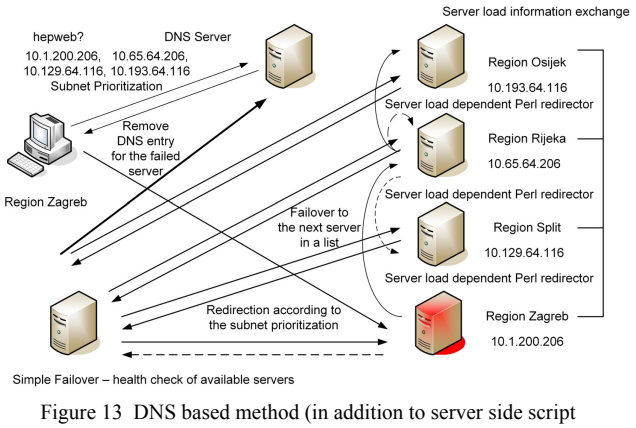


Figure 13 DNS based method (in addition to server side script redirector) and web server failure

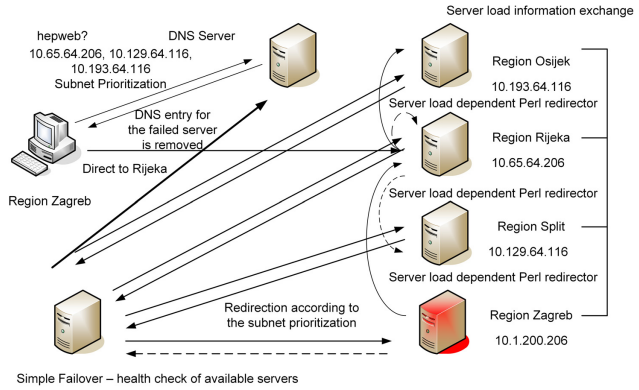


Figure 14 DNS based method with removed entry for failed web server

V. COMPARISON

Advantages, disadvantages and suitability for High Availability implementation is listed in Table VI. Client side script is just an interesting concept that is due the lack of full control of the behavior of client side script (e.g. Java version, security settings etc.) rarely used. Server side script is used much more often as it gives to the server administrator full control of the geotargeting redirection process. Geotargeting enabled load balancers are also not so common because such solutions are still expensive, offer much higher throughputs then necessary for most applications and often require the help of supplier contractors for the configuration. When high availability is required server side script and geotargeting enabled load balancer require dual installations to remove a single point of failure. DNS based method has clear advantage due to inherent internet browsers behavior in the absence of HTTP reply from the requested IP address, when confronted with multiple A-records defined for the same web address. Method is very easy to implement. The DNS based method may be combined with server side scripts installed on all web servers to allow a more balanced redirection to remaining available web servers. Failover behavior can be further enhanced by introducing dynamic changes to the DNS for failed web servers, hence removing even 20s delay for the failover to the operating web server.

TABLE VI COMPARISON OF METHODS

Method		Advantage	Disadvantage	High Availability
Client side script		Simple	Lack of full control on the client side, reveals addresses of the infrastructure	Combine with DNS method
Server side script		Simple, full control of redirection	Two urls, one for redirector, one for web content (use of virtual web server)	Dual server
Geotargeting enabled load balancer		High throughput	Expensive, often needs a help of specialized staff	Dual load balancer
DNS based method	Basic	Direct, Simple	20s delay failover, unequal load on remaining servers, reveals addresses of the infrastructure	Inherent, multiple A-records for same address
	With server side script	Direct, equal load on remaining servers	20s delay failover, reveals addresses of the infrastructure, medium complexity	Inherent, multiple A-records for same address
	With dynamic DNS entries	Direct, immediate failover to remaining servers*	Unequal load on remaining servers, reveals addresses of the infrastructure, medium complexity	Inherent, multiple A-records for same address
	With server side script and dynamic DNS entries	Direct, immediate failover to remaining servers*, equal load on remaining servers	Reveals addresses of the infrastructure, somewhat complex	Inherent, multiple A-records for same address

*access is immediate after the DNS record of the failed server is removed

VI. CONCLUSION

Several geotargeting redirection methods that redirect clients within a corporate WAN have been briefly reviewed. These methods are limited to cases when user IP address corresponds to the user's location. DNS based method is particularly attractive due to its simplicity and low cost. If combined with the second layer of redirection, using server side scripts (that consider availability and loads of regional servers) and dynamic DNS, this method provides quite smart geotargeting redirection. When high availability is not of paramount importance the server side geotargeting redirection script installed on one server can be used. Load balancing devices may be used in rare cases when very high throughput is required. Client side script is mentioned for the completeness despite of its disadvantages.

REFERENCES

[1] S. E. Cruz, An Investigation into Better Techniques for Geo-Targeting, MSc Thesis, June 2002, MIT, Massachusetts; USA

[2] Inya Youn, B. L. Mark and Dana Richards, "Statistical Geolocation of Internet Hosts, Proc. 18th Int. Conf. on Computer Communications and Networks",3-6 Aug.2009,San Francisco, CA

[3] J. Bendale and J. R. Kumar, "Review of Different IP Geolocation Methods and Concepts", International Journal of Computer Science and Information Technologies, Vol.5(1),2014, pp.436-440

[4] E. Katz-Bassett et all, "Towards IP Geolocation Using Delay and Topology Measurements", Proc. 6th ACM SIGCOMM conference on Internet measurement, pp. 71 - 84, New York, NY, USA, 2006

[5] L. Zhu, "Client Side Load Balancing for Web 2.0 Applications", Internet: http://ap-project.org/English/Article/View/9/ [March. 3, 2015]

[6] B. D. Foy, "Geolocation in Perl", Dr. Dobb's, 1 Jan 2005, Internet: http://www.drdobbs.com/web-development/geolocation-in-perl/184416182 [March 3, 2015]

[7] T. Bourke, Server Load Balancing, O'Reilly & Associates, 2001

[8] C. Kopparapu, Load Balancing Servers, Firewalls and Caches, John Wiley, 2002

[9] D. Miljković, Methods for Load Balancing for Application Servers, 16th HROUG, 18-22 October 2011, Rovinj, Croatia

[10] P. Tenerillo, Why DNS Based Global Server Load Balancing (GSLB) Doesn't Work, Part II, Tenerillo Inc., 2004, Internet: http://www.tenereillo.com/GSLBPageOfShameII.htm [March 3, 2015]

[11] P. Tenerillo, Overview of DNS Caching In Browsers, Addendum to "Why DNS Based GSLB Doesn't Work", Tenerillo Inc., 2004, Internet: http://www.tenereillo.com/BrowserDNSCache.htm [March 3, 2015]

[12] P. Tenerillo, "Why DNS Based Global Server Load Balancing (GSLB) Doesn't Work", Tenerillo Inc., 2004, Internet: http://www.tenereillo.com/GSLBPageOfShame.htm [March 3, 2015]

[13] D. Hart, "Load Balancing and Round Robin DNS", Adanac Software, Internet: http://www.adanacsoftware.com/articles/Load%20Balancing%20and%20Round%20Robin%20DNS.pdf [March. 3, 2015]

[14] D. Miljković, "Geografski distribuiran iAS NLB cluster", 10th HROUG , 18-22 October 2005, Umag, Hrvatska

[15] D. Miljković, "Review of Cluster Computing for High Available Business Web Applications", MIPRO, GVS, 26-30 May 2008, Opatija, Croatia

[16] "Using DNS for failover using multiple A records", Stackoverflow, [Online forum comment], Internet: http://stackoverflow.com/questions/5319649/using-dns-for-failover-using-multiple-a-records [March. 3, 2015]

[17] SimpleFailover, Internet: http://www.simplefailover.com/ [March 3, 2015]

Geographically Dispersed Cluster of Web Application Servers on Virtualization Platforms

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko miljkovic@hep.hr

Abstract - For business critical web applications high availability is of paramount importance and clusters of application servers are a viable solution. To achieve even higher availability cluster nodes should be geographically dispersed, i.e. deployed at multiple sites in two or more datacenter premises of a larger business system. In this paper solutions for geographically dispersed clusters of application servers installed on virtualization platforms are described. Most of the time geographically dispersed clusters are overprovisioned and as such are very expensive, considering cost of servers, network, maintenance and software licensing. Virtualization and the choice of appropriate solution can reduce this cost. Advantages and disadvantages of particular solutions are analyzed.

I. INTRODUCTION

Modern corporate businesses like banking, finance, public utility companies, internet sales, travel tickets and hotel reservations are increasingly dependent on highly available business web applications. Business-critical applications require maximum uptime and as such are excellent candidates for various clustering solutions, from conventional single site clusters to geographically dispersed clusters, [1,2]. Clustering can provide high availability and ensure continuous service to users. With cluster ownership (instead of using external cloud solutions) corporations remain under full control of its IT resources and can maintain high security of business data.

Web applications use a three-tier architecture, Fig. 1, [1]. Three-tier is a client-server architecture in which the user interface, functional process logic, computer data storage and data access are developed and maintained as independent modules, most often on separate hardware platforms.

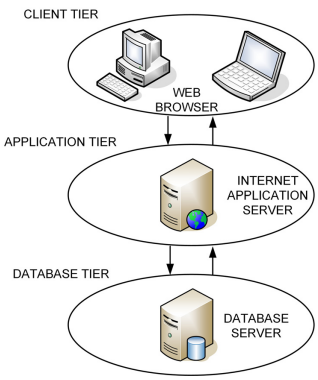


Figure 1 Three-Tier architecture

The Three-Tier architecture has the following tiers:

1.) Client Tier: client is an application that runs on a personal computer and accesses by means of network a server to perform some operation.

2.) Application Tier (Business Logic Tier, Middle Tier): consists of an application server that contains the bulk of the application logic. Application logic resides in a single tier and can be maintained easily at this one location. The architectural design of the application tier is optimized for server functions including access to a database.

3.) Database Tier: consists of one or more database servers, computers dedicated to database storage and retrieval. Upon request from application server, the database server searches the database for selected records and returns them over the network to middle tier.

This paper deals with clustering at the Middle Tier level. Clustering at Database Tier is not covered here.

II. HIGH AVAILABILITY

Let's clarify some basic terms and concepts regarding the high availability, uptime etc., [1,3].

Availability (inherent), A_I, is the probability that the system will be up and running correctly at a given time; this accounts for system operating time and corrective maintenance and excludes downtime associated with preventive maintenance, logistics, and administration, (1).

$$A_I = \frac{MTBF}{MTBF + MTBR} \tag{1}$$

MTBF denotes Mean Time Between Failures and MTBR denotes Mean Time Between Repairs.

Uptime is the time during which a computer system is operational and downtime when it isn't. Operational availability, A_O, is the ratio of system uptime and total time, (2).

$$A_O = \frac{Uptime}{Uptime + Downtime} \tag{2}$$

Values for A_I and A_O are always between 0 and 1, (3)

$$0 \le A \le 1 \tag{3}$$

When defining uptime various criteria may be used, e.g. system is up if all users can access and use applications or in another case when 50% (or some other number) of users can be served.

Downtime includes both planed and unplanned downtime. Planed downtime includes scheduled

maintenance activities, addition of hardware and software to improve availability and applying software and system upgrades. Unplanned downtime includes system failures, operating system crashes, application failures due to application bugs, unpredictable events such as viruses or power surges and human errors like unintentional application or system reset.

Highly available systems have set very tight criteria for permissible annual downtime as shown in Table I. If one aims at annual downtime of several minutes, geographically dispersed cluster should be used, [1-5].

TABLE I. AVAILABILITY, ANNUAL DOWNTIME AND GENERAL SOLUTION

Availability	Annual downtime	General solution
99%	87.6 hours	one server
99.9%	8.76 hours	one server, redundant design, very good configuration and maintenance
99.99%	52.5 minutes	cluster
99.999%	5.25 minutes	cluster, geographically dispersed cluster

High Availability (HA) is generally achieved by eliminating single points of failure. This is accomplished by adding redundancy to various parts of a system. When considering HA solution all components employed in process of providing service to user should be considered. This would include application servers, databases, networks (e.g. ring topology) and network equipment (routers and switches), uninterrupted power supply, air-condition, datacenter building safety and datacenter staff skills. In this paper concentration is on application servers.

HA at the level of application servers is solved by some kind of load balancing, often using Network Load Balancing (NLB) clusters. Nodes of a NLB cluster share among themselves common user load (load distribution) but also perform the failover in case of node failure.

Dispersing cluster nodes to two or more locations (multi site clusters) further eliminates the single point of failure. In case of a software or hardware failure at one location service continues to be provided from remaining locations. Such clusters provide maximum protection against physical disaster. For easier understanding concepts of geographically dispersed clusters map with four regional centers is shown in Fig. 2. Conventional geographically dispersed cluster that uses physical servers (no virtualization) is shown in Fig. 3, [1,5].

Figure 2 Four corporate regional centers of HEP (Croatian Electricity Company)

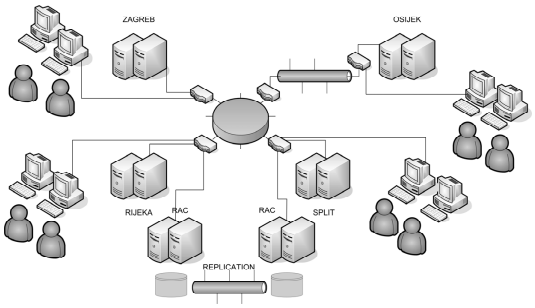


Figure 3 Geographically dispersed cluster

III. APPLICATION SERVER CLUSTERS ON VIRTUALIZATION PLATFORM

A. *Application Server Clustering*

Application server clustering is the use of a group of application servers that coordinate their actions in order to provide scalable, highly available services in a transparent manner. Geographically dispersed virtual clusters are considered in an example of a company with four regional centers.

B. *Virtual Clustering*

Virtual clustering is accomplished by connecting Virtual Machines (VMs) into a cluster, as shown in Fig. 4, [1,6,7]. Virtual machine is software layer which emulates the functionality of a certain machine or processor on a target machine. Main environments for virtual machines are VMware ESX and ESXi, Windows Hyper-V and recently Oracle VM VirtualBox. Type-1 hypervisors (native or bare metal) are commonly used for this purpose.

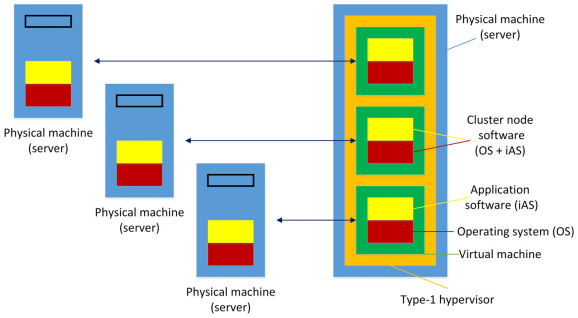


Figure 4 Virtual clustering (with cluster node software OS+iAS)

Clustering concepts considered in this paper use Oracle internet Application Server (iAS) 1.0.2 (could be some other version as well) and Microsoft Windows Server 2008 installed on Hyper-V virtual machines (VMs).

C. *NLB Cluster of Application Servers on Virtualization Platforms*

Application servers installed on virtual machines on the same virtualization platform can be connected in NLB in much the similar way as when installed on physical servers, Fig. 5 [5,8,9]. NLB is configured in Unicast mode, affinity Single (support stateful sessions) and uses virtual network, virtual network interface cards (vNICs) and virtual network switch within a virtualization platform. Each VM use two vNICs, one for the NLB and one for the administration and outgoing NLB traffic.

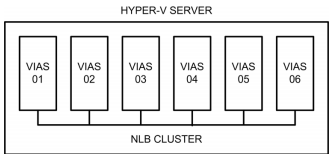


Figure 5 NLB cluster of application servers on a virtualization platform

For redundancy reasons a physical server where virtualization platform resides should be equipped with hard disks in RAID, multiple physical network interface cards (NICs) configured in network teaming and connected to redundant switch(es). Building a NLB cluster of application servers on virtual machines has some advantages. User capacity of the application server is limited by the application server itself and characteristics of an operating system. Single application server can serve few hundreds of users. On a virtualization platform, numerous instances of application server can be installed each on its own virtual machine, all on the same physical server, using less hardware resources and serving several thousands of users.

Different licensing models for application servers often exist (depending on a software provider). Two common models are per user and per CPU licensing. In per user model, licenses are charged according to the number of users. Per CPU licensing model charges are determined based on the number of CPUs on a physical server. The later favors use of virtualization, because multiple instances of application servers installed on virtual machines, are covered by the number of CPUs on a physical server with the virtualization platform (profitable for more than 50 users per CPU). Some providers support licensing per virtual CPU (vCPU) that adds flexibility if not all physical CPUs are needed.

Popular Oracle internet Application Server (iAS) used in considered concepts has very light CPU usage. If installed on a physical server processing power remains heavily underutilized. When application servers are installed on VMs and these VMs clustered using NLB utilization of processing power on a physical server is much better. Each VM with OS and installed iAS uses approx. 4 GB RAM. Even when clustered on the same virtualization platform CPU usage still remains low while RAM usage is the sum of all installed OS+iAS.

Adequate supervisory mechanism (some kind of Sentry) must be install on each virtualization platform. Its role is to supervise the operation of application servers and virtual machines by periodic checks and to ensure that only good instances of iAS/VM are included in a NLB cluster. It can also issue automatic corrective actions on iAS/VM (restart of iAS component or the whole iAS/VM).

IV. GEOGRAPHICALY DISPERSED CLUSTER ON VIRTUALIZATION PLATFORMS

By dispersing cluster nodes to two or more locations protection against the failure of one or more sites is achieved. Virtualization enables better utilization of resources at all location and consequently also at the level of the geographically dispersed cluster as a whole.

A. Integration of Geotargeting Redirection and NLB

Geographic dispersed clusters described here use independent NLB clusters of application servers on virtualization platforms and a single physical server at each location. Independent NLB clusters are integrated into one geographically dispersed cluster using the two level geotargeting redirection, [5,10]. First level is the DNS based redirection that suffice for most situations and is accomplished with multiple A-records for the geographically dispersed cluster web address. Subnet prioritization assigns a user to a regional NLB cluster closest in terms of the NLB cluster IP address. As these addresses are typically regionally assigned, user is directed to the closest (regional) NLB cluster. If that NLB cluster is not available, user browser will after 20s try to access next from the list, eliminating the single point of failure at the level of geotargeting redirection. Example of address space ranges is show in Table II. Addresses in Table III correspond to server side script redirectors and application servers. Same IP address can be used for iAS and redirection script using different address extensions.

TABLE II. EXAMPLE OF CORPORATE ADDRESS SPACE (HEP)

Regions represented by regional center	**IP address range**
Osijek	10.0.0.0 – 10.63.255.255
Rijeka	10.64.0.0 – 10.127.255.255
Split	10.128.0.0 – 10.191.255.255
Zagreb	10.192.0.0 – 10.255.255.255

TABLE III. MULTIPLE A-RECORDS FOR THE SAME WEB ADDRESS

Cluster address	**Multiple A-records in DNS (server side Perl script redirectors)**		
hepweb	IP address of NLB cluster	*url* for server side Perl script	*url* for applications (redirected from Perl script)
	10.193.64.116	*hepwebos*	*hepwebos/applications.html*
	10.65.64.206	*hepwebri*	*hepwebri/applications.html*
	10.129.64.116	*hepwebst*	*hepwebst/applications.html*
	10.1.200.206	*hepwebzg*	*hepwebzg/applications.html*

Exceptions and redirection to remaining sites in failover (or overload) scenarios are handled by the second level of redirection performed by server site scripts installed at all nodes of each NLB cluster. NLB clusters exchange their statuses among themselves. Each NLB cluster advertises its current user load. Based on this info the second level redirection can make smart redirecting decisions when the NLB cluster at some location is overloaded. Please note that NLB in described solutions is limited to a particular location and not used across all locations, this is the role of the geotargeting redirection.

B. Two Site Solution

In this solution servers with virtualization platforms are installed at two regional centers, Fig. 6. Users are directed using the two level geotargeting redirection (DNS and server side script) to the NLB cluster located at the closest regional center. This is a basic, two site, geographically dispersed cluster with both locations serving users simultaneously. It is HA solution where, when one location fails, remaining location immediately

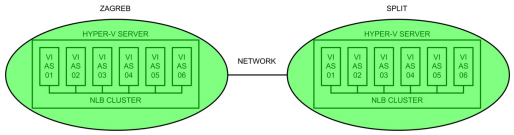


Figure 6 iAS cluster on two Hyper-V platforms on two locations, both location in normal operation, user is directed to the nearest location

takes all load, Fig. 7. If the failure lasts longer, A-record of the failed location can be temporarily revoked from DNS.

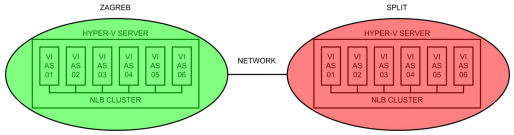


Figure 7 iAS cluster on two Hyper-V platforms on two locations with failure at location Split, all users served in Zagreb

Capacity of the application server cluster at each location must be dimensioned to serve additional users. To avoid too generous overprovisioning at each location for the rare case of accepting load from the failed site, slight degradation in performance and user experience may be acceptable (e.g. greater response time for some reports). There should be a compromise between the overprovisioning and performance in case of the failover.

C. *Four Site Solution*

More redundant solution is to use four geographic locations. Physical servers with virtualization platforms are installed at four regional centers, Fig. 8. All clusters at all locations should be up and running. Users are again directed using the two level geotargeting redirection (DNS and server side script) to the NLB cluster located at the closest regional center. In case of a failure at a particular location, all remaining sites accept users from region with failed cluster, Fig. 9. A-record belonging to the failed location can be temporarily revoked from the DNS.

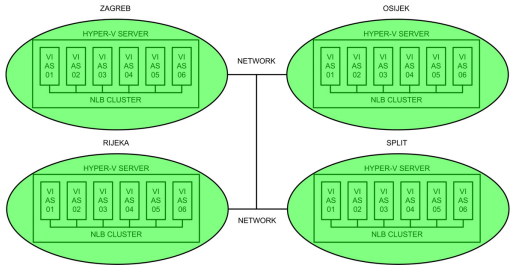


Figure 8 iAS cluster on four Hyper-V platforms and four locations, all four locations are active, user is directed to the nearest location

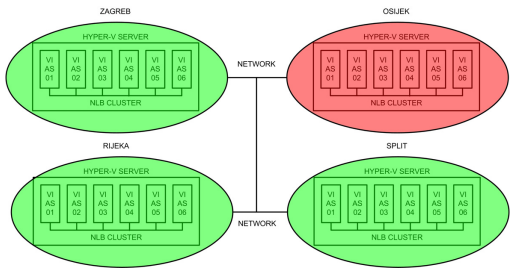


Figure 9 iAS cluster on four Hyper-V platforms and four locations, one location failed, users form Osijek region directed to remaining locations

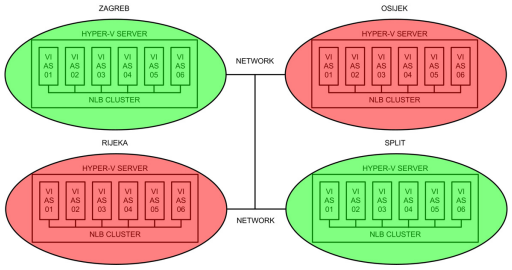


Figure 10 iAS cluster on four Hyper-V platforms and four locations, two locations failed (Rijeka and Osijek), users form regions Rijeka and Osijek directed to remaining locations

In case of failover load at remaining sites will typically increase for about 33% and this doesn't require too much overprovisioning. With failure of a second site, Fig. 10, some degradation in user performance and experience should be tolerated. If there are failures at three sites, Fig. 11, the remaining site should accept only fraction of all users doing most important tasks.

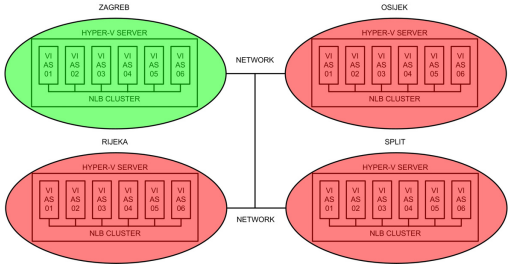


Figure 11 iAS cluster on four Hyper-V platforms and four locations, three locations failed (Rijeka, Osijek and Split), all users directed to the remaining location

Because it uses four locations with four physical servers this solution is expensive, particularly from the point of software licensing (physical servers for providing this solutions are generally far less expensive).

D. *Two Sites Up of Four Site Solution*

Servers with virtualization platforms are installed at four locations. However, only two NLB clusters are in service at any given moment, Fig. 12. This is chosen for licensing reason. Licenses (based on the number of CPUs on physical servers) are charged only for two active servers. Users are directed using the two level geotargeting redirection (DNS and server side script) to the closest NLB cluster. Two A-records are used for cluster address. In case of failure of a NLB cluster at a particular location additional NLB cluster is activated from remaining sites. Failover is almost immediate while failback is performed outside normal working hours. After failover A-record of the failed NLB cluster is replaced in the DNS with the A-record of activated NLB cluster. Choices for two initially active locations are the best equipped datacenters with remaining two acting as spares. Spare locations can be in two states, either with servers turned off or with servers turned on but with VMs turned off. In either case application servers on virtual machines should be brought into operation. In case of turned off servers remote power on command may be issued from remaining operating cluster. This should be followed by a short period (about a

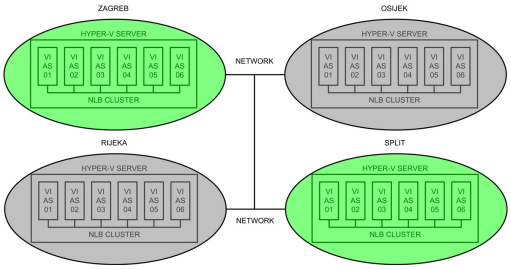


Figure 12 iAS cluster on four Hyper-V platforms and four locations with only two locations active (Zagreb and Split)

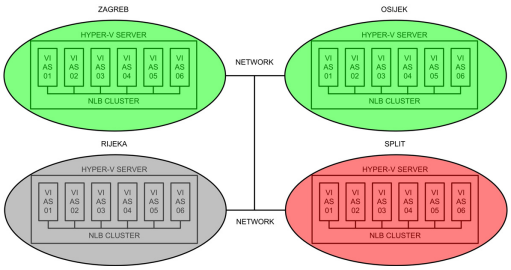


Figure 13 iAS cluster on four Hyper-V platforms and four locations with only two locations active (Zagreb and Osijek) after reconfiguration (failure at location Split)

minute) during which started site update its applications code and application sever configuration files. If licensing conditions are more relaxed, potential failover sites may have its servers powered on all the time and only virtual machines are raised when necessary. Normal operation of cluster is illustrated in Fig. 12, and failover in Fig. 13.

V. OVERPROVISIONING FOR FAILOVERS

NLB clusters must be dimensioned for failovers. They should be able to serve additional users in case of failure of one or more locations or restart of node within a cluster.

A. Levels of Overprovisioning

Overprovisioning of resources can be done at two levels: at the level of individual application server and at the level of virtualization platform.

1) Level of Internet Application Server

Internet application server on each node may be dimensioned to handle more users than needed in normal operation of cluster, Fig. 14. This may increase RAM memory requirements but also requires some changes of software settings on the application server and OS. Number of active application servers on virtualization platform remains the same during failover.

2) Level of Virtualization Platform

Necessary disk (in RAID) and RAM memory space is reserved for import of VMs from other locations, Fig. 15. Once imported VMs should be shut down and kept ready

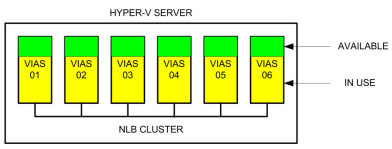


Figure 14 Available resources provided on each application server

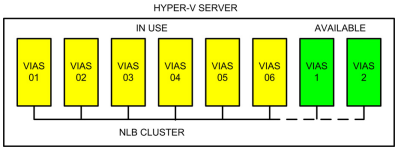


Figure 15 Additional resources for import of VMs

for failover situations. Number of active application servers (and hence virtual machines) on a virtualization platform is increased during the failover.

B. Failovers

Most difficult situation is encountered when the cluster has only two locations with two virtualization platforms. In normal situations each location serves approximate 50% of users (assuming similar loads across regions). If one location fails all user should be served by the remaining location. It would theoretically require 100% overprovisioning reserved at each location. This may be wasting of resources because real life situations often don't require such generous overprovisioning. In most cases a failed site would recover quickly (within several minutes). Many users connected to a failed site are idle users and in these several minutes they will not reestablish their sessions and hence will not overload the remaining site, as illustrate in Fig. 16 and 17.

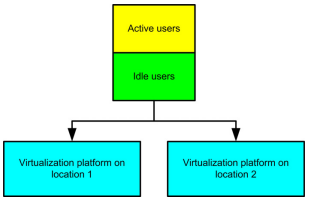


Figure 16 Normal operation on two virtualization platforms

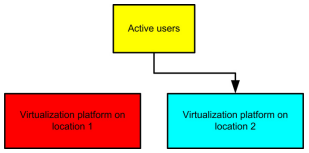


Figure 17 Failover of active users (idle users not reconnecting)

In reality lower numbers like 70-80% overprovisioning may suffice. Also some limited degradation of service (in terms of longer response times) may be acceptable for relatively short period until failed location recovers. Longer outages of location would require full overprovisioning, Fig. 18. When there are four locations, any single failure of location is far less

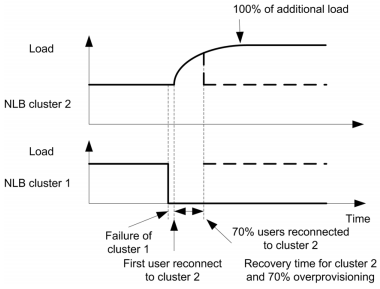


Figure 18 Users reconnect to another NLB cluster

TABLE IV. FULL OVERPROVISIONING REQUIREMENTS FOR VARIOUS NUMBERS OF FAILED LOCATIONS

Solution	Failed locations		
	1	2	3
Two site	100*	-	-
Four site	33*	100*	400**
Four site - two sites up	0	0	100*

*If failed location is quickly recovered (order of minutes) lower numbers (not full overprovisioning) may suffice, **Not feasible

serious. Overprovisioning of 33% will fully cover such situations (again in reality less then this is necessary). Additional location failures would theoretically require 100% overprovisioning, and again lower values should be adequate in real life. Full overprovisioning requirements (assuming equal number of users coming from each region) for each solution and the number of failed locations are listed in Table IV.

C. *Migrations, Replicas and Export/Import*

Migration of failed VMs to a new location is typically not possible once the original location failed. Copies must be made in advance before failure of any location occurs.

Replica of VMs can be easily used among two or more virtualization platforms at the same location and for disaster recovery. All VMs (including replicas) at the same location can be included in the same NLB. Replicas are not so suitable for NLB clusters at different locations due to assigned different address ranges and alternate IP address option must be used. However alternate IP address option doesn't solve problem of NLB settings when replicas are at different locations.

Simple export/import of virtual machines may be used instead of replicas, Fig. 19. This must be done before failures occur. Configuration of each iAS depends on numerous configuration files and these should be updated in when the imported virtual machine at different location is powered up. Imported virtual machines can be easily configured in NLB at new location and after configuration turned off till necessary. Main usage of resources for additional VMs is necessary disk space for each imported VM. Memory usage is nil till the VM is turned on.

Question remains, isn't it simpler to have few more VMs connected in NLB at each location and initially provide 33% overprovisioning instead of complicating with export/import of VMs from all locations. Reliability likes simplicity.

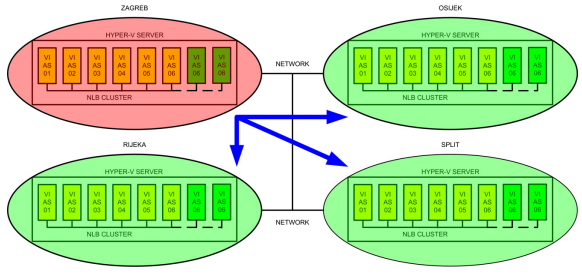


Figure 19 Export/Import of VM among locations, this must be done before failure of the location occurs

VI. COMPARISON

Comparison of described solutions regarding the availability, user capacity, hardware and software licensing cost is shown in Table V. All solutions provide redundancy through geographic dispersion to more locations. Two site solution has advantage of lower cost, but disadvantage of lower availability (still high enough for many applications) and lower user capacity. Four site solution has advantages of very high availability and user capacity, but disadvantage of high hardware and software licensing costs. Four site, two site up, is a compromise.

TABLE V. COMPARIONS OF PRESENTED SOLUTIONS

Solution	Availability	User capacity	Hardware costs	Software costs
Two site	**	**	**	**
Four site	****	****	****	****
Four site - two sites up	***	**	****	**

VII. CONCLUSION

For highly available business critical applications it may be necessary to use geographically dispersed clusters. Virtualization platforms enable use of several application servers installed on separate virtual machines and connected in a NLB cluster, all on the same physical server. This provides better hardware utilization and lower hardware cost, but what is more important, use of per CPU licensing model assures significantly lowered software licensing costs. Availability is further increased by integration of NLB clusters into geographically dispersed cluster using two level geotargeting redirection (DNS based and server site script) and eliminating physical servers as a single points of failure. Three solutions for geographically dispersed cluster that use just one physical server at each location have been presented and analyzed. Each approach is suitable for different requirements on availability and user capacity but also different budgets for hardware and particularly software. The sparing concept of geographically dispersed cluster with some locations powered down till required that combines HA and low software licensing costs may be a good compromise.

REFERENCES

[1] D. Miljković, "Review of Cluster Computing for High Available Business Web Applications", MIPRO, GVS, 26-30 May 2008, Opatija, Croatia

[2] Oracle Application Server High Availability Guide 10g Release 2 (10.1.2), Oracle, February 2007

[3] N. Stojanović, Teorija pouzdanosti tehničkih sistema, BIGZ, Beograd, 1987

[4] Five Levels of High Availability, Compaq, 1998

[5] D. Miljković, "Geografski distribuirani iAS NLB cluster", 10. HROUG konferecija, Umag, 2005

[6] K. Hwang, J. Dongarra and G. Fox, Distributed and Cloud Computing: From Parallel Processing to the Internet of Things, China Machine Press, 2012

[7] Setup for Failover Clustering and Microsoft Cluster Service, VMWare, 2013

[8] Overwiew of Network Load Balancing, Internet: http://technet.microsoft.com/en-us/library/cc725691.aspx [20 March 2015]

[9] D. Miljković, "iAS Cluster on Hyper-V for Capacity and Availability", 19th HrOUG, 14-17 October 2014, Rovinj, Croatia

[10] D. Miljković, "Methods for Load Balancing For Application Servers", 16th HrOUG, 18-22 October 2011, Rovinj, Croatia

Geolocation and data type impact on SQL query execution time using different database approaches

K. Šimenić, I. Lukić and M. Köhler
Faculty of Electrical Engineering, Department of Computer and Software Engineering
Kneza Trpimira 2b, 31000 Osijek, Croatia
ksimenic@etfos.hr, lukic@etfos.hr, mkohler@etfos.hr

Abstract - Picking right database approach or connection – Application Programming Interface in PHP used to achieve connection to database server, together with database and table definition affects query execution time. Servers are located in different geographical positions affecting round-trip time (RTT). Four different database connections are used together with queries consisted of same data trying to point time difference using PHP and MySQL database server. Most common data structures are tested both for time difference and needed storage space. Combining data types and choosing a good database connection in terms of simplicity, security and time needed for execution, lead towards saving space and time. Time and storage problems often occurs when operating with large amount of data which will be case in this paper. Also, providing facts and results, best database connection and table optimization in default database configuration are given. Later optimization can be achieved in terms of choosing different database engines for specific needs.

I. INTRODUCTION

Millions of database queries using different database connections are being run every second on web browsers. There are many types of database servers and their configurations. By configuration we assume: type of MySQL Server, number of concurrent connections, default character set and others explained in database configuration documentation [1]. Nowadays, the collection of data from tables and the providing of results based on it should not only be correct; it must be as quick as possible too. Choosing the best way to access data, by using one of the four ways to connect to MySQL database in PHP are tested in this paper, in terms of query execution time, code simplicity and data safety is discussed. Similar work with Java Query Language is presented in [2] and query optimization in [3]. The geographical location of a database server has a great impact on both the execution time and the database connection which is also presented.

Operating with large amount of data necessitates large storage capacity. The optimization of fields with different data structures affects both execution time and needed storage space. How to choose the most appropriate field data structure is based on the type of data stored for later usage. In this paper, the focus is on character-based structures such as character (CHAR), variable-length character (VARCHAR) and TEXT. The same principle can be transferred to numeric based structures.

Good database server configuration results in almost no difference between database connections. One example of good table definition, or choosing proper cell structures, engine and character set, results is presented in terms of inserting specific data. Finest optimization is achieved by defining good table cell data structure which is strictly defined by the type of data needed to be stored. Basic examples of choosing proper cell structures are discussed and all methods of inserting and collecting data between the user and the database are shown.

II. CONNECTION TYPES OVERVIEW

Four database connections (PHP API – Application Programming Interface [4]) are tested, two of which are object-oriented (OO): PDO or PHP Data Objects and MySQL improved (MySQLi) OO ; and two procedural: MySQLi procedural and MySQL all of them described in literature [4 and 5]. Using PDO or MySQLi over MySQL provides the solution for basic security vulnerabilities such as SQL injection [6] using escape strings and binding parameters. PDO already has a built-in function to quote strings, while in MySQLi this should be done in code. An improvement of MySQLi over MySQL is mostly in the field of security, but it also has an OO version which is more and more used nowadays because of code simplicity. MySQL and MySQLi only support the MySQL database driver, while the PDO supports up to 12 different database drivers including MySQL. Another advantage of PDO is "named parameters". It is much easier to bind specific values to a character based parameter name rather than to numerically based names which follow a specific order of placement in a code. All security and code advantages and disadvantages result in different database insertion query time presented in section III.

TABLE I. SERVERS' GEOGRAPHICAL LOCATION AND RTT

Name	Geographical position	RTT [ms]
gimli.velv.hr	Croatia, Varazdin	22
core-design.info	France, Roubaix	42
directorycentral.com	USA, Newark	115

The list of servers used for testing of database insertion time differences (section III) and presentation of good database configuration (section V) is shown in table 1. The first two servers have a basic database server configuration, while the third one, located in the USA, is well-configured owing to its specific usage for commercial purposes. The goal is to show how good database configuration can minimize the differences between database connection methods. Time invested in such a configuration should be proportional to the usage of the specific web place or more of them located on the same server.

Aside from database connection methods and good database server configuration, good column structure definition also affects not only query execution time while selecting or updating data, but required storage capacity as well. In section IV, average execution time is tested using all four connections with SELECT statement on tables pre-populated with data. Four different tables are used, each one for specific character structure populated with the same data earlier. All methods are implemented and tested in PHP (HTML web page) on XAMPP localhost server on a laptop with an Intel(R) Core(TM) i5-3230M CPU@2.60GHz, 6GB of main memory and Thomson(R) TWG850 router on a 6 Mbit/s downstream and 0.5 Mbit/s upstream connection.

III. QUERY TIME DIFFERENCES BETWEEN CONNECTIONS

One of the four SQL fundamental data manipulation language (DML) statements is INSERT statement. INSERT statement is used for inserting rows of data into a specific table in a database. Sometimes, there is a need for inserting large amounts of data inside a specific table or more of them. The tests were conducted by inserting 50000 rows of data using four different connections two times on two different servers in order to get a uniform execution time. In addition, total execution time diagram is presented as well as a table showing the average time difference from RTT.

A. *Insertion*

In Figure 1, the results shown represent the total execution time for every 12500 rows of data on Croatia server. It is obvious that PDO is the slowest among all four connections, while both MySQLi connections are almost equal in best execution time. In Figure 2, the same results are shown for France server. Differences in the execution time are smaller because the time for code execution stays the same and has a smaller impact when compared to RTT. France server definitely shows the most expected values where MySQL, because of its simplicity and lack of security (which needs time), shows the best results while both OO connections are basically the same. Results presented in Table 2 show the average time for all connections on both servers.

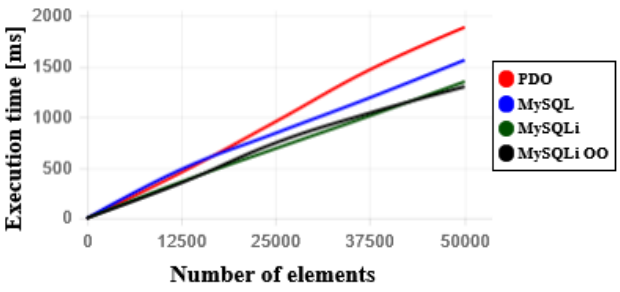


Figure 1. Total execution time on Croatia server

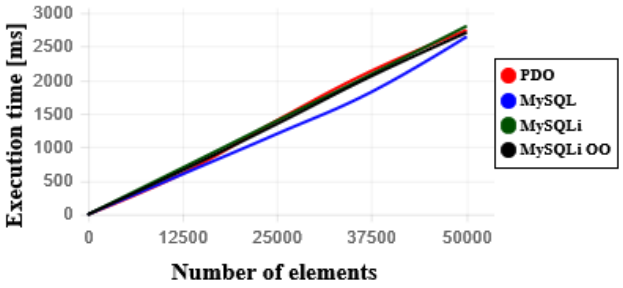


Figure 2. Total execution time on France server

TABLE II. AVERAGE TIME OF EXECUTION

Method	Average time [ms] CRO	Average time [ms] FRA
PDO	37.68	54.90
MySQL	31.20	52.94
MySQLi	26.92	56.11
MySQLiOO	24.28	54.18

B. *Difference*

Average time difference from RTT with percentage difference from RTT is presented in Table 3. The data show the time needed for data processing on the local machine (web page server or localhost server machine power) and on the database server, which depends on database configuration, machine power and response time. Machine power is defined with processor power, RAM type and capacity and type of disk units used for data storage. The variations in collected data on both servers are caused by physical differences between the machines combined with other requests that were currently processed.

TABLE III. AVERAGE TIME DIFFERENCE

Method	Average difference [ms] CRO	%	Average difference [ms] FRA	%
PDO	15.68	71.27	15.95	39.87
MySQL	9.97	45.35	20.39	8.16
MySQLi	8.91	40.49	10	25
MySQLi OO	1.89	8.63	21.16	25.90

IV. PROPER CELL DEFINITON

Storing character-based data, such as first and last names is present in almost every database. Optimizing table cells for specific usage affects execution time. Sometimes, indexing and searching by primary and unique keys is impossible. Time needed for data retrieval (SELECT statement) by using WHERE condition on non-indexed character-based cells varies, which depends on the cell definition as well as the conditions outlined in section IIIB. The same principle for insertion can be applied to SELECT and every other DML statement. As mentioned earlier, the primary focus is on character-based structures. MySQL servers used in tests support CHAR(X), VARCHAR(X) and TEXT fields and those structures were tested. X represents the maximum number of characters that can be stored. The difference between CHAR(X) and VARCHAR(X) is that CHAR always stores X number of characters (if provided with less, it fills the string with blank spaces until X characters is reached) while VARCHAR(X) can hold up to X characters but stores provided characters only. For example, storing "ABC" in CHAR(5) would look like 'ABC--' (where '-' represents a blank space) and in VARCHAR(5) the result would be 'ABC'. CHAR(X) is good for storing a fixed number of characters, e.g., md5 hashed passwords or data that differs in one or two characters from X, while VARCHAR(X) is good for storing data that can differ in many characters e.g. first or last names, but should be of reasonable length (20-30 characters). In comparison to VARCHAR(X), which is limited to 8000 characters, TEXT can store up to 2GB of data with no significant difference in the way data is being stored. VARCHAR and CHAR supports only ASCII values while TEXT can store Unicode characters too.

Four different tables, one for each data structure including two for CHAR, are populated with the same data. The first CHAR table has CHAR(10) structure and another has CHAR(15). There are 5 columns (name1-name5) which specify one row of data; and all have the same structure. One million names were randomly generated and inserted into table (first five names – first row, second five – second row), creating a total of 20000 rows. The SELECT statement used for testing is:

```
SELECT * FROM table_name WHERE name1 = "Robert"
```

The results presented in Table 4 show the average time of 1000 identical query executions with all four connections on Croatia server. All four results are rather close, but CHAR(X) results worsen as X grows. On the other hand, TEXT and VARCHAR(X), no matter how large X is, and because of the identical way of storing data are practically the same and change only when data with more or less characters is inserted.

Not only does the execution time worsen, but also larger capacity is needed to store CHAR because it fills strings with blank space characters. In Table 5, storage capacity after populating of tables is presented. The problem is that the small differences (sub-1 MB) in size are rounded to specific values by the table information schema. Still, it is obvious that CHAR takes up more space for the same amount of data; contrary to VARCHAR and TEXT.

TABLE IV. SELECT STATEMENT TIME OF EXECUTION

Type	Average time [ms]
CHAR(10)	33.28
CHAR(15)	34.67
VARCHAR(X)	36.02
TEXT	34.35

TABLE V. STORAGE CAPACITY

Type	Capacity [MB]
CHAR(10)	2.51
CHAR(15)	2.51
CHAR(20)	3.51
VARCHAR(X)	1.51
TEXT	1.51

V. SERVER CONFIGURATION ADVANTAGES

The prospects of good database server configuration are multiple. When the differences between the database connections are small, it allows the owner to provide a more flexible and complete service. The users could adapt their code to their liking worrying about performance issues. In Figure 3, the total inserting time for all four methods is presented. In table 6, the average time for execution is presented and it is visible that there is almost no difference in query execution time among all methods.

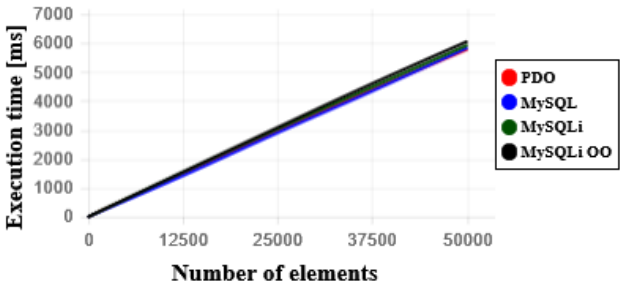


Figure 3. Total execution time on American server

TABLE VI. AVERAGE EXECUTION TIME

Method	Average time [ms]
PDO	115.73
MySQL	116.74
MySQLi	118.52
MySQLi OO	121.16

The average time difference for all methods, presented in table 7, is almost under 5%. One of the reasons is the connection via "level3" network, which reduces the number of points in network from source to destination.

Strong and high-quality connection altogether with a powerful machine and a good table definition will result in the smallest possible execution time. Having the database on the same machine as the web page, or minimizing the number of hops through the network decreases both the RTT and the total execution time.

TABLE VII. AVERAGE TIME DIFFERENCE

Method	Average difference [ms]	%
PDO	0.73	0.63
MySQL	1.74	1.51
MySQLi	3.51	3.06
MySQLi OO	6.16	5.36

Choosing a proper database engine and defining a good table will, as presented, affect storage capacity and time for query execution in a positive way. This is not particularly relevant when speaking of web places connected to databases with small amount of data e.g. few users, blog posts or similar. Extracting large amounts of data, e.g., an administrator's page for user administration on a web page with a large amount of users will be significantly faster in data retrieval and updating if the necessary table is properly defined.

VI. DISCUSSION AND CONCLUSION

In this paper, a basic overview of how database connection and RTT affects total execution time for inserting statements and how SELECT statement execution time and needed storage capacity depends on character data structures is provided. There are many factors that affect query execution time. The most fundamental ones are the physical attributes of the machine and RTT. From a coding point of view, a good database server configuration, the most appropriate database storage engine and a good table definition affects not only execution time but storage capacity needed as well. Most of the conclusion can also be applied to DML statements and data type. Complex statements with given conditions will be executed faster if the columns are indexed either by primary or unique keys.

The connections are largely similar in execution time. PDO improves as RTT rises, while MySQL is almost always the fastest type of connection because of its simplicity. Both MySQLi connections are acting similar regardless to RTT.

Using PDO allows for an easy switch if/when database driver is changed and the same security level as MySQLi makes it the most reasonable type of connection to use. However, MySQLi provides all the neccessary actions for MySQL database driver; and with similar OO interface it is still a good option.

Depending on the amount of data stored, the frequency of usage and the number of users, the right connection type, table definition and server configuration should be implemented. Optimization is a slow and painful process of learning and tracking what is best for use. Purchasing and installing the right equipment for its intended use will positively affect query execution time and this must be considered when creating a database environment.

ACKNOWLEDGMENT

We thank Filip Stojić for providing us with the servers to conduct our tests on.

LITERATURE

[1] Atlassian official database configuration documentation https://confluence.atlassian.com/display/DOC/Database+Setup+For+MySQLi [accessed 12 April 2015]

[2] V. Krishna S. Nerella, S. Surapaneni, S. K. Madria, T. Weigert, "Exploring Query Optimization in Programming Codes by Reducing Run-Time Execution", *COMPSAC*, 2010, 2010 IEEE 34th Annual Computer Software and Applications Conference (COMPSAC 2010), 2010 IEEE 34th Annual Computer Software and Applications Conference (COMPSAC 2010) 2010, pp. 407-412, doi:10.1109/COMPSAC.2010.48

[3] M. Jarke, J. Koch, "Query Optimization in Database Systems", ACM Computing Surveys (CSUR), Volume 16 Issue 2, June 1984, Pages 111-152, doi:10.1145/356924.356928

[4] Overwiev of connection types in PHP (PHP database connection API) http://php.net/manual/en/mysqli.overview.php [accessed 12 April 2015]

[5] Overwiev of PHP Data Objects connection type http://php.net/manual/en/intro.pdo.php [accessed 12 April 2015]

[6] L. K. Shar, H. B. K. Tan, "Defeating SQL Injection", *Computer*, vol.46, no. 3, pp. 69-77, March 2013, doi:10.1109/MC.2012.283

[7] K. Tatroe, P. MacIntyre, R. Lerdorf, foreword by M. Bourque, *"Programming PHP"*. 3rd ed., O'Reilly Media, February 2013.

[8] S. M. M. "Saied" Tabagbogbi, H. E. Williams, *"Learning MySQL"*. O'Reilly Media, November 2006.

[9] PHP offical website 2001-2015. Offical PHP docuemntation [online].Available from http://php.net/[accessed 08 February 2015]

[10] WWW Consorcium offical website 2015. Offical W3C web page, web page standards and documentation [online]. Available from http://www.w3.org/ [accessed 08 February 2015]

[11] MySQL official website 2015. Offical MySQL development documentation including standards [online]. Available from http://dev.mysql.com/doc/ [accessed 08 February 2015]

A novel versatile architecture for Internet of Things

Luka Milić* and Leonardo Jelenković**
* University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
luka.milic@foi.hr, leonardo.jelenkovic@fer.hr

Abstract – This paper presents an overview of contemporary architectures for Internet of Things and then introduces a simple novel architecture. The main goal of proposed architecture is to remain simple but still applicable to any Internet of Things environment. Specific applications of IoT should be implemented in application layer, using proposed architecture as backbone. Proposed architecture isn't yet fully defined, but ideas on which is based are well defined and should provide straightforward design and implementation.

I. INTRODUCTION

Internet of Things (IoT) is a relatively new paradigm of connecting anyone, anything, anywhere and anytime. Despite being relatively fresh, the idea is much older. Even though the term IoT is first mentioned by Kevin Ashton in 1999, similar networks were discussed through the 90s. Also, IoT does not necessarily include new technologies, but often only a new use of the old or current technologies, like Radio Frequency Identification (RFID) and Wireless Sensor Networks (WSN).

However, only recently, in the current decade or so, IoT has been studied more intensively, as the technology has been rapidly evolving to enable the IoT. IoT has gone beyond Ashton's original idea of connecting *things* in supply chains to the *Internet* and grew to occupy almost every part of human lives. It has been successfully applied to, for example, climate monitoring, transport and road safety, home automation and building monitoring, health care, supply chain, agriculture and rural development, border security and military application, etc.

IoT is commonly defined as a network of many resource-constrained nodes connected to local networks and then, through the higher layers of IoT platform, connected to everyone and everything. A typical IoT architecture looks as illustrated by Fig. 1. IoT is commonly divided into three layers, possibly with two sub-layers each, as seen in the picture.

The lowest layer, encompassing sensors and the physical world, is called the perception layer. The most prominent technologies in the perception layer are RFID for identifying and sensing things, WSN for sensor communication, and lightweight wireless communication protocols for WSNs and other types of networks. ZigBee seems to be the current state of the art protocol for such cases. Bluetooth is another protocol, but not as adapted for resource-constrained devices. Recently, there has been a lot of effort put into Internet Protocol version 6 Low-Power Personal Area Networks (6LoWPAN) and Constrained Application Protocol (CoAP) for communicating on higher layers. For things less constrained by power or performance, sometimes even their full counterparts Internet Protocol version 6 (IPv6) and Hypertext Transfer Protocol (HTTP) are used.

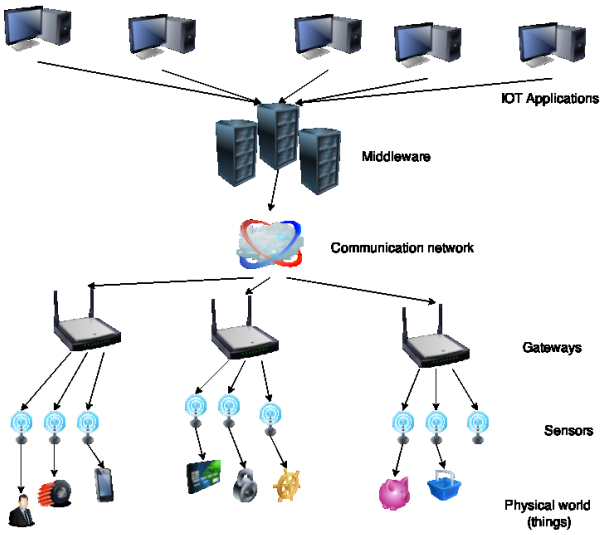


Figure 1. A typical IoT architecture

On the network layers, consisting of gateways communicating through the broadband network, there have been suggestions to use mobile networks, the Internet or even completely new networks. However, recently the use of the current Internet has been taken for granted.

On the application layer, here consisting out of IoT middleware and user applications, the research is extremely diverse. There are many unique or less unique approaches proposed by researchers in order to connect presumably at least billions of things in a working network.

Still, a significant amount of research that has been put into IoT neither means the field has been sufficiently explored nor that satisfying global solutions have been found. A common person still does not know what IoT is, and the deployment of IoT is only in nascent stages, with the little that has been deployed varying in technology and coverage. Also, the researchers have mostly been

concentrated on building IoT for specific cases, effectively building small isolated networks of things, not Internet of Things but rather Intranets of Things. We didn't find adequate architecture proposals in rest of the research, architecture that could be used as base for all IoT applications. Therefore we designed one architecture that could be used as base for implementing IoT.

The rest of the article is organized as follows. In Section II the state of the art is presented and similar conclusions drawn. In Section III our new architecture for IoT is proposed. In Section IV the conclusions and future work are laid out.

II. Related Work

We will now try to summarize main research directions in IoT architectures. Usually, first ones referenced are architectures backed up by global groups. Some of these architectures will be presented here, but it should be noted that there is not much scientific research based on those architectures and we do not think that they have necessary features for a global IoT. Two ongoing Seventh Framework Programme (FP7) projects will also be presented, IoT-A and OpenIoT, as they are European Union (EU) projects relevant to the local scientific community, but also very large projects with important investors.

First architecture is the EPCglobal project [1]. It is promulgated by the EPCglobal organization, successor of the Automatic Identification (Auto-ID) Labs where the term IoT was coined. The GS1 architecture offers three sets of standards. Those are standards for creating a variety of codes for things, standards for representing these codes by graphics and RFIDs, and standards for manipulating the data captured by reading these codes. It is more concerned with creating business models like supply chains than creating a global network.

The second architecture is the Machine Type Communication (MTC) [2] by Third Generation Partnership Project (3GPP). MTC is a set of standards for communication between different things and the core network. It is focused on implementing the communication through Long Term Evolution (LTE) networks and mobile phones. No upper architecture seems to be emerging. Paper [3] presents a potential one, putting servers and applications over it. Communication is rather complex, requiring a lot of telecommunication knowledge in order to push messages from one side to another.

The third architecture is the Machine to Machine (M2M) [4] done by European Telecommunication Standards (ETSI). M2M aims to create applications, built upon service capabilities, in both the upper layer and the gateways, but also the things, which do not need a gateway in that case. It does so by defining three sets of interfaces. These are M2M application interface, device application interface, and M2M to device interface. There does not seem to be concrete technologies defined.

The fourth chosen commonly referenced architecture is the Near Field Communication (NFC) [5] put forward by the NFC Forum, mainly Nokia. This system is used only for identifying and capturing things. It is already built into many smartphones. It does not seem to aspire to building higher layers, but only uses RFID to communicate with sensors and let them interpret the data.

The most successful of these global groups seem to be the Internet Engineering Task Force (IETF). They have constructed two protocols, the more successful of which is the Constrained Application Protocol (CoAP). Recently, it has generated many research papers and its architectures will be presented a bit later. CoAP alone is not complete architecture. It is a counterpart for the Hypertext Transfer Protocol (HTTP) for resource-constrained devices.

The second protocol is the Extensible Messaging and Presence Protocol (XMPP), an even higher-layer protocol family for near real-time communication. An example system in [6] builds a basic platform for bidirectional communication based on publish-subscribe and request-response.

Currently, various visions of IoT seem to converge in implementing IoT using the current Internet. There are architectures like MobilityFirst [7], which envision a complete overhaul of the current network. The researcher developed a new vision or the internet, with the main goal being extreme mobility of the members, which they see as the ideal environment for the IoT. In the paper they specify exactly how IoT is combined with the network. Still, visions of replacing the current internet seem more and more far-fetched.

Searching for more implementable global architectures, the first thing visible is the huge work targeted on creating small, isolated IoTs, here dubbed "Intranets of Things". After a brief search on, for example, Institute of Electrical and Electronics Engineers (IEEE) Explore, many narrow niches like electronic commerce, smart supermarkets, rural patient tracking smart gas tanks, smart scenic sites, smart food, electronic identities and much more pop out. A good architecture needs to be flexible, meaning it can be used for many use cases, unlike those. Only architectures intended for a global IoT will be further analyzed. Also, there has been much work in order to completely fixate the IoT to the current Internet. Host Identity Protocol [8] has been used for IPv4 and IPv6 Low-Power Personal Area Networks (6LoWPAN) for IPv6. The paper referenced deals with address translations.

Some common architecture proposals will now be presented. The first is Service Oriented Architecture (SOA). An example in [9] outlines a network of service entities. These are connected with services, actually general data, being sent through channels or links. It is just a sketch of a potential model and its behavior.

SOA is borrowed from the Semantic Web, and so is Representational State Transfer (REST). Actinium [10] is an example of a RESTful runtime container for scriptable IoT applications. It uses CoAP for communicating in a real network. The user needs to implement applications in a RESTful fashion to communicate via CoAP, and still seems to only get an Intranet of Things. As already said, CoAP is a HTTP counterpart. But even if HTTP is great, binding to four layers of protocols does not have to be.

Another Semantic Web technology is the Simple Object Access Protocol (SOAP). Paper [11] defines a

things management protocol based on it. It relies on Web Services Description Language (WSDL) to define three types of operations, getting information, getting next information and setting information. It does not provide a higher layer.

The use of the ontologies is presented in [12] in order to refine the crude data received from things. Having defined a data model ontology in Web Ontology Language (OWL) and described data sources, data providers, and the data, and also defining concrete ontologies, an algorithm is provided in order to integrate them.

Another CoAP proposal is [13]. It is focused on the EPC and creates a network using ZigBee on the lower layer and CoAP on the higher, also utilizing Universal Plug and Play (UPnP) for discovery mechanisms.

Paper [14] implements SOA through CoAP with the help of peer-to-peer (P2P). A new P2P protocol is put forward in order to create such networks, used for their Distributed Hash Tables (DHT). All in all it seems rather complex, and we believe that if the Internet of the 80's started with a complex architecture, there would be no Internet of today.

There is also a notion of message-oriented architectures. Paper [15] discusses a message-oriented middleware as a set of devices, which can receive, send, store and forward messages between devices, acting as a temporary database with different channels acting like dynamically created pipes.

Having mentioned P2P, reference [16] presents a system built especially upon it. It uses the DHT for identifying and locating things, and also uses other nodes for publishing services. It is still rather abstract, since it is just a helper network.

P2P is just another way of dealing with scalability. Special systems with a focus just on the scalability have emerged.

Some work is focused on treating IoT as a database. This system [17] creates three storage layers as an IoT abstraction and uses database indexes to navigate it. The idea is expanded with more indexing types and supporting infrastructure.

A common buzzword in the scalability is the self-organization. For example, architecture [18] is based on the nature-inspired algorithm of endocrine systems. It uses so-called hormones to, not interfering with each other, communicate between things and slightly modify their behavior. It does not consider that the things will be joined in local networks.

Another architecture is [19], which presents a middleware platform built upon immobile WSNs, which and then successfully spatially grouped and included to the global network.

Sometimes, there is talk about autonomic architectures, which is not connected with self-organization, but seems to emphasize things cannot work with IP. Paper [20] outlines an architecture based on multiple planes with their agents and protocols to abstract the IP in the communication.

Supposed to complex networks, there are systems for only helping existing networks. For example, Virtual Sensor Editor [21] integrates nodes from different networks using graphical component programming.

There are also security architectures built that way, on top of the networks, which seems to introduce more overhead and effort that just including security.

For example, [22] presents a security system based on the HIP Diet Exchange (DEX), extending the key management protocol by a number of ideas like giving the attacker increasingly difficult puzzles. It is also expanded to other key management protocols, but nothing else.

We believe that, conversely, security should be one of the basic features of the IoT.

Also, simplicity is another requirement. Only when our homes and offices get IoT people will care more about it, and that will not be possible with a complicated network.

Finally, some complete architecture proposals will be presented. Paper [23] shows a system with extremely thin client nodes, where the servers are used for presenting their data. It also presents compressed TCP and compressed HTTP, which are used for communication. No security seems to be emphasized, even though things need to implement four layers.

Paper [24] seems to be a large project for a complete architecture with proprietary physical protocols and IPv6, but it is still a work in progress. Gateway servers are envisioned to be used to connect things to the Internet.

Paper [25] sees IoT as a network of gateways without intermediate servers. Not too much is detailed, but it appears to be intended for Intranets of Things. Still, it is modeled to be modular and scalable.

Paper [26] shows a system based on a SOA spin-off called Event-Condition-Action (ECA). Even though envisioning a complete architecture, only the platform intended for making triggers upon the ECA rule and communication with neighboring layers are detailed.

Reference [27] utilizes a similar system like ours, based on M2M and using HTTP with REST on upper layers. It also enables discovery, connections with non-smart things, associating metadata and actuator control.

OpenIoT project [28] develops an open source middleware for getting information from sensor clouds, without having to worry about what exact sensors are used. The OpenIoT architecture is comprised by seven main elements that belong to three different logical planes. In the Utility-App Plane – the application layer, there are three modules: Request Definition – for users to define requests to the IoT, Config & Monitor – for administrators to configure and monitor the middleware and things, and Request Presentation – for presenting responses from the IoT. In the Virtualized Plane – the service layer, there are three modules too: Scheduler – for handling requests to the IoT, Cloud Data Storage – for a cloud database of the IoT, and Service Delivery & Utility Manager – for

delivering responses from the cloud and tracking metrics of the data. In the Physical Plane – the physical layer, there is only one module: Sensor Middleware – for communicating with the things.

IoT-A [29] creates an architectural reference model for the IoT. Four things are identified as building blocks of the model. The vision summarizes the rationale for providing an architectural reference model for the IoT. Business scenarios & stakeholders are the drivers of the architecture work. The IoT Reference model provides the highest abstraction level for the definition of the IoT-A Architectural Reference Model. The IoT Reference Architecture is the reference for building compliant IoT architectures. The Reference Model mainly defines five things as abstractions. The entity model defines Physical Entities (roughly Devices), Virtual Entities (roughly Services), and Users. The data model defines information types. The functional model defines layers, lower to higher, as Device, Communication, IoT Service, Virtual Entity, IoT Business Process Management, and Application. The communication model defines communication layers, higher to lower, as Data, End to End, ID, Network, Link, and Physical. The security model defines security features and their layering. The Reference Architecture mainly defines functionalities each mentioned actor has to implement. An extensive list of desired functionalities is given. Also, there are UML Use Case Diagrams, UML Sequence Diagrams, and interfaces, defined.

III. OVERVIEW OD PROPOSED ARCHITECTURE

Proposed architecture could be roughly divided into four layers:

1. Things layer
2. Communication layer
3. Data storage and control layer
4. Application layer.

Similar division is presented on most IoT architectures. However, operations on those layers and communication between them differentiates architectures. Fig. 2 presents an overview of presented architecture. Short description of concepts for each layer follows.

Things in things layer have their own operation mode we don't model. We assume they need to communicate with layer 3 to fulfill their purpose. Since things could be various, from very simple battery operated devices to powerful computers, we don't define multi-layer communication stack – communication mechanism is broadcast on physical layer. Other things might catch that broadcast message and broadcast it further. We don't propose special communication protocol on physical layer, any current or future will work. We just need them to carry a message with IoT data (described later).

Communication layer provide connectivity between things and data storage and control layer (layer 3). We assume that gateways could be used in this layer to catch message that things broadcast and forward it (using TCP/IP or some application protocol) to layer 3. Gateways will also forward messages in other direction: from layer 3 to devices.

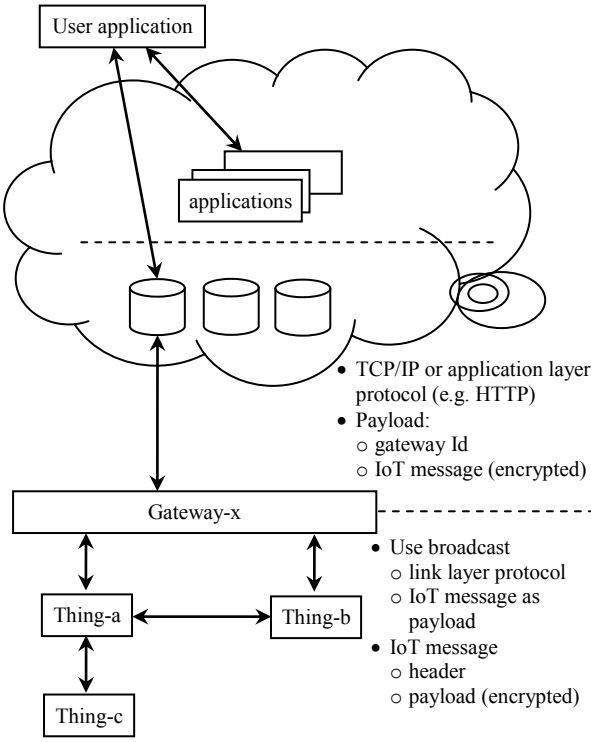


Figure 2. Overview of proposed architecture

Data storage and control layer contains databases with information about devices (and how to reach them - through which gateways), users owning devices, past exchanged information with devices (both received from devices and send to devices), and triggers for control. Triggers could be used to start some action when an event occurs. Events could be simply related with sending and receiving data from one thing or a complex trigger that involves evaluation of received data, combination of data in databases, periodic events (periodically request update from things) and similar. Action on triggers could vary, from simple data retrieval request from things to a start of complex operation implemented in application layer.

Application layer is where things with their data and operations are used for creating some benefit for a end user, such as sensor readings, command dispatch to remote devices and similar. We don't design this layer, we perceive that most required operations in this level can be created using described lower layers we define.

Things with Internet connectivity could act as gateways (directly communicate with layer 3) or connect with gateway using TCP/IP. There is no need to use broadcast if a thing is smarter and can detect internet connectivity, or connectivity with gateway. For connection to gateway it doesn't have to use full TCP/P stack – physical layer protocol could be used, just without broadcast.

Using broadcast in layer 1 looks terribly ineffective, choking the network with packets. However, there are a number of facts helping here. Firstly, the packets are supposed to be short and rare. Secondly, constant ping-pong of the packets must be mitigated using a Time to Live (TTL) mechanism, analogous to the one used by the Internet Protocol (IP). Thirdly, each packet has a local-network identifier. A packet that ends up in a different

local network could simply be dropped. Fourthly, if a thing receives a packet meant for it, it will naturally not forward the packet. Fifthly, as every packet has its identifier so that already forwarded packet can also be dropped.

Some gateways can have a subset of layer 3 functionality, enabling operating with things from local network only, in which applications could use local gateway instead of layer 3 from cloud. Such gateways could still communicate with full layer 3 implementation in cloud (and forward a copy of messages to them).

IV. IoT MESSAGE FORMAT AND SECURITY ISSUES

IoT messages must contain relevant information - payload (per thing specific) but must also contain communication data - header. Header should contain:

a) protocol identifier

b) time-to-live (TTL)

c) thing's identifier (e.g. physical layer address)

d) packet identifier

e) network identifier

f) payload length.

Identifiers could be used to manage messages. A thing can process messages sent to it, broadcast all others or just ones that belong to same network. Packet identifier could be use to manage packet flow (e.g. discard duplicate packets).

Payload could be in various formats but for maximal extensibility we propose simple string, carrying a message (that could be in e.g. Java Simple Object Notation - JSON) with values.

Payload must be encrypted with thing's password (known to thing and layer 3 only). Advanced Encryption System (AES) or similar algorithm could be used to provide sufficient security on this layer.

Gateways could communicate with things on layer 1 with previously described principles. For communication with layer 3 in cloud, standard TCP/IP or some application layer could be used (like HTTP). In a simple implementation, the UDP protocol could be used, where an UDP packet will contain IoT message forwarded from layer 1 or for layer 1. This message could be encrypted with gateway's password (known to layer 3 and gateway only) which could solve security issues on this level.

Security of communication protocols between applications (layer 4) and layer 3 is out of scope of this proposal. For example, some application layer protocol could be used (e.g. SSL).

Management in layer 3 must ensure that only authenticated users (their applications) could manipulate with data and devices that belong to them or they are given appropriate privileges by owners or administrators.

V. COMPARISON WITH IoT-A AND OPENIOT

In this section we compare presented architecture with architectures from projects IoT-a and OpenIoT, as most recent and probably most refined architectures for IoT today.

Comparing proposed architecture with the OpenIoT, the biggest difference is that proposed architecture is incomparably simpler. Proposed architectures is concerned about protocols for communicating with things, protocols for communicating with servers in cloud, definitions of gateways and definitions of a cloud. Also, the Sensor Middleware in the OpenIoT relies heavily on running their web server X-GSN written in Java. In proposed architecture the local network with arbitrary link protocols are used, inside of which is encapsulated our own simple protocol, while gateways speak with the cloud over an arbitrary network layer.

Comparing the architecture with the IoT-A, it is seen that IoT-A defines no physical architecture. Gateways, for example, are only briefly mentioned saying they can possibly be put in the middle to help with the communication. We, on the other hand, put forward actual physical architecture and actual communication. IoT-A only describes what needs to exist, but not how it has to be implemented. Same goes for describing IoT-A communication. Our system is also much simpler (even though ontologically are similar). There are still things, services, and users, here, but the vast amount of requirements listed as functionalities is not considered appropriate for a nascent system. We have only considered basic things needed for a simple IoT.

VI. CONCLUSION

After screening the literature, we didn't found architecture that accomplishes ours requirements of simplicity and generality. We therefore conceptually designed simple, secure and general architecture.

On the lowest level (level 1) contemporary link layer protocols should be used with broadcasting mode for transmission of an IoT packet between things and gateway, which has a communication role in our model (level 2). The IoT packet consists of IoT header and IoT payload which should be encrypted with a contemporary symmetric algorithm using thing's password. Gateways should be used for forwarding messages to and from databases and services located in the cloud (layer 3). Users should observe and control their things with their applications (agents) directly or via the services in the cloud (level 4).

We think that proposed architecture looks promising as a foundation for implementation of tomorrow's IoT. There is a lot of work ahead to prove its feasibility and usefulness with an example implementation.

REFERENCES

[1] EPCglobal project, http://www.gs1.org/epcglobal, (last accessed 1.4.2015.).

[2] Machine-Type Communications, Third Generation Partnership Project, http://www.3gpp.org/, (last accessed 1.4.2015.).

[3] S. Yang, X. Wen, W. Zheng, and Z. Lu, "Convergence architecture of Internet of Things and 3GPP LTE-A network based on IMS", Mobile Congress (GMC), Shanghai, pp. 1-7, 2011.

[4] Machine to Machine Communications, European Telecommunication Standards, http://www.etsi.org/technologies-clusters/technologies/m2m, (last accessed 1.4.2015.).

[5] Near Field Communication, http://nfc-forum.org, (last accessed 1.4.2015.).

[6] S. Bendel et al., "A service infrastructure for the Internet of Things based on XMPP", Pervasive Computing and Communications Workshops (PERCOM Workshops), San Diego, pp. 385-388, 2013.

[7] J. Li, Y. Shvartzshnaider, J. Francisco, R. P. Martin, and D. Raychaudhuri, "Enabling Internet-of-Things services in the MobilityFirst Future Internet Architecture", IEEE Int. symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM), San Francisco, pp. 1-6, 2012.

[8] K. Toumi, M. Ayari, L. A. Saidane, M. Bouet, and G. Pujolle, "HAT: HIP Address Translation protocol for Hybrid RFID/IP Internet of Things communication", Int. conf. on Communication in Wireless Environments and Ubiquitous Systems: New Challenges (ICWUS), Sousse, pp. 1-7, 2010.

[9] A. Zakriti and Z. Guennoun, "Service entities model for the internet of things: A bio-inspired collaborative approach", Int. conf. on Multimedia Computing and Systems (ICMCS), Ouarzazate, pp. 1-5, 2011.

[10] M. Kovatsch, M. Lanter, and S. Duquennoy, "Actinium: A RESTful runtime container for scriptable Internet of Things applications", Int. conf. on Internet of Things (IOT), Wuxi, pp. 135-142., 2012.

[11] G. Dai, "Design and implementation on SOAP-based things management protocol for internet of things", Intelligent Control and Automation (WCICA), Beijing, pp. 4305-4308, 2012.

[12] C. Fan et al., "A scalable Internet of Things Lean Data provision architecture based on ontology", GCC Conference and Exhibition (GCC), Dubai, pp. 553-556, 2011.

[13] H. Hada and J. Mitsugi, "EPC based internet of things architecture", RFID-Technologies and Applications (RFID-TA), Sitges, pp. 527-532, 2011.

[14] D. Tracey and C. Sreenan, "A Holistic Architecture for the Internet of Things, Sensing Services and Big Data", Cluster, Cloud and Grid Computing (CCGrid), Delft, pp. 546-553, 2013.

[15] Z. Peng, Z. Jingling, and L. Qing, "Message oriented middleware data processing model in Internet of things", Computer Science and Network Technology, Changchun, pp. 94-97, 2012.

[16] F. Andreini, F. Crisciani, C. Cicconetti, and R. Mambrini, "Context-aware location in the Internet of Things", GLOBECOM Workshops (GC Wkshps), pp. 300-304, 2010.

[17] G. Ying, J. Huang, and L. GuanYao, "A tag indexing method in the Internet of Things", Electronics, Communications and Control (ICECC), Ningbo, pp. 681-685, 2011.

[18] Y. Ding, Y. Jin, L. Ren, and K. Hao, "An Intelligent Self-Organization Scheme for the Internet of Things", Computational Intelligence Magazine, volume 8, issue 3, pp. 41-53, 2013.

[19] P. Malo, B. Almeida, R. Melo, K. Kalaboukas, and P. Cousin, "Self-Organised Middleware Architecture for the Internet-of-Things", Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), Beijing, pp. 445-451, 2013.

[20] G. Pujolle, "An Autonomic-oriented Architecture for the Internet of Things", Modern Computing, Sofia, pp. 163-168, 2006.

[21] J. Zhang et al., "Supporting Personizable Virtual Internet of Things", Ubiquitous Intelligence and Computing, Vietri sul Mere, pp. 329-336, 2013.

[22] R. Hummen, H. Wirtz, J. H. Ziegeldorf, J. Hiller, and K. Wehrle, "Tailoring end-to-end IP security protocols to the Internet of Things", Network Protocols (ICNP), Goettingen, pp. 1-10, 2013.

[23] S. Bae, D. Kim, M. Ha, and S. H. Kim, "Browsing Architecture with Presentation Metadata for the Internet of Things", Parallel and Distributed Systems (ICPADS), Tainan, pp. 721-728, 2011.

[24] N. W. Bergmann and P. J. Robinson, "Server-based Internet of Things Architecture", Consumer Communications and Networking Conference (CCNC), Las Vegas, pp. 360-361, 2012.

[25] V. Parviainen, A. Yin, A. Romu, and R. Virkkala, "EfiIoT: An efficient software architecture for internet of things", Industrial Electronics and Applications, Singapore, pp. 709-712, 2012.

[26] S. R. Bhandari and N. W. Bergmann, "An internet-of-things system architecture based on services and events", Intelligent Sensors, Sensor Networks and Information Processing, Melbourne, pp. 339-344, 2013.

[27] S. K. Datta, C. Bonnet, and N. Nikaein, "An IoT gateway centric architecture to provide novel M2M services", Internet of Things (WF-IoT), Seoul, pp. 514-519, 2014.

[28] OpenIoT – FP7 project, http://openiot.eu (last accessed 1.4.2015.).

[29] IoT-A – FP7 project, http://www.iot-a.eu (last accessed 1.4.2015.).

Design and Development of the Software Solution for Analysis and Acquisition of the High Voltage Circuit Breakers Dynamic Resistance Measurement results

mr.Kerim Obarčanin dipl.ing.el.*, mr. Adnan Šečić.dipl.ing.el ** and prof.dr.Nijaz Hadžimejlić,dipl.ing.el. ***
* IBEKO Power AB, Stockholm, Sweden, kerim@dv-power.com
** IBEKO Power AB, Stockholm, Sweden, adnan@dv-power.com
*** IBEKO Power AB, Stockholm, Sweden

Abstract - This paper presents design and development process of software solution for the control, acquisition and analysis of the dynamic resistance measurement results for high voltage circuit breaker analyzer devices. Five years' experience in algorithms and software development for the circuit breaker analyzer control and data acquisition and circuit breaker analyzer itself has been summarized in this paper.

I. INTRODUCTION

The design of modern high-voltage puffer type SF6 gas circuit breakers (HVCB) is based on the switching of two parallel contact sets. The first one are the low-resistance contacts – main contacts which are specifically designed to carry the load current without any excessive temperature rise. The second set, tungsten-copper arcing contacts, operate at breaker opening following the main contact part. The electrical arc starts after the separation of the arcing contacts. The tungsten-copper material is designed to carry the arc until it is cleared at the next current zero-crossing value. [1][2]

Measurement of the dynamic resistance of the contact system (DRM) can be used as a tool to diagnose the condition of the arcing contacts. The method consists of applying high DC current through the contact system, measuring voltage drop on the contacts and calculating the resistance waveform during open or close sequence of the HVCB. [3] Relevant information related to contact anomalies, wear and/or misalignment can be obtained by analyzing resistance waveform together with the waveform of the motion of the contact system.

Instruments used to perform measurement of the dynamic resistance of the HVCB are known under the name of Circuit Breaker Analyzer and Timers (CAT). Usually, those are portable field devices with the possibility of generating high current up to 500A and performing measurement with very high sampling rate. [3] DRM Test instrument is used to inject a very high DC current through the contact system, measure voltage drop and current as well as motion of the contact system and calculate the resistance samples. All measurements are obtained with sampling rate of 0.05ms (or higher) and forwarded to the software application system for presentation, interpretation and analysis. [4]

Design and implementation of the software solution has to meet strict requirements for control and data acquisition, follow hardware implementation rules and provide a user with a rich set of features such as report creating, results comparison, appearance adjustment etc.

II. REQUIREMENTS

Design and implementation of this software suite was guided by following these requirements:

- Software Specification Requirements (SRS) defined prior to the development
- Hardware design & implementation specification and features that must be incorporated and supported by software
- Customers feedback

Following those requirements and guidance, the core features of this software are:

- Communication with the instruments - send, handshake and receive using defined communication protocol and firmware command set
- Presentation of obtained results in a graphical and numerical form
- Saving the results in the file format or into the internal database
- Tools for analysis (cursor, results comparison, graphical tools for zoom, pan, adjust operations etc.)
- Report creating, saving and exporting with the customization feature
- Test plan creation prior to going on the field testing and pass/fail tests of the obtained or saved numerical results

III. APPLICATION ARCHITECTURE

Application vertical architecture is structured in the 4 layers:

- Application layer (APP)
- Business Level Logic (BLL)
- Data Access Layer (DAL)
- Communication Layer (COM)

The APP layer contains implementation of the user interface and events. BLL layer is responsible for linking APP layer with DAL and COM layer and creating methods that respond to the events from the APP layer. It also uses DAL and COM layers to perform requested operation.

The DAL layer implements methods related to any kind of work with database and/or storing or exporting results or test plan templates to a file. It is horizontally divided in two groups: database methods and file methods.

The COM layer implements methods for sending and receiving data from the device. It is organized in the logical groups according to the communication protocol selected to work with the certain CAT device being used in a specific test environment

IV. USER INTERFACE ARCHITECTURE

There are three user interface groups: test settings frontend, results viewer frontend and test plan frontend. The test settings frontend must include the following parameter adjustments:

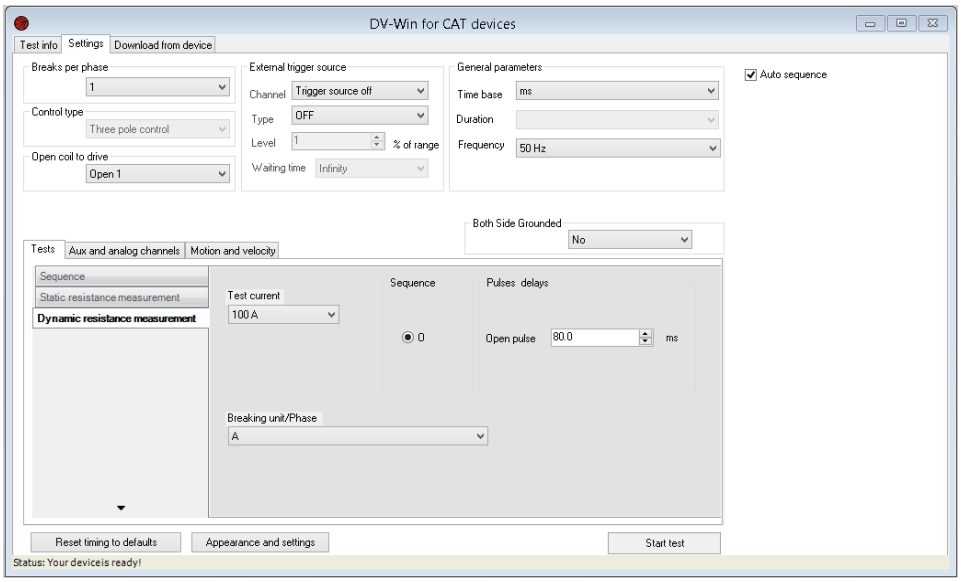


Figure 1. - DRM settings adjustment frontend

APP layer horizontal architecture is organized in the following groups:

- Test settings group
- Results viewer
- Results comparison
- Report generator
- Test plan interface

BLL level contains implementation of the functions and groups of functions such as algorithms, appropriate response to the events triggered at the APP level etc. It is organized in the form of abstraction layers going from the device, device family and specific circuit breaker analyzer within one device family.

- Time base / frequency / resolution
- Open coil to drive
- Test current value
- Sequence selection (C or O)
- Sequence pulse timing (close or open pulse)
- Motion transducer channel
- Motion transducer type (linear or digital)
- Motion transducer unit (mm, inch, degree)
- Motion transducer transfer function type and parameters (linear or nonlinear)

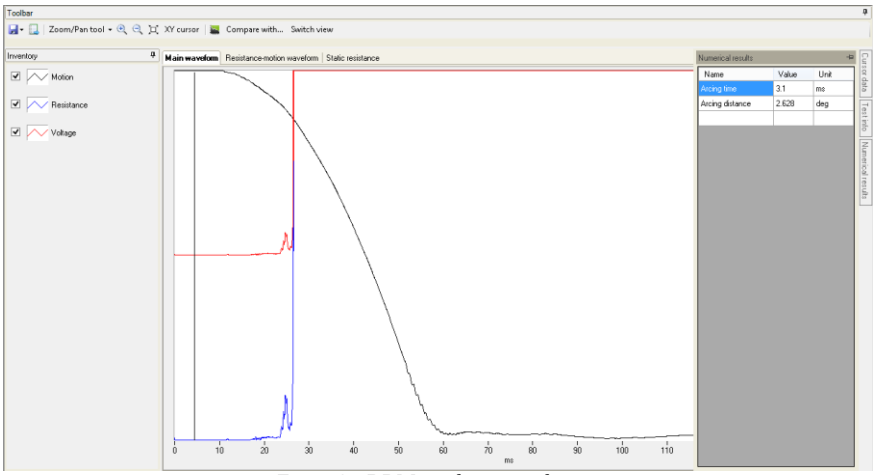


Figure 2 - DRM results viewer layout

Final layout of the frontend is shown in the Figure 1. It is organized in two sections that are related to the DRM testing: "Tests" and "Motion and Velocity". Tab "Tests" contains parameters such as sequence selection, sequence timing etc. Tab "Motion and Velocity" contains parameters related to the motion transducer.

Results viewer contains the following segments:

- Voltage, motion and resistance graphs
- Resistance-motion graph
- Numerical results (static resistance, overlapping time and distance)
- Tools for analysis (vertical cursors, XY cursor, pan and zoom option , comparison with another DRM result)
- Tools for storing results (save to a file or database, save as image or copy graph to a clipboard, create a report)

Layout of the result viewer is shown in the Figure 2.

V. GRAPHICAL PRESENTATION OF THE RESULTS

Main guideline is to provide users with "different flavors" of the graphical results presentation with a high customization ability. For that reason the graphical presentation is provided in two different forms:

- Default view
- Compact view

Default view is the one where voltage, motion and resistance waveforms are presented on one waveform area with invisible Y axis. This view is used to look for intersection in time of certain waveforms since waveforms are auto scaled to the full screen and overlaid.

Compact view is more appropriate if a user would like to read values of certain points in time since in the compact view the waveforms are divided in its individual waveform areas and Y axis is visible.

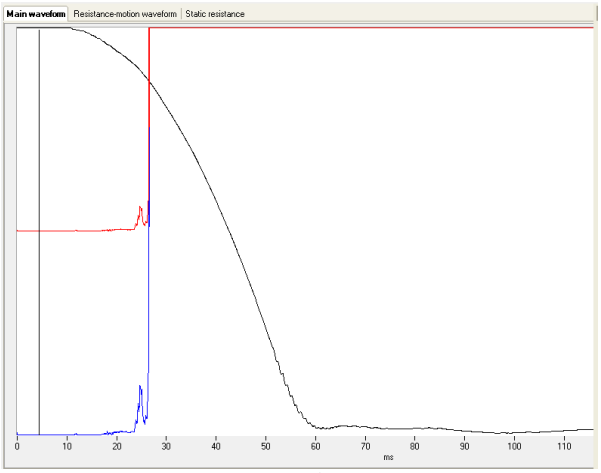


Figure 3 - Default view

It is easy to switch between these two views "on the fly". All tools including all type of cursors are active in both view modes. Customizations such as changing waveform plot color, width, waveform area color, background grid etc. are applicable to both view modes.

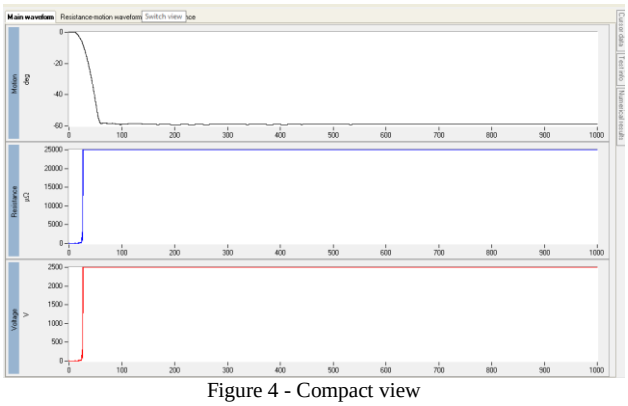


Figure 4 - Compact view

VI. RESULTS COMPARISON INTERFACE

DRM results comparison consists of main graph comparison, resistance-motion graph comparison as well as the static resistance measurement comparison

All waveform plots are overlaid with the possibility of panning and zooming. Main interface is shown in the Figure 5.

Cursor data panel is created according to the presentation in the single result viewer and it is populated with signals from both of the results.

In the inventory panel the user is able to select the results to work on, modify its appearance, show or hide it or pan it along X or Y axis. For improved comfort of the user, there are built in tools such as manual zoom or pan, XY cursor which provide X and Y value for each sample using a mouse pointer, fine X axis zoom in and out as well as an appearance adjustment for each result.

In the appearance adjustment, parameters such as title, line width, line color and style for a single channel in each of the results can be adjusted. However, general parameters such as waveform background color, grid appearance, color and cursors color can be adjusted as well.

Table 1 – HVCB data sheet

Type	SFE 11/ 18 – G
Manufacturer	Energoinvest
Number of breaks per phase	1
Rated voltage	123 kV
Rated short circuit current	31.5 kA
Rated short circuit duration	1 s
Operating mechanism	Hydraulic
Control type	Three pole control
Medium	SF6

As can be seen in the Figure 6, the moment when main contacts separate is the moment when the resistance changes (increases) suddenly for the first time (23. 6 ms). The moment when the arcing contacts separate finally is the moment when the resistance goes to infinity (26.6). The time difference (3 ms) between these two moments is the arcing contact overlapping time.

Based on these two moments and motion curve the arcing contact wipe is calculated (2.53 deg). Since motion is measured with rotary transducer, it is expressed in degrees. When this is converted into contact motion, it

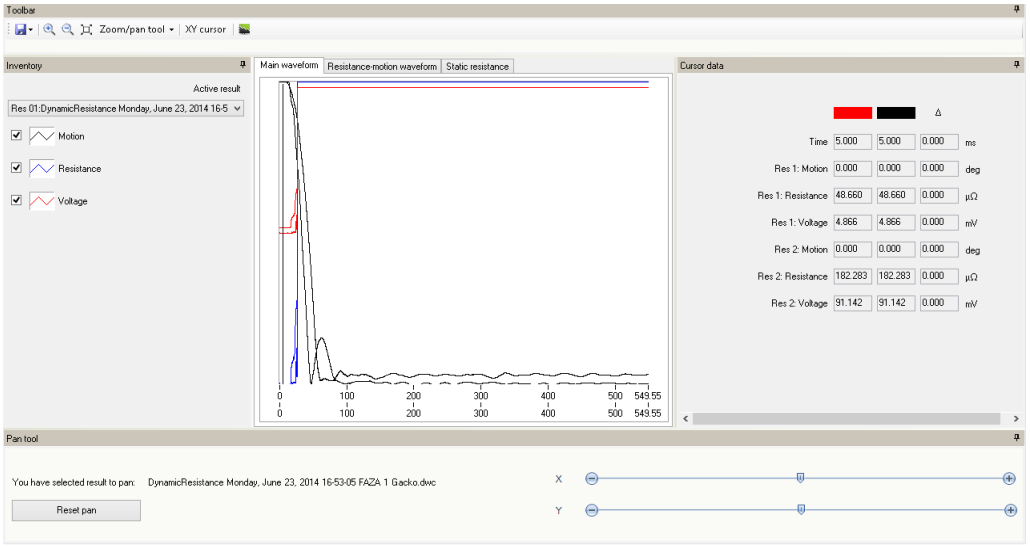


Figure 5 - DRM results comparison

VII. EXAMPLE

For the purpose of the demonstration, a DRM test has been performed on an Energoinvest SFE 11/ 18 – G 123 kV HVCB.

As a measurement tools it has been used circuit breaker analyzer and timer test instrument with the software solution described in this paper. For the analysis purpose, graphical and numerical results have been considered. The dynamic resistance measurement has been performed during trip operation.

Obtained result is shown in the Figure 6.

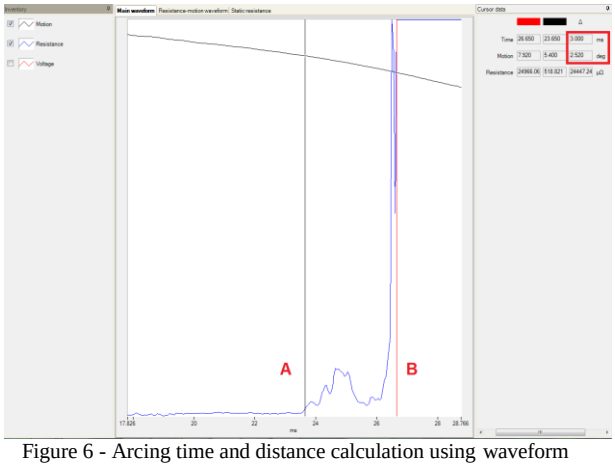


Figure 6 - Arcing time and distance calculation using waveform and cursors

corresponds about 7 mm.

Also, from the waveform of the resistance it can be concluded that there are no abnormalities in the term of outliers or sudden peaks during the arcing contact engagement (A-B in the Figure 6). Furthermore, in the period prior to the A point in time where the contact are still, there is no increase in the voltage and resistance waveform what indicates that the contact force and contact surface in the closed position of the HVCB is good.

These parameters indicate there is no serious damage or wear at the arcing contacts.

VIII. CONCLUSION

During the project development, each iteration has been evaluated according to the customer response and wishes. However, initial project pillars: “make it reliable, accurate, international and enormously ease to use” has been guidance in each iteration and release.

Since high voltage circuit breakers are tested mostly in the field under various weather conditions, testing procedure has to be quick and simple. That is the main reason why software solution user interface simplicity and frontend components arrangement is highly important

IX. REFERENCES

[1] G. Ruben, “High Voltage Circuit Breakers - Design and Application”, Marcel Dekker, Inc, 2002.

[2] M. Kapetanovic, “High Voltage Circuit Breakers”, Published by Faculty of Electrical Engineering Sarajevo, 2011

[3] DV Power, “*Dynamic contact resistance measurement (DRM) - Application Note”*, Lidingo, Sweden: IBEKO Power AB, 2012

[4] A. Secic and B. Milovic, "Dynamic Resistance Measurement Method with high DC Current," in *The Circuit Breaker Test & Maintenance Workshop*, Palm Desert, CA, 2014.

[5] R. Ostojic, J. Levi, "Use of micro ohm meter as a power source for DRM testing on dead tank circuit breakers," in *80th International Conference of Doble Clients*, 2013.

[6] A. Secic, N.Hadzimejlic, M. Brorsson, “Safety Improvements in Testing Procedures for High and Medium Voltage Circuit Breakers with Both Sides Grounded”, in The 11th Annual Euro TechCon 2013 Conference and Expo, 2013, Glasgow, United Kingdom

Data Collector Service – Practical Approach with Embedded Linux

M. Saari*, P. Sillberg*, P. Rantanen*, J. Soini* and H. Fukai**
* Tampere University of Technology/Department of Software Engineering, Pori, Finland
** Keio University/Faculty of Policy Management, Fujisawa-shi, Japan
mika.saari@tut.fi

Abstract - Nowadays embedded systems are one of the most important application areas in information technology. Embedded systems are often used in life critical situations, where reliability and safety are more important criteria than performance. This paper presents a data collector service that has been developed based on embedded Linux, which operates as a key element in a larger intelligent alarm system. The target of this study was to test out how well a cost-efficient single-board computer could be used to gather sensory data, and how this data can be provided for the client over the public Internet. The paper describes the data collector service currently in use and its functionality and also gives a concrete example of how to utilize a microcontroller with an embedded Linux distribution. The paper presents one solution on how to utilize embedded systems for managing and controlling conditions in buildings and also environmental conditions in a smart and cost-effective way.

I. INTRODUCTION

Sensors are commonly utilized components including various kinds of warning and alarm systems. With the huge development of sensor technology, it has been possible to create minuscule, reasonably priced components, sensors and controllers with ultra-low power consumption. These kind of components are enabled for the rapidly improvement of the sensor networks and therefore it can be seen how sensor network implementations have been applied in numerous different fields of operation. A sensor network consists of single sensors, with the purpose of sensing the surroundings and to forward the collected data. There is a great variety of sensors available and their classification is based mainly on the features of the information collected. These kinds of physical features include temperature, humidity, brightness or air pressure.

In this research, one particular subject was to sense the condition changes in indoor spaces and data collection and entry related to their authentication as well as transmitting them forward. The objective of the research was to construct an automatic service – i.e. a data collector service – for collecting condition data that would, in turn, enable the collection and transmission of condition data for a backup system that exploits and analyzes measurement data as flexibly as possible in real-time. The starting point was the development of a system optimally attending to an independent, selected backup system for collecting, recording and transmitting condition data. The aim was to study and develop a solution that was as simple, reliable, cost-effective and easy to maintain as possible for the defined purpose. In this case, it was decided to implement the solution by adapting a microcontroller with embedded Linux for data collection and distribution. Embedded systems usually adopt embedded Linux as the operating system because of the numerous economic and technical benefits – the Linux kernel sources are well structured so that CPU-specific code is easy to find and is minimized. A prototype for realizing the service in question, which will be presented in detail in this paper, was created during the KiiauData research project in 2014.

The study presented here is part of the intensive collaboration between the Tampere University of Technology (TUT) in Finland and Keio University in Japan. The Global Environmental System Leaders (GESL) Program, ongoing in Keio University's Shonan Fujisawa Campus (SFC), and the Alert system for detecting anomalous situations developed as one of its outcomes serve as the background of the research. One main part of the alert system under development is the *data collector service* which was developed in collaboration by the university partners in TUT Pori, enabled by the ongoing KiiauData (Smart analysis of property systems data) project. One of the main aims of this two-year (2013-2014) project, funded by TEKES [1], was to study potential new technologies for managing and controlling conditions in buildings in a smart way. The expected results of the project will enable providers of products and services in the built environment to form wider and more automated solutions both for new breakthroughs and recognized problems in the smart built environment. The data collector service presented here is one concrete example of the studied and piloted solutions produced during the joint project.

Related research in this specific area – i.e. utilization of a microcontroller with embedded systems – has been conducted by Rakesh et al. [2], for example, who have introduced a system which implements an embedded system for monitoring wireless sensor nodes and a camera installed inside a building for security surveillance. Toshniwal and Conrad [3] have studied how to make a cost-effective network-based sensor monitoring system which is portable for various applications. They have developed Linux-based systems based on desktop architecture with a sensor package, and also another system which used an embedded Single Board Computer

(SBC) together with sensors. In addition, Cheng and Shen [4] have introduced a wireless sensor network communication terminal based on embedded Linux. Voinescu et al. [5] describe a device which can work as a network connection to a single board computer (BeagleBone or similar), where the target was to make an easy-to-use wireless networking device. Sawant et al. [6] studied a device that is capable of making file manager operations with two USB flash drives. Their study gives basic knowledge of using of an ARM-based embedded Linux and touch screen. Banerjee et al. [7] proposed and implemented the design of a secure sensor node prototype. They built the prototype using a single board computer (Raspberry Pi in this case), accelerometer, and Bluetooth dongle. The above-mentioned studies deal with the same research area and have a very close connection to the specific research topic presented in this paper.

The following section (Section 2) briefly describes the background system and Section 3 gives a detailed explanation of how its first part – the data collector service – has been carried out. Section 4 includes a discussion and suggestions for future research on the topic and finally section 5 summarizes the study.

II. BACKGROUND – THE INTELLIGENT ALERT SYSTEM

The basis of the study is the alert system under development in Keio SFC. With this planned system it is possible to collect environmental data for different purposes. The system will make environmental sensing easier in various places by using small sensors, and there are many ways to utilize this collected data. One of the intended applications for using the system is for detecting anomalous situations. The aim of such a system is to handle environmental sensor data collected from multiple locations. In practice, it is not easy to understand the problems inherent in a given place just by looking over a graph of sensor data. Therefore, an alert system is need for interpreting the environmental changes in the space in question and associated problems.

This intelligent alert system consists of three (3) main parts shown in Fig. 1: *the first part* is the collection of environmental sensor data via the new *data collector service. The second part* is the detection of anomalous situations by utilizing the sensor data. *The third part* is sending the alert to where the situations were detected.

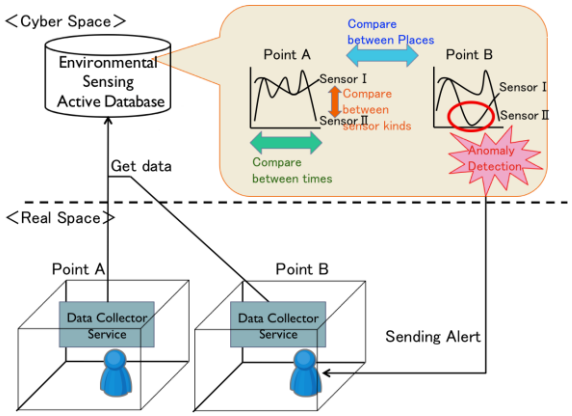


Figure 1. Overview of the Alert system

This particular system targets indoor spaces utilized by the public such as offices, meeting rooms, stations, trains, etc. In this case, the variables being sensed include temperature, light, and humidity. The system automatically collects the sensor data from sensors placed in various places, and analyzes changes over time. As environmental sensor data is collected for a given place over a long time-span, the results become more useful, not only by evaluating the results at one location but also by comparing the results from the sensor data from many locations. Each type of sensor data has four (4) feature values: location information, time, data type, and sensing value. The alert system registers the collected sensing data with an *active database* system in real time. The active database (upper left corner in Fig. 1) automatically reacts in response to detected state change rules that have been pre-defined by the users of the alert system. If an anomalous situation is detected as specified in the active database's rule set, the database system *sends alerts* to the anomalous location. In this case, the role of the active database is to support rule definition, compare data between sensors, compare similar data at different time-points, and compare between different sensor types.

This paper deals with the first part of the alert system, i.e. the *data collector service* (lower left corner in Fig. 1), and does not describe the Intelligent Alert System as a whole. The paper presents the features and architecture, hardware and software components of the developed service, and also the physical connections between the components. The structure of the developed software used for data collection is also described.

III. DATA COLLECTOR SERVICE – IMPLEMENTATION

The architecture of the *Data Collector Service* is shown in Fig. 2. The purpose of the service is to measure light, temperature, and humidity. The sensors used are shown in Fig. 2 (right side). The data can be collected continuously, up to six times a minute. The collected data is then provided as a service for clients, who access the service using the Internet (left side of Fig. 2). In our use case the clients can also be other service providers.

The physical device itself is placed in a public space – TUT's laboratory in this case – to collect data. Attention had to be paid to the physical size of the device as a device smaller than an ordinary PC is easier to install in a public area. In this implementation, the device must also have Ethernet or Wi-Fi capabilities in order to be remotely accessible. These two requirements lead to the use of a device with embedded Linux. The embedded Linux in this context means Linux that can run on ARM-based processors. The embedded Linux devices often have database and web server capabilities, or they can be easily added afterwards. The size of the device's internal mass memory is not critical as long as the device has peripheral ports for external flash memory or Secure Digital (SD) cards.

The integral component of the Data Collector Service, the BeagleBone Black [8], is a low-cost, high-expansion focused SBC using an ARM Cortex-A8 based processor. It can host a Linux operating system and has a 10/100 Ethernet connection and a microSD connector, with 512

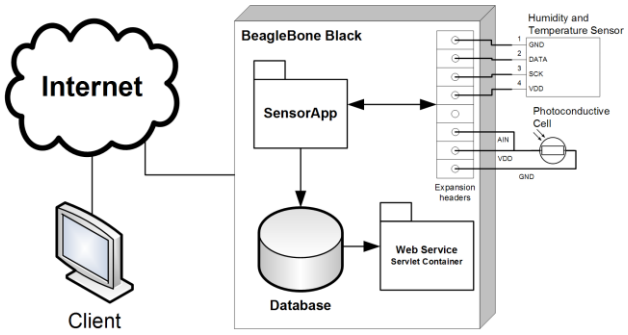


Figure 2. System Architecture of Data Collector Service

MB system memory and 2 GB of embedded MultiMediaCard (eMMC) memory.

The BeagleBone Black has two expansion headers, labeled P8 and P9, which allows the integration of BeagleBone electronics projects [8, 9], and in this research the features of these expansion headers were used. The sub system developed utilizes GND (Ground), 3.3V and 1.8V power, two GPIO (General Purpose Input/Output), and AIN (analog input) pins to drive light, humidity, and temperature sensors.

A. BeagleBone Black with Embedded Linux

The BeagleBone board comes with a pre-installed operating system called Ångström Linux**,** which is a stable and user-friendly distribution for embedded devices and is categorized in the Embedded Linux category. [10] Despite being designed for embedded devices, the Ångström Linux has many of the capabilities that can be found in other full-fledged Linux distributions, such as the X11 windowing system and a substantial amount of software packages in its package repositories.

Some basic Linux server hardening configurations were made to the Ångström Linux for more secure network operation. For example, the system time management was changed to use ntpdate [11], the X11 service was disabled, direct root user access was removed and a new basic user was created which could be used for remote Secure Shell (SSH) access. Python programming language was used together with an Adafruit-BeagleBone-IO-Python library, to utilize the I/O operations of the sensors [12].

The Data Collector Service itself does not utilize the data it collects, as this was to be done by a remote computer with more computing capacity. Thus the Data Collector Service only collects the data and serves it over the Internet for use or as input to the next part of the alert system. We chose to store the data to persistent memory so it could be later accessed by one or more clients. In this case there were no special requirements on how, when and how often the data should be delivered to the remote computer, so the decision was to implement relatively simple server software utilizing the client-server architecture. This way the consumer of the sensor data can decide the most convenient update cycle.

The Data Collector Service provides the data to the clients over a representational state transfer (REST) HTTP/GET [13] interface. Clients may access the interface for historical data from a chosen time interval, or if they choose to, poll periodically for the newest data. By using a short enough polling interval, it would be possible to get near real-time data from the service within the limitations of the computing performance of the BeagleBone platform. The number of new data points per minute would be limited by how often the sensors are read on the Data Collector Service.

The server software was deployed on an Apache Tomcat web server [14] and the sensor readings stored in a MySQL database [15]. In Fig. 2, these two components are called *Web Service* and *Database*, respectively. A third major software component called *SensorApp* also runs on the *BeagleBone Black* platform. The task of the *SensorApp* is to communicate with the *Expansion headers* which are used to drive the physical sensors attached to the BeagleBone Black.

The *SensorApp* and the *Web Service* running on top of the servlet container were written by the project team. The *Database* and *Servlet Container* were written by a third party. Ångström's own package manager provided MySQL, while Apache Tomcat and Java Virtual Machine (Oracle's Java in this case) were installed using their latest available installation packages for Linux operating systems.

Fig. 2 also illustrates the directions of data flow which happen between the components. The *Web Service* can only read data from the *Database*, while the *SensorApp* has read and write access to the device's GPIO headers in order to operate the sensors, and it also directly writes the collected data into the *Database*. For the completeness of the REST interface, the full CRUD (Create, read, update and delete) operations could have been implemented through the interface, but that would have increased the total code complexity of the system. In addition, it would have raised security issues, such as the malicious removal or modification of the collected data. Unauthorized access could be mitigated by the use of access control such as passwords or certificates. In the end, the prime interest was on collecting the data, so keeping outside access as "read only" was the most effective method in terms of computational capacity and code complexity.

Fig. 3 shows the architecture of the physical device with the sensors attached. BeagleBone expansion header

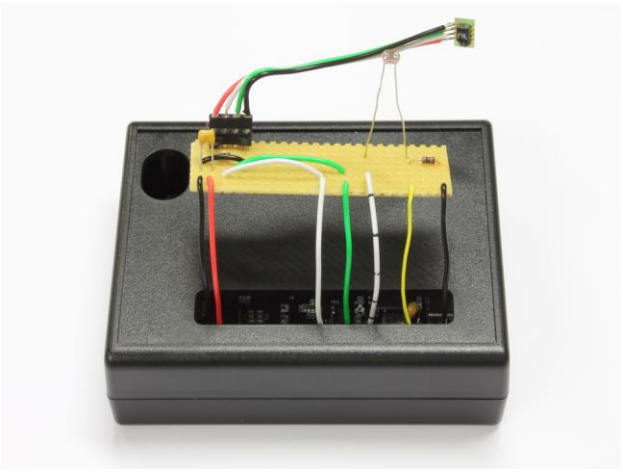


Figure 3. Sensor device used for Data Collector Service

P9 was used to connect the sensors – a photoconductive cell (NSL-19M51 [16]) and humidity and temperature sensor (SHT11 [17]). The photoconductive cell is connected by using a typical application circuit [18]. The humidity and temperature sensor is connected to the BeagleBone by using a datasheet application circuit [17].

B. System Functionality

The web service provides the collected data through one read-only interface. The interface can be accessed by using a simple HTTP GET request. The default query without any parameters returns the ten latest data points. By using different parameters (such as begin_date, end_date, limit and paging) one can request a desired data set from the service. Fig. 4 shows the sequence of accessing the data from the web service. The *Client* connects to the device's *Web Service* interface, which in turn retrieves the requested data from the *Database*. The data is then marshaled into XML format by utilizing JAXB (Java Architecture for XML Binding) annotations and sent back to the client.

As the data is stored to the device's persistent memory, the data can be requested when needed and as often as needed. However, due to the limited computing resources of the platform, a request for a large data set may take a long time or even fail. The code for the web service was not specially written nor optimized for this use case, but merely as a generic proof-of-concept implementation. The marshaling of the XML output at least could have been done in a more memory-efficient way, or a different approach such as JSON serializing could have been used instead. An often-used paradigm is to limit the maximum measurement count to a known safe figure, and use the paging parameter to retrieve the rest of the results. The current implementation appears to cap at around 40000 measurements (or about five days of collected data) on the device, while a desktop machine was able to double that amount (both are with the default Java VM settings). Requesting any more than the aforementioned number of measurements either causes serious degradation of performance, or results in an out-of-memory situation because of the limited memory capabilities of the BeagleBone platform.

The sensor data is collected in a separate process to the web service. Fig. 5 is an illustration of one loop of the sensor reading process. Each loop corresponds to a single measurement point. The humidity and temperature sensor used has its own built-in circuit, and the measurements can take up to 80 or 320 milliseconds at the default accuracy (12 bits for humidity and 14 bits for

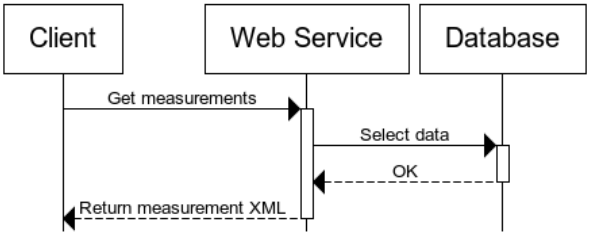


Figure 4. Sequence diagram for accessing the data from the web service

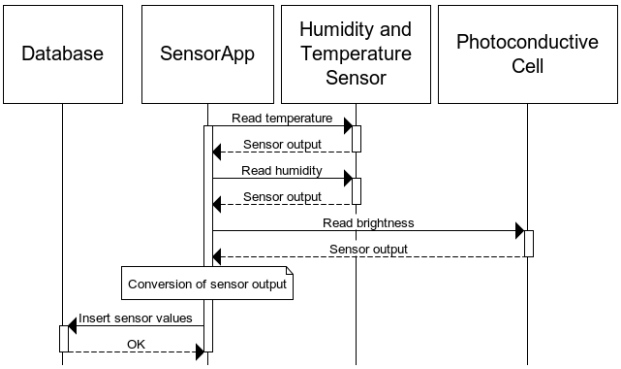


Figure 5. Sensor device used for Data Collector Service

temperature). Also, in order to avoid excess self-heating of the sensor, a maximum of one measurement per second at 12 bit accuracy should be made. [17] These limitations set a theoretical maximum of 15 measurements per minute at default accuracy (when the communication with the sensor is not taken into account). When a lower accuracy (20ms/8bit and 80ms/12bit) is used, a maximum of 60 measurements per minute can be achieved. For this application, it was decided to use the default accuracy and an interval of six measurements per minute.

As can be seen in Fig. 5, first the application reads the raw temperature value for the combined *humidity and temperature sensor*, and then pauses for a while to let the sensor cool down, after which the raw humidity value is read from the same sensor. Then the raw value from the *photoconductive cell* is read. The raw signal values have to be converted before they can be stored to the *database* by using the conversion formulas provided by the manufacturer. After the database has been updated, the application enters sleep mode to attain the desired measurement interval.

IV. DISCUSSION & FUTURE RESEARCH

This paper presents a *data collector service*, which utilizes BeagleBone Black development board with an embedded Linux distribution. The goal was to experiment how well a cost-efficient SBC can be used to gather sensory data, and how this data can be provided to the client over the public Internet. This goal was reached successfully, and the designed system was tested and found to work as planned. Nevertheless, the development process raised several improvement ideas, which could be realized in the future.

One of the issues is the packaging of the sensor system. The current version was a prototype version, and consequently the focus was on making the system functional, both by testing the sensor connections and readings, as well as benchmarking the functionality of the REST API and the software components. This limited the practical usability of the system, for example, making it unfit for use outdoors. In addition, the current software consists of various libraries, programming languages and components, and is as such slightly tricky to install. The software components could be packaged into a single application for easier installation. There is also the possibility to release the source code as an open source release.

The chosen sensor components could also be improved. The serial interface of SHT11 offers good power efficiency, but it cannot be addressed by standard I2C (Inter-Integrated Circuit) protocol, which would make programming tasks easier. In this use case the system is always provided with a continuous power supply and thus better programmability would be a major asset in future studies. In our case, the limited amount of sensor devices to be deployed makes the cost of the single board computer largely irrelevant, though it should be noted that there are other SBCs that are slightly cheaper, but still offer reasonable computing performance. One popular choice is the RaspberryPi (e.g. [19]), which would work – specification-wise – equally well in this use case. Both of these boards offer excellent extension capabilities and can be expanded with additional sensors. The addition of multiple sensors raises another issue, which has not been studied in this research. It is unclear how well the board and the developed system would cope with a very large number of sensors. Also, in this system, all sensors are located very close to the actual board, and thus, the signal degradation can be thought negligible, but this is not necessarily true in large-scale monitoring systems. One example of this kind of system would be the monitoring of an entire apartment complex, where sensors are physically located very far apart and connected to the board by long wires or cables. In this case, it would be possible to deploy a multiple sensor system, but simply using multiple sensors with a single board is a more cost-efficient solution.

In addition to the aforementioned improvement ideas, our future research will focus on topics only briefly discussed in this paper, such as the utilization of the collected data using the designed REST API in various end-user applications.

As can be seen, there are many ways to continue and improve this study. However, the current service produces real-time data regularly and reliably for the benefit of the main system. The prototype developed has proven to be stable and reliable in practice. Work on building the final alarm system is currently ongoing at Keio University.

V. SUMMARY

The paper introduced a prototype system created for sensor data collection and transmission. The presented data collector service is part of a larger alarm system and provides sensor data for the main system. The aim of the study was to test how a single board computer can be used to gather sensory data and how this data can be provided to clients over the public Internet. The paper presented the features and architecture of the developed service, the used hardware and software components, the physical connections between the components, and also the structure of the software. The paper gives a concrete example of how to utilize a microcontroller with an embedded Linux distribution.

REFERENCES

[1] Finnish Funding Agency for Technology and Innovation, Tekes, http://www.tekes.fi/en. Retrieved February 6th, 2015.

[2] V. S. Rakesh, P. R. Sreesh, and S. N. George, "An improved real-time surveillance system for home security system using BeagleBoard SBC, Zigbee and FTP webserver," In India Conference (INDICON), 2012 Annual IEEE, pp. 1240–1244, December 2012.

[3] K. Toshniwal, and J. M. Conrad, "A web-based sensor monitoring system on a Linux-based single board computer platform," Proceedings of the IEEE SoutheastCon 2010 (SoutheastCon), pp. 371–374, March 2010.

[4] X. Cheng, and F. Shen, "Design of the wireless sensor network communication terminal based on embedded Linux," 2011 IEEE 2nd International Conference on Software Engineering and Service Science, pp. 598–601, July 2011.

[5] A. Voinescu, D. Tudose, and D. Dragomir, "A lightweight, versatile gateway platform for wireless sensor networks," In Networking in Education and Research, RoEduNet International Conference 12th Edition, pp. 1–4, September 2013.

[6] T. Sawant, B. Parekh, and N. Shah, "Computer independent USB to USB data transfer bridge," 6th International Conference on Emerging Trends in Engineering and Technology, pp. 40–45, December 2013.

[7] S. Banerjee, D. Sethia, T. Mittal, U. Arora, and A. Chauhan, "Secure sensor node with Raspberry Pi," Impact-2013, pp. 26–30, November 2013.

[8] G. Coley, BeagleBone Black System Reference Manual. 2014.

[9] M. Richardson, Getting Started with BeagleBone. Sebastopol, CA: Maker Media, 2014.

[10] L. Merciadri, and K. Koen, Angstrom Manual. 2010.

[11] D. L. Mills, ntpdate - set the date and time via NTP, http://doc.ntp.org/4.2.6p5/ntpdate.html. Retrieved February 6th, 2015.

[12] J. Cooper, "Setting up IO Python library on BeagleBone Black," https://learn.adafruit.com/downloads/pdf/setting-up-io-python-library-on-beaglebone-black.pdf. Retrieved February 6th, 2015.

[13] R. T. Fielding, Architectural Styles and the Design of Network-based Software Architectures. Doctoral dissertation, University of California, Irvine, 2000.

[14] The Apache Software Foundation, Apache Tomcat, http://tomcat.apache.org. Retrieved February 6th, 2015.

[15] Oracle Corporation, MySQL, http://www.mysql.com. Retrieved February 6th 2015.

[16] L. Curvan, Data Sheet NSL-19M51, TO-18 Open Plastic Encapsulated. 2008.

[17] SENSIRION AG, Datasheet SHT1x. 2011.

[18] PerkinElmer Optoelectronics, Photoconductive Cells and Analog Optoisolators (Vactrols®). 2001.

[19] V. Vujović, and M. Maksimović, "Raspberry Pi as a wireless sensor node: performances and constraints," 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) , pp. 1247–1252, 2014.

Overview of ICT Technologies for the SMILE concept of Adaptive Stimulating Envirvonment

Ana Guseva
Innovation dooel,
Skopje, Macedonia
Email: ana.guseva@innovation.com.mk

Marjan Gusev, Sasko Ristov
Ss. Cyril and Methodius University, FCSE
Skopje, Macedonia
Email: {marjan.gushev, sashko.ristov}@finki.ukim.mk

***Abstract*—Several ICT technologies are used in the SMILE project for StiMulating IntelLEctual activity with adaptive environment.**

In this paper we give an overview of state of the art of these technologies and discuss possibilities of their integration for realization of a SMILE prototype.

***Index Terms*—Stimulating intellectual activity, Adaptive environment, Social Networks, Tablets, Self-health Care, Avatars.**

I. Introduction

Having in mind that by 2025 more than 20% of Europe's population will be over 65, including a rapid increase in the number of citizens over 80 years old, dealing with the process of aging through facing all challenges and obstacles it imposes, has become one of the greatest social and economic issues of European societies. Normal, optimal and pathological aging are all acknowledged as types of aging in literature, and equally demand further development of different services provided to senior citizens.

Online social networking services, along with IPTV, smart TVs and Internet TVs have been become the most popular and fastest growing services in the recent years.

SMILE concept stands for realization of the project StiMulating IntelLEctual activity with adaptive environment. In this project we plan a promising approach to develop an innovative solution delivered as software as a service that will act as Self-Care Social interactive TV. It includes the integration of social and networking services with interactive TV functionality into one consolidated user profiled service. Such an integration presents difficulties in that their utilisation is based on the fact that the older users are trained to use a TV as passive experience. However, in order to benefit from social TV, users will have to actively interact with the content as well as to communicate and collaborate with others, including self-care systems with physicians or subordinate care givers.

Hence, the selection of integrated services and applications is non-obvious, since it needs to gauge between the comfort of passive content consumption and active contribution. SMILE project promotes the concept of how to integrate interactive multimedia technologies, link interactive TV with a set of sensing units on one side and on the other, how to convenience the user to accept Self-care social services.

The main goal of the SMILE project is to build a prototype solution that integrates several technologies into an interoperable solution using open standards. The activities will include analysis, design, interoperability specification, development of software components, their integration in a software as a service solution, development of application interfaces and adapters to existing popular personalised medical devices, personalised hardware user devices, health related services, social media and third party mind stimulating services.

The innovative character of the prototype solution is in the integration of existing technologies with the goal to support independent living. Besides deploying several components into an integrated and scalable cloud-based solution, the project will also include activities for testing its functional and non-functional performances and product evaluation, including validation, usability and user satisfaction.

The final outcome is the realisation of a commercialisation strategy with a business plan and model; sustainability and market research with awareness raising campaigns, dissemination of results and early product marketing. All planned project activities are aimed at intellectual cognitive stimulation, social inclusion and personalised health-care system increasing independence and quality of life.

In this paper we give an overview of current technologies used for the SMILE project.

The paper is organized as follows. Background information is given in Section II. Section III describes the newest social networking services as tools. The new self helath-care management systems are presented in Section IV. The concept of Avatars and several projects are elaborated in Section V, while Section VI describes the tablets. Section VII discusses the SMILE project. The conclusions and directions for future work are presented in Section VIII.

II. Background

Fig. 1 presents the overall architecture of the Smile concept. The system integrates smart TVs or IPTV with personal computers into everyday life of the elderly with goal to establish better environment for independent living. Integration of technologies such as social networks, adapted for elderly, use of tablets as new human interface specially adapted for elderly, and avatars for presentation of a human assistant in self health care is crucial for the concept.

A typical scenario of use of this system is based on limited capabilities of elderly to communicate via small smartphones

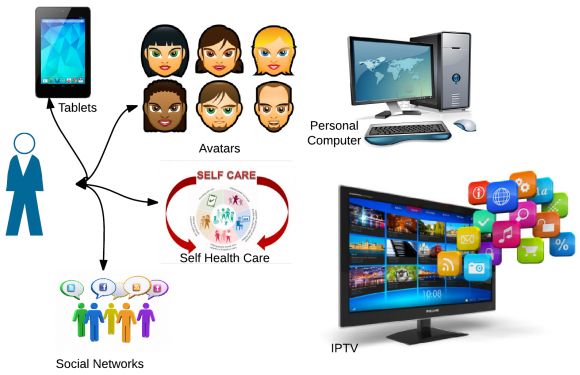


Fig. 1: Architecture of the SMILE concept

and smart TVs, so the idea is to adapt the technology to us of tablets and big TVs as a new adapted human computer interface. The built in camera of the smart TV enables monitoring if the person is in the room. The system offers a welcome message, and list of recommendations supported by the self-care system. In addition any other person (from a list of friends) who is also in the system will be presented in the TV and the person can establish a communication by voice activation or using a tablet with appropriate use of big icons and relevant touchpad.

IPTV can present content which is required to support mental activities of the elderly and keep them in good mental heath. Increased interaction is in the essence of using tablets and Smart TV systems. Besides the self health care recommendations and social network interactions, the system offers interactive content that keeps the elderly mentally active. Most of these stimuli activate corresponding centres and ensure a good mental shape of elderly avoiding content, which they just stare at, without any mental activities at all.

III. Social Networks

Put simply, social networking is a way for one person to meet up with other people on the Internet. A social networking service is a platform to build social networks or social relations among people who, for example, share interests, activities, backgrounds or real-life connections. People use social networking for meeting new friends, finding old friends, or locating people who have the same problems or interests they have, called niche networking [1].

Most social network services are web-based and provide means for users to interact over the Internet, such as e-mail and instant messaging and they incorporate new information and communication tools such as, mobile connectivity, photo/video/sharing and blogging [2]. Online community services are sometimes considered as a social network service, though in a broader sense, social network service usually means an individual-centered service whereas online community services are group-centered. Social networking sites allow users to share ideas, pictures, posts, activities, events, and interests with people in their network [1].

The main types of social networking services are those that contain category places (such as former school year or classmates), means to connect with friends (usually with self-description pages), and a recommendation system linked to trust. Popular methods now combine many of these, with American-based services such as *Facebook*, *Google+*, *YouTube*, *LinkedIn*, *Instagram*, *Pinterest*, *Tumblr*, and *Twitter* widely used worldwide; *Nexopia* in Canada; *Badoo*, *Bebo*, *VKontakte* (Russia), *Delphi* (also called Delphi Forums), *Draugiem.lv* (mostly in Latvia), *Hi5* (Europe), *Hyves* (mostly in The Netherlands), *iWiW* (mostly in Hungary), *Nasza-Klasa*, *Soup* (mostly in Poland), *Glocals* in Switzerland, *Skyrock*, *The Sphere*, *StudiVZ* (mostly in Germany), *Tagged*, *Tuenti* (mostly in Spain), and *XING* in parts of Europe; *Hi5* and *Orkut* in South America and Central America; *Mxit* in Africa; and *Cyworld*, *Mixi*, *Orkut*, *renren*, *weibo*, and *Wretch* in Asia and the Pacific Islands [1].

Social networking tools are increasingly the object of scholarly research. Scholars in many fields have begun to investigate the impact of social-networking sites, investigating how such sites may play into issues of identity, privacy [3], social capital, youth culture, and education [4].

While the popularity of social networking consistently rises, new uses for the technology are frequently being observed. At the forefront of emerging trends in social networking sites is the concept of ”real-time web” and ”location-based.” Real-time allows users to contribute contents, which is then broadcast as it is being uploaded - the concept is analogous to live radio and television broadcasts. *Twitter* set the trend for ”real-time” services, wherein users can broadcast to the world what they are doing, or what is on their minds within a 140-character limit. *Facebook* followed suit with their ”Live Feed” where users’ activities are streamed as soon as it happens. While Twitter focuses on words, *Clixtr*, another real-time service, focuses on group photo sharing wherein users can update their photo streams with photos while at an event. *Foursquare* gained popularity as it allowed for users to ”check-in” to places that they are frequenting at that moment. *Gowalla* is another such service that functions in much the same way that Foursquare does, leveraging the GPS in phones to create a location-based user experience [1].

IV. Self Health-Care Measurement Systems

The research in the information and communication technologies area has helped the development of automated concepts to be used in personalized healthcare. This implies that we are entering the e-Health era. Newly developed self healthcare systems are characterised by new ways of healthcare delivery through a broad range of teleservices. However, providing ubiquitous access to high-level healthcare for everyone, anytime, anywhere requires the real integration of the various technological platforms [5]. One of today’s most promising essential new technologies in ICT is most certainly IPTV (Internet Protocol Television) which changes the perspective

of the way end users consume TV services, but also serves as a platform for new ideas and concepts in other areas, such as medicine, science and industry. It is considered as a technology that brings significant contribution in different scientific areas, as well as technology assisted living, mainly personalized self health-care systems based on e-Health diagnosis and monitoring. Such personalized system is defined by Mlakar et al. [6] which present UMB-SmartTV, a novel platform that combines traditional IPTV services with multimodal-based non-IPTV services (automatic speech synthesis, speech recognition ambience control). It is important to note that any service within IPTV in this platform can be operated by using speech. The core of this platform is IMS, which is combined with a distributed DATA architecture in order to merge IPTV services and other non-IPTV services. This systems consists of four main blocks, described as follows. The first block, named Content core, represents an application server for content production and presentation. The second block is named IMS core. In the context of IPTV system, IMS represents multimedia service platform architecture. The third block is named Multimodal services. It actually represents the novel contribution of this platform- a distributed system intended for providing non-web speech-based multimodal services (automatic speech recognition, automatic text-to-speech synthesis, virtual assistants, home automation, personalization, etc.). Finally, the fourth block is named Environment controller, which controls several devices in the household (either by using TV remote, keyboard, mobile device, or speech)[6]. Another example of contributing to the designing of personalized health care systems is given by Pustivsek et al. [7]. Their research focuses on the development of a personal health telemonitoring application integrated into the IPTV home communication environment. The application enables monitoring of blood pressure and body weight, with rather common sensors and weight scales that have built-in communication interfaces, thus taking medicine to a whole new level. The application also supports on-line medical interviews. The user interface is manageable by a simple TV remote-control. Furthermore, enhancing advancement in the field of medicine is the goal of the FP7 EU-funded Vital Mind project [8], [9]. Here, instead of improving the diagnostics (remote measurement of blood pressure and weight), the goal is increasing the lifespan of the elderly by enabling them to actively participate in mind fitness exercises while watching TV by using a combination of cognitive psychology, the TV medium and advanced interactive ICT. This is done by new methods for users control, production of Authoring and Production tools for IDTV STB and the development of usage of USB Flash-Device-UFD as an alternative to a broadcast delivery system [9].

Komninos et al. [10] have introduced HealthPal, which is an intelligent, powerful and adaptive health monitoring dialogue-based system, which aims at assisting the elderly people control and manage their health condition on a regular basis, without relying exclusively on their doctors or caretakers. The implementation of the HealthPal system is based on a combination of advanced of-the-shell medical engineering, PDA, WLAN, mobile computing and natural language text and speech processing technologies integrated into a common, open and intelligent medical suite, in order to guarantee that the user's communication with the system is efficient, convenient, and interactive and that it does not pre- requisite any technology skills on behalf of the user.

V. Avatars

Avatars have been defined in [11] as online manifestations of self in a virtual world, and are designed to enhance interaction in a virtual space. This definition is extended by adding that avatars allow the user to take on a visible persona within a virtual world, providing them the opportunity to engage in surreal and imaginary experiences that transcend the actual world in which they live. Hemp et al [12] define avatars as the most conspicuous online manifestation of people's desire to try out alternative identities or project some private aspect of themselves. The word 'Avatar' originally described the incarnation of the Hindu god Vishnu and was popularized in its cybersense by Neal Stephenson in his 1992 cult novel Snow Crash. Not only does this term describe complex beings belonging to a shared virtual reality, but also any visual representation of a user in an online community. For example, more than seven million people have created Yahoo avatars, simple but personalized cartoon-like characters used as pictorial signatures in activities ranging from instant messaging to fantasy sports. [12]. Gusev et al. [13] explores the design technology of avatars in order to increase the lifespan of the elderly by providing the active aging of the newly developed personalized self health-care system.

Besides for entertainment purposes, avatar concepts can also make a huge difference in the area of social networking, education, as well as personalized self-care.

Researchers in their discussion of online object-oriented environments, claim that virtual environments have the potential to enhance social interaction. They comment that the anonymity of interacting in the virtual world via a programmable avatar often enables individuals to communicate and express themselves in ways they may have been incapable of doing otherwise. Katz and Rice [14] found evidence that cyber-relationships facilitated by the use of virtual environments generally proved to be robust and long lasting. Similar claims have been promoted by Antonacci et al. [15] in their discussion paper on the use of avatar virtual worlds in education. They state that virtual worlds hold considerable potential as powerful mediums for learning. Consistent with previously conducted research, they define the existence of benefits in three key areas:

1) Virtual worlds give users the ability to carry out tasks that could be difficult for them in the 'real world' due to constraints, including cost, scheduling or location;
2) Virtual worlds' persistence provides continuing and growing social interactions, which can serve as a basis for collaborative education;
3) Virtual worlds adapt and grow to meet user needs [15].

Also, engagement in virtual environments enables students to experience learning opportunities that would not normally have been easily accessible, including role playing, operating simulated equipment, designing and building things, or creating simulations of physical or procedural processes.

Furthermore, Oestreicher et al. [16] introduce iBot2000, which is a "virtual interactive robot that can display different characters, so that it can look like a robot or a human being; designed differently to attract various audiences, whether for corporate use, children or academia". Some of it's advantages include attractive presentation of learning content, inviting to repeatedly studying content for intensified learning of details, learning at any place with web access, customisation of teaching to specific groups of ages, overcoming certain language barriers, etc. Nevertheless, this robot has its disadvantages, such as its complexity that may present a number of adoption problems to interested universities. Also, different learning styles in respecting to different students' individualities suggests an additional problem for instance, such virtual environment may be for some students too distracting and could cause a negative effect on their pursuit of a good degree. This is expected to be overcome in the the future [16].

Morie et al. [17] state that one of the main benefits of virtual worlds is their persistent nature. Because of the activeness of the virtual space, scientists are enabled to place virtual sensors where needed to record various forms of data. This data can be used to enhance VW forms of telehealth therapy. These system can be further customized to benefit patient-therapist needs. According to Morie et al. [18] Virtual Worlds (VWs) are not the same as Virtual Reality, because of the fact that they provide much easier accessibility, as all a user needs is a computer and an Internet connection. VWs are connected social spaces, so that many hundreds or thousands of users can be concurrently online. VWs also allow for highly focused forms of environments that can be specially built and targeted to assist in any healthcare process that is delivered by them. Finally, almost all VWs provide the means for a participant to create and customize their avatar with which they will navigate the world and interact with others. These characteristics of VWs allow them to make inroads to the healthcare field as an emerging type of "telehealth care." Traditional telehealth care provides several proven advantages that complement traditional, face-to-face health care, including accessibility from rural locations, remote patient monitoring, reduction in need for travel, fostering a positive client-patient relationship, providing enhanced continuity of care, reduction of hospitalization, as well as overall savings in cost of care, even with increased contact [18].

One example of personalized healthcare avatars is defined by Morie et al. [17]. They introduce "a novel method for engaging an amputee in rehabilitation activities delivered through an advanced telehealth mechanism using virtual worlds". This system places the avatar, being the graphical, embodied representation of the user, as the basic, most essential element through which each user accomplishes all socialization in the newly developed Virtual World (VW). Any avatar can be customized or personalized to a range of appearances, from cartoony to fairly realistic. Avatars can even be made to look like animals, fantastic creatures, or robots. By each particular user having their own avatar, all VW participants interact avatar to avatar. The VW allows hundreds of users to be logged on at the same time. This system functions in the following way: the data containing the newly developed virtual world is stored on server-side computers. A user downloads a software client that runs locally on his computer, synchronizing to the data server in a continuous real time process [17].

Finally, elderly people can benefit from the integration of avatar concepts in telemedicine. An example of this is presented in [19] where it is stated that "older adults benefit from information and communication technology solutions and services in the Ambient Assisted Living (AAL) domain". The aim of the AALuis project is to offer a practical solution for adapting user interfaces and services to changing needs and wishes of older adults in a flexible way by providing various devices and I/O modalities. This research describes the approach for the integration of an avatar into the generated user interfaces. The user interfaces aim to provide consistent look and feel for different services to interact using the user's preferred modality [19].

VI. Tablets

Conceptualized in the mid-20th century and prototyped and developed in the last two decades of that century, tablets became popular in 2010. A tablet computer, or simply tablet, is a mobile computer with display, circuitry and battery in a single unit. Tablets are equipped with sensors, including cameras, microphone, accelerometer and touchscreen, with finger or stylus gestures replacing computer mouse and keyboard. Tablets may include physical buttons, e.g., to control basic features such as speaker volume and power and ports for network communications and to charge the battery. An on-screen, pop-up virtual keyboard is usually used for typing [20].

A. Features

Tablet features include [20]:

- Hardware
  - High-definition, anti-glare display
  - Wireless internet connectivity (usually with Wi-Fi standard and optional mobile broadband)
  - GPS satellite location
  - Front- and/or back- facing camera(s) for photographs and video
  - Lower weight and longer battery life than a comparably-sized laptop
  - Bluetooth for connecting peripherals and communicating with local devices
  - Docking station: Keyboard and USB port(s)
- Software
  - Mobile web browser
  - Reader for digital books, periodicals and other content

  - Downloadable apps such as games, education and utilities
  - Portable media player function including video playback
  - E-mail and social media
  - Mobile phone functions (messaging, speakerphone, address book)
  - Video-teleconferencing
- Data Storage
  - On-board flash memory
  - Ports for removable storage
  - Cloud storage options for backup and syncing data across devices

Besides a touchscreen and keyboard, some tablets can also use these input methods:

- Accelerometer: Detects the physical movement and orientation of the tablet. This allows the touchscreen display to shift to either portrait or landscape mode. In addition, tilting the tablet may be used as an input (for instance to steer in a driving game).
- Ambient light and proximity sensors, which help distinguish between intentional and unintentional touches.
- Speech recognition
- Gesture recognition

Tablets, like conventional PCs, run multiple operating systems (though dual-booting on tablets is relatively rare). These operating systems come in two classes, desktop-based and mobile-based (”phone-like”) operating system. Desktop-based tablets usually are thicker and heavier, require more storage, more cooling and give less battery life, but can run desktop applications such as Photoshop in addition to mobile apps and have more ports, while mobile-based tablets are the reverse, only run mobile apps.

B. Use

Sony Mobile revealed its Xperia Z2, proudly describing it as ”the world’s slimmest and lightest waterproof tablet”. ”It is perfect if you want to take it into the bathtub, to the beach, to the pool”, Sony Mobile official Almos Szabo said at the Japanese manufacturer’s stand in the world’s largest mobile industry fair.

Yet, according to Informa, only 20% of tablets are connected to the mobile network, the rest relying on wifi. During the last iPad launch, Apple showed a promotional video with thousands of different ways that people can use their iPad including mountain bikers, surfers, and doctors. I think that is how they would like to show the iPad is being used, but it is mainly in the home, streaming media, streaming movies said Julian Jest, analyst at the technology research house Informa. The classic profile of a tablet user, is actually a person lounging on the sofa watching videos or playing electronic games. In 2012, a poll by Google showed that only 21% of users took their tablets out of the home.

VII. Discussion

To the best of our knowledge so far, there are no integrated commercial solutions, which can bring social network on IPTV, along with health self-care. The SMILE concept could make the European ICT industry more competitive in the forthcoming years. Modularity and interoperability of the SMILE components, as well as, wide exploitation of each of the deployed technologies, guarantee the scalability and a long-term influence of the integrated SMILE solution for the European market. The technologies to be integrated into the new system are widely available across Europe, and will be accessible in a gender-neutral and accessible way, without ethical implications.

The SMILE project aims at design and development of the following measurable outputs:

- Demonstration multimedia box with specially developed remote control that integrates the pro-posed hardware and software solutions into an intuitive, user friendly and highly sophisticated interactive system;
- An interoperable platform with wireless communication to the medical sensors that will measure relevant health parameters, process the results and store in a cloud-based environment, and interact with physicians, which provide medical advises;
- An avatar environment for increased interactivity between users and the prototype system;
- A notification system with reminder capabilities, interchanging information with social media, calendars, contacts;
- A solution as a service that integrates newly designed interactive user interfaces and self-care system with social media, including contacts, comments, blogs, photos, communication,
- Hardware and software specification for cross-platform interoperability, including mobile phones, tablets, external medical devices, and third party services;
- Application Programming Interface (API) for remote control integration both in hardware and software;
- Demonstration and evaluation of the prototype;
- Dissemination of results, raising awareness, and commercialisation strategy including sustainability.

These outputs promote independence, prevent or delay the deterioration of wellbeing resulting from ageing, illness or disability and delay the need for more costly and intensive services.

Our approach is therefore to prevent older people, which are particularly vulnerable to social isolation or are alone due to loss of friends and family, mobility or income. Social isolation and loneliness impact upon individuals’ quality of life and wellbeing, which effects health and increase dependence on health and social care services. The interventions to tackle social isolation or loneliness include befriending, mentoring, Community Navigators, social group schemes. People using befriending or Community Navigator services are less lonely and less socially isolated.

Some studies recorded that older people who were part of a social group intervention had a greater chance of survival than those who had not such service. Therefore the aim of the project is to develop interoperable ICT solutions with services to reduce social isolation or loneliness, strong partnership arrangements also to ensure the self-care services. In order to achieve a given objective the research needs should be carried out on actions that include different genders, populations and localities. There also exists an urgent need for controlled trials that incorporate all standardized services for life quality with minimised costs.

The overall project readiness level is positioned as TRL 7 (system prototype demonstration in operational environment), since it realises all intermediate stages of observing basic principles, specifying technology concepts and designs, experimental proof of concept, validated technology in lab environment, followed by validation and demonstration in relevant and operational environments. It follows the whole life cycle of product development: from idea, concept analysis and research, design, development, verification and validation, performance and security evaluation, application demonstration, market analysis, business modelling and commercialisation strategy, sustainability analysis, product marketing and dissemination of results, determining a path from lab to market, analysing possibilities for third parties involvement in an open standard product realisation.

VIII. Conclusion

So far there are no published articles or projects that are targeted at the same goal to develop a self-care social interactive TV system as a Service, which is the aim of the SMILE project. There are a lot of projects and papers about the relevant technologies, however not the integrated approach, which is the key innovation in this project.

We conclude that using the newest information communication technologies could be customised to be used by the older people in order to improve their life, without being isolated by the environment, loneliness, as well as to ensure the self-care services. This paper analyzes and presents several technologies in this direction.

References

[1] Wikipedia. (2007) Social networking service. [Online]. Available: http://en.wikipedia.org/wiki/Social_networking_service

[2] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, pp. 210–230, 2007.

[3] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks (the facebook case)," in *ACM Workshop on Privacy in the Electronic Society (WPES)*, 2005.

[4] D. Boyd, "Why youth (heart) social network sites," *MacArthur Foundation Series on Digital Learning - Youth, Identity, and Digital Media Volume (ed. David Buckingham)*, 2007.

[5] G. Graschew, T. Roelofs, S. Rakowsky, and P. Schlag. Virtual hospital and telemedicine for telementoring of the health workforce. [Online]. Available: http://webistem.com/psi2009/output_directory/cd1/Data/articles/000025.pdf

[6] I. Mlakar, Z. Kačič, D. Zimšek, and M. Rojc, "A novel ims based iptv platform for integrating multimodal technologies," in *Proceedings of the 2013 International Conference on Systems, Control and Informatics*, 2013, pp. 364–370.

[7] M. Pustišek, E. Stojmenova, L. Zebec, and D. Kervina, "Bringing health telemonitoring into iptv based ami environment," *Previous Publications of the AMEA Association*, 2013.

[8] Vital Mind. (2011). [Online]. Available: http://www.vitalmind-project.eu/

[9] B. Veselinovska, M. Gusev, and T. Janevski, "Overview of current trends in IPTV related FP7 projects," in *Information, Communication and Energy Systems and Technologies (ICEST) 2013 48th International Scientific Conference on*, 2013, pp. 111–114.

[10] A. Komninos and S. Stamou, "Healthpal: an intelligent personal medical assistant for supporting the self-monitoring of healthcare in the ageing society," *Proceedings of the UbiHealth*, 2006.

[11] G. Falloon, "Using avatars and virtual environments in learning: What do they have to offer?" *British Journal of Educational Technology*, vol. 41, no. 1, pp. 108–122, 2010.

[12] P. Hemp, "Avatar-based marketing," *Harvard business review*, vol. 84, no. 6, pp. 48–57, 2006.

[13] M. Gusev, S. Ristov, and A. Guseva, "Avatar concepts in self healthcare systems," in *Prociidings of ICEST 2014 - XLIX INTERNATIONAL SCIENTIFIC CONFERENCE ON INFORMATION, COMMUNICATION AND ENERGY SYSTEMS AND TECHNOLOGIES*, Jun. 2014, pp. 45–48.

[14] J. E. Katz and R. E. Rice, *Social consequences of Internet use: Access, involvement, and interaction*. MIT press, 2002.

[15] D. Antonacci, S. DiBartolo, N. Edwards, K. Fritch, B. McMullen, and R. Murch-Shafer, "The power of virtual worlds in education," *ANGEL Learning White Paper*, 2008.

[16] K. Oestreicher, J. Kuzma, and D. Yen, "Avatar supported learning in a virtual university," *Worcester Journal of Learning and Teaching*, no. 4, pp. 1–8, 2010.

[17] J. F. Morie, C. E. Lathan, A. Skinner, E. Chance, D. Rajpurohit, and K. Haynes, "Using virtual world activities for amputee rehabilitation," *Journal of Research and Practice in Information Technology*, 2012.

[18] J. Morie, E. Haynes, E. Chance, and D. Purohit, "Virtual worlds and avatars as the new frontier of telehealth care," *Studies in health technology and informatics*, vol. 181, p. 27, 2012.

[19] C. Mayer, M. Sili, M. Gira, M. Morandell, S. Fagel, A. Hilbert, C. Schüler, and I. Cernei, "Avatar enriched user interfaces for older adults," in *GLOBAL HEALTH 2013, The Second International Conference on Global Health Challenges*, 2013, pp. 1–4.

[20] Wikipedia. (2011) Tablet computer. [Online]. Available: http://en.wikipedia.org/wiki/Tablet_computer

Android-Powered Web Server for Volunteer-Based Environment Sensing

D. Skvorc, M. Orsic and S. Gros
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
dejan.skvorc@fer.hr, marin.orsic@fer.hr, stjepan.gros@fer.hr

Abstract - Many people today are using smartphones and similar forms of mobile computers in every aspect of their lives as a standard and undetachable personal outfit. Besides the significant amount of computing power, storage capacity, and communication capabilities, these devices are equipped with a set of embedded sensors. This makes them not only sufficient for being a reliable on-site assistants, but also powerful enough to deliver sensory information collected from the environment to the remote users through the Internet. Having an open community of smartphone users involved as data sources into the environment sensing program reduces the need for a dedicated sensory infrastructure preinstalled throughout the environment, yet providing a satisfactory level of territory coverage. The necessary prerequisite is, however, that users can contribute effortlessly, with minimum or no overhead compared to a normal smartphone usage.

In this paper, we present a prototype implementation of an Android-powered web server for autonomous capture and delivery of sensory information from a mobile device to the remote users. A modular software framework is described for seamless extension of the server with support for new types of sensors as new models of devices appear on the market, while still providing basic level of interoperability with these sensors until the server gets upgraded.

I. INTRODUCTION

Since around 2005, when user-generated content entered the mainstream usage on the Internet [1], many popular web services owe their success to the people's addiction to share their work with others. *Wikipedia*, *Youtube*, *TripAdvisor*, different social networking sites, discussion groups, and blogs are all examples of web services where the content delivered to the users is not created by the service providers themselves, but by the open user community [2]. Even before, the *BOINC* infrastructure [3] that was originally developed to manage the *SETI@home* project [4], has based its operation on the Internet users willingness to share the idle time of their desktop computers for mass-scale data processing used to cure diseases, study global warming, discover pulsars, and do many other types of scientific research. All of these systems and services are examples of volunteer-based distributed computing where individual users contributes to the community either with their computing resources, intellectual work, or creativity in content creation.

While the benefit derived from user-generated content and volunteer-based computing for the service providers and users is clear, the benefit to the contributor is less obvious. The motivation for contributing with user-generated content or other types of computational and intellectual resources ranges from altruistic, to social, to materialistic [1]. Due to the strong dependence on and high value of volunteer-based activity, many sites use various models to encourage users to participate in their ecosystems as contributors [5, 6]. Models based on implicit incentives do not reward users with material goods, but allow them to build social status within the community. These models target people who feel good as active members of the community. Examples of social popularity metrics are number of likes someone gains on his/her *Facebook* post, number of *Twitter* followers, or number of pageviews of uploaded *Youtube* video. On the other hand, models based on explicit incentives refer to tangible rewards. These can either be expressed in direct financial payment, but also through vouchers, coupons, frequent traveler miles, etc. Example is the crowdsourcing-powered Internet marketplace *Amazon Mechanical Turk*, where individuals and businesses publish human intelligence tasks that computers are currently unable to do and pay monetary payment to users for their completion [7].

In our research, we apply the concepts of volunteer-based computing to build an infrastructure for environment sensing. Traditionally, environment sensing systems required dedicated sensory equipment preinstalled throughout the environment. Except being costly for initial installation and maintenance during the operation, such an approach also suffers from territory coverage, which is limited to locations where service operator has installed the equipment. There is, however, a lot of underutilized sensory equipment available in existing and emerging wearable devices, smartphones being one of the most representative ones. Many people today are using smartphones and similar forms of mobile computers in every aspect of their life as a standard and undetachable personal outfit. Besides the significant amount of computing power, storage capacity, and communication capabilities, these devices are equipped with a set of embedded sensors, such as GPS receivers, cameras, microphones, accelerometers, compasses, thermometers, barometers, etc. This makes them not only sufficient for being a reliable on-site assistants to their owners, but also powerful enough to deliver sensory information collected from the environment to the remote users through the Internet.

The idea behind the volunteer-based environment sensing program is to utilize sensory, computational, and

communication resources of ordinary smartphones to collect sensory information from the device's physical surrounding and to deliver them to the interested end users or applications through the Internet. Although the expected reliability and accuracy of such information is inferior to those provided by professional dedicated equipment, their availability, low cost of system installation and maintenance, and extensive territory coverage due to the widespread use of smartphones makes the concept interesting for non-mission-critical applications. However, since the concept is highly dependent on the voluntary contribution of open community of smartphone users, the necessary prerequisite is that users can contribute with as little additional effort as possible, compared to a normal smartphone usage.

In this paper, we present a prototype implementation of an Android-powered web server for capture and delivery of sensory information from a mobile device to the remote users. Once started, the web server autonomously reads the information from device sensors and delivers them to the connected clients. Since it does not require any user intervention during the operation, even passive or lazy users are able to contribute as much as most eager ones.

Since different types of client applications require different rates of delivery of sensory information, the web server supports two modes of operation. Using the standard HTTP protocol [8, 9], the information from device sensors is available to a client application on request. The application receives a snapshot of the sensor readings at the time of placing a request. This enables the application to have the information delivery rate under control. On the other hand, the web server also supports the communication using the emerging *Websocket* protocol [10, 11, 12]. This enables the application to establish a permanent communication channel with a remote device and to receive updates as soon as sensor readings change on the device.

Smartphone-powered web servers are terminal components of a more comprehensive distributed computing infrastructure for volunteer-based environment sensing. This infrastructure also includes core components for volunteer location tracking, search, discovery, etc. However, these components are out of scope of this paper.

The rest of the paper is organized as follows. In Section 2, we briefly describe existing projects whose goal is to transform an ordinary smartphone into a web server. Section 3 describes the device-agnostic sensor server architecture that enables the design of a single version of web server transparent to the variations in devices' hardware equipment. In Section 4, we present a web-server self-configuration procedure used to autonomously generate the URI scheme and the REST interface based on the available set of sensors on a given device. Section 5 describes the development model used to upgrade the web server with support for new types of sensors. Section 6 concludes the paper.

II. Related Work

Having a web server running on a smartphone is not a new idea. There are several general-purpose web server implementations available for Android and other operating systems for mobile devices.

NanoHttpd [13] is a lightweight HTTP server designed for embedding in Java applications. It fully implements the features of HTTP 1.1 protocol version and supports the *Websocket* protocol with low memory overhead. Since *NanoHttpd* is written in Java and is available as open source software, the integration with Android applications is seamless. The project shares similarities with our implementation regarding the support for both HTTP and *Websocket* protocols. However, *NanoHttpd* is a skeleton web server implementation that enables application developers to add content and functionalities to the server as necessary for specific application. The web server we present in this paper is integrated with Android sensor framework with strong focus on a modular architecture that enables easy extension of the server with support for new types of sensors.

Jetty [14] is a Java-based web server and servlet container intended to run on standalone server machines. It supports HTTP, *Websocket*, and SPDY protocols. *I-Jetty* [15] is a port of *Jetty* to the Android platform. It is implemented as an Android service [16, 17]. Once started and configured, it continues to run in the background, listening on a specified port. Application-specific functionalities are exposed to remote clients through Java servlet technology [18]. *I-Jetty* is similar to our implementation in terms of function and smartphone user experience. They both run as an Android service in the background, thus enabling the smartphone owners to perform other tasks on their devices without interruption.

DroidPHP [19] provides a lightweight Android web server which runs PHP and MySql based applications. It enables developers to create fully functional PHP applications based on dynamic web pages and run them on an Android-powered device. With PHP being the most frequently used server side language, *DroidPHP* provides a powerful Android server platform for creating web applications. The project demonstrates the capability of mobile Android-powered devices to execute tasks normally expected from a larger and more powerful standalone server computers. This underpins our hypothesis that contemporary smartphones and similar wearable devices are powerful enough and useful to act as mobile web servers.

III. Device-Agnostic Sensor Server Architecture

Android-powered smartphones and tablets are equipped with various built-in sensors for retrieving data about device motion, orientation, and environmental conditions. While the diversity of underlying hardware produced by different vendors is satisfactory hidden from the application layer by the Android operating system and corresponding API, the number and types of sensors still vary significantly from device to device. Moreover, new types of sensors emerge as new models of devices appear on the market. Since this type of diversity is inherently

present and cannot be hidden from the application layer, we designed a device-agnostic sensor server architecture independent of the number and types of hardware sensors available on a specific device. This architecture enables web server to operate with all available sensors from any existing and future device, without the need to customize the internal server organization to any specific device. The main objective of the proposed architecture is to provide:

- minimal set of functionalities for any type of sensor, even for those that do not exist yet, but will emerge with new types of devices
- full set of functionalities for the selected types of sensors

A. *Integration of Android Sensor Framework and the Web Server*

In Android operating system, data collected from sensors are made available to the applications through Android sensor API [20]. Android sensor API is designed to enable the inspection of available sensors on a particular device and access to all of the device sensors in a uniform way. In addition, it allows for customization of data handling logic for particular sensors to specific needs of an application. Main classes and interfaces exposed to applications through Android sensor API are: *SensorManager*, *Sensor*, and *SensorEventListener*.

SensorManager is used to retrieve the list of sensors available on a specific device. It is also used to register listeners to sensors, which are then triggered once a sensor's state changes.

Sensor represents a default class that is used to access a single sensor, regardless of its type. It contains the information about the sensor type, name, vendor, power usage etc. *Sensor* class enables uniform access to any type of sensor and is thus independent of any hardware- or device-specific configuration.

SensorEventListener is an interface used for receiving notifications from the *SensorManager* when the state of a particular sensor has changed. To get notifications from sensor in an application, application developer implements custom event listener that implements the *SensorEventListener* interface. Custom listeners receive raw sensor data from the hardware sensors represented as a list of floating point values. Customization of these data into an appropriate application-specific format is up to the application developer.

Fig. 1 shows a common pattern of integration of Android sensor framework into an application and emphasizes how we use this pattern to design a device-agnostic sensor server architecture. Using the *SensorManager* class, application developer registers a sensor-specific listener with a selected type of sensor. Sensor-specific listener acts as a bridge between hardware sensor and the application. On one hand, sensor-specific listener implements the *SensorEventListener* interface to be able to receive notifications when sensor state changes. On the other hand, it allows application developer to implement the internal logic for customized handling of the data received from the sensor. Sensor listeners are registered with sensors during the application run-time, usually during the application startup.

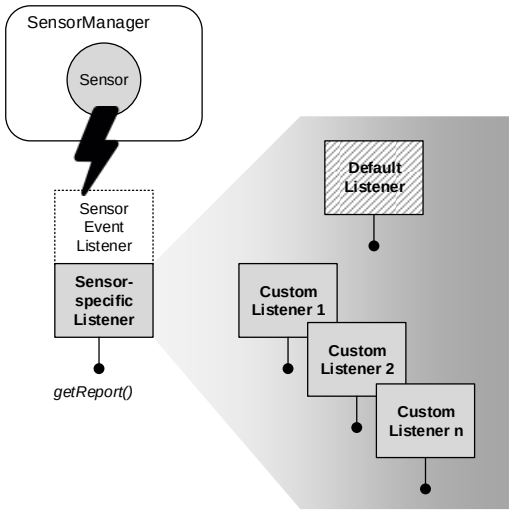


Figure 1. Implementation of the device-agnostic sensor server architecture using a common Android sensor framework design pattern

Using this common pattern of accessing Android sensor framework, we designed an architecture that supports both existing and emerging types of sensors that will appear in the future. For each type of sensor that is common in already released devices, we provide a specific implementation of custom sensor listener. Each of these listeners receives the raw data from the hardware sensor and reformats them according to the semantics of a specific sensor. The web server then uses the *getReport()* method to read the latest sensor readings and sends the data to the remote client. To support new types of sensors that will appear in the future, we provide a default sensor listener. It is registered with all sensors of unknown type. Default sensor listener is a generic type of listener that receives the raw data from the hardware sensor, but does not do any sensor-specific data formatting. Data are delivered to the client in raw format and is up to the client to interpret the semantics of the sensor data.

IV. Self-Configurable Web Server for Android Sensor Framework

Device-agnostic sensor server architecture described in Section 3 enables the development of a single version of the web server that is ready to run on any type of Android-powered device, regardless of the available sensory equipment. In addition, such architecture enables the web server to configure its interface to the remote clients according to a specific capabilities of the target device. The procedure of a web server self-configuration is shown in Fig. 2.

A. *Sensor Discovery and Listener Setup*

Initialization of the web server begins with sensor discovery. The Android operating system keeps tracks of the state of each sensor, as well as their types, power consumption, etc. This information is available to applications via *SensorManager* class. At the startup, the web server uses the *SensorManager* to retrieve the list of available sensors on a device it is running on.

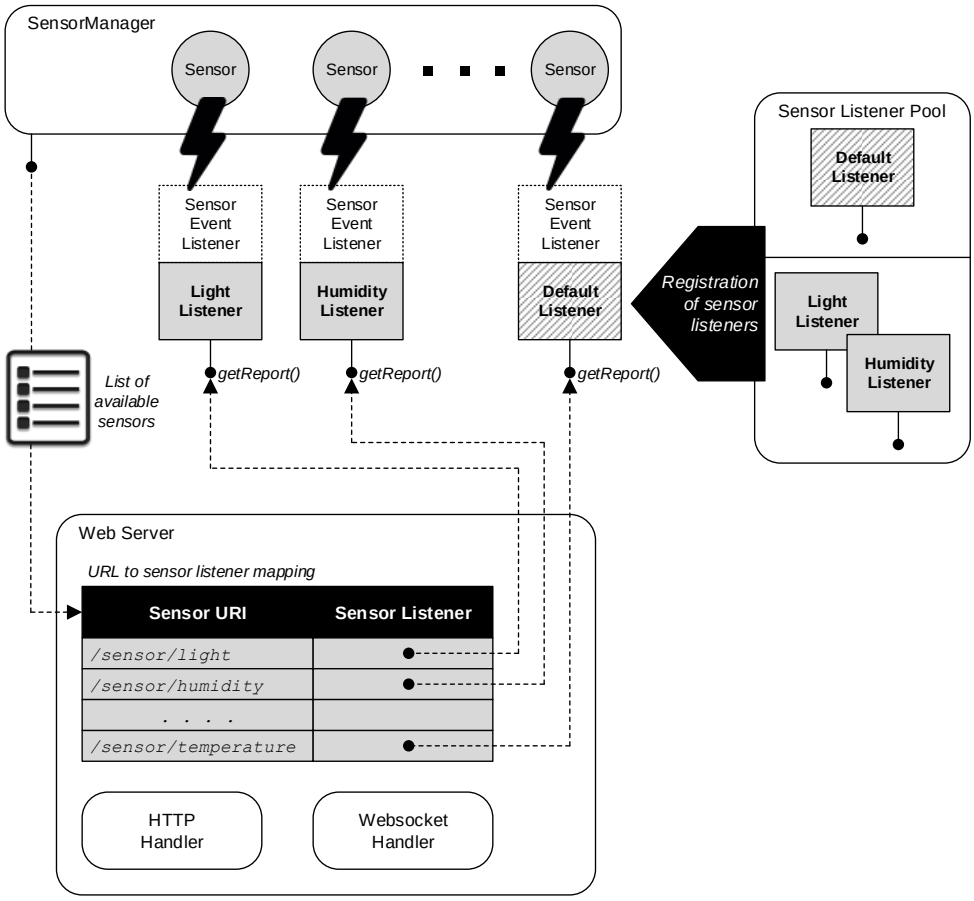


Figure 2. Self-configured web server for Android sensor framework based on the device-agnostic sensor sever architecture

SensorManager delivers a list of *Sensor* objects, each of which represents a particular hardware sensor.

The next step is to register a listener with each sensor in order to be able to access sensor data. Web server iterates through sensor list retrieved from *SensorManager* and inspects the type of each sensor. Then it searches through the pool of sensor listeners to check if there is any sensor-specific listener assigned to the given type of sensor. If so, it registers an instance of the listener with the sensor. If no sensor-specific listener is found in the pool, the web server registers an instance of the *DefaultListener* with the sensor. The role of the *DefaultListener* is to provide a minimum functionalities necessary to expose raw sensor data of unknown sensors to the web server. At the end of this procedure, each hardware sensor has associated listener with public *getReport()* method exposed through which the web server retrieves the sensor data when clients ask for them.

The example shown in Fig. 2 illustrates the scenario in which the web server is aware of the existence of light and humidity sensor, but does not yet have support for top-notch device that comes equipped with a temperature sensor. Therefore, it instantiates specialized listeners to handle light and humidity sensors, but uses the default listener for temperature sensor. The light- and humidity-related data sent out to the clients will be formatted according to their semantics, while the data read from the temperature sensor will be sent out in a form of raw floating point values. Until the web server gets upgraded with full support for handling temperature-related data, this mode of operation provides a minimal necessary support for remote clients to retrieve these data.

B. *URI Scheme Construction*

Final step in web server initialization is to build a URI scheme for exposing sensory data to remote clients through a REST interface. Each sensor needs to be associated with a URI through which remote clients can access sensor data. Since the number and types of sensors vary from device to device, the URI scheme is also different on different devices.

For particular device, the URI scheme is generated dynamically during the web server startup based on the list of available sensors retrieved from *SensorManager*. Based on the type of a particular sensor in sensor list, the web server constructs a URI for each sensor and maps the URI to the instance of the sensor listener registered with that type of sensor during the sensor discovery phase.

C. *Dual Protocol Support*

Sensory readings are formatted into HTML snippets and delivered to the remote clients using standard HTTP protocol. Responses are generated using HTML templates, which get filled with requested sensory data. Although the sensory data representation format may be adapted to the specific needs of particular applications, HTML is chosen as a default format in order to be easily displayed in clients' web browsers.

Using HTTP protocol to read sensory data is satisfactory for sensors whose state stays stable for a relatively long period of time, or for applications that do not require frequent updates of sensor state. However, there are situations where a constant feed of data is crucial for correct operation of an application. For example, if a client is interested in getting info on temperature or atmospheric pressure, a single HTTP request followed by the server's response may be enough for a period of few hours because temperature and pressure are changing slowly. This, however, does not hold true if client is interested in using device's motion sensors for gathering GPS coordinates to calculate speed or routes. Frequent delivery of updated sensory data using HTTP protocol makes communication inefficient due to the fact that each update requires new HTTP request-response cycle.

To make the communication more network-efficient, the proposed implementation of the web server also supports the *Websocket* protocol. Using the *Websocket* protocol, a permanent network connection is established between the client and the server, so a stream of data can flow through the network continuously. To get updated sensor reading, the client does not need to send new HTTP request to the server for each update. Instead, the server pushes the data to the client when they changes on the mobile device. *Websocket*-based communication enables near real-time updates of sensory data in client applications. Besides being network-efficient, *Websocket*-based communication also saves device's battery life because no need for repeating client's HTTP requests significantly reduces the amount of network traffic received through a device's wireless network interface.

V. Web Server Upgrades to Support New Types of Sensors

Proposed architecture enables seamless introduction of new modules if the web server needs to be upgraded to provide full support for new types of sensors that appear with new models of Android devices (e.g. radiation sensor). To support a new type of sensor, we need to develop new sensor-specific listener that replaces the default listener.

Fig. 3 shows the object-oriented design pattern used to develop new sensor-specific listener. The listener is developed by subclassing the *DefaultListener* class. Since the *DefaultListener* already implements the *SensorEventListener* interface defined in Android sensor API, each new subclass is compatible with Android sensor framework to receive notifications about sensor state changes. This is achieved by exposing the *onSensorChanged()* method, which is invoked by the Android sensor framework. Subclassing the *DefaultListener* allows the developer to redefine the raw sensor data handling and formatting by overriding the *getReport()* method. Once the object instantiated from new developed listener class gets registered with device's *SensorManager*, the web server will use this object instead of the *DefaultListener* to retrieve sensor data. This, in turn, enables the web server to get the sensor data in their sensor-specific, rather than raw form, and to deliver them to the clients preformatted.

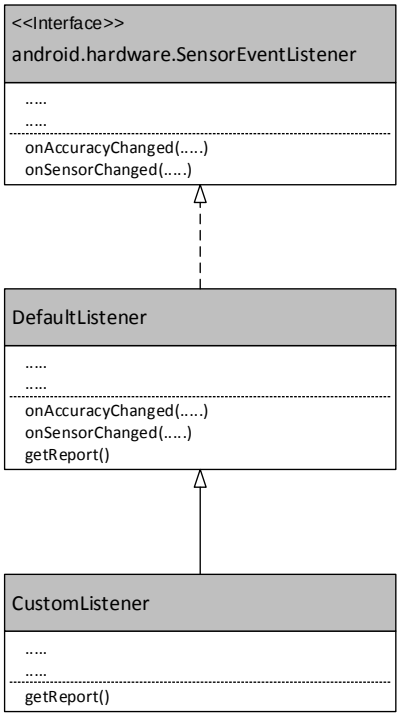


Figure 3. An excerpt of the class diagram that enables the extension of the web server with custom sensor listeners

VI. Conlusion

In this paper, we present the concept of exposing a smartphone's sensor framework to the remote users or applications through the Internet. To make sensor data available for remote access, we present an implementation of a web server for execution on Android-powered mobile devices. The purpose of the web server is to operate as a component of a volunteer-based environment sensing infrastructure where smartphones are used as terminal devices to collect and deliver sensory data from their surroundings to the remote clients.

Each individual sensor is exposed to clients through REST API via HTTP and *Websocket* protocol. The web server is implemented as an Android service, which is a special type of Android application component that can perform long-running operations in the background. This enables that smartphone owners can contribute to the environment sensing program, while at the same time regularly and unobtrusively using other applications on their devices. Since the number and types of available sensors vary from device to device, special attention is paid to the modular web server architecture that enables dynamic creation of URI scheme based on the available set of sensors, as well as default data handling for unknown types of sensors. However, the architecture is modular and enables seamless integration of sensor-specific data handling modules as new types of sensors are introduced with new models of devices.

References

[1] M-F. Moens, J. Li, and T-S. Chua, Mining User Generated Content, Chapman and Hall/CRC, 2014

[2] A. Lih, The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia, 1st ed, New York: Hyperion, 2009

[3] BOINC, http://boinc.berkeley.edu/

[4] E. Korpela, D. Werthimer, D. Anderson, J. Cobb, and M. Leboisky, "SETI@home – massively distributed computing for SETI", Computing in Science & Engineering, vol. 3, no. 1, pp. 78-83, January-February 2001

[5] F. Smadja, "Mixing financial, social and fun incentives for social voting", in Proceedings of WEBCENTIVES'09, Madrid, Spain, 2009

[6] J. Chamberlain, U. Kruschwitz, and M. Poesio, "Motivations for participation in socially networked collective intelligence systems", in Proceedings of CI2012, MIT, Cambridge, MA, USA

[7] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk", Judgment and Decision Making, vol. 5, no. 5, pp. 411-419, June 2010

[8] D. Gourley et al., HTTP: The Definitive Guide, Sebastopol, CA: O'Reilly Media, Inc., 2002

[9] R. Fielding et al., Hypertext Transfer Protocol – HTTP/1.1, http://tools.ietf.org/ search/rfc2616

[10] V. Wang, F. Salim, and P. Moskovits, The Definitive Guide to HTML5 Websocket, New York, NY: Apress, 2013

[11] I. Fette and A. Melnikov, The Websocket Protocol, http://tools.ietf.org/search /rfc6455

[12] I. Hickson, The Websocket API, http://www.w3.org/TR/ websockets/

[13] NanoHttpd – the tiny, easily embeddable HTTP server in Java, http://nanohttpd.com/

[14] Jetty – servlet engine and HTTP server, http://eclipse.org/jetty/

[15] I-Jetty – Jetty for Android, https://github.com/jetty-project/i-jetty

[16] Android API Guides: Application Fundamentals, http:// developer.android.com/guide/components/fundamentals.html

[17] Android Reference: Services, http://developer.android.com/guide/ components/services.html

[18] Java Servlet API, http://docs.oracle.com/javaee/6/api/javax/ servlet/Servlet.html

[19] DroidPHP – a small, fast and lightweight open source Web Server for Android, https://github.com/DroidPHP/DroidPHP

[20] Android API Guides: Sensors Overview, http://developer.android.com/guide/topics/sensors/sensors_overview.html

Estimating diver orientation from video using body markers

Ivor Rendulić, Aleksandar Bibulić, Nikola Mišković
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
Email: ivor.rendulic@fer.hr
Email: aleksandar.bibulic@fer.hr
Email: nikola.miskovic@fer.hr

***Abstract*—Diving is a dangerous activity during which diver's safety is significantly compromised, both due to unpredictable environment and dependence on technical systems for life support. Introducing the concept of a cognitive autonomous underwater vehicle (AUV), that is capable of guiding, observing and assisting the diver at all times, requires precise positioning of the AUV relative to the diver. While diver orientation can be measured locally on the diver and transmitted to the underwater vehicle using acoustic communication channels, the problems of limited bandwidth and communication delay can influence the quality of AUV positioning. Another approach, which is considered in this paper, is to use remote sensing techniques, i.e. mono or stereo camera on board the AUV, to determine orientation of the diver based on markers placed on the divers shoulders. As an initial step, laboratory experiments were conducted where orientation of two physically coupled spherical markers is determined by using mono and stereo camera, and compared to the ground truth orientation obtained from an inertial measurement unit. The obtained results prove the concept in laboratory conditions, with the limitations imposed on the distance of markers relative to the camera.**

I. Introduction

Diving is an extremely dangerous activity not only because of the hazardous and unpredictable underwater environment in which divers operate, but also because of the fact that divers lives depend on technical equipment such as breathing regulators. One of the main objectives of CADDY project is to introduce an autonomous underwater vehicle that will act as cognitive robotic diving buddy, thus significantly reducing risks inherent to diving activities (see Fig. 1). This autonomous vehicle will act at the same time as a buddy "guide", leading the way for the diver; buddy "observer", looking out for the diver; and buddy "slave", helping the diver execute specific tasks. In order to achieve these functionalities, the vehicle has to perceive the diver (determine diver's position and orientation) in order to be able to position itself around the diver while at the same time keeping the diver within its field of view.

This work is supported by the European Commission under the FP7-ICT project "CADDY - Cognitive Autonomous Diving Buddy" Grant Agreement No. 611373.

Fig. 1. A diver and a remotely operated underwater vehicle working together.

Underwater environment imposes many constraints on technical systems. Estimating orientation of a person, which can easily be done on dry land with inertial measurement sensors, becomes a non-trivial task. Sensor has to be placed in a waterproof casing and the data transfer at distances longer than a few meters has to be done either by acoustic communication or by wired link. Given the very high price of an acoustic communication system and many constraints caused by having a wire connected to a diver, it is reasonable to explore other remote sensing options.
In this paper we address the problem of estimating orientation and distance of a diver for the purpose of positioning an autonomous underwater vehicle (AUV) which serves as a cognitive diving buddy. As an alternative to having a sensor mounted on the diver and needing to transmit the data to our vehicle we are using video cameras to record the diver. Our vehicle is equipped with a stereo and a mono camera, both of which were used for the task.
To cope with often poor visibility issues, caused by water turbidity or low lighting, we are using highly visible body markers to mark predefined body parts of a diver. Altho-

ugh diver detection and distance estimation can be done in various conditions even without the markers, estimating the orientation is much more tricky if a distinct marker (e.g. mask or air tank) is not visible in the image. As a starting point in diver orientation estimation, this paper focuses on laboratory experiments on dry land. Even though this setup significantly simplifies the situation which is expected in the underwater environment, this step is crucial for determining the proof of concept. Separate approaches by both mono and stereo camera are taken to calculate the orientation and distance based on the detected spatial positions of the markers.

The paper is organized as follows. Sections II and IIi describe the principles of using mono and stereo camera for determining positions of visual markers, while Section IV describes how orientation is determined from the obtained data. Section V focuses on the experimental setup and reports results from laboratory experiments, while the paper is concluded with Section VI.

II. Mono Camera

The first approach that was used in estimating the orientation was by tracking the spatial position of two distinct visual markers with an ordinary mono camera.

It is clear that mono cameras have a major limitation in tasks that include estimating distances from objects. It is impossible to give a good estimate of depth from a single image and without any prior knowledge. In [1] authors use supervised learning to train a model for depth estimation in a single still image. Other approaches include estimating depth from video of a moving camera based on optical flow [2], [3].

For the problem of determining diver orientation, we are using a much simpler solution of estimating depth from the perceived size of a familiar object. Given a reference size of an object in pixels at a known distance, we are estimating the change of distance from the changes in size.

If an object is recorded at two different distances z_1 and $z_2 = sz_1$, the ratio of object areas in pixels in the two recordings P_{1pix}/P_{2pix} is equal to the ratio of areas A_2 and A_1 captured by the camera.

$$\begin{aligned} P_{1pix}/P_{2pix} &= \frac{P/A_1}{P/A_2} = \frac{A_2}{A_1} = \frac{a_2 b_2}{a_1 b_1} \\ &= \frac{k(2z_2 \operatorname{tg}(\alpha/2))^2}{k(2z_1 \operatorname{tg}(\alpha/2))^2} \\ &= \frac{z_2^2}{z_1^2} = \left(\frac{sz_1}{z_1}\right)^2 = s^2 \end{aligned}$$

In the formula above, P_{1pix} and P_{2pix} are the sizes of the recorded object in pixels when the object is recorded at distances z_1 and z_2 from the camera. A_1 and A_2 are areas of the field captured by the camera at distances z_1 and z_2. s is the ratio of distances z_2/z_1, α is the horizontal field of view angle and k is the height-to-width aspect ratio of the image $k = b/a$. All values are depicted on Fig. 2. Given that the ratio of object area sizes is inversely proportional to the square of the ratio of distances, square roots of object area sizes or any linear measure of objects (such as height, width or radius) are linearly inversely proportional to the ratio of distances. This enables us to estimate the unknown distance to the object by measuring it's size and knowing size at a reference distance.

$$z_{curr} = z_{ref}\sqrt{\frac{P_{ref,pix}}{P_{curr,pix}}}$$

Also, several measurements, such as area, width and height, can be combined to give a better estimate.

For the lateral movement from the camera two approaches are considered. Firstly, pixel size at a given distance can also be estimated from a pixel size at a reference distance and estimated depth of an object. However, it is reasonably to assume that the markers mounted on a human body (e.g. on the shoulders) maintain constant mutual distance l. This enables the calculation of relative positions of the markers, given the previously calculated depths z_1 and z_2.

$$\begin{aligned} t &= \sqrt{\frac{b_{pix}}{b}} \\ t &= \sqrt{\frac{(x_{2,pix} - x_{1,pix})^2 + (y_{2,pix} - y_{1,pix})^2}{l^2 - (z_2 - z_1)^2}} \\ x_2 &= (x_{2,pix} - x_{1,pix})/t \\ y_2 &= (y_{2,pix} - y_{1,pix})/t \end{aligned}$$

In the equations above, t is the ratio of distance measured in pixels and actual distance. It is calculated as a ratios of distance b, marked in 3. The numerator, b_{pix}, is the distance in pixels, and the denominator, actual distance b, is calculated from the previously calculated depth difference between the two markers. Based on the ratio t marker locations can be transformed from pixels to millimeters.

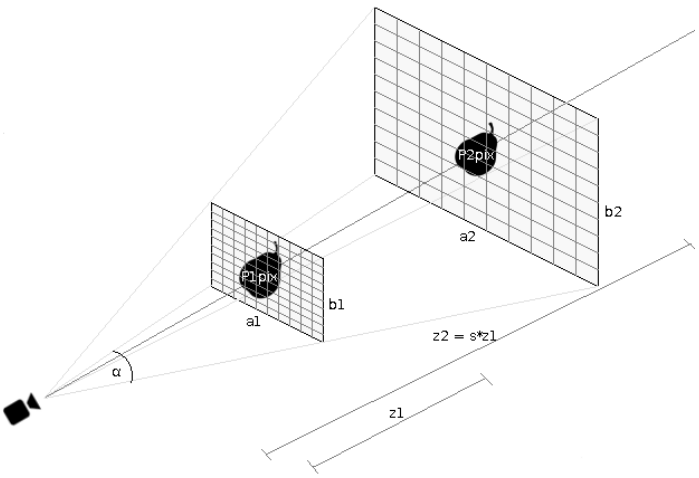


Fig. 2. Visualization of object recording at different distances.

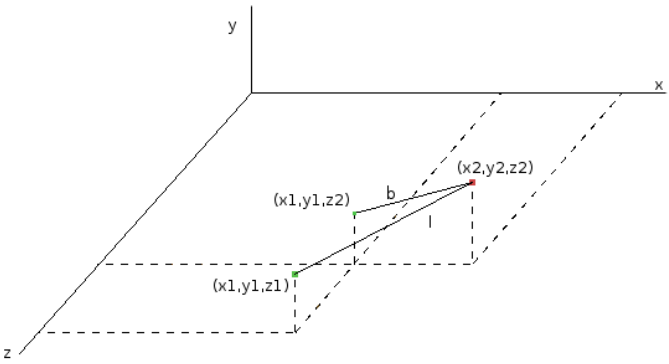


Fig. 3. Visualization of object recording at different distances.

III. Stereo Camera

In the second approach, orientation was estimated by using stereo camera. Stereo vision system is used to determine depth from two or more images taken at the same time from horizontally shifted cameras.
The calculation of depth includes the following steps of processing pairs of stereo images:

1) Calibration of a stereo camera system
2) Stereo image rectification
3) Search for correspondence points in stereo pair
4) Calculating distance from disparity

The process of calibration is to determine the estimation of the intrinsic parameters for each camera and estimation of the translation and rotation of the second camera relative to the first one. This is accomplished using a calibration plane with an asymmetric checkerboard. Multiple pairs of images of a calibration pattern is needed to calibrate the system. Detailed explanation of calibration process are omitted but can be found in in [4].
Estimated parameters are used to rectify a stereo pair of images. Rectification is a transformation process used to project two or more images onto a common image plane, which reduces the 2D stereo correspondence problem to a 1D problem (see e.g. [5]).
Now corresponding point in left image can be found in the same row in the right image. Points chosen are centers of the markers. Disparity is calculated as distance between corresponding pixels in the left and right image as

$$d = |x_{1pix} - x_{2pix}|$$

where x_{1pix} is x coordinate of the corresponding pixel in the left image and x_{2pix} is x coordinate of the corresponding pixel in the right image. Now we can tell that the geometry of the stereo camera system is known, meaning that x, y and z coordinates of the markers can be calculated as

$$z_p = T * \frac{f}{d}$$
$$x_p = x_{lpix} * \frac{f}{d}$$
$$y_p = y_{lpix} * \frac{f}{d}$$

where T is the distance between two cameras, f is the focal length, x_{lpix} and y_{lpix} are the coordinates of the coordinate system from the left camera.

IV. Orientation From Spatial Coordinates

After obtaining the spatial coordinates of both markers by either of the methods described in previous sections, orientation of the diver (e.g. assuming the markers are mounted on his shoulders) can be calculated by simple trigonometry. For representation of the orientation Euler angles yaw and roll are used.

$$\psi = \text{arc tg}(\frac{z_2 - z_1}{x_2 - x_1})$$
$$\phi = \text{arc tg}(\frac{y_2 - y_1}{\sqrt{(z_2 - z_1)^2 + (x_2 - x_1)^2}})$$

Here ψ is the yaw angle around the z-axis and ϕ is the roll angle around the x-axis. The position of the markers on a diver is shown in Fig. 4, along with the rotation axes. Calculation of all three angles cannot be done from only two markers and would require a third one.

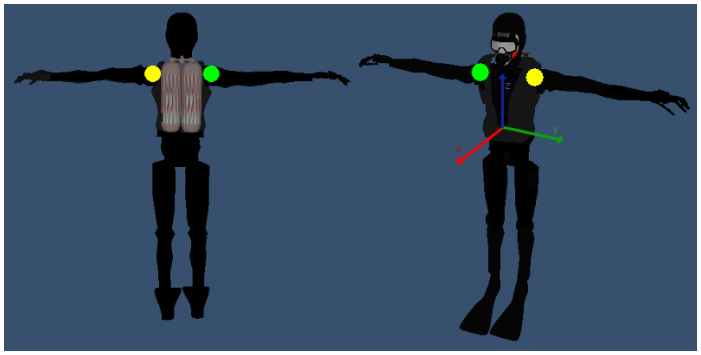


Fig. 4. Position of the markers on a diver and coordinate system of the diver for calculating orientation.

V. Experimental Setup and Results

To test the accuracy of the proposed systems an experimental setup consisting of two items (visual markers) mounted on opposite ends of a pole was used. An IMU sensor was used to provide reference orientation during the recording with a mono and a stereo camera. The setup is shown in Fig. 5.

Fig. 5. Experimental setup for visual orientation estimation.

A. IMU Reference Orientation

A Pololu MinIMU-9 was used to obtain reference orientation. It includes three 3-axis sensors, measuring acceleration, angular velocity and magnetic field. All sensors were sampled at 50Hz and time-stamped. The independently measured nine degrees of freedom enables the calculation of absolute orientation in space. Data from the three sensors is processed by a complementary filter described in [6] in order to get a precise estimate.

B. Visual Marker Detection

1) Mono Camera: Video acquisition was performed with an ordinary computer web camera in VGA resolution (640x480) for the mono orientation estimation experiment. Frames were captured at average of 15Hz and time-stamped to allow direct comparison to the IMU reference orientation.
Template- and color-based detection were used to obtain the positions (in pixels) of the x and y coordinates of the markers, and to give an estimate of the marker size in pixels.
Color-based detectors were used to simply distinguish pixels matching the marker colors in the image and find the areas that are most likely corresponding to the markers. This method allows for simple size estimation by counting the pixels that are matched to each marker.
For the template-based detection, OpenCV template matching library was used on a multiple scales of the marker image and the results were used to estimate the position and depth of the markers.

2) Stereo Camera: For the stereo video acquisition a monochromatic Point Grey Bumblebee XB3 was used. The camera has three sensors distanced at 12cm apart, allowing three separate stereo pairs to be formed for a single shot. Frames were taken at 5Hz and also time-stamped for comparison to the IMU.
The markers used in stereo vision test were spherical, so the detector used Hough Circle Transform [7].

Table I
Standard deviations of differences between orientations obtained by IMU and visual sensing.

	$< 1m$	$3m$
mono camera roll	$\sigma = 3.58°$	$\sigma = 10.54°$
mono camera yaw	$\sigma = 8.16°$	$\sigma = 24.14°$
stereo camera roll	$\sigma = 2.53°$	$\sigma = 7.67°$
stereo camera yaw	$\sigma = 4.01°$	$\sigma = 12.23°$

C. Results

In this section we present the results of the visual orientation estimation.

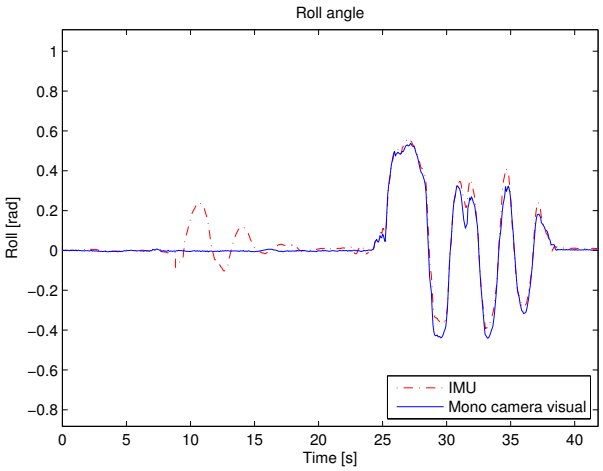
1) Mono Camera: Tests with the mono camera were conducted by tracking two almost-spherical markers whose diameter is approximately $10cm$. As expected, the depth estimation based on size change is heavily influenced by the distance to the object. The tests were conducted at two distances.
First, the setup was moved at distances between 0.5 and 1.0 meter from the camera. Results from those tests are visible in Fig. 6. The orientation obtained from the visual tracking is very accurate and closely follow the one obtained with the IMU. Then, the same markers were used at a larger distance from the camera. This impairs the ability of the marker detector to give a good estimate of the size, as well as limits the resolution of the detected marker size in pixels. The results shown in Fig. 7 show the orientation calculated when the setup was at slightly over 3.0 meters from the camera.
Standard deviation from the reference orientation obtained with IMU is shown in Table I.

2) Stereo Camera: Unlike the mono-based orientation estimation described above, the stereo based one does not suffer as much on longer distances from the accuracy of object size and position detection.
Tests were conducted with the same spherical markers as for mono camera. The distance between the markers and the camera was $1m$ in the first trial and $3m$ in the second one. Results of two trials are shown in Fig.8 and Fig.9.
Standard deviation from the reference IMU orientation is given in Table I.

VI. Conclusions and future work

This paper presented the results of laboratory experiments that were designed to determine two angles of orientation (yaw and roll) of physically coupled spherical markers by using mono and stereo camera. These markers will at a later stage be mounted on the divers shoulders, and will be used to determine diver's orientation relative to the autonomous underwater vehicle equipped with the cameras.

The initial results have shown that, at least in controlled laboratory conditions, both mono and stereo camera can be used to determine orientation. Since calculation of yaw and roll angle requires distance of the object to the camera, the experiments were carried out at two distances.



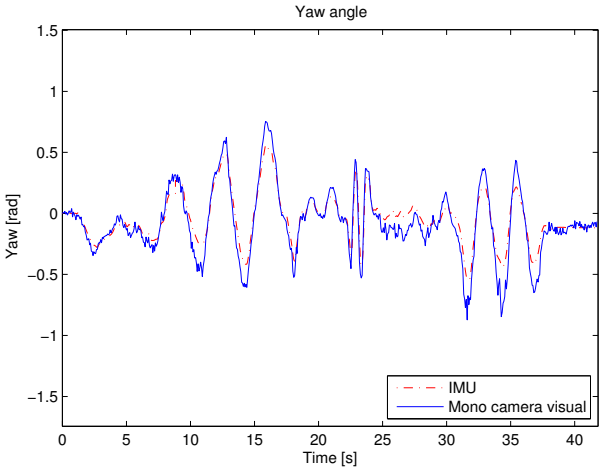
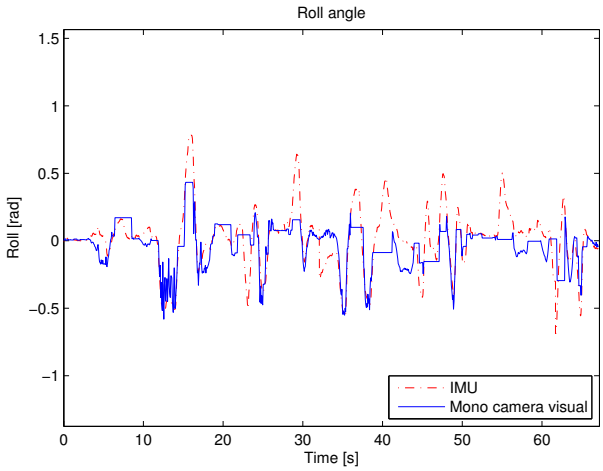


Fig. 6. First trial with mono camera, movement induces changes in both yaw and roll angles. Distance from the camera $< 1m$



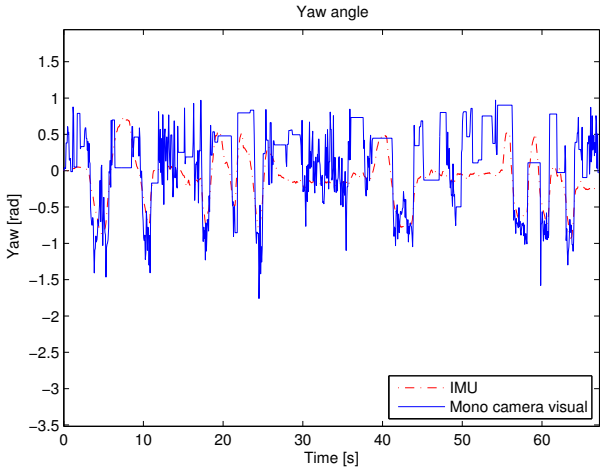


Fig. 7. Second trial with mono camera, movement induces changes in both yaw and roll angles. Distance from the camera $\approx 3m$

Ground–truth was set with measurements from an inertial measurement unit.

We have shown that results obtained by mono camera are very good at small distances, but are heavily deteriorated with increasing distance. On the other hand, measurements from the stereo–camera are less influenced by distance, mostly due to more precise distance measurements, keeping the standard deviation of measurements below 10°.

Future work will include using the experimental setup underwater, in laboratory conditions. After that, the markers will be mounted on the diver and full field experiments will be conducted, in order to test the robustness of image processing in real conditions, with low visibility.

References

[1] A. Saxena, S. H. Chung, and A. Y. Ng, "3-d depth reconstruction from a single still image," *International journal of computer vision*, vol. 76, no. 1, pp. 53–69, 2008.

[2] B. Shahraray and M. K. Brown, "Robust depth estimation from optical flow," in *Computer Vision., Second International Conference on.* IEEE, 1988, pp. 641–650.

[3] W. Kruger, W. Enkelmann, and S. Rossle, "Real-time estimation and tracking of optical flow vectors for obstacle detection," in *Intelligent Vehicles' 95 Symposium., Proceedings of the.* IEEE, 1995, pp. 304–309.

[4] P. Hillman, "White paper: Camera calibration and stereo vision," *Lochrin Terrace, Edinburgh EH3 9QL, Tech. Rep*, 2005.

[5] C. Loop and Z. Zhang, "Computing rectifying homographies for stereo vision," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1. IEEE, 1999.

[6] S. O. Madgwick, A. J. Harrison, and R. Vaidyanathan, "Estimation of imu and marg orientation using a gradient descent

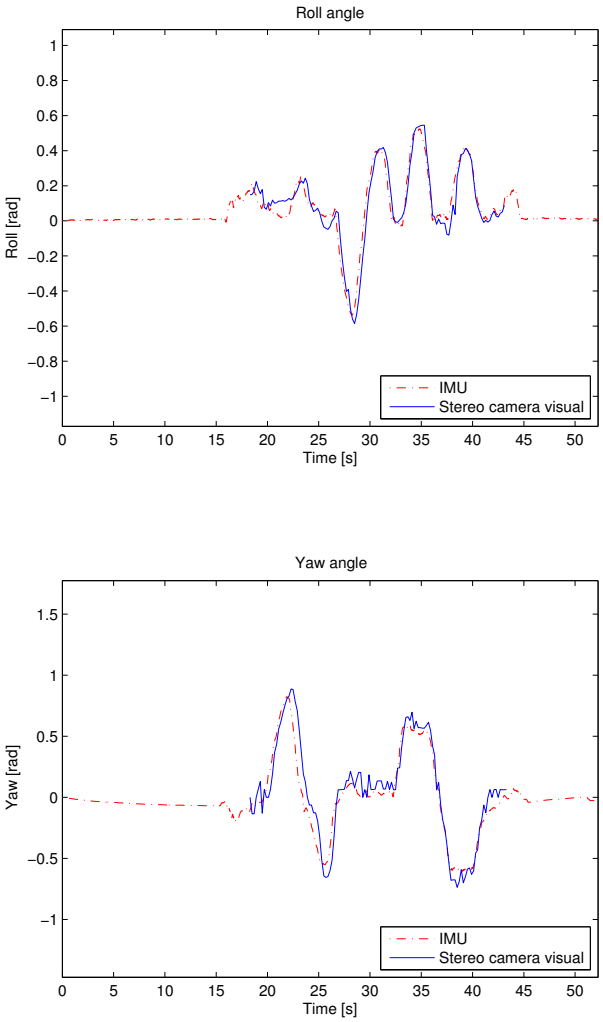


Fig. 8. First trial with stereo camera, movement induces changes in both yaw and roll angles. Distance from the camera $\approx 1m$

algorithm," in *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–7.

[7] M. Smereka and I. Dulęba, "Circular object detection using a modified hough transform," *International Journal of Applied Mathematics and Computer Science*, vol. 18, no. 1, pp. 85–91, 2008.

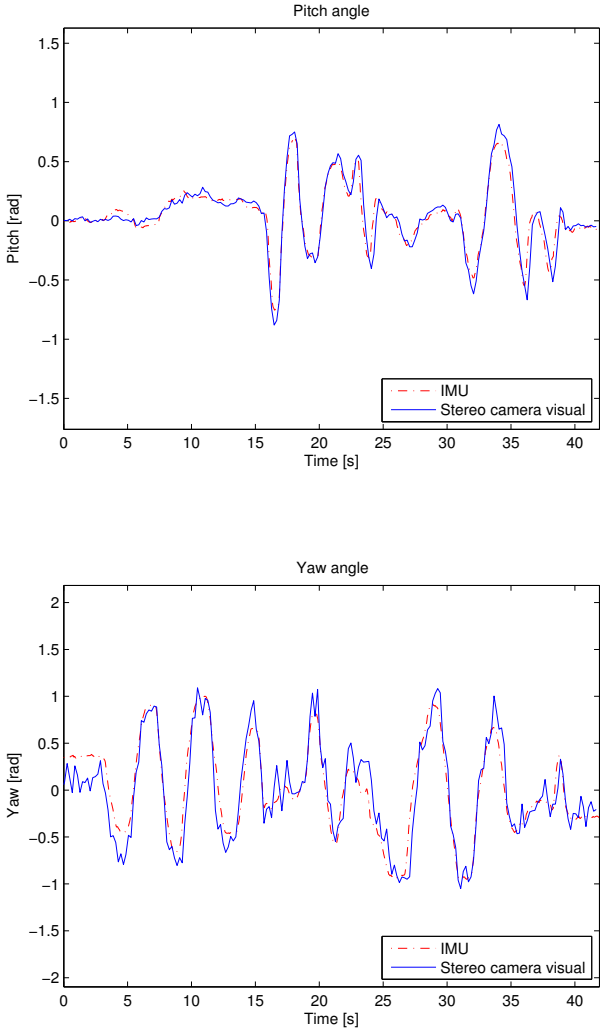


Fig. 9. Second trial with mono camera, movement induces changes in both yaw and roll angles. Distance from the camera $\approx 3m$

Technologies for Distributed Flight Control Systems: a Review

M. Šegvić, K. Krajček and E. Ivanjko
University of Zagreb/Faculty of Transport and Traffic Sciences, Zagreb, Croatia
miroslav.segvic@computer2cockpit.com, karolina.krajcek@fpz.hr, edouard.ivanjko@fpz.hr

Abstract - This paper reviews the state of the art (SOTA) in distributed flight control technologies using publicly available, scientific and technical publications. Technological developments, such as embedded computing and microelectromechanical systems are enabling advanced aerospace-oriented distributed systems. Distributed systems are comprised of a large number of simple elements, each with its own sensing, actuation and control, in order to obtain the desired behaviour. These systems have the potential to be more economic than a centralized system due to the simplicity of individual components, and possibility of using the same production unit for a different role within the system. The challenge with these systems is the coordination amongst nodes comprising a distributed control network. Benefits of distributed architecture are increased robustness trough redundancy and inherent fault tolerance. The SOTA summary comprises a description of challenges in the design of flight control systems with a distributed structure, technologies currently used in flight control systems and also technologies not specifically related to distributed flight control but applicable for the design of future flight control strategies. Described systems and technologies are represented with examples of real systems including swarms of small Unmanned Aerial Vehicles and distributed networks for Fault Detection and Isolation.

I. INTRODUCTION

Recent advancement in aviation requires more and more sophisticated control systems. New control systems are needed for both, air traffic control and aircraft flight control systems. One of the new approaches are distributed control systems which are examined in this paper. The aim of this paper is to provide a survey of technologies developed and deployed for distributed flight control. The need for distributed control systems arises from the demands of today's modern aircraft which contain many subsystems. All subsystems have to work in an optimal way to ensure that all security and economic constraints are fulfilled. To ensure that, new smart sensors and actuators are used. Such smart elements contain a local embedded computer (controller) with data processing and communication abilities comprising a distributed control system. In such a system a significant amount of data processing is done in local embedded controllers and the master control unit has a global overview of the whole system.

There are three possible types of system architectures for a control system in general: (i) Central architecture; (ii) Distributed architecture; and (iii) Federated architecture. A central architecture uses a centralized hardware and a centralized software framework. One computer is used for several subsystems. As all control hardware is centralized, the environment can be controlled very well [1]. Also the maintenance of these systems is easy. All calculations are also centralized. The distributed architecture uses a distributed hardware and distributed software framework. All calculations are finding place in the sensors and the results are transmitted. A central control unit does not exist and all subsystems have to communicate with each other [1]. The federated architecture is a compromise between the central and distributed architecture. It uses a distributed hardware and centralized software. There are more subsystems than in the case of centralized hardware, but fewer than in the case of distributed hardware [1].

A centralized control approach for Flight Control System (FCS) requires a large amount of electrical cables originating at the flight control computer and ending at actuators and control surfaces as shown in the Fig. 1. The issue of larger mass and complexity of centralized FCS, along with the susceptibility of servo control signals and sensory wiring to noise originating from surrounding electrical systems, are the main technical reasons for the development of distributed FCS.

Recent effort in the field of FCS design is directed towards the improvement of safety, efficiency and overall cost [2]. Airbus patented a robust control system that does not need a large amount of cabling and is based on transmitting calculated signals over the data bus in the form of command messages [3]. The idea of the Airbus invention is to use a distributed FCS, organized around a communication bus in which the control function and some monitoring functions are performed remotely at the actuators.

This paper is organized as follows. In the second section, the problem of flight control is described. The third section describes distributed FCS, and proposed methods for system decentralisation and design. The fourth section reviews emerging technologies related to FCS, such as MEMS (microelectromechanical systems) based systems, unmanned aerial vehicle (UAV) swarm control, distributed fault detection and isolation (FDI) and optimisation. Paper ends with a conclusion.

II. THE PROBLEM OF FLIGHT CONTROL

The problem of aircraft flight control is in creating appropriate control inputs to the aerodynamic control surfaces. An aircraft can have primary, secondary and

auxiliary control surfaces. Primary flight control surfaces are required to safely control an aircraft during flight and include ailerons, elevators and a rudder (Fig. 1). Secondary control surfaces are intended to relieve excessive control load and include trim and spring tabs. Auxiliary control surfaces improve the aircraft performance characteristics and include flaps, spoilers, speed brakes and slats. These control surfaces can be controlled by a pilot or by an autopilot as given in Fig. 1.

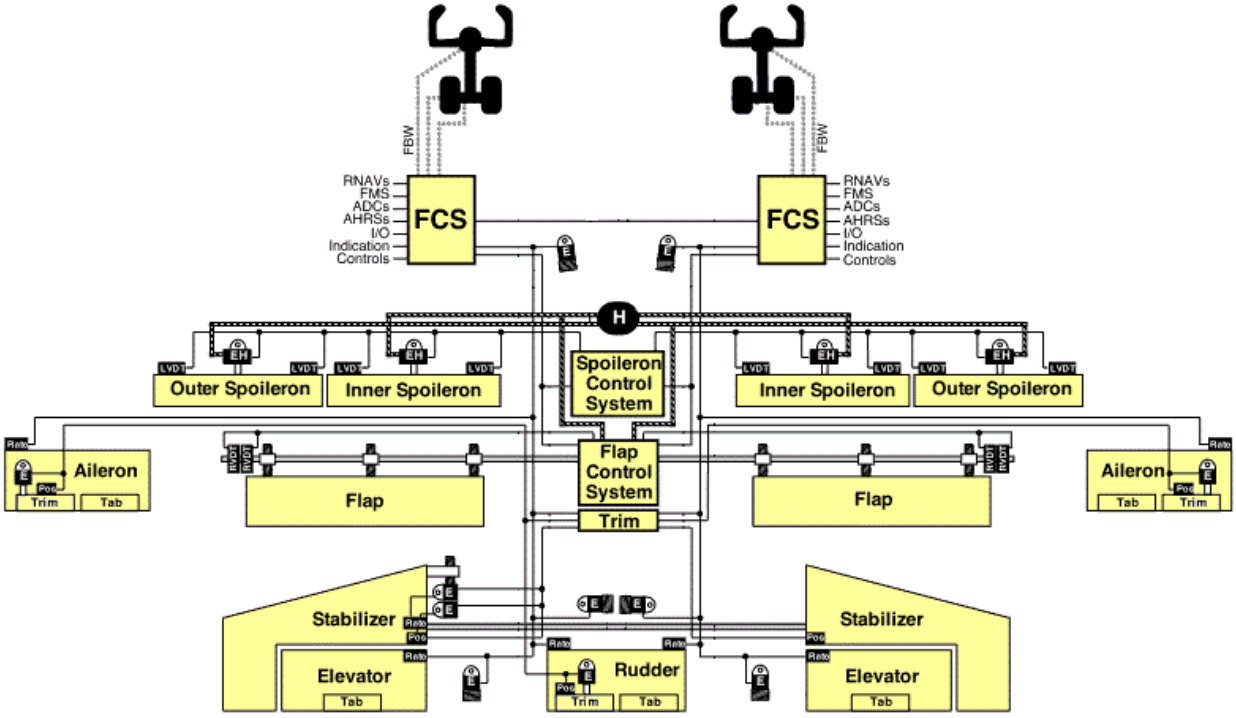


Fig. 1. Aircraft flight control surfaces and architecture [26]

The aim of all mentioned control surfaces is to control the airplane height, speed (V), roll (p), pitch (q) and yaw (r) rates, and moments (L, M and N) as presented in Fig. 2.

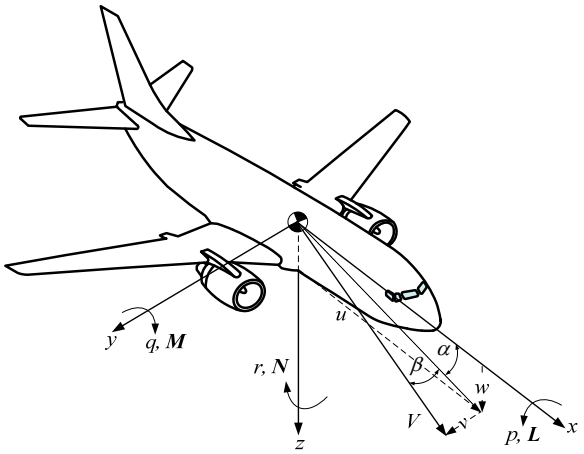


Fig. 2. Body fixed frame with aerodynamic angles, pitch rates and moments

On a modern high speed aircraft, the pilot's control joystick and pedals are not directly linked with the control surfaces. Appropriate systems are in between and augment the pilot control commands applied on the joystick and pedals. Large aircrafts have systems that convert pilot control commands into electrical signals. They are sent via electrical or optical cables to various actuators which can be electromechanical, pneumatical, electrical or hydraulic devices. Actuators then move the appropriate control surfaces to the desired direction. Today's sensors, actuators and the pilot are part of an electronically controlled system or fly-by-wire (FBW) system that enables full-time surface control utilizing advanced control laws. Control commands from the pilot or autopilot are adjusted to improve the stability characteristics of the airplane.

When the time scale of aircraft dynamics is too small that the airplane can be manually managed or when the pilot workload is demanding, it is necessary to have automatic flight control implemented. The automatic flight control system (AFCS) has two separate control subsystems. One subsystem is required for guiding the vehicle's center of mass along a specific trajectory and the other for controlling the aircraft attitude by rotating it around its center of mass. The first control subsystem type is known as the autopilot. The second control subsystem type is known as the stability augmentation system (SAS) and control augmentation system (CAS).

Autopilots are usually segregated into a lateral and a vertical mode. Typical autopilots in lateral mode are: roll angle hold, turn coordination and heading hold/ VOR hold and in vertical mode: pitch attitude hold, altitude hold, speed/Mach hold and automatic landing. Typical CAS functions are: roll rate, pitch rate, normal acceleration and lateral/directional control. SAS serves as a roll, pitch and yaw damper.

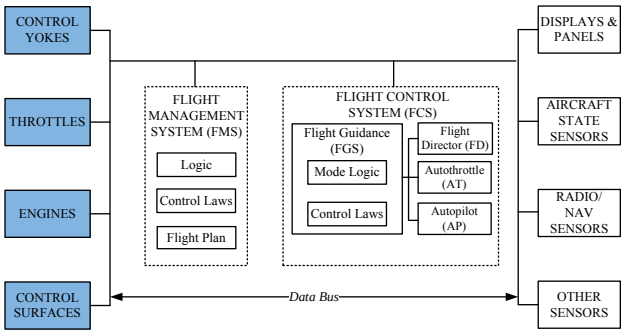


Fig. 3. High level avionics architecture in modern airliners and business jet airplanes [25]

Modern avionics architecture is comprised of many individual systems. Two key functions are the Flight Management System (FMS) and FCS. The FCS is composed of a Flight Guidance System (FGS), Flight Director (FD), Autothrottle (AT) and Auto-Pilot (AP). The FGS is a software function that generates roll and pitch values, and also speed commands used to control the aircraft.

III. AIRCRAFT DISTRIBUTED CONTROL SYSTEMS

Decentralized/distributed control has been a control approach of choice for designing large scale systems for decades. Decentralization assumes splitting the unmanageable complex problems into manageable smaller sub problems, so that the resulting solutions solve the original overall problem.

A. State of the Art in current aircraft distributed control systems

Continuously increasing requirements for aircraft and air transport safety along with operational demands for reliability, performance, efficiency and costs, are shifting the focus of recent development to distributed systems. The massive voting architecture proposed by Airbus [4] suggests to allocate the task of control laws and logic between flight control computers and control surface actuator nodes as shown in Fig. 5. Flight control computers and actuator nodes are connected via an advanced data communication network developed by Airbus. Flight control computers calculate the control laws and proprietary commands for control surface actuator nodes, which are then broadcast as messages over the communication bus. Actuator nodes are equipped with flight control remote modules, and perform massive voting upon receiving the messages from many flight control computers. The massive voting architecture resides upon digital communication technologies. New smart actuator technologies are explored for particular system application. Fault handling in the system proposed from Airbus is resolved within the actuator nodes. A high degree of fault detection as well as a fault location is demonstrated, both due to the large number of nodes [5].

A distributed FCS architecture is presented also for accessing fault handling and redundancy managing on the JAS39 Gripen [6]. The proposed system included 16 nodes. Various simulations showed that distributed sensor nodes meet fault detection coverage of 99% for both transient and permanent faults. The proposed system used a triggered multi master broadcast bus with Time division multiple access communication. As a result, the failure on any node cannot jeopardize communication by sending data outside the dedicated time slot, resulting in a fail silent system.

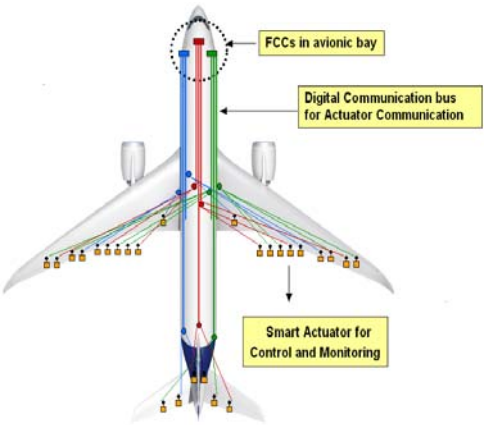


Fig. 4. Fully distributed FCS architecture [5]

Power line communications (PLC) have been proposed for distributed aircraft control systems in [7]. The PLC communications approach eliminates the need for digital data bus wiring by modulating the data on power cables that are installed between the flight control computer and control surface actuators.

Although technology is promising and widely used in other applications, vehicular control systems are not usually installed with PLC systems. For aircraft's FCS, there are many requirements that make implementation of PLC difficult, such as using negative return wires on the power bus instead of chassis return as usual. The problem arises from selective frequency fading or multipath fading. Furthermore, as a general system design safety rule requires, primary and secondary flight surfaces must remain independent. More than one network must be used to reduce wiring, and also for tail surface reliability. Communication speed requirements for various standards must be met, and also to ensure reliability with a given number of remote units. From many other aspects, PLC has to be further developed for aircraft use and its usability is yet to be explored.

Decentralisation is entering other aircraft subsystems, with the development of larger and more complex aircraft. Smart components are proposed for a decentralised fuel management system [8] and microcontrollers are embedded in the pumps, valves and sensors (Fig. 5).

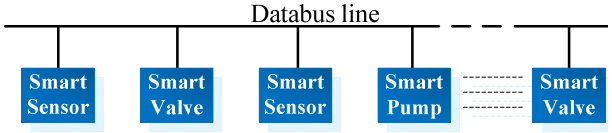


Fig. 5. Architecture of the distributed fuel control system

Proposed system components make its own decisions during various fuel operations, depending on the performed action. They share a time-triggered bus for communication. When a smart component reaches a decision, it transmits it over the bus. For safety, all system components retain a copy of the state vector that describes

the system state. Laboratory and real scale testing have been performed proving such a distribution is possible and that the new system can be adaptable to faults.

B. Challenges of distributed system design

In order to create a control system, the control system designer first has to establish the structure of the whole system, its subsystems and their interconnections. Then control inputs and measured outputs are determined. Local controllers are designed using available input and outputs to satisfy the overall system stability and performance. Described procedure is known as decentralized control design.

Properties characterizing complex dynamical systems are: a large scale of system equations, different types of uncertainties of system models, information structure limitations, and communication delays [9]. The problem of high dimensionality can be solved by decomposing the system into subsystems with their respective interconnections. A complex system can be decomposed in two ways. One way is to physically decompose the system according to its physical properties. That can be challenging as there are many systems where it is difficult to identify weak couplings. The other approach is a numerical decomposition of the system. It effectively addresses the problem of high dimensionality [9].

Control of large scale complex systems requires efficient design methods and algorithms whose implementation needs minimal information exchange among local systems [9]. Recent growth of communication networks intensifies the move of traditional system theory to communication networks. This is particularly suitable for complex real world systems such as transportation and large manufacturing systems, the internet or embedded control systems. Wireless networks offer significant advantages but also introduce some drawbacks compared with wired networks. The advantages include low cost of operation, ease of installation and reconfiguration, natural reliability, robustness, and adaptability enabling in avionics a flight-by-wireless system [10]. Drawbacks are time delays, packet loss, finite capacity, and data flow problems.

Digital networked dynamic system design can be analysed from the position of:

- Control over a network covers the design of feedback strategies adapted to existing control systems where control is performed via unreliable communication links;
- Control of networks that is concerned with providing a defined level of performance to network data flow;
- Multi agent systems study how network architecture and interactions between different network components affects global control goals.

Interconnected systems have been classified according to different principles and system structures into disjoint, overlapping and hierarchical decompositions, as well as decompositions for time scale, singular symmetric systems [9]. A basic modern theory problem in interconnected systems is the approach to timing. The timing problem is in finding the communication frequency of the networked-based feedback loops for which system stability and performance is satisfied. Typical phenomena concerning network performance are: random delays, packet dropouts, quantization and data rate limits.

We can separate time driven and event driven feedback schemes, regarding to the data transmission approach. The periodic sampling of the continuous time system characterizes the time driven approach. This approach generates redundant communication within the feedback loop. That results in an effort to reduce the communication traffic only to necessary data. The event driven system communication is triggered by the occurrence of an event to engage a sample realization. As a result, the event triggered control relieves the communication network and also saves energy.

IV. Emerging technologies related to flight control systems

A. MEMS based FCS

The MEMS is a "batch-fabricated integrated microscale system that; converts physical stimuli, events, and parameters to electrical, mechanical and optical signals and vice versa; performs actuation, sensing and other functions; comprises control, diagnostics, signal processing and data acquisition features" [11] .

MEMS sensors are very promising in aerospace technology because of their small size (the overall dimensions are a few millimetres), low power consumption (fractions of a watt), high reliability (operation life up to 120000 h) and extremely low cost [12]. Although MEMS sensors lack the accuracy of conventional sensors, it is being improved continuously. For example, the drift of the first MEMS gyros put into mass production in 1995 was greater than 500 deg/h according to [12]. Currently, it is reduced to below 1.0 deg/h [13], and it is predicted to go down to 0.1 deg/h [12].

Large manufacturers are developing MEMS based gyros for production to replace a low and medium accuracy laser and fiber optic gyros. For example, the company Northrop Grumman LITEF GmbH is producing a small sized inertial reference system LCR110 with MEMS accelerometers for civil aviation purposes. Current development lies in bringing the performance of mass produced MEMS sensors to a level required for use in aircraft navigation systems, and in designing MEMS based devices for a wide range of aircraft. MEMS sensors are known to feature zero bias and scale factor depending on temperature, gyro output depending on linear overloads, high noise components, and sensor output parameters practically linearly related to power supply instability. Performance improvements are based on a number of circuit, software, and design modifications.

MEMS sensors are particularly important for the development of UAV flight control systems. The flight control system enables an UAV to fly autonomously based on a series of preplanned waypoints while transferring real-time data of mission targets to ground control stations. Commercial UAV flight control systems

are available by companies like MicroPilot and UAV Flight Systems.

UAV flight control systems must provide at least altitude hold control and global positioning system navigation, in order to perform UAV missions. Using several UAV agents to perform a mission quickly and more efficiently as a team is a problem of distributed control of UAV swarms.

B. UAV swarms

UAV are a rapidly evolving technology with a variety of military applications such as reconnaissance, target and decoy combat missions [14]. A variety of civil and commercial applications can be performed also. Certain surveillance applications such search and rescue can be accomplished most effectively by teams of cooperating autonomous agents, i.e. swarms of UAVs.

Development of reliable autonomous navigation algorithms for individual UAVs was a technological prerequisite for development of cooperative team technologies and frameworks. Recently explored approaches that deal with multi-agent collaboration significantly rely on agent-to-agent communication. Ad-hoc wireless technologies are explored for line of sight communications, primarily due to the availability of low cost of the shelf technologies/parts [15]. Using the communication infrastructure, a team of UAVs receives a list of mission tasks and collaborates to autonomously and efficiently assign the tasks among themselves.

Multi-level control systems are designed for centralized mission command centres. Levels for mission control, agents monitoring and task execution monitoring comprise mission command center control. Such an approach reduces the amount of training for the mission operator, relieving him of monitoring and task organizing duties.

Completely decentralized and autonomous system enables the fleet of UAVs to accomplish missions beyond the range of any centralized command station [16]. Agents within such a system are performing all of the collaboration and control processing on the aircraft. Also they perform decentralized task allocation for a dynamic mission via on-board computation and ad-hoc aircraft-to-aircraft communication.

Determining the stability properties of multi-agent teams is rather challenging, but can be verified in restricted scenarios. This problem is addressed using methods such as model predictive control [17] and receding horizon control [18]. Future work includes development of platforms with described properties within larger airframe UAVs, all in an effort to increase flight time and support more complicated missions [19].

C. Fault Detection and Isolation

According to [20], a fault is „an unpermitted deviation of at least one characteristic property (feature) of the system from the acceptable, usual, standard condition“ and a failure is “a permanent interruption of a system’s ability to perform a required function under specified operating conditions”. Large distributed systems like automotive systems, jet engine control, and teams of unmanned vehicles are comprised of many individual components and sensors. It is very difficult to diagnose faults in interconnected distributed systems using a centralized FDI architecture, mostly due to constraints imposed by computational abilities and communication limitations. Therefore, a distributed FDI architecture is developed extensively in recent years by [21] and [22].

Operations of distributed systems rely significantly on sensory health. A sensor fault may lead to poor control performance and compromise stability of the control system. Sensor information error can propagate to other subsystems trough interconnections, affecting their performance, and also causing difficulties pinpointing and isolating the faulty unit. Therefore, sensor fault diagnosis is a critical issue in distributed interconnected control systems. “A fault-tolerant control system is capable of controlling the system with satisfactory performance even if one or several faults, or more critically, one or several failures occur in this system” [23].

Distributed FDI architecture is proposed by [21]. For each subsystem a local FDI component is designed. The FDI component uses local measurements and communicated information from neighbouring FDI components associated by interconnection. Each local FDI component consists of a fault detector estimator and adaptive fault isolation estimator. Fault detector estimator monitors the local subsystem under normal operating conditions. If a sensor fault is detected, fault isolation estimator isolates subsystem where the fault has occurred. There is an inherent trade-off between robustness and fault detectability. The importance of adaptive thresholds for distributed sensor FDI is demonstrated, as they ensure robustness with respect to interactions among interconnected subsystems and modelling uncertainty.

D. Distributed optimisation

Aerospace systems control can be described as a complex problem. Approaches from computer science are introduced to aerospace systems in order to cope with the large complexity of such systems. One of such approaches is reinforcement learning which is based on selecting actions, receiving rewards, and updating settings for future actions.

Another promising approach for distributed aerospace systems is the Collective Intelligence (COIN) framework [24]. Collective in this case is “a system of adaptive computational agents with a specified system level performance criterion” [24]. COIN techniques have already been applied to a many of distributed optimization problems such as network routing and data collection by autonomous vehicles. Collective systems address the optimization task as tasks distributed among agents which represent system variables. The variables can represent a specific control surface position on an aircraft. The collective’s solution is generated when agents select actions and receive rewards depending on the system objective. The actions can be commands relayed to a control system. Equilibrium for the multi agent system is reached when the agents can no longer improve their rewards by selecting a different choice of actions. The COIN research puts a focus on local reward

functions which are also referred to as private utilities, as the connection to the increase of the convergence rate is clearly proven [24]. Probability Collectives (PC) theory extends the COIN framework and handles constraints more explicitly as it is necessary for aerospace systems. PC theory concentrates on how the system agents update the probability distributions on range of their possible actions instead of focusing on the joint action created by sampling those distributions.

The collectives approach is specifically designed for systems that are distributed. The use of probabilities removes any issues associated with discrete actuation. Because noise and disturbance are often characterized using probabilities, the collectives approach is naturally suited. Collectives can use locally available information, or alternatively include information from other agents and can use a data broadcast from a centralized source. Formulating the problem of control as a distributed optimization problem makes these systems benefit from the application of machine learning algorithms, statistics, gradient-based methods for optimization, and game theory strategies [24].

V. CONCLUSION

Distributed FCSs present a significant leap in the evolution of aircraft FCS architectures. Novel technological advances in areas of embedded computing and communication, machine learning, and multi agent systems control continue to push FCS design towards distributed systems. Although there are some demonstrations of distributed systems for aircrafts, they are mostly analysed from the aspect of fault detectability and identification. However, distributed control systems should be further explored to find the final optimal way how the control law calculation can be decentralised at the same time fulfilling all safety criteria. For example, some systems offer voting mechanisms for identical flight control computers and nodes, to achieve redundancy.

Authors of this review conclude that other ways of decentralisation, possibly across hardware architecture boundaries between different control surfaces should be investigated, and that the possibility of decentralised decision making needs to be further examined.

LITERATURE

[1] Wikibooks. (2014, October) Embedded Control Systems Design/Aviation. [Online]. http://en.wikibooks.org/wiki/Embedded_Control_Systems_Design/Aviation

[2] T. Jones, "Large transport aircraft: Solving control challenges of the future," Annual Reviews in Control, vol. 38, no. 2, pp. 220-232, 2014.

[3] M. Fervel, J. J. Aubert, and A. Maussion, "Distributed flight control system," US 8,600,583 B2, December 3, 2013.

[4] P. Goupil, "AIRBUS state of the art and practices on FDI and FTC in flight control system," Control Engineering Practice, vol. 19, no. 6, pp. 524-539, 2011.

[5] M. Sghairi et al., "Distributed and Reconfigurable architecture for flight control system," in Proceedings of 28th Digital Avionics Systems Conference, Orlando, FL, 2009, pp. 6.B.2-1 - 6.B.2-10.

[6] K. Ahlstrom, J Torin, K. Fersan, and P. Nobrant, "Redundancy management in distributed flight control systems: Experience and simulations," in the 21st Proceedings on Digital Avionics Systems Conference, vol. 2, Irvine, USA, 2002.

[7] J. O'Brien and A. Kulshreshtha, "Distributed and remote control of flight control actuation using power line communications," in 27th Digital Avionics Systems Conference, IEEE/AIAA: St Paul, USA, 2008, pp. 1 -12.

[8] J.M. Giron-Sierra, S. Cifuentes, J. F. Jimenez, M. Seminario, and C. Insaurralde, "Aircraft Fuel Management Reconfigurable System with Smart Components for Distributed Decision," in Proceedings of 10th Intl. Conf. on Control, Automation, Robotics and Vision, Hanoi, Vietnam, 2008, pp. 1534-1539.

[9] L. Bakule, "Decentralised control: Status and outlook," Annual review in Control, vol. 38, no. 2, pp. 22-71, 2014.

[10] R. K. Yedavalli and R. K. Belapurkar, "Application of wireless sensor networks to aircraft control and health management systems," J Control Theory Appl, vol. 9, no. 1, pp. 28–33, 2011.

[11] S. Lyshevski, "Smart flight control surfaces with microelectromechanical systems," IEEE Transactions On Aerospace And Electronic Systems, vol. 38, no. 2, 2002.

[12] A. G. Kuznetsova, Z. S. Abutidzeb, B. I. Portnova, V. I. Galkina, and A. A. Kalika, "Development of MEMS Sensors for Aircraft Control Systems," Gyroscopy and Navigation, vol. 2, no. 1, pp. 59–62, 2011.

[13] D. Arch. (2012, December) MEMS Journal, Inc. [Online]. http://www.memsjournal.com/2010/12/high-performance-mems-gyroscopes-current-status-and-emerging-trends.html

[14] P. Scharre. (2014, October) Robotics on the Battlefield Part II, The Coming Swarms. The Center for New American Research, 2014.

[15] T.X. Brown, S. Doshi, S. Jadhav, and J. Himmelstein, "Test bed for a wireless network on small UAVs," in Proceedings of AIAA 3rd Unmanned Unlimited Technical Conference, 2004, pp. 20-23.

[16] R. R. McCune, G. R. Madey, "Swarm Control of UAVs for Cooperative Hunting with DDDAS," Procedia Computer Science, vol. 18, 2537 – 2544, 2013.

[17] A. Richards and J. How, "Decentralized model predictive control of cooperating UAVs," in 43rd IEEE Conference on Decision and Control, vol. 4, 2004, pp. 4286 - 4291.

[18] E. Franco, T. Parisini, and M.M. Polycarpou, "Cooperative Control of Distributed Agents with Nonlinear Dynamics and Delayed Information Exchange: a Stabilizing Receding-Horizon Approach," in Proceedings of 44th IEEE Conference on Decision and Control and the European Control Conference, 2005.

[19] A. Ryan et al., "A Modular Software Infrastructure for Distributed Control of Collaborating UAVs," in AIAA Guidance, Navigation, and Control Conference and Exhibit, Keystone, Colorado, 2006, p. 10.

[20] R. Isermann, Fault-Diagnosis Systems, An Introduction from Fault Detection to Fault Tolerance. Berlin Heidelberg: Springer-Verlag, 2006.

[21] Q. Zhang and X. Zhang, "Distributed sensor fault diagnosis in a class of interconnected nonlinear uncertain systems," Annual Reviews in Control, vol. 37, no. 1, pp. 170-179, 2013.

[22] R. M. G. Ferrari, T. Parisini, and M.M. Polycarpou, "Distributed Fault Detection and Isolation of Large-Scale Discrete-Time Nonlinear Systems: An Adaptive Approximation Approach," IEEE Transactions on Automatic Control, vol. 57, no. 2, 2012.

[23] G. J. J. Ducard, "Fault-Tolerant Flight Control and Guidance Systems for a Small Unmanned Aerial Vehicle," ETH ZURICH, Zurich, PhD Thesis, 2007.

[24] S. R. Bieniawski, "Distributed Optimization and Flight Control Using Collectives," Stanford university, USA, PhD Thesis, 2005.

[25] A. C. Tribble, S. P. Miller, and D. L. Lempia, "Software Safety Analysis of a Flight Guidance System," in Proceedings of 21st Digital Avionics Systems Conference (DACS), 2002, p. 75.

[26] J. Wilson. (1996) ASTECH Engineering. [Online]. http://www.astech-engineering.com/systems/avionics/aircraft/flightcontrolsdesign.html

The Computer Architecture and Hardware Description Language

A.S. Burlakov, A.E. Khmelnov
Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences,
Irkutsk, Russian Federation
alex_burlakov@bk.ru

The paper introduces CAHDL, the Computer Architecture and Hardware Description Language. CAHDL is aimed to specify processor instructions and computer hardware formally. Although CAHDL supports RTL modeling to some extent, it is does not describe digital circuits. CAHDL is designed first to specify CPU and hardware behavior, rather than their structure. CAHDL can be applied in simulators, compilers, disassemblers, testers and other programs requiring specification of computer architecture. In the end of the paper we observe how CAHDL is applied in computer simulation. The paper brings some CAHDL syntactic constructions as examples and explains their semantics. Parsing methods are also briefly reviewed.

I. INTRODUCTION

There are many hardware description languages like Verilog HDL [1] or VHDL [2] aimed to specify integrated circuits. These languages work on three levels of abstraction: on structural level, on RTL level and on behavioral level. Although CAHDL supports RTL modeling to some extent, it works on behavioral level. The main purpose of the language is to provide a high level specification of a computer architecture which could be used in software.

The CAHDL is an object oriented imperative language with declarative constructions. It operates the high level constructions like processor, memory, keyboard, etc. The syntax of CAHDL bases on C++ ISO/IEC programming language. There are also some syntax constructions taken from Verilog and from technical literature [3-6]. The source [7] was used in the development of the syntax of description of arguments of processor instructions.

The CAHDL was successfully applied in the DCU32INT program (Delphi Compiled Units Disassembler) [8] alongside with FlexT specification language [9]. In the end of the paper we observe the application of CAHDL in a computer simulator. This simulator was used in testing of the CAHDL itself and in debugging specifications written in CAHDL.

II. SYNTAX

Computer Architecture and Hardware Description Language is an object-oriented programming language designed for describing computer architecture. The syntax was influenced by C++ and Verilog and reminds C++ to some extent. However, there are special syntactical constructions, which cannot be easily realized in any existing programming language, even by C macros.

The language supports data types of objects, integers and strings. They may be used in scalar variables, arrays, constants and references. There are special mechanisms of *nesting* and *concatenation* of integer variables (see below) which are implemented by the use of references.

Besides the standard assignment operators (=, +=, -=, *=, /=, &=, |=, %=) and program flow control operators (*if, for, while*), the language provides operators working with different bits. These operators considerably facilitate the process of describing of computer architecture [10].

The most interesting of the operators are extract, concatenation and branching operators. Extract operators can extract a separate bit <*bit_number*> or a sequence of bits <*higher_bit..lower_bit*>. We don't use the Verilog form of bits extraction with square brackets, because they are reserved for arrays. An operator of concatenation (colon character) concatenates two or more operands. A branching operator *case* works similar to the visual basic *case* operator.

In order to adjust an environment, the specialist has to operate with objects of corresponding types. Memory, processor, monitor are all some types and they are represented in a description as some device classes. The language supports an encapsulation mechanism. A class may encapsulate fields and methods. The language also supports an inheritance mechanism. It allows to access public fields and methods of a parent class. Class fields and methods could be hidden if they are located in a *private* section. Like in C++, if a field is located in *protected* section, it is only available in derived classes. Listing 1 represents a fragment of the description of the most complicated computer device – CPU (Central Processor Unit).

```
class CCpuIx86 : public CCpu
{
private:
  CMemoryIx86 m_memory;
  // general purpose registers
  bits<32> EAX, EBX, ECX, EDX;
  bits<32> ESI, EDI;     // index
  bits<32> ESP, EBP;     // stack
  bits<32> EIP;          // program
  bits<32> EFLAGS;       // flag
  bits<16> CS,SS,DS,ES,FS,GS;
  bits<16>& AX = EAX<15..0>;
  bits<8>& AH = AX<15..8>;
```

```
  bits<8>& AL = AX<7..0>;
  bits<8>& BH = EBX<15..8>;
  bits<8>& BL = EBX<7..0>;
  bits<16>& BX = BH:BL;
  // SP length would be set automatically
  bits& SP = ESP<15..0>;
  bits<1>& CF = EFLAGS<0>;
  bits<1>& ZF = EFLAGS<6>;
public:
  bits& RNs(bits<3> N, bits<1> W){
    if(W)    case(N){
      maskb 0:0:0 then return EAX;
      maskb 0:0:1 then return ECX;
      maskb 0:1:0 then return EDX;
      ...
    }
  }
  void Process(){
    bits<3> Ns, Nd;
    bits<1> W;   // ? dword : word
    unsigned short opcode = m_memory.ReadByte(EIP);
    case(opcode){
      maskb 1:0:0:0:1:0:1:W:1:1:Nd:Ns
        then MOV(R(Ns,W),R(Nd,W));
      maskb 0:1:0:0:Nd:0          then INC(R(Nd));
      maskb 1:e                   then PUSH(DS);
      maskb 1:f                   then POP(DS);
      ...
    }
  }
  ...
};
```

LISTING 1. EXAMPLE OF IA-32 PROCESSOR ARCHITECTURE DESCRIPTION

The listing 1 is an incomplete description of IA-32 processor architecture. This description is represented as a *CCpuIx86* class. In the body of the class are registers and a memory type object (see Listing 5). General purpose registers, index registers and stack pointer registers are represented as integer variables 32 bits length. Segment registers are 16 bits integers.

The reference type realizes mechanisms of nesting and concatenation. These mechanisms make it possible to declare AX register as a nested variable of EAX register. AH and AL registers are nested in AX register. At the same time, a variable containing other variables may be declared as a concatenation (colon character) of these variables. Thus, in the example, BX is declared as a concatenation of AH and AL registers.

It is notable, that if the size of the variable is not defined, then it would be set automatically, as it is done for SP register. In that case, the length of SP register would be 16 bits pointing to the first 16 bits (from 0 to 15) of ESP register.

The carry and zero flags are defined as a single bit variables pointing to zero and sixth bits of EFLAGS respectively.

It is worthy to discuss the *case* brunching operator. The keyword *case* was chosen to designate for branching operator because of its similarity to the Visual Basic *case* operator. It works not only with constants as C++ *switch* operator does, but also with expressions and bitmasks. An example of the *case* operator could be found in listing 1, in the *Process* and *RNs* functions. In the row

```
maskb 1:0:0:0:1:0:1:W:1:1:Nd:Ns
```

within the *Process* function, the *case* statement works as a predicate of arity 3, where 3 is a number of entries of variables in the concatenation sequence. This feature of *case* operator is used for representation of tabular information, such as processor instruction specification. The syntax of *case* operator is as following:

```
CASE '(' expr ')' '{'
  case_body_list
'}' ';'

case_body_list: case_body_row
| case_body_list case_body_row
;

case_body_row:
  WHEN expr1 THEN stmt_list1
  | MASKB binary_mask_conc      THEN stmt_list2
  | MASKO octal_mask_conc       THEN stmt_list3
  | MASKH hexadecimal_mask_conc THEN stmt_list4
;
```

LISTING 2. CASE SYNTAX

The keyword *when* is followed by an *expr1*, the result of which will be compared to the result of *expr*. In case of matching, the *stmt_list1* will be executed. After the keywords *maskb*, *masko*, *maskh* follows a concatenation of numbers written in binary, octal or hexadecimal format, but without prefixes '0' or '0x' for octal and hexadecimal numbers respectively. If there is a variable in a concatenation sequence, it would be interpreted as a number, rather than a variable. For instance, in a row

```
maskb 1:0:0:0:1:0:1:W:1:1:Nd:Ns
    then MOV(R(Ns,W),R(Nd,W));
```

LISTING 3. IA-32 MOV COMMAND IN CAHDL

after the keyword *maskb* follows a sequence of binary formatted bits and variables, joined by the concatenation operator. When this concatenation is compared to a number, these variables may be interpreted as any value, which does not exceed the length of a corresponding variable. The length of *Ns* and *Nd* is 3 bits. It is known from their declarations (see listing 1).

A concatenation may be considered as a bitmask. If a mask matches some required number, then the variables in a concatenation sequence would be assigned with values so that the result of concatenation would be equal to the value required. Instead of using of C++ statements like in example below

```
// expr – is a case argument
// look listing 2
Ns = expr & 7;
Nd = (expr >> 3) & 7;
if((expr >> 6) & 3!= 3)
  return;
W = (expr >> 8) & 1;
if((expr >> 9) & 127!= 69)
  return;
MOV(R(Ns,W), R(Nd,W));
```

LISTING 4. FRAGMENT OF IA-32 MOV COMMAND IN C

we write this instruction in two lines of code (see listing 3).

It is notable, that different formats of a digit representation cannot be mixed in a single row, i.e. *maskb* cannot be followed by digits in a hexadecimal format. Moreover, if we write some digits after *maskh* keyword, these digits would be interpreted as hexadecimal ones. For example, an instruction of concatenation *maskh 1:1* would be interpreted as number 17 rather than number 3.

Abstract classes are used as an interface of basic devices. Like in other programming languages, in CAHDL we cannot create objects of abstract classes. In order to create a device we need to define its behavior in a class inherited from the base abstract class. In the example above (see listing 1) the class *CCpuIx86* inherits fields and methods of a virtual abstract *CCpu* class, hence it must implement all *CCpu* abstract methods, one of which is the *Process* method.

Other computer devices like random access memory or hard disk drive may be described in a similar way. There we need to define an inherited class, which will implement all the abstract methods. Let's see how it would look like for a memory class.

```
class CMemoryIx86 : public CMemory
{
private:
  // the size of memory in bytes
  const RAM_SIZE = 0x80000000;//2Gb

public:
  CMemoryIx86(){…};
  unsigned char ReadByte(unsigned long address){…};
  void WriteByte(unsigned long address
                , unsigned char value) {…};
  …
}
```

LISTING 5. MEMORY CLASS IN CAHDL

The code in listing 5 is self-explanatory. There we define memory class as one derived from *CMemory* class. In order to set a memory size, we just change the value of *RAM_SIZE* constant.

If we need to track memory changes, we would have to edit code for *WriteByte* method. It is also possible to change CPU class methods in order to track register values. (Tracking values changes method is widely used in debugging technics.) Besides memory tracking we may even change the logic of some methods, accordingly to our demands.

III. PARSING

The description of a computer's architecture is stored in a simple text file. When the simulator is started, the function *yyparse* begins to parse the description text. The body of *yyparse* function is generated automatically by Bison and Yacc programs, according to the language grammar description[11]. The CAHDL has some syntactical constructions, which could not be reduced by a single lookahead symbol, thus the GLR algorithm was preferred to LALR(1). The main difference between LALR(1) and GLR is that GLR can backtrack more than one parsing node if the production is not matched [11,12]. In the body of *yyparse* function we build a parsing tree. In the nodes of the tree are objects of different classes. However, all these classes have the same base class.

The grammar of the language developed is context-sensitive. The aim of context-sensitivity is to simplify code understanding. Context-sensitivity deals only with the representation of numbers. In general, hexadecimal numbers must follow '0x' prefix. If we use a hexadecimal number in the *case* construction within *maskh* section, then the prefix must be omitted. In case of ambiguity, when an identifier may be interpreted as both a number and a variable of the same name, the number would be preferred due to the fact that it is easier to rename the variable rather than represent a number in another form.

For example (see listing 1), *PUSH* operation in *Process* method would be executed, if only the *opcode* variable matches mask 0x1e. In this case, the symbol *e* would be interpreted as a number 14, rather than an "e" named variable. This means that *opcode* must be equal to 0x1e.

The language supports modality at a level of files. This modality is similar to one used in C++. In order to include another file of description, we have to place it in quotation marks following an *#include* keyword.

Bison and Yacc generate a parser. This parser is not an executive file, but a bunch of source files. This source code is included into the simulator's project. When the embedded parser founds a grammar mistake in a description file, it alerts an error outputting some debug information and then it terminates the program's execution. Now, the debugging information is quite simple and reveals only position of a mistake, but does not tells, what kind of mistake there was. It is supposed to improve this description debugging technics in the later releases of CAHDL.

IV. APPLICATION

The CAHDL was used in Delphi Compiled Units Disassembler [8]. (This tool is aimed to parse *.dcu files and represent their structure in terms of Delphi statements and functionality in terms of assembler code.) The CAHDL was also applied in computer simulator to specify computer architectures. This simulator is observed below.

V. THE SIMULATOR'S STRUCTURE

The simulator has the user interface, framework, virtual machine and parser. The text of architecture specification is located in an outer plaintext file. The emulator works as follows: when user starts the application and selects the hardware to create, the specifications of the hardware are parsed into internal representation then the specified memory image is loaded into memory and the emulation starts.

The structure of the simulator can be represented as follows:

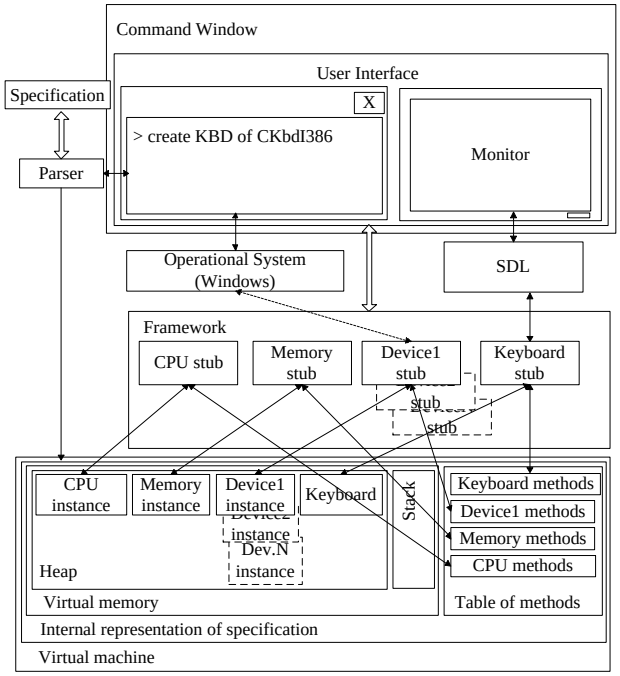


FIGURE 1. SIMULATOR'S CLASS HIERARCHY

The emulator framework is a set of stub computer devices. It complies with the von Neumann architecture. There are processor, memory, I/O devices, etc. Since all the devices are stubs, they must be overridden by child classes in the text of specification. The interaction between the stub objects and their child class instances is based upon the event driven programming. The stub objects call their overridden methods. If a method is absent, then the corresponding stub method is called.

The simulator consists of two windows. The main window is console another is graphical.

A. Console window

Console window is used to adjust the environment. There the user inputs a path to an image file to upload, a path to a file with specification of architecture and a command that starts simulation.

B. Graphical window

Graphical window is created by SDL library [13][13] only when the simulator starts, because the resolution of the screen and other settings are set only when the architecture description is already parsed and interpreted. This window is aimed to simulate the computer's screen. Graphical window may also work in a textual mode. In this mode, all the symbols are represented as small bitmaps. Graphical window handles keyboard buttons messages, mouse movements and mouse button clicks. In order to synchronize two windows an infinite cycle is used rather than a threading mechanism, because it makes the debugging of simulator more convenient.

C. Periphery devices

Simulated periphery devices interact with a user by the standard Windows API functions. For example, when a mouse button is clicked within bounds of SDL simulated window i.e. the simulated screen, a message loop handles the message and dispatches it to a mouse device object. Mouse object interrupts processing of the simulated operational system. When the interruption is handled, the processor goes on processing the operational system, where the cursor's position is already shifted.

VI. INITIALIZATION

The simulator's environment is being initialized only when a text of a computer's architecture description is parsed into a set of objects of different classes. Every syntax construction has its own class. For example, there are classes storing objects of type of class, function, variable, integer, operation, etc. All the objects of these classes are stored hierarchically. For instance, a variable may belong whether to a function or to a class. A function may only belong to a class. A class of an upper level contains all the lexical classes.

The initialization proceeds in two steps. Firstly, the class and function headers are parsed. Secondly function bodies are traversed. Two traversals are needed to be able to refer to a method which is declared in the text below.

VII. SIMULATION

When the simulator's environment is properly configures, there may start the application to do simulation of the required architecture. In order to start simulation, the specialist is provided with the console interface where he has to input a path to an image-file, which would be loaded into the simulator's memory.

When the simulation stars, the program creates a screen object by the use of SDL library (see Simulators structure). Then the main loop is started to handle device messages (mouse movements and clicks, keyboard buttons pressed, etc.).

VIII. CONLUSION

The paper describes the Computer Architecture and Hardware Description Language. The syntax of the language is based upon the C++ and Verilog syntax. Some examples of usage of the language were given. The CAHDL can be used for development of the programs, which require information about computer's architecture, such as compilers, decompilers, emulators, etc.

The CAHDL was tested on two computer architectures I8080 and PDP11. They were specified by 600 and 2000 lines of code respectively. At this time a specification of the real mode of I386 architecture is being developed. The estimated number of lines of the specification code is 5000. Considering that the language is still being improved, the number of lines of code may decrease significantly.

REFERENCES

[1] IEEE Standard Verilog® Hardware Description Language, http://inst.eecs.berkeley.edu/~cs150/fa06/Labs/verilog-ieee.pdf

[2] IEEE Standard VHDL Language Reference Manual, http://edg.uchicago.edu/~tang/VHDLref.pdf

[3] Brusnetzov N.P. Microcomputers. – Moskow: "Nauka", home editing of physical and mathematical literature, 1985. UDC 519.6

[4] Myers G. Advances in computer architecture. 2nd edition. Willey – interscience publication; 1982.

[5] Brumm P. 80386 A Programming and Design Handbook. TAB Professional and Reference Books. 1987.

[6] Tanenbaum A. Structured computer organization. 5th edition. PH PTR; 2006.

[7] Ramsey N., Fernandez M. "Specifying Representation of Machine Instructions", ACM Trans. Program. Lang. Syst. Vol 19, No 3. 1997, pp. 492-524.

[8] DCU32INT - Delphi Compiled Units Disassembler, http://hmelnov.icc.ru/DCU/index.ru.html

[9] Data description language FlexT: Flexible types for description static data, http://hmelnov.icc.ru/FlexT/FLEXT_CSCC99.htm

[10] A.S. Burlakov. "Specification Language of Computer Simulator Virtual Hardvare", Vestnik of Irkutsk State Technical University. Vol. 12, 2014, pp. 12-17

[11] Bison – GNU parser generator, http://www.gnu.org/software/bison/

[12] Donald E. Knuth. "On the Translation of Languages from Left to Right", Information and control 1965, Vol 8, pp. 608-639.

[13] Simple DirectMedia Layer – Homepage, http://www.libsdl.org/index.php

Heterogeneity Impact on MPSoC Platforms Performance

Nikolina Frid, Danko Ivošević and Vlado Sruk
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
nikolina.frid@fer.hr, danko.ivosevic@fer.hr, vlado.sruk@fer.hr

Abstract – Embedded systems have become an integral part of High Performance Computing (HPC) due to their appealing energy and resource consumption characteristics. Required performance goals can be achieved only by deploying the application on a heterogeneous platform. The established approach of designing a custom made FPGA architecture platform targeting particular application is challenged by novel multiprocessor heterogeneous platforms built using existing of-the-shelf embedded CPUs. The challenge is to exploit all the available parallelism and heterogeneity to design a time and cost efficient, but also a reusable solution which will meet the performance goals.
In this paper the heterogeneity of parallel SoC system is evaluated through different processor cores and memory configurations usage examination. The design space exploration undertaken relies on several hypotheses concerning design concepts relations. Since system operating frequency is crucial, but not the only design performance success parameter, the importance of operating processor data path is emphasized as it conducts the application mapping approach. We show that, depending on the different heterogeneous elements configurations used within some multicore SoC solution, the suitability of particular processor cores usage varies on application type, but with achieved performance comparable to application-specific custom generated hardware.

Keywords: Embedded Computer Systems, Heterogeneous Computing, FPGA, High Level Synthesis, Design Space Exploration, Software partitioning and mapping,

I. INTRODUCTION

Broad and versatile application of embedded computer systems has marked the last decade of their usage. Due to the wide variety of applications these systems are used for, their design can pose many challenges. The design requirements for embedded systems involve meeting desired performance goals while observing numerous other constraints ranging from development cost and time, energy efficiency, final product cost to reliability concerns and more [1][2].
Choosing the optimal architecture platform for a particular application often proves to be the biggest design issue. In order to satisfy the above enlisted design requirements, often a heterogeneous multiprocessor platform must be used. Still the design will not benefit from the platform heterogeneity unless the software application is properly tuned to exploit available hardware potential. Common approach is to choose a certain platform based on one or more of the following criteria: price, availability, earlier experience or legacy status, and then manually optimize the application software for the target platform.
The alternative is the Design Space Exploration (DSE), a process to find out near-optimal system architectures for a given application, considering the design constraints and objectives. It involves task partitioning and mapping, architecture and processing element selection, and performance estimation before a hardware prototype is built [3]. The process starts with application specified using a model of computation, description of architecture instance and a set of constraints for the overall system. Those inputs are fed to mapping engine which performs partitioning and mapping of software to hardware. Finally, a performance analysis is conducted to determine whether the output design meets given constraints.
During DSE process, the platform can be built either entirely out of-the shelf components (embedded CPUs, RAM memories etc.) or custom made FPGA components can be used to build the part or even an entire architecture platform targeting particular application. High level synthesis (HLS) [4] denotes a set of methodologies that generate custom hardware units from high level specification in automated manner. The idea is to help designers to build and verify custom piece of hardware in short time by writing algorithmic specification. Usually, the specification language is C or some of its derivate such as C++ or SystemC, and the output is synthesizable RTL level (VHDL, Verilog) circuit description. FPGA is the target implementation device as the most popular prototyping platform with good software support for logic synthesis process. It is suitable for generating dedicated high performance components that will be used as coprocessors for design functionalities selected for acceleration, and during last decade a number of academic and commercial tools appeared in design community and on the market.

II. RELATED WORK

Lately, many research groups have tried to optimize and automate parts or the entire DSE process. The main two points of focus have been: (1) code analysis and partitioning and (2) task to processor mapping.
Early work in this area embraced the approach to take the application, divide it into tasks on a very high level and try to come to best near-optimal (NP-hard problems) task-to-processor mapping scheme in a reasonable

amount of time [5][6]. The scheduling algorithms used were modified versions of integer linear programming.
More recent work has focused on optimizing both the application and the mapping scheme, while automating the most part of this process. The most successful at the time were the Daedalus tool [8], developed at University of Amsterdam and MAPS tool [7] developed at the research group at RWTH Aachen University. While MAPS features a more advanced compiler and focuses on adapting the application to the given platform, Daedalus provides better support for platform customization and is able to produce custom cores in FPGA, which MAPS is not. However, both tools rely on virtual platforms for performance estimation, which are not cycle-accurate by definition and do not reflect the implicit processor data path features. They are also not able to work with a loosely defined platform definition i.e. choose a certain number from a pool of components.

III. Problem Definition & Hypothesis

As stated above, DSE application-to-processor mapping process currently relies on virtual platform simulations for performance estimation or manual intervention of the system designer based on professional experience. However, virtual platform simulations do not take inherent features of a certain platform, like inner data path architecture, into account and thus do not correctly estimate the time for a certain task to execute on an element of a platform. Consequently, any approach relying on this type of performance estimation can potentially fail to exploit the full potential of heterogeneous platforms.
The goal of this research is to explore the extent of impact of platform heterogeneity on mapping efficiency regarding performance goals. The rest of this paper will try to give an answer to whether it is possible to partition the application and map it to heterogeneous platform to achieve performance results comparable to a custom hardware platform built using fully automated HLS approach. The custom hardware platform has been chosen as a reference point because currently it is the most common architecture platform type used for critical tasks that require significant speedup.
At the beginning of this research three hypothesis have been established.
Hypothesis 1: Inherent features of a certain platform architecture significantly affect the performance of execution of specific types of operations. The time needed for a processor to execute a certain operation does not only depend on the operating frequency, but also on the inner data path architecture, which can indeed have a major impact on the execution duration. Consequently, the suitability of two processors with a somewhat similar architecture (e.g. RISC type) can be proved between different types of operations, assuming the same external memory environment.
Hypothesis 2: In order to be able to determine which processor is more suitable for a certain task (in broader sense then operation) it is enough to examine the performance of "basic tests" i.e. execution of operations typical for a certain task (e.g multiplication and trigonometry vs. bitwise operations) in a given memory environment.
Hypothesis 3: By examining the results of processor performance for "basic tests" in a given memory environment and applying efficient task-level parallelization of the application, it is possible to achieve an overall performance comparable to custom built hardware platform using HLS method. It is assumed that basic tests were chosen to accurately represent the prevailing types of operations in each task in the application of interest.
In the rest of this paper the experiments conducted in order to prove these three hypothesis are described.

IV. Experimental Setup & Results

In first part of this section the evaluation platform and test bench are described in detail. Later the results of conducted experiments are presented and discussed.

A. Evaluation platform

All experiments were conducted on the ZC706 evaluation board which provides a hardware environment for developing and evaluating designs targeting the Zynq®-7000 XC7Z045-2FFG900C AP SoC. The ZC706 evaluation board provides features common to many embedded processing systems, including DDR3 SODIMM and component memory (1GB), a four-lane PCI Express® interface, an Ethernet PHY, general purpose I/O, and two UART interfaces.
Two types of RISC processors were used:

1. The ARM Cortex-A9 processor, a popular general purpose choice for low-power or thermally constrained, cost-sensitive devices. The processor is a mature option having been introduced in 2008, and remains a very popular choice in smartphones, digital TV, and both consumer and enterprise applications enabling the Internet of Things. Zynq system integrates two ARM cores.
2. MicroBlaze™, a 32-bit RISC Harvard architecture soft processor core with highly flexible architecture, plus a rich instruction set optimized for embedded applications, delivers the exact processing system needed at the lowest system cost possible. This industry-leader in FPGA-based soft processors features advanced architecture options like AXI or PLB interface, Memory Management Unit (MMU), instruction and data-side cache, configurable pipeline depth, Floating-Point unit (FPU), and much more.

B. *Test bench*

According to the previously stated hypothesis three different test cases have been prepared.

Two test cases were adopted from MiBench [9] and represent "basic" test cases which aim to determine the behaviour of each type of processor in performing different types of computation.

1. CubicTB test solves the third-degree polynomial equation, and serves to measure performance in performing typical mathematical operations (addition, multiplication, basic trigonometry).
2. SqrtTB finds the square root of an integer with all the calculations being the base-two analogue of the square root algorithm learned in grammar school. Bit-wise operations are the prevailing type in this test case with absolutely no multiplications or divisions.

The above two test cases have been executed on twelve different architecture configurations. A single core of the ARM processor was used in the following configuration: operating frequency @ 667MHZ, 32 KB L1 cache and 512 KB L2 cache with two different memory mapping schemes:

- both instructions and data stored in local DDR3 SDRAM memory operating at 533 MHz (*local*),
- instructions stored in local memory and data was stored in shared DDR3 SDRAM operating at 200 MHz (*mixed*).

MicroBlaze processor was used in three different architecture configurations: (1) 3-stage pipeline @ 100 MHz, (2) 5-stage pipeline with hardware multiplier and barrel shifter @ 100 MHz, and (3) 5-stage pipeline with hardware multiplier and barrel shifter @ 200 MHz. The 200 MHz clock frequency is the maximum achievable frequency on PL side of Zynq system and thus is the reason why both shared DDR and MicroBlaze are not able to operate at higher speed. As was the case for the ARM processor, all three configurations of MicroBlaze have been tested with three different memory mapping schemes:

- both the instructions and data stored in local BRAM memory (*local*),
- instructions stored in local BRAM memory, and data in shared DDR memory operating at 200 MHz (*mixed*),
- both instructions and data stored in shared DDR memory operating at 200 MHz (*shared*).

The performance results are reported in TABLE I. and visually compared in Figure 1 and 2.

The results clearly show that the ARM processor greatly outperforms MicroBlaze for CubicTB in all memory settings. On the other hand, SqrtTB results show that

TABLE I. CUBICTB AND SQRTTB EXECUTION TIMES

Target processor	Memory configuration	Execution time [s]	
		CubicTB	***SqrtTB***
ARM [1]	Local*	4,74E-06	1,69E-06
	Mixed**	6,15E-05	2,58E-04
	Shared***	*N/A*	*N/A*
MB100 [2]	Local*	2,21E-03	2,74E-05
	Mixed**	5,20E-03	1,67E-04
	Shared***	7,94E-02	8,02E-04
MB100+ [3]	Local*	5,77E-04	1,54E-05
	Mixed**	2,83E-03	1,61E-04
	Shared***	1,85E-02	4,08E-04
MB200+ [4]	Local*	2,56E-04	8,74E-06
	Mixed**	1,20E-03	9,27E-05
	Shared***	1,06E-02	2,31E-04

[1] ARM® Cortex™-A9 CPU at 667 MHz
[2] XILINX® MicroBlaze™ CPU at 100 MHz
[3] XILINX® MicroBlaze™ CPU at 100 MHz with HW multiplier and barrel shifter
[4] XILINX® MicroBlaze™ CPU at 200 MHz with HW multiplier and barrel shifter
* Both instructions and data are stored in local memory
** Instructions are stored in local memory but data is stored in shared memory
*** Both instructions and data are stored in shared memory

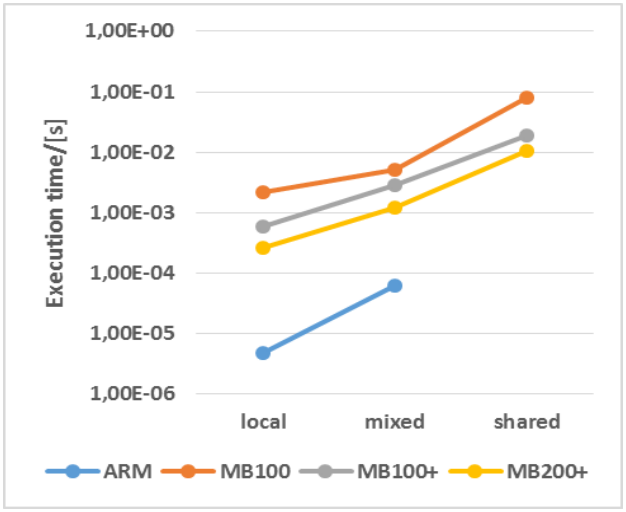


Figure 1. CubicTB performance comparison

Figure 2. SqrtTB performance comparison

MicroBlaze delivers much better performance than ARM for bitwise operations when accessing shared memory is required. Additionally, modifying MicroBlaze by adding two more stages in pipeline, hardware multiplier and a barrel shifter delivers up to four time better performance at the same clock speed.

As for the memory configuration, it can be clearly stated that the best performance is always achieved when everything is stored locally. However, this is often not feasible, especially in multicore systems when processor need to exchange data and thus the mixed memory configuration results should be considered in further discussion.

At this point it can be deduced that ARM would be a better choice for tasks consisting of mathematically more complex operations, while MicroBlaze would be a much better choice for simple and bitwise operations.

For the third test case JPEG [10] compression algorithm was chosen since it consists of several tasks with different prevailing types of operation which should be a good test of previously deduced assumption about processor affinity.

The JPEG compression algorithm has 6 consequential steps with YUV 3-component image model given as input. After the initial extraction of 8x8 pixel block from the original image, every pixel is shifted by -128 so that the whole range of DCT is used. The shift is followed by a DCT transformation which broadens the pixel data range to 12 bits, and scales back to 8 bits per pixel when quantization is performed. Next, data is arranged in a zig-zag manner so that the coefficients with higher value come first which is important for the efficiency of the following step – the Huffman entropy coding. Finally, coded data is formatted and written to a .jpeg file according to Standard. The JPEG compression flow is shown in Figure3.

In the original standard DCT requires 64x64 passes through the pixel block and its core includes 3 multiplications per each step thus taking up more time than all other parts of the JPEG compression together. Through exploration of the specific properties of DCT, special algorithms were devised that help achieve significant speedup. In this paper the AAN algorithm [11] for scaled 1D DCT (8 points) was chosen for implementation and testing. Single pass of this algorithm requires 5 multiplication, 29 additions and 16 2nd complements. Two passes are needed for transformation of one block of pixels. Since it is completely obvious that

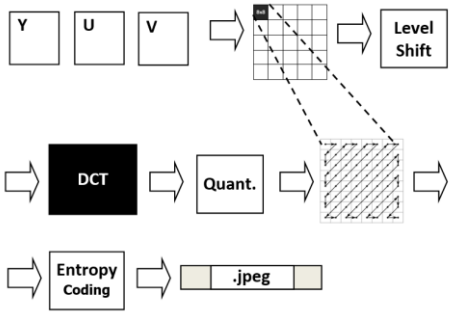


Figure 3. JPEG algorithm flow

MB is no match for ARM for original DCT, this algorithm was chosen as a more interesting alternative which could prove to be suitable enough for MicroBlaze to be able to compete in performance with ARM.

From the previous twelve architecture configurations only three that have previously shown best performance results were chosen:

1. ARM with local memory access,
2. ARM with mixed memory configuration,
3. MicroBlaze at 200 MHz with 5-stage pipeline, hardware multiplier and barrel shifter and mixed memory configuration.

It must be noted that, as previously stated, the first configuration where ARM communicates only with local memory is not a viable solution for any multicore system but it has been put here for comparison purposes.

The obtained performance results for a picture size of 40 blocks are presented in TABLE II. . As expected from the previous test cases, MicroBlaze outperformed the ARM processor in mixed memory mode in *shift*, *Huffman coding* and *create_image* tasks in which simple and bitwise operations prevailed. These results give strong evidence in favour of the first and the second hypothesis.

C. Case Study – JPEG performance optimization

Each of the previous test cases was executed on a single processor in order to obtain performance results for each task and determine the affinity of each type of processor to a certain type of tasks. The ultimate goal of these research is to show how to most efficiently use available platform resources based on processor affinity to certain tasks in order to achieve performance comparable to

TABLE II. JPEG EXECUTION TIMES

Target configuration	**Execution time [s]**						
	Create_blocks	***Shift***	***DCT***	***Zig-zag***	***Huffamn coding***	***Create _image***	***Total***
ARM local	3.42E-04	2.72E-04	1.85E-03	2.13E-04	4.31E-03	1.28E-04	7.11E-03
ARM mixed	2.12E-02	1.90E-02	9.28E-02	1.91E-02	7.55E-01	3.26E-02	9.40E-01
MB200+ mixed	2.02E-02	1.85E-02	9.41E-02	1.85E-02	6.70E-02	3.25E-03	2.22E-01

TABLE III. JPEG ON CUSTOM HARDWARE - EXECUTION TIMES

Target configuration	Execution time [s]						
	Create_blocks*	***Shift***	***DCT***	***Zig-zag***	***Huffamn coding***	***Create_image****	***Total***
solution1	2.12E-02	5.08E-05	1.73E-03	2.16E-05	2.17E-04	3.26E-02	5.59E-02
solution2	2.12E-02	5.08E-05	1.62E-03	2.16E-05	2.17E-04	3.26E-02	5.57E-02
solution3	2.12E-02	5.08E-05	2.32E-05	2.16E-05	2.17E-04	3.26E-02	5.41E-02

*These tasks were executed on ARM

custom made FPGA architecture platform targeting this particular application - JPEG.

For this purpose, a custom made architecture, which would serve as a reference point, has been designed on Zynq platform. It consists of a single core ARM processor and custom made hardware designed using HLS method with the source code previously used and implemented in FPGA. The ARM processor performs the first and the last task in JPEG flow (receiving/sending data from/to memory and pipelining it to FPGA) while the other tasks are completely executed in hardware. In order to ensure equal testing environment, mixed mode memory was used. Three levels of optimization of DCT function were applied and resulted in three solutions:

- solution 1 – no optimization
- solution 2 – first level of DCT pipelined
- solution 3 – DCT transform loop pipelined.

The results are presented in TABLE III. and, as expected, the third solution achieves best performance results.

After obtaining performance results from a custom made architecture, the next step was to devise an application partitioning and mapping scheme on the same Zynq platform, which would deliver results comparable to the previous ones. Zynq platform provides two ARM A9 cores with the surrounding programmable logic which can be used to instantiate a number of soft-core MicroBlaze processors. In this research the maximum possible number of soft processors was targeted. Due to routing and timing limitations the maximum possible number of MicroBlaze instances was ten. By observing the timings obtained for each type of processors in previous experiments for a mixed memory environment, using a combination of LPP[13] and LPT[12] mapping and scheduling heuristics for the same picture size as in the previous cases the following average execution time has been achieved: 9.18E-02. It is important to point out that it is only 1.7 time slower than the best time achieved on custom hardware platform and provides strong arguments in favour of the third hypothesis.

TABLE IV. RESOURCE UTILIZATION

Target configuration	Resource utilization [%]				
	LUT	***FF***	***MUX***	***BRAM***	***DSP***
config1*	16	7	13	70	4
config2**	20	7	1	13	58

* multicore platform
** custom hardware platform

Resource utilization for multicore system is compared with resource utilization in custom hardware system in TABLE IV. The percentage of utilized resources is rather similar, and while custom hardware platform uses more DSP resources, the multicore platform uses more BRAM resources. However, it is important to mention that while custom hardware platform can be used only for the application for which it was designed, while a multicore platform is much more versatile and same configuration (with a different mapping scheme) can be used for a various set of applications thus improving the reusability of the solution.

V. DISCUSSION & FUTURE WORK

The results presented in the previous chapter provide strong evidence in favour of all three hypothesis established at the beginning of this research. The key concept of this approach is the possibility to gain insight into platform architecture specifics by using a basic set of tests and use that knowledge to develop an efficient mapping scheme for a more complex application. What is more, this approach, unlike expert knowledge, can be more easily automatized and integrated into a mapping tool targeting heterogeneous embedded multiprocessor platforms. It is important to note that the custom made hardware cores are often irreplaceable for accelerating parts of the application, especially in hard real-time systems, and in the future research devising a heuristic which will determine when a part of the application must be "moved" to hardware should be considered. Certainly, many more types of architectures and applications need to be studied in order to develop a viable mapping heuristic. Finally, much more emphasis needs to be put on reliability end energy consumption parameters to be able to provide solutions which will fully keep the pace with current demands for embedded computer systems.

REFERENCES

[1] H. Park, H. Oh and S. Ha, "Multiprocessor SoC Design Methods and Tools," *IEEE Signal Processing Magazine*, vol. 26, pp. 72 – 79, November 2009.

[2] R. Leupers, L. Thiele, X. Nie, et al.," Cool MPSoC Programming", in *Proc. of the Design, Automation and Test in Europe Conference (DATE 2010)*, Dresden, Germany, pp. 1488–1493, EDAA, Mar. 2010.

[3] K. Keutzer, S. Malik, A. R. Newton, J. M. Rabaey and A. Sangiovanni-Vincentelli, „System-Level Design: Orthogonalization of Concerns and Platform-Based Design", IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems, Vol. 19, No. 12, December 2000.

[4] P. Coussy, D.D. Gajski, M. Meredith, A. Takach, "An Introduction to High-Level Synthesis," Design & Test of Computers, IEEE , Vol.26, No.4, pp.8,17, July-Aug. 2009.

[5] J. Lin, A. Srivatsay, A. Gerstlauer, and B. L. Evans, "Heterogeneous multiprocessor mapping for real-time streaming systems," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 1605 – 1608.

[6] K. Vivekanandarajah, and S. K. Pilakkat, "Task Mapping in Heterogeneous MPSoCs for System Level Design," in *Proceedings of 13th IEEE International Conference on Engineering of Complex Computer Systems*, Belfast, UK, 2008, pp. 56 – 65.

[7] R. Leupers, J. Castrillon, MPSoC "Programming using the MAPS Compiler", in *Proceedings of the 15th Asia and South Pacific Design Automation Conference (ASP-DAC '10)*, Taipei, Taiwan, pp. 897–902, Jan. 2010,

[8] M. Thompson, H. Nikolov, T. Stefanov et al., "A Framework for Rapid System-level Exploration, Synthesis, and Programming of Multimedia MPSoCs", *Proceedings of the 5th International Symposium on Hardware/Software Codesign CODES+ISSS'07, Salzburg, Austria, 2007*, pp. 9-14.

[9] M. R. Guthaus, J.S. Ringenberg, D. Ernst, et al., "MiBench: A free, commercially representative embedded benchmark suite", in *Proceedings of IEEE 4th Annual Workshop on Workload Characterization"*, Austin, TX, 2001., pp. 3-14.

[10] G. K. Wallace, "The JPEG Still Picture Compression Standard"Communication of the ACM, vol 34., Issue 4, ACM Press, Ne w York, NY, USA, pp.30-44 , 1991.

[11] Arai Y., Agui T., Nakajima M., "A Fast DCT-sQ Scheme for Images", Trans. IEICE, Vol. E71, No.11 pp. 1095-1097, 1988.

[12] Zadrija and V. Sruk, "Mapping algorithms for MPSoC Synthesis" in *Proceedings of MIPRO 2010, 33nd International Convention*, Opatija, Croatia, 2010, pp. 624 - 629.

[13] N. Frid, V. Sruk. "Critical Path Method Based Heuristics for Mapping Application Software onto Heterogeneous MPSoCs", *in Proceedings of MIPRO 2014, 36th International Convention*, Opatija, Croatia, 2014, pp. 1030 – 1034.

Problem-oriented on tasks Data Mining parallel processor

D. Adzinets*, D. Lavnikevich*, M. Lukashevich*, Y. Seitkulov** and M. Tatur***

*Belarusian State University of Informatics and Radioelectronics / Faculty of Computer Systems and Networks, Minsk, Belarus

**L.N.Gumilyov Eurasian National University/ Faculty of Information Technology, Astana, Republic of Kazakhstan

*** "Intellectual Processors Ltd", Minsk, Belarus

adzinets@i-proc.com
lukashevich@bsuir.by
haff@midgard.by
erj@mail.ru
tatur@i-proc.com

In this paper the clustering, classification, regression, and associative search are viewed as formal tasks Data Mining. Objects presented in the form of multi-dimensional numerical vectors of informative features are initial data. For each of the tasks exist many known algorithms of solving, but they have fundamental differences. For example, the first group is based on calculation of the distance between the images in a specific metric, the second group is based on approximation of regression or discriminant function, and others are based on analysis and processing of math graphs. These differences are a cause for very narrow application of known neuro-computers and associative processors. At the same time, the universal parallel processors (multi-core CPU and GPU) have objective limitations in accordance with the Amdahl's and Gustafson's Laws, and aren't optimal for solving Data Mining. We propose an approach that unifies hardware and software implementation of the tasks of regression, classification, clustering and associative search through the unification of different algorithms in a unified mathematical model. This opens the way to create of hardware universal core of computation for all tasks Data Mining and Knowledge Discovery, and hence to the creation of a new type of cognitive or intelligent computers.

INTRODUCTION

Modern intelligent data analysis presents a wide spectrum of mathematical methods and algorithms for classification, clustering, associative search and other kinds of problems including evaluating efficiency of such systems. Some works are even reveal main constraints of these methods applications and introduce possible ways to overcome it [1]. But still practical accomplishments in the field of intelligent systems are quite small: actually implemented systems only scarcely can be called "intelligent" while cost and efforts required are often unacceptable.

Most of applications are not bound to a single formal task of classification, clustering or some other. Instead there is a problem of presenting target task as a composition of formal intelligent data analysis subtasks. This resulting solutions are often represent a straightforward application of one particular method known to developer or weakly predictable mix of different algorithms and heuristics. And very rarely there is a scientifically and technically well-founded composition of methods and algorithms that will allow to achieve balanced and efficient solution for the problem.

This disbalance between theory and practice becomes even more noticeable in development of high performance hardware for intelligent data processing. It is only natural since specialized hardware design preceded by algorithmic preparations and thus requires much more efforts. Typical example of this is fate of neurocomputers and neurochips which were expected to have bright future. But in reality it became apparent that in addition to technical constraints (such as number of connections and simultaneously active neurons) neural model doesn't cover all methods and algorithms needed to solve actual tasks. As a result it looks like neurocomputers do exist but there are no customers ready to use them in practice. The only exception for this is when some electronics manufacturers provide simple DSPs under the name of neurochips since in both cases weighted sum is used as the massively parallel operation.

From the other side there are technological tendencies like evolving multicore CPU, GPU, supercomputers, clusters and distributed processing systems which are in fact unified parallel computational architectures. They create additional rivalry for specialized intelligent processing hardware.

It rises a natural question: what can be presented as an alternative to "hard" neuron, associative, semantic processors from one side and to universal parallel computers from the other? Maybe problem-oriented "intelligent" or "cognitive" computers designed for a wide spectrum of intelligent data analysis problems?

UPPER LEVEL FUNCTIONS AND ARCHITECTURE

From the very beginning we will limit our research to the level of computing technologies that are accessible in our country (and also Russian Federation) and to reasonable price category. This means that we will avoid self-organizing synergetic structures, 3D-chips, nano-,

quantum- and other highly technological types of computers.

The main target of this work is to propose practical approach which will lead to the creation of an effective integrated hardware/software solutions as a base for creating intelligent systems. Ideas presented in this work are developed from authors' previous results (mathematical models of fuzzy classification, semantic processing and their implementation) [2-4] in direction of further unification.

This work proposes original architecture of a problem-oriented machine which lies on top of a usual parallel or specialized computer. Architecture includes next parts: central processor (supervisor), parallel co-processor (hardware acceleration unit) and library of intelligent data processing programs (Fig. 1).

Complex of CPU and co-processor presents well-known and widely used approach, while key feature is library of algorithms. Note: particular details of computing system implementation may differ and in this work are not considered.

Proposed conceptual approach implements principle of upper-level program control. It is understood that actual intelligent data processing task may be presented as a composition of clustering, classification, regression, associative search, semantic processing and other kinds of mathematical problems including statistical ones. For each problem there is a known set of algorithms and most of them are actually enumeration methods. Which means that computational complexity grows in a non-linear way with amount of input data. Each of these algorithms is implemented in form of program optimized for specific hardware implementation of parallel processor. In addition to it library includes methods of evaluating these algorithms' efficiency.

Main intelligent processing host-program runs on CPU and operates through subprograms which present data analysis and auxiliary algorithms from library plus set of commands not included in the library. Fig. 1 shows

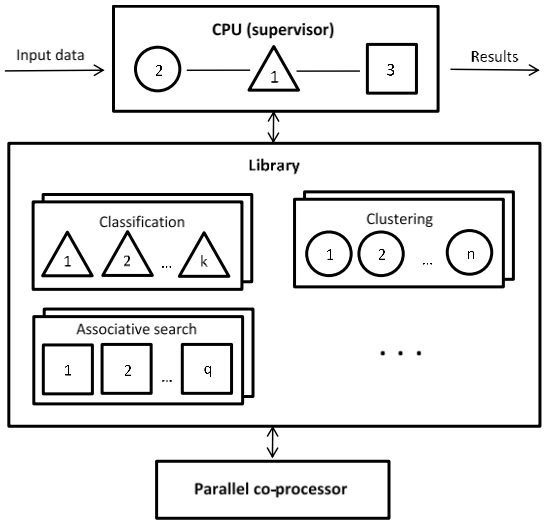


Figure 1. Problem-oriented machine architecture.

simplified example of such program. It consists of clustering algorithm №2, classification algorithm №1 and associative search algorithm №3. This program also implements input data receiving, acceleration on parallel co-processor and delivering computation results.

From functional point of view presented computing system reminds software complexes like Weka [5]. The main difference is in use of parallel co-processor (for example GPU cluster) and algorithms adaptation for specific architecture.

In case of an original co-processor implementation higher degree of architecture-algorithm coherency can be expected. And thus efficiency of whole system is expected to be higher.

MASS CO-PROCESSOR OPERATIONS AND COMPUTATIONAL MODEL

There are three linked problems that should be addressed during design of new parallel co-processor architecture:

- Choosing mass operations functions which will be appropriate for specified set of Data Mining problems;
- Designing processing element (PE) structure and schematics (hardware basis);
- Developing protocol and implementing hardware/software communication between processing units and host processor.

Naturally, the main criteria is unification of library algorithms implementation within framework of some computation model.

Experience in design of several DSP and neurocomputer architectures shows that there is sense in taking weighted sum as the mass operation. This arithmetical operation is used by many Data Mining algorithms [6, 7] and often implemented in hardware. But current level of development in electronic technologies allows to reach mass operations of higher complexity.

For instance distance function can be used in such operation. There are works in this field related to similarity measures for text classification [8] and common metrics in real-valued spaces [6, 9]. They have some differences which should be considered in computation model.

Most often used distance metric is Euclidean distance since it possesses homogeneity in all directions. But from computational point of view this function is really heavy and difficult to implement in hardware. Manhattan and Chebyshev distances are examples of easier and faster metrics. They use only sum and max operations correspondingly. But the same time they are less homogenous than Euclidian metrics. Also there are other operations that may be useful for some fuzzy classification and clustering solutions such as membership function and informative features of critical importance. Some of mentioned features were already tested on generalized identification model [2] and were

implemented in one of the specialized co-processor variants [3].

HARDWARE PLATFORMS EFFECTIVENESS COMPARISON

It is common to understand efficiency as a ration of quality to cost (or other type of resources). In relation to computing it is usually ration of performance to cost of hardware. From the experience in neural computers efficiency evaluation we know, that measurement of the productivity in flops will not be quite correct. Because of it different approaches were suggested, like counting number of modelled synapses (multiplication plus summing), or neurons (same plus threshold function) in time. A number of works in the field by default assume fully parallel functioning of neural networks. They, however, usually overlook one important question in this case: How many neurons can really work simultaneously at each moment of time? Experimental modelling of neural networks using parallel computers shows that Amdahl-Gustafson's law is still effective for parallel implementation of neural networks. I.e. increasing number of PE does not lead to linear growth of computational performance as a whole. It is only natural since parallelization during modeling touches not only parallel architecture but also implemented model.

What kind of efficiency can we expect from problem-oriented cognitive computer? To answer this question let us first point out and comment on parameters which influence efficiency. This includes single task processing time, amount of input data, processing precision and number of processing elements. Naturally, total cost is integral parameter which characterizes machine efficiency but in this work we will not discuss this aspect. Instead we will think of number of processing elements as resources spent on achieving specific computation performance.

Size of input (processed) data is a value defined by a number of informative features, patterns, classes and/or clusters. For simplicity's sake we will discard influence of data representation on amount of calculations.

Computation precision (in our case it is decision making reliability) is a value which is defined through a method and algorithm by dataset representativity, decision making criteria etc. For simplicity's sake we will discard influence of data representation on computation precision.

Number of PE directly characterizes parallelization potential of computation process and indirectly, as was stated above, cost of the whole system.

Computation time of specified task (clustering, classification, associative search or their composition) is defined as interval from the moment of passing input data until result retrieving. This parameter can be measured both in absolute and conventional (model tick) units. In the first case efficiency evaluation will be affected by hardware platform technical characteristics. In the second case – by model time definition. Naturally, performance will be measured not in Giga-, Tera- flops and not in Kilo- or Mega- neurons, but in tasks of clustering/classification per time. The total computation time will consist of parallel co-processor computation time and overhead time costs related to: loading/retrieving data, operations preparation and other code that can't be effectively paralleled.

Mutual influence of mentioned parameters characterizes intelligent computation efficiency. Fig. 2 shows curves of such dependencies for different hardware platforms.

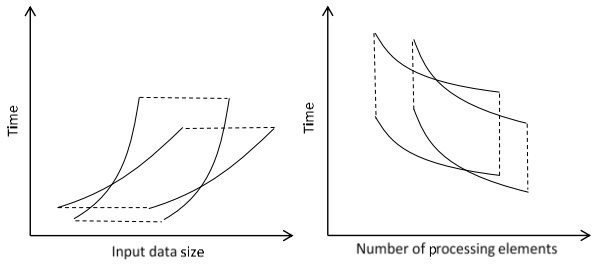


Figure 2. Dependency of task execution time from (a) input data size; (b) number of processing elements

Fig. 2a illustrates expected form of exponential dependency caused by enumerative character of data analysis problems. Actual curve slope may significantly vary depending on quality of algorithms implementation with respect to specific hardware platform. Graph on Fig. 2b demonstrates Amdahl-Gustafson's law according to which performance of parallel machine cannot grow linearly with increasing number of processing elements. But using proposed approach in cognitive computers design, slope of this curve can be significantly lowered. Both graphs consider other parameters (precision, data size and number of processing elements) to be constant.

On both graphs similar curves form range in space where may lay some third curve. Its actual position will be defined by technical characteristics of hardware platform. It is important to mention that curves which describe efficiency of different alternative platforms may intersect and this intersection coordinates can be very useful during computing system design stage.

CONLUSION

This work considers such Data Mining problems as classification, regression, clustering, associative search and others. Input data presented by multidimensional real-valued vectors of informative features. Each problem has a wide set of algorithms with their differences. For instance some are related to calculating distances in defined metrics, others – to approximating discriminant or regression function. All these differences cause narrow application field for known neurocomputers and associative processors.

This work proposes approach for hardware/software solutions unification with regard to Data Mining techniques through bringing different algorithms to single hardware platform. Here platform can be presented by both universal parallel machine (multicore CPU, GPU, supercomputers) and specialized one. This opens path to designing hardware processing core common for a set of Data Mining and Knowledge Discovery problems, and

thus to creating new type of computers. Proposed computation model shows more versatility in comparison to neuron, associative, semantical and some other models since it is oriented on higher level of mathematical tasks representation. Because of it term "cognitive" or "intelligent" is used for defining this type of machine.

Suggested approach allows to perform comparative evaluation of alternative hardware platforms used in actual intelligent systems. For this purpose functional unit should be presented by known Data Mining problem (classification, clustering, associative search and others).

This work has been supported by grant of MES of RK 0314/ГФ3-14-ОТ .Young specialists who take part in this investigation accept the methodical assistance from EU program TEMPUS in framework of project "Centers of Excellence for young RESearchers" No 544137-CERES.

REFERENCES

[1] K. J. Cios, W. Pedrycz, Roman W. Swiniarski, and L. A. Kurgan, Data Mining: A knowledge Discovery Approach. Springer, November 2007.

[2] D. Adzinets, M. Lukashevich, S. Bairak, M. Tatur, "Synthesis and analysis of classifiers based on generalized model of identification", Advances in intelligent and soft computing, vol. 71, pp. 529-536, Springer, 2010.

[3] M. Tatur, "Problem-oriented processors for the solving of classification tasks", Journal of Information, Control and Management Systems, vol. 11, №2, pp. 155-164, 2013.

[4] N. Verenik, F.Girel, Y. Seitkulov, M. Tatur, A. Razhkova, "Cognitive information processing based on a parallel processor", in proceedings to 10th int. conf. on Digital Technologies. Zilinf, Slovakia, pp. 367-371, 2014.

[5] Weka 3 – Data Mining with Open Source Machine Learning Software in Java, [online] 2015, http://www.cs.waikato.ac.nz/ml/weka (Accessed: 3 February 2015).

[6] Charu C. Aggarwal, Data Classification: Algorithms and Applications. Data Mining and Knowledge Discovery Series, 2014.

[7] A. Quinn, A. Stranieri, J. Yearwood, "Classification for accuracy and insight: a weighted sum approach", in proceedings to the 6th Australasian Conference on Data Mining and Analytics, vol. 70, pp. 203-208, 2007.

[8] P. Sowmya, V. Sushma, T. Manasa, Lakshmi, "Different similarity measures for text classification using knn", IOSR Journal of Computer Engineering, vol. 5, №6, pp. 30-36, 2012.

[9] Kilian Q. Weinberger, Lawrence K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification", Journal of Machine Learning Research, vol. 10, pp. 207-224, 2009.

A Model of Speed Profiles for Urban Road Networks Using G-means Clustering

Tomislav Erdelić [1], Silvija Vrbančić [1], Lovro Rožić [1]
[1] University of Zagreb, Faculty of Transport and Traffic Sciences
tomislav.erdelic@fpz.hr

Abstract - In this paper, we present a method for computing speed profiles by processing GPS data collected by vehicles in an urban area. The vehicles were tracked during a five year period on the road network of the capital city of Croatia (Zagreb). Traffic congestions in Croatia appear almost exclusively in urban areas, therefore Zagreb was chosen for this study as it is by far the largest city. The profiles for the roads were computed for each day of the week, where each day was segmented into five minute intervals. As there are no congestions during the night, the free flow speed for the roads was determined by averaging vehicle speeds which were recorded during that time. The speed profiles were clustered using G-means, a variant of the k-means clustering algorithm, to reduce storage space and to categorize roads observed in future research based on their profile. By applying this algorithm we reduced the number of total speed profiles by more than 90%. The profiles were developed as part of the SORDITO project, with the goal of developing algorithms for vehicle route optimizations.

I. INTRODUCTION

With the advent of large scale data gathering and handling, prediction of travel time in urban areas has become an important tool in business as much as in personal life. GPS receivers typically produce a log that contains a time stamp when data were recorded; measured speed, longitude, latitude and heading. Such data have been used to observe various regularities regarding traffic, such as traffic congestions, and has been used in route planning applications. Today, GPS data are widely used as a source of information regarding traffic regularities [1,2] and general mobility patterns [3,4], because of the ubiquitous presence of GPS technology in various mobile devices, including mobile phones. This makes GPS data relatively easy to gather when compared to alternate means [5]. However, data gathered from GPS devices do have drawbacks: there are instances where the signal is weak (e.g. when driving in tunnels or between buildings) and there are inaccuracies which add a certain error to the data [6]. In order to minimize such errors a constant stream of data is needed to correct possible measurement errors. Real time analysis of such data, however, requires a significant and complex infrastructure as the traffic data can come from a number of different vehicle types such as personal, delivery or public transportation vehicles, and many heterogeneous GPS devices which might record data differently.

Coupled with the amount of data needed to get an accurate description of traffic in real time, this presents a serious challenge. On the other hand, historical data are much easier to analyze as they are static, and can be easily preprocessed. In this work, we have developed speed profiles using historical data gathered by taxis and delivery vehicles over a course of five years on the road network of the capital city of Croatia (Zagreb). There are many factors which determine travel time besides just geographical locations, such as road quality, weather conditions, road accidents and roadwork etc. However, although linked to certain predictable parameters, accidents and closures are not predictable on a smaller timescale and are therefore not usable as historical data. In contrast, congestions in urban areas are mostly regular as they are the result of daily migrations of commuters, either going to or returning from their workplace. It is therefore reasonable to assume that speed profiles for the roads with the most traffic during the day are highly dependent not only on the time of day, but also on the day of the week, as most people work only on weekdays and not on weekends. This has prompted us to develop a model taking these differences into account. It was determined that the profile differences were indeed significant when comparing not only workdays to weekends, but all days of the week to each other.

In order for the resulting profiles to be used in practice, one has to consider the space requirements. One possible usage of the resulting profiles is solving the Time Dependent Shortest Path Problem (TDSPP) [7], which requires frequent lookup and calculation of travel times. With that in mind, it is unfeasible to store every single profile for a whole city. Moreover, the TDSPP needs to be solved in real time, possibly on a machine with limited resources available, such as a mobile phone with navigation software. Using the G-means algorithm described in [8], we have developed categories which represent sets of profiles and have reduced the number of needed profiles by over 90%.

II. DATASET PROCESSING

Our main goal was to compute speed profiles for roads on a digital map. A brief overview of the data set and a method of calculating average speed for roads, based on the time of day, is given. At the end of the section, the coverage of roads with computed speed profiles is presented.

A. Dataset

The GPS data (approx. 373 million records) were recorded during a five year period between August 2009 and October 2014 by 1937 tracked vehicles which passed

This work is part of the SORDITO project, which is funded in part by the European Regional Development Fund.

Figure 1. GPS records matched to a part of the city; faster roads are shown in green, slower roads in red

through the road network of Zagreb, which measures approximately 2900 kilometers. GPS receivers were installed by the partner company Mireo Inc. on taxis and delivery vehicles in Zagreb. A small portion of the recorded data is shown in Fig. 1, where each record is depicted by a single dot. Data collection and preprocessing was done by the partner company as well, so along with the common data every record had the following information: a note whether the vehicle was driving or not, corrected longitude and latitude matched to the road on which the vehicle was driving, matched course, speed and distance covered by the vehicle since the previous GPS signal. The recorded data were stored by month and the unique number of the receiver installed on the vehicle. This way, a chronological route of a tracked vehicle could easily be obtained.

The traffic map of Croatia is represented by a network of road-links. A road-link (from this point link) is, in most cases, a segment of road between two intersections or between traffic lights. Every link on the map has a unique ID, longitude and latitude of the first and last point on link and length in meters. For some larger roads such as highways, each direction is represented by a separate link. Longitude and latitude of all historical data were matched to an appropriate link.

Before the profiles could be calculated, some additional processing was needed. Records that were recorded when the vehicle was turned off (approx. 5 % of total stored records), as well as records that did not have longitude and latitude matched to a road (approx. 4 % of total stored records) were discarded. Speeds that were higher than 200 km/h were considered as outliers and were also discarded.

B. Computing speed profiles

The developed method for calculating speed profiles takes into account the time that passes until the vehicle exits the link, or more precisely, the time needed for the vehicle to appear on the next link in its route. The main idea of this approach is the fact that most of the speed drops occur towards the end of the link, on intersections and traffic lights. Taking into account data which appear only on a single link would disregard the time needed to traverse the intersection.

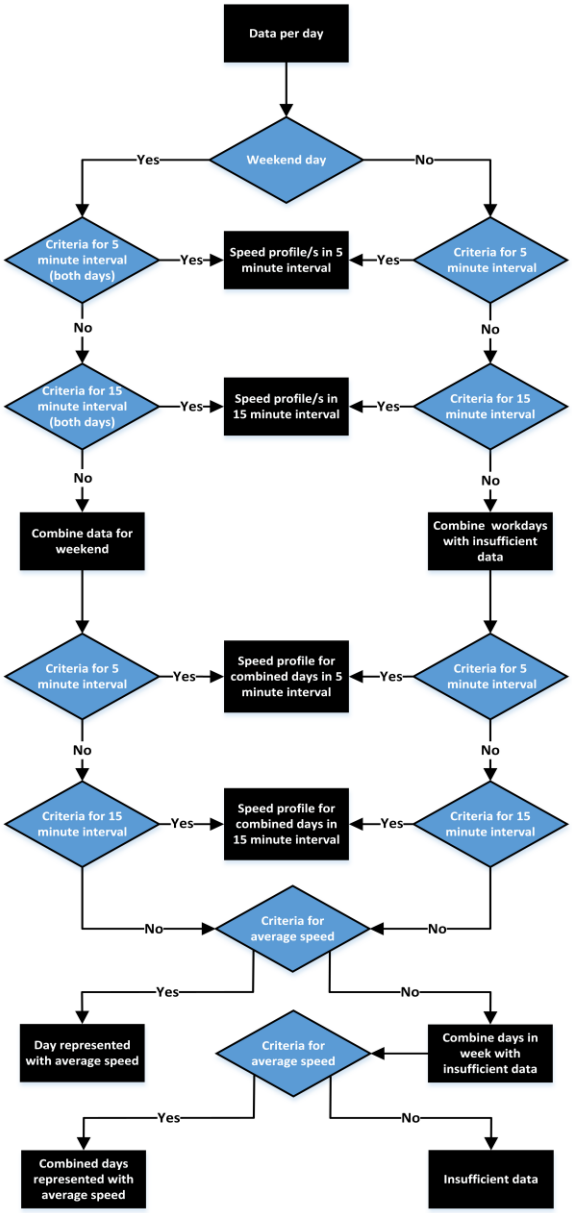


Figure 2. Flow chart depicting steps of data analysis

Free-flow speed is "practically determined as the average speed of passenger cars in conditions of low traffic flow rates" [9]. As there were no explicit data available on free-flow speeds of roads in Zagreb, free-flow speeds were estimated using available nighttime speeds. Nighttime speeds were used as there are no congestions during the night and traffic moves freely. As we do not expect there to be any difference in nighttime speeds for various days of the week, in order to use as much data as possible and maximize precision, the free-flow speeds were calculated using combined nighttime data from every day of the week. In our case, nighttime is the time interval between 22 h and 5:30 h.

Each day of the week was divided into five minute intervals, resulting in a total of 288 intervals per day. For a given link, speeds in nighttime intervals (intervals 0 to 66, and 264 to 287) were all set to the free-flow speed. The speeds for the rest of the intervals were computed based on vehicle trajectories through the link. For a given vehicle,

an entry point on the link was determined and the first signal which appeared outside of the given link (i.e. the next link on the vehicle route) was designated as the exit point. Although matched speed from the GPS record is very precise when the time differences between consecutive data are small, this was not the case with the used historical data as the time gap could vary from a couple of seconds when a vehicle was moving, up to five minutes when the vehicle was stationary. However, using the distances recorded by the GPS receiver between the entry and exit points, and the recorded time difference, it was possible to calculate the speed. The speeds were then included in all intervals between the starting and exiting time, inclusively. At the end of the process speeds in all intervals were averaged.

Although our goal was to calculate speed profiles in five minute intervals, for some links there were insufficient data for each of the 288 intervals to be calculated precisely. The procedure for each workday (as described in Fig. 2) is as follows: an attempt is made to calculate the profiles, trying respectively to generate five or fifteen minute intervals. If these attempts fail some predetermined criteria, the data for each workday which failed are combined and the procedure is again applied to the combined data. Finally, if the procedure fails again, only an average speed is calculated. The same procedure is then repeated for the weekend. If both workday and weekend profiles fail to meet the criteria, the data are combined and average speed is designated as the profile for every day of the week. The criteria used to determine whether there were enough data to calculate a profile was the following: the average number of data per time interval had to be larger than 4, and the percentage of intervals which did not contain at least one datum had to be less than 30%. In order to use the average speed as a profile, the number of data on a given link had to be larger than 7.

The problem with the computed data is the possibility of sudden changes between two adjacent time intervals, which could result in a significant change of predicted travel time. With that in mind, the profiles were smoothed using a cubic smoothing spline [10]. A cubic smoothing spline is a function f which minimizes the following expression:

$$p\sum_{j=1}^{n} w(j)\left|y(j)-f(x(j))\right|^2+(1-p)\int_{x_1}^{x_n}\left|\frac{\partial^2 f(t)}{\partial t^2}\right|^2 dt \quad (1)$$

where x is a vector of the given discrete time sequence and y is a vector of speed profile values. A cubic smoothing spline f at the time interval $x(j)$ is an approximation of the value $y(j)$ for j ranging from 1 to n, where n is the number of discrete time intervals. The expression $|y(j)-f(x(j))|$ represents the error between the speed profile y and smoothed speed profile f. The integral ranges over the smallest interval containing all the entries of x. The value p is the smoothing parameter which varies from 0 to 1, as the smoothing spline changes from the regression line to a natural cubic spline. For uniformly spaced data, the smoothing factor p is chosen as:

$$p=\frac{1}{1+\frac{h^3}{0.6}} \quad (2)$$

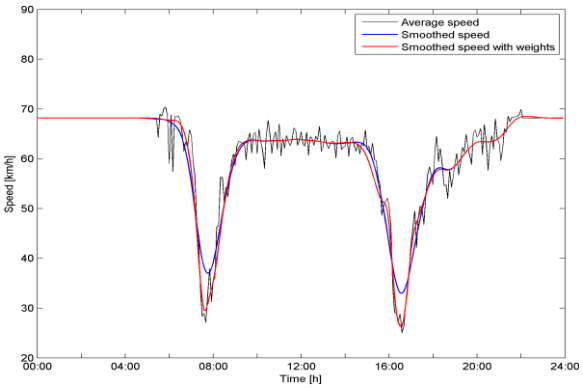


Figure 3. Example of simple and weighted profiles

where h denotes spacing of data sites. The value w denotes weights in the error measure, as a vector of nonnegative entries of the same size as x, with the default value of 1.

In order to apply the smoothing spline to fifteen minute intervals, each interval was interpreted as three separate five minute intervals.

We noticed that the times in which congestion appears are the two hour intervals in the morning (approx. 7:00-9:00 h) and in the afternoon (approx. 15:30-17:30 h). In those cases the described smoothed representation was not a good congestion indicator (Fig. 3). For that reason, additional weights were given to points in intervals between the mentioned times. This process ameliorates the effect of the smoothing spline which tends to average values across the whole domain, and does not affect data which are close to the free-flow values and indicate low congestion traffic.

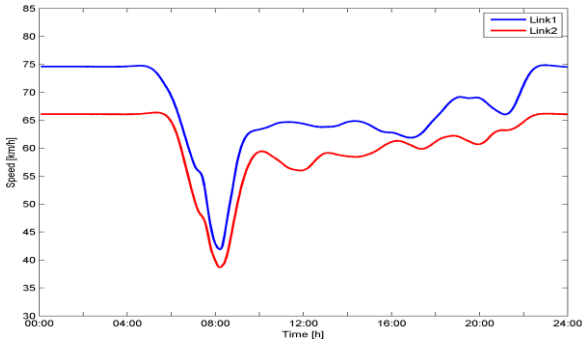


Figure 4. Absolute speed comparison

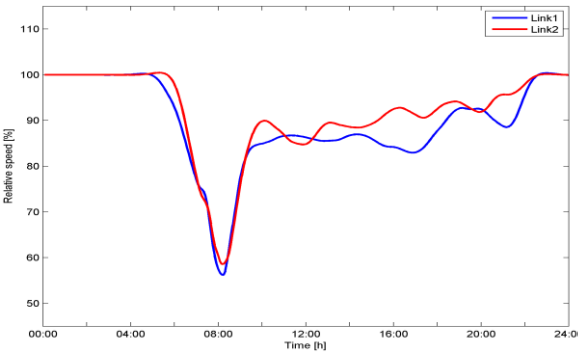


Figure 5. Relative speed (slowdown) comparison

TABLE I. PERCENTAGE OF COVERED URBAN ROADS

	Type of road	Type 1	Type 2	Type 3
1.	Highway	100.00	97.68	86.54
2.	Avenue	97.66	96.03	87.77
3.	Fast road	100.00	96.81	75.15
4.	Main road	99.99	96.34	76.82
5.	Main streets	98.51	74.15	30.70
6.	Local street	98.77	33.79	5.98
7.	Small streets	86.00	3.52	0.80

Finally, in order to compare traffic behavior on links which have different free-flow speeds but show a similar congestion pattern, speed profiles were normalized using free-flow speed. This allowed us to group links based on the displayed traffic pattern and not on absolute speed, as speed varies significantly between links and could therefore impact clustering precision. Fig. 4 and Fig. 5 depict the difference between profiles using absolute and relative speed values, respectively.

C. Profile coverage of Zagreb

Links in Zagreb are organized in seven categories by type of road, e.g. highway, state roads, country roads, local roads, streets and etc. This categorization was provided by Mireo Inc. However, those categories are insufficient for speed profile analysis as it was noticed that there are a lot of different types of speed profiles within each category. Link types can be used for comparison and analysis of behavior between different types of roads after speed profile clustering.

In the road network of Zagreb, GPS signals were matched to total number of 28496 links out of a total of 34111 links present in the area of Zagreb (84% of links). For those links a total number of 40821 speed profiles and 105731 free-flow speeds were computed. For 25818 speed profiles there were insufficient data and no profiles were computed. Table I shows the percentage of covered distance by type of road for three types of comparison. The percentage is calculated as a sum of lengths of found links relative to the total length of links in the corresponding category. Type 1 represents links found in Zagreb represented with free-flow and speed profiles. Type 2 represents links where every workday is represented by a speed profile, and type 3 links that have every day of the week represented by a speed profile.

III. G-MEANS

As it was mentioned in the introduction, the number of profiles of a city network is too large to be used effectively for various calculations. Furthermore, storing profiles which are similar or close to identical is redundant. One would also want to know what kind of traffic congestion could be expected in different times of day considering different locations in the city.

Many algorithms for data size reduction were developed over the years. Since 1967 the k-means algorithm [11] has been used in clustering methods on a daily basis. In general, k-means attempts to generate k clusters by trying to keep the distance between clusters as big as possible while minimizing the distance of points inside clusters from their respective cluster centroids. Although it is widely used, the k-means algorithm requires defining the number of clusters in advance, which could lead to problems such as overfitting and generating empty clusters. For speed profile clustering, the number of centers is not intuitively obvious, and an a priori determination of parameter k could cause results not complying with the actual data.

Hamerly and Elkan [8] suggested the G-means (Gaussian-means) algorithm which is based on k-means, but with the added improvement of not requiring the number of clusters as a parameter. G-means computes a number of clusters based on the normal distribution of data within clusters. More precisely, the algorithm starts with one center and checks whether data around that center follow the normal distribution. If the answer is positive, the data are clustered around one center, otherwise, two new centers are initialized and the procedure continues recursively until the data around all clusters are normally distributed.

For normality testing the Anderson-Darling normality test was used [12]. The Anderson-Darling (A-D) test statistic A^2 is defined as:

$$A^2 = -N - S, \tag{3}$$

where N is the sample size and

$$S = \sum_i^N \frac{2i-1}{N} \left[\ln\left(F(x_i)\right) + \ln\left(1 - F(x_{N+1-i})\right)\right]. \tag{4}$$

The vector $\left(x_{i_i}\right)_{i=1}^{N}$ represents sorted data in ascending order with zero mean and standard deviation equal to 1, and F is the normal cumulative distribution function (CDF). According to [12], the modification A_*^2 is computed as follows:

$$A_*^2 = A^2\left(1 + \frac{0.75}{N} + \frac{2.25}{N^2}\right). \tag{5}$$

The P-value is then estimated based on critical values given in [12].

A. Determining the Number of Clusters

In this subsection the G-means algorithm will be described. The first step in G-means is defining the initial cluster center C and applying the k-means algorithm starting with the computed center C. Two new child centers are then defined as it was described in [8]:

- The main principle component s and eigenvalue λ of the data are computed.
- A new variable m is defined as $m = s\sqrt{2\,\lambda/\pi}$.
- Child centers $c_{1,2}$ are then initialized as

$$c_{1,2} = C \pm m. \tag{6}$$

After initializing two new centers, the k-means algorithm clusters data around center C with parametars $k = 2$ and initial centers $c_{1,2}$. For simplicity, two centers chosen by k-means are also named $c_{1,2}$.

In the next step the hypothesis that the data around C follows the normal distribution is checked by the Anderson-Darling normality test. The data around the center C are labeled as Y. To make the normality test simpler to apply, Y is reduced to one dimension:

- A new variable v is defined, $v = c_1 - c_2$,
- Y is then projected onto v:

$$Y^*: =< Y, v > /||v||^2 \tag{7}$$

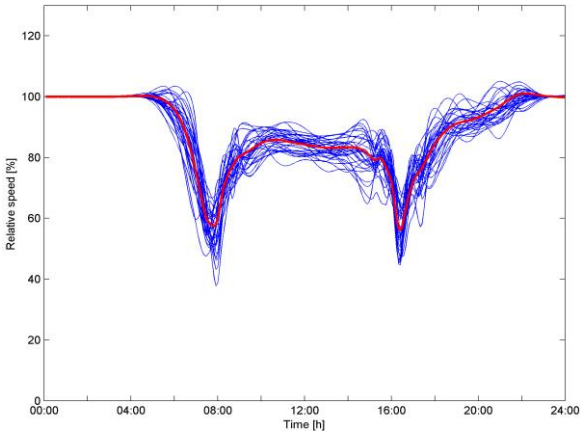


Figure 6. Example of one cluster created with G-means

A couple of transformations on the data are done before the Anderson-Darling test statistic is computed:

- The mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of Y^* are estimated and Y^* is standardized.

$$Y^+ = \frac{Y^* - \hat{\mu}}{\hat{\sigma}}. \quad (8)$$

- Elements in Y^+ are sorted in ascending order as it is required in A-D test.

After A_*^2 is calculated and p-value p is estimated, the conclusion is as follows:

- If $p \geq 0.0001$, the hypothesis is not rejected and the original center C is a true cluster center,
- If $p < 0.0001$, the hypothesis is rejected, C is discarded and elements in Y are clustered in two groups with centers c_1 and c_2.

The described procedure is repeated recursively until elements of each cluster are normally distributed. It should be noticed that multiple normality A-D tests will be preformed, so for m performed independent tests and significance level α there will be $m \times \alpha$ significant results only by chance [13]. Therefore, the significance level α must be preserved. For this reason, a correction of significance level α was done. As it was used in [8], the Bonferroni adjustment [13] on Type I error was applied, and corrected significance level was $\alpha = 0.0001$.

Our goal in this work was efficient and quick speed profile clustering, therefore the performance of the G-means algorithm was not compared to other clustering algorithms that were described as less efficient in the literature. For example, authors in [8] compared k-means which uses Bayesian Information Criterion (BIC) significance level as cluster criterion and they showed that the G-means is considerably better in all aspects.

IV. RESULTS

Using G-means on 40821 calculated speed profiles 770 clusters were created, which is a reduction of 98.11 %. An example of a cluster is shown in Fig. 6. Table II shows cluster size distribution of speed profiles per determined clusters. The coefficient l denotes the number of speed profiles in one cluster, while the number of clusters with respect to l is denoted by *NC*.

TABLE III. CLUSTER SIZE DISTRIBUTION

l	≤ 5	$5 < l \leq 20$	$20 < l \leq 50$	> 50
NC	13	122	299	336

TABLE III. ANALYSIS OF CLUSTERS BY THE TYPE OF URBAN ROADS

CAT	N	N_{90}
1.	201	82
2.	377	204
3.	683	421
4.	666	399

In order to calculate variation inside clusters, an average correlation coefficient between each cluster element and corresponding cluster center was computed. Distribution of average correlation coefficients between cluster elements and corresponding cluster centers is shown in Fig. 7. The averaged correlation coefficient is 0.82.

To compare clusters, Euclidean distances were computed. Average Euclidean distances between cluster elements and corresponding centers are 146.89, which means that average distances between daily intervals inside one cluster is about 0.74. The average Euclidean distance between cluster centers is 379.78.

Table III shows the analysis of the top four types of urban roads presented in Table I. The first column represents the observed type (CAT) of urban road for

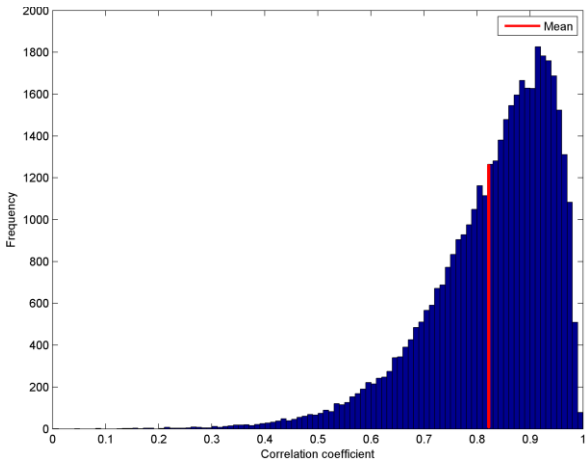


Figure 7. Distribution of correlation coefficients

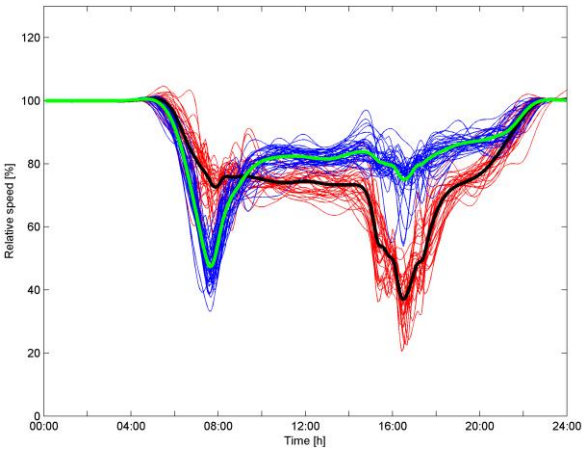


Figure 8. Two clusters representing morning and afternoon speed drops

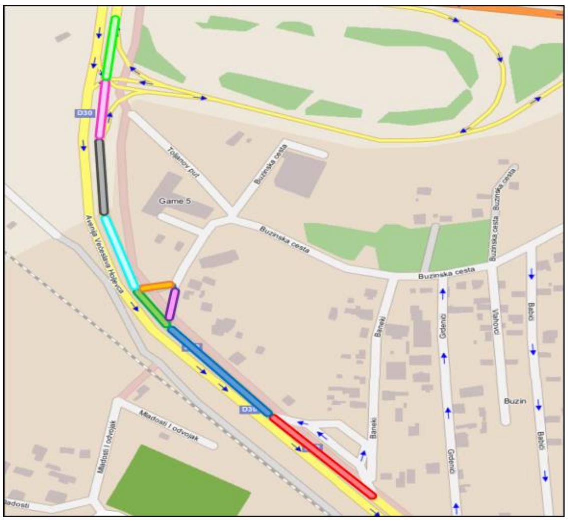


Figure 9. Links found in cluster on Fig. 6

which the speed profile was calculated. The second column represents the number of clusters which contain all profiles from the respective category. For example, all highway profiles are contained in 201 clusters.

The third column represents the number of clusters covering 90% of the profiles of the respective category.

Fig. 6 shows 36 profiles contained in a cluster, where 3 speed profiles correspond to urban roads of type 2 (avenue), 24 for roads of type 3 (main roads), and 11 for roads of type 7 (local streets). This shows that links with different category of urban road can have similar speed profiles. A group of links which are represented by that cluster are shown in Fig. 9. Each link is represented with a different color.

In Fig 8. two clusters are displayed. The differences between those clusters are notable: in one cluster (green line), the morning speed drop is more pronounced while the same is true for the second cluster (black line) for the afternoon period. We can say that those two clusters represent links which have regular daily migrations of commuters.

V. Conclusion

In this work, we have presented a method to compute speed profiles for the road network of a whole city by using GPS data as the source of information. The proposed method of calculating speeds on links takes into account the time loss on intersections. The processed data were used to compute speed profiles for most of the major roads of a large city network. The resulting 40821 profiles were smoothed with weighted cubic smoothing splines in order to better approximate traffic during congestions. The G-means algorithm was used to reduce the size of the dataset in order to make its usage in various applications feasible.

As a possible goal for future research, other factors which influence traffic behavior can be studied, for example, weather conditions, holidays, school break and tourist season.

Acknowledgment

The authors would like to thank our colleague Juraj Fosin for great mentoring and numerous advices regarding the research.

References

[1] A. Bazzani, B. Giorgini, R. Gallotti, L. Giovannini, M. Marchioni, S. Rambaldi, “Towards congestion detection in transportation networks using GPS data”, The Third IEEE International Conference on Social Computing (SocialCom-2011), Boston(USA), 9-11 October 2011, pp. 1455-1459, (2011).

[2] A. Bazzani, B. Giorgini, S. Rambaldi, R. Galloti, L. Giovannini, “Statistical laws in urban mobility from microscopic GPS data in the area of Florence”, Journal of Statistical Mechanics: Theory and Experiment, vol. 05, pp. P05001, (2010).

[3] Y. Zheng, L. Liu, X. Xie, “Learning transportation mode from raw GPS data for geographic applications on the Web”, Proceeding WWW '08 Proceedings of the 17th international conference on World Wide Web, pp. 247-256, (2008).

[4] L. Liao, D.J. Patterson, D. Fox, H. Kautz, “Building Personal Maps from GPS Data“, IJCAI MOO05, (2005).

[5] A. Gühnemann, R.P.Schäfer, K.U. Thiessenhusen, P. Wagner, „Monitoring Traffic and Emissions by Floating Car Data“. Working Paper ITS-WP-04-07. Institute of Transport Studies, The University of Sydney, Australia, (2004).

[6] T. Kos, I. Markezic, J. Pokrajcic, „Effects of multipath reception on GPS positioning performance“ in ELMAR, 52nd International Symposium ELMAR-2010, 15-17 Sept. 2010, pp. 399-402, (2010).

[7] S. E. Dreyfus, “An appraisal of some shortest-path algorithms, Operations Research, 17(3):395-412, (1969).

[8] Greg Hamerly, Charles Elkan, “Learning the *k* in *k*-means”, Department of Computer Science and Engineering University of California, San Diego La Jolla, California 92093-0114, (2004).

[9] I. Lovrić, D. Cvitanović, D. Breški, “Modelling free flow speed on two lane rural highways in Bosnia and Herzegovina”, Promet – Traffic&Transportation, Vol. 26, 2014, No. 2, pp. 121-127, (2014).

[10] G. Wahba, “Spline models for observational data”, SIAM, Philadelphia, (1990).

[11] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297, (1967).

[12] Ralph B. D'Agostino , "Tests for the Normal Distribution", D'Agostino, R.B. and Stephens, M.A. Goodness-of-Fit Techniques. New York, (1986).

[13] J.M. Bland, D.G. Altman, “Statistics notes: Multiple significance tests: the Bonferroni method”, BMJ. 1995;310:170, (1995).

Optimisation of factory floor layout using force-directed graph drawing algorithm

T. Kanduč* and B. Rodič*
* Faculty of Information Studies, Novo mesto, Slovenia
tadej.kanduc@fis.unm.si

Abstract - In this paper we tackle the problem of factory floor layout optimisation in a furniture factory. The goal of optimisation was the reduction of manufacturing costs, specifically reduction of the costs of product transport between machines by reducing the product travel distance and consequently the amount of labour. Limiting factors are the costs of relocating the machines to the optimal positions. We have analysed the current state of the manufacturing system by developing a detailed discrete event simulation (DES) model. The DES model was then used to develop optimisation methods and verify optimisation scenarios. Finally, from output results of the simulation we developed a novel heuristic method, based on force-directed graph drawing algorithms which has produced significantly improved floor layouts.

I. INTRODUCTION

Analysis and optimisation of complex manufacturing systems can be a demanding and time-consuming work since the processes are interwoven and impossible to analyse separately. Processes are usually also too complex to be modelled and optimised with exact mathematical methods. Methods more suited for modelling complex manufacturing systems include discrete event simulation (DES) modelling, which despite its ease of use can provide enough details to understand and analyse the logistic processes in a manufacturing system.

Construction of a DES simulation model requires that the data that describe the manufacturing processes are obtained, extracted and analysed in structured form. Integration of simulation software, auxiliary applications and databases with the model is necessary in order to have one coherent system that operates accurately despite changes in manufacturing processes or variations of scenarios in work orders.

System or process optimisation can be performed in two ways: by changing model parameters, or by changing the model structure. Changing the parameters is often implemented by built-in optimisation methods within the simulation tools, while modifications of model structure are not. In our case the optimisation involved changing the factory layout, which was modelled by a network model, which is part of structure of the model.

Optimisation through modification of model structure can be performed by constructing several versions of the model and input data (i.e. scenarios) and comparing simulation results. To accelerate the development of model versions and scenarios one can construct algorithms that build or modify simulation models according to model input data. This is especially useful in cases of large simulation models and if the model versions are prepared by an optimisation algorithm. Automated model building and modification however requires that the model structure can be modified with an algorithm, without manual interventions. To this end we have developed an automated model construction algorithm, which will be extensively described in our further publications. The focus of this paper is on the novel heuristic optimisation method.

In the paper we present the novel factory layout optimisation method, based on force-directed graph drawing. The method was developed in the course of manufacturing process optimisation project in a furniture company. Our goal was to investigate how the layout of machines on the factory floor affects the efficiency and costs of manufacturing processes. Furthermore, the objective was to develop an optimised layout that would be implemented in real life. In [1] we focused on optimising the total distance the manufactured products need to travel on the floor. In this paper we extend the problem by minimising two dependant criteria. The considered criteria are the total costs of one-time machine relocation and the labour costs in transport of products between the machines as a result of changes in machine relocation.

A. Problem situation

The factory floor contains approximately 140 machines of different types – from simple carpeting workstations to expensive multi-purpose CNC machines. Company catalogue contains more than 30,000 different products and semi-finished products. Every item is manufactured according to the prescribed bill of materials (BOM) and its technical procedure. In BOM, required semi-finished products and materials to manufacture the given product are listed. The technical procedure data describe the sequence of operations in the manufacturing of the given product. The number of operations per finished product ranges from 3 to 20 with the average of 8 operations. The technical procedure data include a group of equivalent machines (GEM) for each operation and standard machine setup and machine operation duration times. The times are listed as single numbers, not as distributions. Compound products are manufactured by joining smaller semi-finished products also produced in the factory.

Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

Production scheduling is based on customer orders and their priorities and performed using the Preactor scheduling system. Our simulation model uses schedule generated by Preactor or prepared by factory planners, without modification. Typically, there are several open customer orders in active production schedule. Products are manufactured in series ranging from approximately ten to several hundred pieces with a typical series containing approximately 30 pieces. For every operation, unfinished products are stored at input pallets or carts next to the machine. Machined products are stored on the output palette or carts until the entire series of products has completed the current operation. The series is then moved to the next location (machine) manually. Therefore the total number of required transport workers depends on the number of simultaneous carts in transition between machines. The assumption in the optimisation process was that shorter transport routes will result in shorter transport worker tasks, increasing the worker availability and potentially reducing the required number of transport workers.

The Preactor database defines GEM for each operation from the technical procedures. Most operations can be performed on several machines, either because there are several identical machines available, or because the operation can be performed on different multi-purpose machines. During the manufacture, a machine is dynamically selected from the GEM for the entire series of a product, according to machine preference and machine availability. Typically a machine that is currently least loaded within the GEM is selected. Hence the manufacturing process can be referred to as "flexible manufacturing". The simulation model has to reflect this flexibility and model the machine groups, group selection and machine selection process for each technical procedure (i.e. each product).

The manufacturing process involves a large set of different products and monthly variations of open orders. Therefore, developing a static simulation model that would cover all available (i.e. 30,000) products and their technical procedures is not realistic. Instead, the simulation model is built automatically from a model template by an algorithm that uses the database of technical procedures and the database of currently open orders to place only the necessary machines in the model. The model reads the technical procedures and BOMs dynamically during the simulation to adjust machine model parameters and assembly of products.

B. Optimisation goal

The costs of manufacturing can be reduced by decreasing the need for labour in the transport of products between machines through better machine placement, i.e. factory layout. However, relocation of a machine is a difficult and costly measure and disrupts the manufacturing process. Therefore it makes economic sense to move a machine only if the relocation will considerably reduce product travel distance and consequently the need for labour. Namely, relocating the machines is associated with additional expenses: moving the machine m_i costs g_i amount of currency. Presumably, good candidates for relocation are machines with high product flows.

Various model statistics are measured during the simulation runs. Most important in this part of the project is the flow of products $\tilde{f}_{ij}$ between a pair of machines m_i, m_j, $i = 1,2, \dots, N$. Flow $\tilde{f}_{ij}$ represents the total amount of volume of products that was directly transported between these two machines. From the product flow $\tilde{f}_{ij}$ we straightforwardly compute cost flow f_{ij}, i.e., cost to move all products between the two machines for distance of 1 m. The distances between the machines are yet not known and are obtained from the optimisation of the factory floor layout. To reduce the overall costs we need to solve the following optimisation problem

$$\min_{\{p_1,p_2,\dots,p_N\}} \left(\sum_{\substack{i,j=1 \\ i \neq j}}^{N} f_{ij} \cdot d(p_i, p_j) + \sum_{i=1}^{N} g_i \right), \qquad (1)$$

where p_i represents position of machine m_i and d is a distance functional. If we neglect the costs g_i and restrict the positions p_i to a predefined grid, the problem simplifies to well-known quadratic assignment problem (QAP). The latter is NP-hard optimisation problem and exact optimisation methods are successful only for smaller number of machines, usually around $N < 30$. In our case, where $N = 140$, the exact methods are not feasible and we need to apply heuristic methods instead, which in practice return a near optimal solution.

C. Previous research (review of literature)

Simulation is commonly used for the evaluation of scenarios [2], [3], [4]. However, the models developed with the visual interactive modelling method (VIM) are usually manually constructed through careful analysis of the real-life system and communication with process owners. Automated model development is more common with methods that allow easier and more standardized formal description of models, e.g. Petri nets [5], [6]. Automation of model construction and adaptation can importantly facilitate the development of models of complex systems [7], [8] and generation of simulation scenarios.

Several papers deal with factory layout optimisation, with paper [9] stating that multiproduct enterprises requires a new generation of factory layouts that are flexible, modular, and easy to reconfigure. Evolutionary optimisation methods are often proposed due to problem complexity [10]. Layout optimisation problem is identified as hard combinatorial optimisation problem and the simulated annealing meta-heuristic resolution approach is proposed to solve to problem [11]. A novel particle swarm optimization method is proposed by [12] for intelligent design of an unconstrained layout in flexible manufacturing systems.

Factory layout design optimisation is further discussed in [13], [14], [15]. Authors [13] propose a new facility layout design model to optimise material handling costs. Sources [14] and [15] propose genetic algorithm based solutions to respond to the changes in product design, mix and volume in a continuously evolving work environment.

The layout optimisation problem is similar to well-known quadratic assignment problem (QAP) [16], [17]. But the problem in its original form does not consider including fixed expenses to move the facilities (machines)

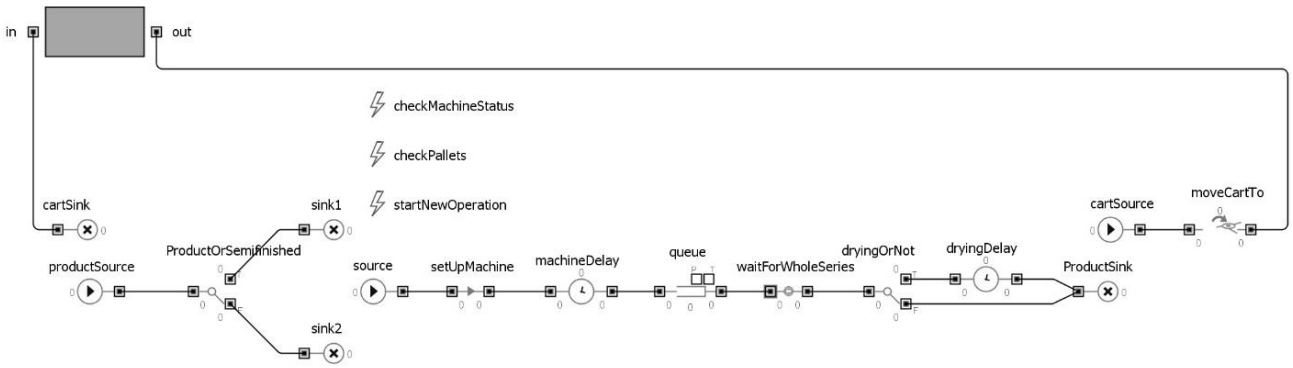


Figure 1. Schematic of a single machine in DES model.

to new locations as it is in our case. Hence QAP algorithms in the original form cannot be used to solve our optimisation problem.

There are numerous methods for graph drawing and class of force-directed methods are one of the most commonly used in practice due to their simplicity and visually appealing representation of the graphs (see [18], [19], [20] and references therein). One of the earliest and still commonly used method is from the authors Fruchterman and Reingold [20].

II. METHODOLOGY

We have used Discrete Event Simulation methodology to develop a simulation model that captures all of the important features of manufacturing processes. The purpose of the model is verification of new manually or algorithmically generated floor layouts. Optimisation of floor layout is conducted in cooperation with experienced manufacturing planners, managers and other experts within the company, and is greatly facilitated by our novel heuristic optimisation method that is employed to generate new layout scenarios, i.e. to search for the optimal layout within a large set of possible layouts.

A. Machine model

Every machine is modelled as a machine block in Anylogic as shown in Fig. 1. On the input of the block, carts filled with products enter the system. Products are sorted according to their type at *productOrSemifinished*. The corresponding sinks, *sink1* and *sink2*, monitor products on input pallets. Once the product is chosen for operation, it is injected at *source*. Blocks *setUpMachine* and *machineDelay* are standard service blocks. Block *waitForWholeSeries* plays a role of output pallet. Products wait there until complete series of products is finished. Some products need to wait at *dryingDelay* according to the technical procedure (paint, varnish drying, etc.). Filled carts are injected at *cartSource* and moved to the next location at *moveCartTo*. Output of the main machine block sends a cart to the input of the next machine.

B. Optimisation problem

In this section we describe the problem of finding the factory floor that minimises the overall costs of machine relocations and costs of workers that move carts between the machines. The problem is to find the optimal mathematical network, in which nodes of the network represent the machines on the factory floor. Weighted edges between the nodes represent transactions between the machines. Additional edges connect the sought network with the fixed old machine positions. Real routes on the floor between the machines are neglected in this case, since it considerably complicates the optimisation problem. The optimisation method should only propose a basic outline of the layout, since the final layout needs to be further tuned by the company experts to meet other less precise criteria.

Factory floor is described as a region Ω in the plane $\mathbb{R}^2$. We will simplify the problem by restricting Ω to the rectangular shape,

$$\Omega = \{(x,y) \in \mathbb{R}^2; x_{min} \le x \le x_{max}, \; y_{min} \le y \le y_{max}\}, \tag{2}$$

where $x_{min}, x_{max}, y_{min}, y_{max}$ represent boundaries of the rectangular factory floor.

Let us denote machines by $m_i,\ i = 1,2,\dots,N$. Position of the machine m_i is described by

$$p_i = \{(x_i, y_i) \in \mathbb{R}^2\}. \tag{3}$$

Each machine takes certain amount of space which can be conveniently described by a metric rectangular-like ball $B_{r_i}(p_i)$ with radius r_i and centre p_i in ∞-norm L_∞,

$$B_{r_i}(p_i) = \{(x,y) \in \mathbb{R}^2; d_\infty((x,y), p_i)\} = \max\{|x - x_i|, |y - y_i|\} < r_i\}. \tag{4}$$

Let p_i' be old position (before relocating) of machine p_i. For every pair of machines m_i and m_j, $i,j = 1,2,\dots,N$, we obtain overall costs $f_{ij} \ge 0$ to move the products between the machines for 1 m. The quantity is obtained as a result of the simulation of the manufacturing processes. The cost is proportional to the amount of products that is transferred between the machines. The cost is also proportional to the considered observed time interval. In practice due to high costs of relocating the machine, the observed time interval for workers' costs should be at least a year.

Manhattan distance d_M between two points p_i, p_j is defined as

$$d_M(p_i, p_j) := |x_i - x_j| + |y_i - y_j|. \quad (5)$$

Manhattan distance is a good approximation of actual route distance between the machines on the factory floor.

Relocation costs for machine m_i from position p_i' to p_i is defined as a value $g_i > 0$ if $d_M(p_i, p_i') \neq 0$ and zero cost otherwise. Costs g_i were obtained from company planners who consider several criteria such as switching on and off the machine, construction work, resupplying costs, cleaning, etc.

The optimisation problem of minimising the total cost is described as

$$\min_{\{p_1, p_2, \ldots, p_N\}} \left(\sum_{\substack{i,j=1 \\ i \neq j}}^{N} f_{ij} \cdot d_M(p_i, p_j) + \sum_{i=1}^{N} g_i(p_i, p_i') \right), \quad (6)$$

where positions p_i must satisfy the conditions

$$B_{r_i}(p_i) \cap B_{r_j}(p_j) = \emptyset \quad (7)$$

for every $i \neq j$ and

$$B_{r_i}(p_i) \subset \Omega \quad (8)$$

for every $i = 1,2, \ldots, N$.

The first conditions states that the regions of machines must not intersect between each other and the second that every machines must lie entirely on the factory floor.

III. RESULTS

The main result of our project described in this paper is the novel force-directed graph drawing optimisation heuristic method that generates new factory floor layouts by minimising overall costs of repositioning the machines and costs of workers that move the carts between the machines. Costs for machine relocation were estimated by the company planners. Workers' costs depend on three parameters: manufacturing process, observed time interval and distances between the machines. The first two parameters are fixed in our optimisation method. The distances are uniquely determined from machine positions. For predefined manufacturing processes we obtain flows of products between the machines. From this data we can directly obtain workers' costs f_{ij} to move the products for 1 m. Taking into account f_{ij}, distances between the machines $d_M(p_i, p_j)$ and relocation costs g_i we obtain the overall transport costs of the manufacturing processes.

1) Force-directed graph drawing algorithm

The heuristic optimisation algorithm for assigning positions p_i to machines m_i is based on force-directed graph drawing methods. Initial position of the machines is chosen randomly but machines with greater product traffic are added to the system earlier. The machines are positioned according to the applied forces in the system and converge to configuration with a local minimum of the overall energy. After the systems converges or the maximum number of iterations is reached, the simulation is finished.

The optimisation was implemented in the system dynamics (SD) methodology in Anylogic software. The method is an extension of the original method that was used recently in QAP of minimising the total distance of machines [1].

Every machine is presented as a node on a plane. To every node n_i we prescribe the corresponding repelling force F_{ij} to all other nodes n_j,

$$F_{ij} = H_{ij}(||p_j - p_i||_\infty) \cdot \frac{p_j - p_i}{||p_j - p_i||_\infty}, \quad (9)$$

where H_{ij} is a positive monotonically decreasing function. For our algorithm we used

$$H_{ij}(r) = \begin{cases} c_i \cdot (r_i - r)^2, & r \leq r_i \\ 0, & r > r_i \end{cases}, \quad (10)$$

where c_i and r_i are distance influence parameters. Repulsive forces keep the nodes away from each other since we want sufficient space between the machines.

For every pair of nodes n_i, n_j we define a weighted edge e_{ij} with weight f_{ij}, i.e., cost to move all products between m_i and m_j for 1 m. Attractive forces between the nodes are defined as

$$G_{ij} = -f_{ij} \cdot I_{ij}(||p_j - p_i||_1) \cdot \frac{p_j - p_i}{||p_j - p_i||_1} \quad (11)$$

where I_{ij} is a positive monotonically increasing function. In our case, I_{ij} is defined as $I_{ij}(r) = d_M(p_i, p_j)$. Attractive forces move the nodes with large edge weights closer to each other.

Attractive force representing machine m_i relocation cost is defined as

$$J_i = \begin{cases} g_i \cdot d_M(p_i, p_i'), & d_M(p_i, p_i') < 5 \\ g_i, & d_M(p_i, p_i') \geq 5 \end{cases}. \quad (12)$$

Note that J_i needs to be a continuous function of the distance, otherwise the optimisation method would not converge to a stable node configuration. Distance of 5 m is a threshold at which moving the machine is still considered.

To keep the nodes inside the prescribed location Ω, we also need to define forces that pull the nodes back to the interior if they are outside the prescribed region Ω,

$$K_i = \begin{cases} 0, & p_i \in \Omega \\ dist(p_i, \Omega), & p_i \notin \Omega \end{cases}, \quad (13)$$

and $dist(\cdot, \cdot)$ is a function measuring the distance between objects.

Screenshot of the running algorithm implemented in Anylogic is shown in Fig. Machines (nodes) are presented as circles on the rectangular factory floor.

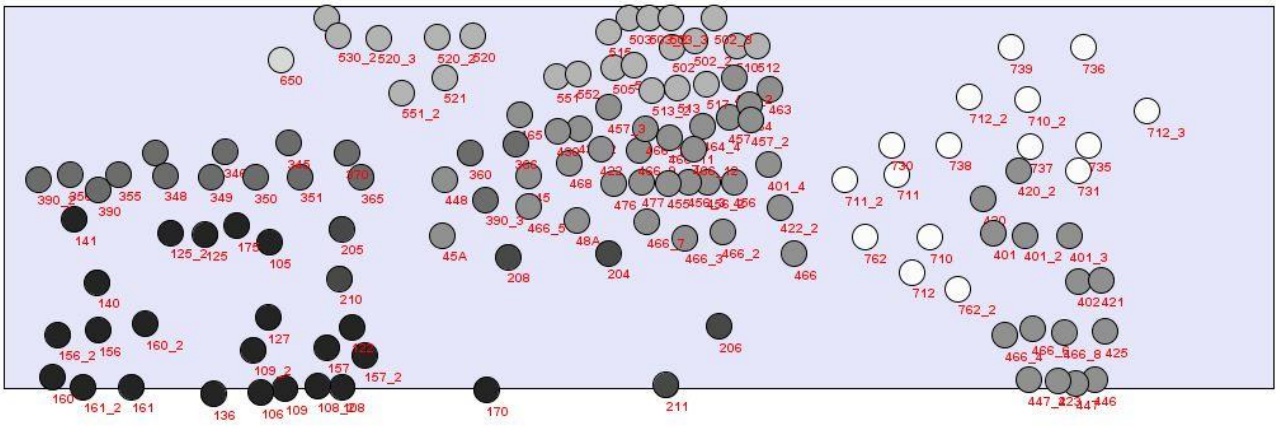


Figure 2. Screenshot of the optimisation algorithm in Anylogic.

For our furniture factory floor layout we tested the algorithm for two observed time intervals: one year and three years. The distances between the machines were optimised by the algorithm to obtain the lowest overall costs. The results of optimisation are presented in Table 1. For both observed time intervals we reduced overall costs of around 7 %.

TABLE I. OVERALL COSTS AND NUMBER OF MOVED MACHINES FOR DIFFERENT FACTORY FLOOR LAYOUTS

Time interval	Layout	Cost	Number of moved machines
1 year	current	50,800 €	0
1 year	optimal	47,400 €	12
3 years	current	149,400 €	0
3 years	optimal	139,700 €	32

Further important results of the project are the automated model construction algorithm and the discrete event simulation model in Anylogic. The model is integrated with a manufacturing scheduling system database and is used for in-depth analysis of the manufacturing process and verification of optimisation scenarios.

IV. CONCLUSION

In this phase of the project the deliverables include the novel force-directed graph drawing optimisation heuristic method, the automated model construction algorithm and the integrated detailed simulation model in Anylogic that communicates with external database files. The model captures all important features of the manufacturing process and servers as an indispensable tool for in-depth analysis and extracting additional data for further optimisation.

The described force-directed graph drawing optimisation heuristic method has been used to generate several new factory layouts which have been validated by the factory planners. The new layouts were significantly better than the current layout and manually produced layouts prepared by the factory planners. The optimisation method allows the company to significantly reduce the product transport costs by decreasing the number of workers performing in-factory transport of products. The reduction considers the costs of machine relocation, i.e. only the most relevant machines are moved.

The factory layouts will be further refined and implemented in the course of ongoing micro-logistic optimisation project. During this project we will further develop the factory model by modelling scenarios, including automatisation of machine servicing using robots and new automated transport system that uses automated guided vehicles (AGVs).

REFERENCES

[1] T. Kanduč, and B. Rodič, "Optimisation of workshop layout in a furniture development company," Croatian operational research review, in press.

[2] M. Kljajić, I. Bernik, and A. Škraba, "Simulation approach to decision assessment in enterprises," Simulation, vol. 75, pp. 199-210, 2000.

[3] R. S. Edis, B. Kahraman, O. U. Araz and M. K. Özfirat, "A facility layout problem in a marble factory via simulation," Mathematical and Computational Applications, vol. 16, pp. 97-104, 2011.

[4] P. Tearwattanarattikal, S. Namphacharoen and C. Chamrasporn, "Using ProModel as a simulation tools to assist plant layout design and planning: Case study plastic packaging factory," Songklanakarin Journal of Science and Technology, vol. 30, pp. 117-123, 2008.

[5] R. Conner, "Automated Petri net modeling of military operations," in IEEE Proceedings of the IEEE 1990 National Aerospace and Electronics Conference - NAECON 1990, vol. 2, Dayton, Ohio, USA, 1990, pp. 624-627.

[6] D. Gradišar, and G. Mušič, "Automated Petri-Net Modelling for Batch Production Scheduling," in Petri Nets - Manufacturing and Computer Science, Pawel Pawlewski, Eds. InTech, 2012, pp. 3-26. http://dx.doi.org/10.5772/48467

[7] A. D. Lattner, T. Bogon, Y. Lorion, and I. J. Timm, "A knowledge-based approach to automated simulation model adaptation," In Proceedings of the 2010 Spring Simulation Multiconference (SpringSim '10). Society for Computer Simulation International, San Diego, CA, USA, 2010, Article 153.

[8] R. Kannan, and H. Santhi, "Automated construction layout and simulation of concrete formwork systems using building information modeling," In Proceedings of The 4th International Conference of Euro Asia Civil Engineering Forum 2013 (EACEF 2013), eds: Djwantoro Hardjito & Antoni, Eds. National University of Singapore, 26-27 June 2013, pp C7-C12.

[9] S. Benjaafar, S. S. Heragu, and S. A. Irani, "Next generation factory layouts: Research challenges and recent progress," Interfaces, vol. 32, pp. 58-76, 2002.

[10] A. Sadrzadeh, "A genetic algorithm with the heuristic procedure to solve the multi-line layout problem," Computers and Industrial Engineering, vol. 62, pp. 1055-1064, 2012.

[11] G. Moslemipour, and T. S. Lee, "Intelligent design of a dynamic machine layout in uncertain environment of flexible manufacturing systems," Journal of Intelligent Manufacturing, vol. 23, pp. 1849-1860, 2012.

[12] M. Ficko, S. Brezovnik, S. Klancnik, J. Balic, M. Brezocnik, and I. Pahole, "Intelligent design of an unconstrained layout for a flexible manufacturing system," Neurocomputing, vol. 73, pp. 639-647, 2012.

[13] K. K. Krishnan, S. H. Cheraghi, and C. N. Nayak, "Facility layout design for multiple production scenarios in a dynamic environment" International Journal of Industrial and Systems Engineering, vol. 3, pp. 105-133, 2008.

[14] J. S. Kochhar, and S. S. Heragu, "Facility layout design in a changing environment," International Journal of Production Research, vol. 37, pp. 2429-2446, 1999.

[15] M. Enea, G, Galante, and E. Panascia, "The facility layout problem approached using a fuzzy model and a genetic search," Journal of Intelligent Manufacturing, vol. 16, pp. 303-316, 2005.

[16] T. Koopmans, and M. Beckmann, "Assignment problems and the location of economic activities," Econometrica, vol. 25, pp. 53–76, 1957.

[17] S. Sahni, and T. Gonzalez, "P-complete approximation problems," Journal of the ACM, vol. 23, pp. 555–565, 1976.

[18] S. G. Kobourov, "Force-directed drawing algorithms", in Handbook of Graph Drawing and Visualization, R. Tamassia, Eds. CRC Press, 2013, pp. 383–408.

[19] M. J. Bannister, D. Eppstein, M. T. Goodrich, and L. Trott, "Force-directed graph drawing using social gravity and scaling," Lecture Notes in Computer Science 7704, 2013.

[20] T. Fruchterman, and E. Reingold, "Graph drawing by force-directed placement," Software-Practice and Experience, vol. 21, pp. 1129–1164, 1991.

Review on Unit Commitment under Uncertainty Approaches

Kristina Jurković, Hrvoje Pandžić and Igor Kuzle
University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{kristina.jurkovic, hrvoje.pandzic, igor.kuzle} @ fer.hr

Abstract – Wind power has already become an important renewable energy resource in many regions of the world. Because of its variability and uncertainty, integration of wind power presents a challenge that, if not adequately addressed, can jeopardize the operational reliability of a power system. Generally, generation unit commitment decisions are made once a day, i.e., the commitment decisions are made 24 or more hours ahead of the actual operation. Taking into account the uncertainty of wind power prediction, these decisions need to provide sufficient flexibility at a minimum price. This paper describes the current practice and analyzes unit commitment formulations available in literature highlighting their advantages and shortcomings.

I. INTRODUCTION

The primary concern in operating an electrical power system is to meet the demand for electricity at all times and under different conditions depending on the season, the climate, and the weather.

Modern power systems are supposed to accommodate large total capacity of distributed, volatile generation, as well as large-scale price responsive demand and electric vehicles which dramatically increases both supply and demand uncertainty [1]-[3]. Because of its variability and uncertainty, wind generation impacts power system operation and can potentially jeopardize its reliability. To deal with the larger uncertainty on the net load (the difference between electricity demand and the output of non-dispatchable generation), power system operators are increasing the reserve margins, thus increasing the regulation cost [4].

In order to minimize the operating cost of non-dispatchable resources, it is essential to derive a computationally effective approach to optimally select the units and their output level to preserve the operational reliability of the system. Unit commitment (UC), one of the most critical decision processes, is an optimization problem that generates the outputs of all the generators in a way that minimizes the system-wide fuel cost. Features included in most modern unit commitment models include generator minimum and maximum generation limits, ramping limits, minimum up and down time constraints, time-dependant start up costs and transmission capacity limits [5]-[8].

The work of the authors is a part of the FENISG – Flexible Energy Nodes in Low Carbon Smart Grid funded by Croatian Science Foundation under project grant No. 7766

During the normal operation, system operator dispatches the committed generation resources to satisfy the actual demand and reliability requirements. In the event of a significant deviation between the actual and the expected system condition, system operator needs to take corrective actions, such as committing expensive fast-start generators, voltage regulation or load shedding, to maintain system security. The main causes of the unexpected events come from the uncertainties associated with the load forecast error, changes of system interchange schedule, and unexpected transmission and generation outages. [9]

Deterministic UC formulation is a traditional solution in which the net load is modeled using a single forecast for each wind farm output and the associated uncertainty is handled using ad-hoc rules, i.e., the generating units are committed to meet the deterministic forecast and the uncertainty is handled by imposing reserve requirements [10]-[13]. Such an approach is easy to implement in practice, but the ad-hoc rules do not necessarily adequately account for this uncertainty. Namely, committing extra generation resources for reserve is economically inefficient, while the power system may still suffer from capacity inadequacy in case of a significant deviation between real-time and expected net load. There is a lot of research on optimizing the reserve requirements based on deterministic criteria [14]-[17]. In [14] a new technique to determine the SR requirements at each period of the optimization horizon is proposed using a cost/benefit analysis. Similarly, in [15] the cost of interruptions is considered when optimizing the scheduling of spinning reserve. In [16] a probabilistic analysis of the reserve requirements is taken into account. The authors of [17] show that reserve requirements cannot be specified a priori without sacrificing the optimality.

A more rigorous approach is incorporating uncertainty in the unit commitment model itself, which is the focus of this review paper. Section II describes stochastic unit UC, Section III robust UC formulation, while Section IV describes interval UC formulation. Section V describes some recent advancement in hybrid UC models that combine the aforementioned formulations. Conclusions are duly drawn in Section VI.

II. STOCHASTIC UNIT COMMITMENT

Stochastic UC is based on probabilistic scenarios. A finite set of scenarios is generated and assigned weight in proportion to their likelihood. Stochastic UC is

formulated as a two-stage problem that determines the generation schedule that minimizes the expected cost over all of the scenarios respecting their probabilities. The commitment decisions are unique over all the scenarios, while dispatch decisions are scenario dependent. Including a large number of scenarios in the model requires computationally demanding simulations. Computational burden of the stochastic UC is dramatically increased with the time horizon as well, which is visualized in Figure 1. Thus, scenario reduction techniques that eliminate scenarios with very low probabilities and aggregate close scenarios are developed [18]. Similar scenarios get aggregated based on a particular metrics, such as their probability, hourly magnitude, or the resulting cost [19]. In [19] the authors did a comparison of scenario reduction techniques for the stochastic unit commitment. A clustering method, k-means is used to partition a given set of scenarios into a given number of clusters. The cluster features similar scenarios and is represented by a scenario with the lowest probability distance from the centroid. The centroid is an average pattern of all the scenarios from the cluster [19] [28]. The forward scenario selection and backward scenario reduction approaches are based on minimizing the Kantorovich distance between the scenarios in the original and in the reduced set [19] [29]. The forward scenario selection approach is used to construct a reduced set containing a desired number of scenarios by iteratively adding a scenario from the original set. Similarly, the backward reduction approach gives a reduced set by iteratively eliminating one scenario from the original set until the desired number of scenarios remains. Importance-sampling scenario reduction technique is used to select the scenarios that best represent the monetary impact of uncertainty on the operating cost [19] [30]. However, insufficient number of scenarios reduces accuracy of the solution and increases its cost. The eliminated scenarios may have great impact on the system, so stochastic UC formulations provide only probabilistic guarantees to the system reliability. It is important to note that the stochastic UC solution contains a certain amount of unhedged uncertainty, i.e. load shedding or wind curtailment in the most extreme scenarios might be cheaper than modifying the schedule to serve the net load over all the scenarios. Due to increased uncertainty in later hours of the time horizon, the amount of unhedged uncertainty increases over time [20]. In order to secure the robustness of the solution, a large set of scenarios is required, which is computationally demanding. Problems to be considered:

- Possible loss of information
- Disregarding the scenarios with comparatively low probability but great impact on the power system
- Availability of data
- Difficulties to identify accurate probability distribution of the uncertainty

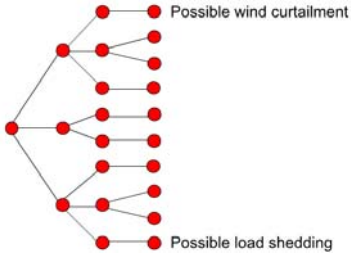


Figure 1. Scenario tree

In [21] the authors consider a set of possible scenarios rather than solving the UC problem for one expected and the worst-case demand scenario. Each of the scenarios is assigned a weight that reflects the possibility of its occurrence in the future. The solution must satisfy the constraint that if two different scenarios *s* and *s*' are indistinguishable at time period *t* based on the available information at time period *t*, the decision made for scenario *s* must be the same as that for scenario *s*'. The constraint is modeled by partitioning the scenario set at each time period into disjoint subsets called scenario bundles. Mathematically, a bundle at time period *t* is represented as a constraint on the decision variables of its scenarios. The objective function is to minimize weighted sum of the objective functions of the smaller problems, i.e. to minimize the expected cost over all of the scenarios. The problem can then be solved using a Lagrangian relaxation type of technique.

In [18] authors present a security-constrained unit commitment algorithm considering the volatility of wind power generation. To capture volatility it is assumed that the wind power is subject to a normal distribution N (μ, σ^2) with forecasted wind power as expected value μ and a percentage of μ as its volatility (σ). The Monte Carlo simulation is used to generate a large number of scenarios subject to the normal distribution. The probability assigned to each scenario is one divided by the number of generated scenarios. To decrease the computational requirement for large number of scenarios, a scenario reduction technique is used. The algorithm is formulated as an optimization problem with the objective function composed of fuel costs and startup and shutdown costs of generating units over the scheduling horizon. The problem is a large-scale mixed-integer non-linear program. The Benders' decomposition is applied to decompose the problem into master problem, feasibility check subproblems, and network security check subproblems. The master problem provides a commitment and dispatch solution that minimizes the operating cost of dispatchable units. The feasibility check subproblems whether the commitment and dispatch solution of the master problem can accommodate the volatility of the wind power in individual scenarios. The paper shows that the physical limitations of units, such as ramping, are crucial for accommodating the volatility of the wind power generation.

In [22] the authors present a stochastic model for the long-term solution of security-constrained UC. Forced outages of generating units and transmission lines are modeled as independent Markov processes, and load

forecasting uncertainties as uniform random variables. Optimization problem is decomposed into deterministic long-term subproblems. A scenario reduction method is used to obtain a tractable solution. A 6-bus system, the IEEE 118-bus system, and an 1168-bus system are used to test the algorithm. The stochastic solution provides more reliable decisions on energy allocation, fuel consumption, emission allowance, and long-term utilization of generating units in comparison to the deterministic UC solution.

III. Robust Unit Commitment

Robust unit commitment formulations require a deterministic set of uncertainty, rather than a probability distribution on the uncertain data. Robust UC model described in [9] is a two-stage model: the first stage finds the optimal commitment decision, while the second stage generates the worst case dispatch cost under a fixed commitment solution from the first stage. The range of uncertainty is defined by the upper and lower bounds on the net load at each time period.

The robust model generates the optimal solution feasible for all realizations of the uncertain data within the given bounds. By minimizing the highest cost over all realizations, the model tends to provide conservative solutions, thus more expensive, which can be adjusted using the budget of uncertainty. The budget of uncertainty is defined as the number of buses that are allowed to deviate from a given central wind forecast in the worst case scenario [9]. The higher the budget of uncertainty the more robust the solution. The value for the budget of uncertainty is not known in advance but depends on the system and the experience. ISO New England (NE) enlisted Lawrence Livermore National Laboratory (LLNL)'s help in determining whether a robust UC would improve system reliability while keeping the operation cost relatively low in the presence of renewable variability [31]. In the study, a comprehensive evaluation of robust UC was conducted. The objectives were to identify the optimal conservatism level to balance the economic efficiency and operational reliability of robust UC solutions, as well as to compare the robust and deterministic approaches.

In [23] authors present a two-stage network constrained robust UC problem introducing a two-dimensional uncertainty set to describe the uncertain problem parameter, allowing the uncertainty correlations among different buses and among different time periods. A bilinear separation approach generates tight lower and upper bounds for the optimal objective value and it is tested for computational efficiency on a 118-bus system. The authors use the Benders' decomposition that includes feasibility and optimality cuts. A case in which the demand at each bus in each operating hour may be uncertain is addressed, and the uncertainties are described by a given polyhedral uncertainty set rather than by the probability distribution.

In [24] a two-stage robust UC model is developed to obtain day-ahead generator schedules where wind uncertainty is captured by a polytopic uncertainty set. The uncertainty set modeling method captures the random nature of wind without any explicit description of the distribution function. The model is also extended to include the demand response strategy. The authors performed a computational study on an IEEE 18-bus system to show that the robust UC model can utilize wind generation and lower overall generation cost.

A robust UC model that takes into account the worst-case scenario of wind power output with deterministic loads during all periods is presented in [25]. This approach distributes the random problem parameters in a predetermined uncertainty set containing the worst-case scenario. Uncertain wind power output in each time period is within an interval defined by its lower and upper bounds which are obtained based on historical data or estimated with a confidence interval. The problem is formulated as a two-stage min-max problem with the objective to minimize the total cost under the worst wind power output scenario. The degree of conservatism is adjusted using the budget of uncertainty, an integer that takes a value between 0 and the number of hours in the time horizon T, to restrict the number of time periods that allow the actual wind power output to deviate from its forecasted value. By adjusting the value of the budget of uncertainty, system operators can control the robustness of the solution. The higher the budget of uncertainty, the more robust the solution. The UC decisions are made at the first stage, while the second stage results in economic dispatch. Wind power generation values are subject to uncertainty, and they are presented by random variables described by the uncertainty set. The authors describe their solution methodology and test the algorithm on a 6-bus and a modified 118-bus system. The wind power uncertainty is additionally hedged using pumped storage hydro units.

IV. Interval Unit Commitment

Interval UC formulations produce a schedule that minimizes the cost of serving the most probable net load forecast while guaranteeing feasibility in the entire uncertainty range that is delimited with upper and lower bounds as in robust unit commitment formulations. Figure 2. shows the central forecast, i.e. the most probable realization to be minimized in the objective function, along with the upper and lower bounds and the transitions between them. The solution is optimal along central forecast while remaining feasible along upper and lower bound. The solutions tend to be conservative because of the steep ramp requirements that need to be satisfied in between all consecutive time periods, as shown in Figure 2. The formulation is computationally more efficient than the stochastic unit commitment formulation because the model can be formulated using three scenarios, i.e. the central forecast and the upper and lower bounds, while the transition constraints are modeled as constraints. The interval UC can also be formulated as a two-stage problem where the optimal solution is found in the first stage and then tested for feasibility in the second stage.

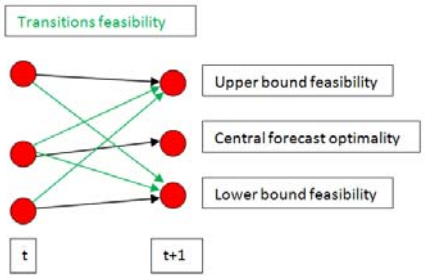


Figure 2. Interval unit commitment

A method for daily UC and dispatch incorporating wind power based on the interval number theory is introduced in [26]. Uncertain wind power generation is represented by a functional interval. The optimal model is first divided into two deterministic mix-integer programming subproblems with the parameters expressed as constants. The interval solutions of the model can be constructed using the solutions of two subproblems. The model is tested on a 30-bus system showing that the proposed method can be used for the unit commitment. The forecasting accuracy has a great impact on the optimal interval of UC.

V. HYBRID UNIT COMMITMENT MODELS

Recently, some authors have developed hybrid models that exploit the advantages and eliminate disadvantages of the models presented in the previous Sections. Such models are unified stochastic and robust unit commitment formulation proposed in [27] and hybrid stochastic/interval unit commitment model proposed in [20].

A. Unified stochastic and robust unit commitment

Stochastic UC formulations face computational challenges due to the large scenario size necessary to secure system robustness, while the robust UC formulations tend to result in conservative, thus expensive solutions. To take advantage of both of the approaches the authors of [27] propose a unified stochastic and robust UC model able to achieve low expected total cost while ensuring the system robustness. The objective function contains stochastic and robust parts that are weighed with scaling factors which can be adjusted by system power operators. The model generates a less conservative solution as compared to the two-stage robust optimization approach and a more robust solution as compared to the two-stage stochastic optimization approach. As in previous two-stage models, at the first stage the day-ahead unit commitment decisions are made. The second stage decides on the dispatch for each scenario for the stochastic optimization part and the worst-case scenario for the robust optimization part. A new parameter is introduced ranging between 0 and 1 to represent the weight of the worst case generation cost. The authors tested their approach on a modified IEEE 118-bus system.

The downside is that this approach employs heuristics to balance the stochastic and robust unit commitment solutions, which may result in suboptimal solution.

B. Hybrid stochastic/interval unit commitment

The authors of [20] propose a model that applies the stochastic formulation to the initial operating hours of the optimization horizon and then switches to the interval formulation for the remaining hours. The switching time is optimized to achieve optimal trade-off between the cost of unhedged uncertainty from the stochastic UC and the security premium of the interval UC. The two formulations are applied sequentially, instead of simultaneously according to their heuristically chosen weights in [27]. The stochastic UC, which is more cost effective but for a computationally tractable numbers of scenario less robust, is applied to the first part of the time horizon during which the wind power output predictions are more accurate. The model then switches to interval UC offering more robust solution for the remaining time period when the uncertainty is also greater.

The authors introduce a day-ahead cost (DAC) which represents the expected operating cost at the day-ahead stage. The DAC of a stochastic UC formulation is a minimum expected operating cost over the set of scenarios, while DAC of the interval UC formulation represents the cost of the schedule that minimizes the cost of meeting the central net load forecast while ensuring the feasibility of the predefined worst-case scenarios. A more conservative uncertainty model results in greater DAC because of a more conservative schedule. The actual operating cost (AOC) is the cost that includes the corrective actions which include redispatch of committed generators and starting up or shutting down other generators to account for deviations between actual and predicted output. This hybrid UC formulation aims to minimize AOC by finding the optimal balance between the day-ahead security cost and the expected cost of uncertainty associated with the day-ahead schedule. The authors present a method to optimally select the switching time and they test their approach on a modified version of the 24-bus system

VI. CONCLUSION

The common goal of UC formulations is to minimize the operating cost, while ensuring sufficient reserve to accommodate real-time realization of uncertainty. The main difference between models is the representation of this uncertainty. The stochastic UC formulation tends to give the most cost effective solution. However, in order to secure the robustness of solution, a large number of scenarios is required which renders this formulation computationally intensive. The robust and interval UC formulations secure a robust solution, but tend to give conservative, thus more expensive solutions. The hybrid models are developed to exploit the best traits of stochastic formulation (being the most cost effective) and robust or interval formulation (securing the robustness).

A research direction in unit commitment under uncertainty will include a combination of the existing methods that will try to exploit their respective

advantages. However, an important issue is the scenario generation and reduction techniques, which is a field of the ongoing research.

Also, in the scientific community it is still not clear how to validate the solutions obtained using different techniques. Although Monte Carlo simulations are the most common technique for classification of the solutions, it is sensitive to its input parameters (historical errors). This method should also be re-examined and improved.

REFERENCES

[1] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE Power Energy Mag*., vol. 7, no. 2, pp. 52-62, Mar./Apr. 2009.

[2] J. G. Vlachogiannis, "Probabilistic constrained load flow considering integration of wind power generation and electric vehicles," *IEEE Trans. Power Syst*., vol. 24, no. 4, pp. 1808-1817, Nov. 2009.

[3] Y. Wang, Q. Xia, and C. Kang, "Unit commitment with volatile node injections by using interval optimization," *IEEE Trans. Power Syst*., vol. 26, no. 3, pp. 1705-1713, Aug. 2011

[4] J. Kiviluoma et all., "Impact of wind power on the unit commitment, operating reserves, and market design," in *Proc. of IEEE Power and Energy Society General Meeting 2011*, San Diego, California, July 2011, pp. 1-8

[5] H. Pandžić, T. Qiu, and D. Kirschen, "Comparison of state-of-the-art transmission constrained unit commitment formulations," in *Proc. of IEEE PES General Meeting 2013*, Vancouver, Canada, July 2013, pp. 1-5.

[6] J. M. Arroyo and J. Conejo, "Optimal response of a thermal unit to an electricity spot market," *IEEE Tran. Power Syst*., vol. 15, no. 3, pp. 1098-1104, Aug. 2000.

[7] D. Rajan and S. Takriti, "Minimum up/down polytopes of the unit commitment problem with start-up costs," Jun. 2005, *IBM Research Report*

[8] C. K. Simoglou, P. N. Biskas, and A. G. Bakirtzis, "Optimal self-scheduling of a thermal producer in short-term electricity markets by MILP," *IEEE Tran. Power Syst*., vol. 25, no. 4, pp. 1965-1977, Nov. 2010.

[9] D. Bertsimas, E. Litvinov, X. A. Sun, Z. Jinye, and T. Tongxin, "Adaptive robust optimization for the security constrained unit commitment problem," *IEEE Trans. Power Syst*., vol. 28, no.1, pp. 52-63, Feb. 2013

[10] Y. V. Makarov, C. Loutan, M. Jian, and P. de Mello, "Operational impacts of wind generation on California Power Systems," *IEEE Trans. Power Syst.,* vol 24, no. 2, pp. 1039-1050, May 2009.

[11] Y. G. Rebours, D. S. Kirschen, M. Trotignon, and S. Rossignol, "A survey of frequency and voltage control ancillary services. Part I: Technical features," *IEEE Trans. Power Syst*., vol. 22., no. 1, pp. 350-357, Feb. 2007.

[12] Y. G. Rebours, D. S. Kirschen, M. Trotignon, and S. Rossignol, "A survey of frequency and voltage control ancillary services. Part II: Economic features," *IEEE Trans. Power Syst*., vol. 22, no. 1, pp. 358-366, Feb. 2007.

[13] ERCOT *Methodologies for determining ancillary service requirements,* 2012. [Online] Available at: www.ercot.com/meetings/wms/keydocs/2004/0819/WMS08192004-3.doc

[14] M. A. Ortega-Vazquez and D. S. Kirschen, "Optimizing the spinning reserve requirements using a cost/benefit analysis," *IEEE Trans. Power Syst*., vol. 22, no. 1, pp. 24-33, Feb. 2007.

[15] M. A. Ortega-Vazquez, D. S. Kirschen, and D. Pudijanto, "Optimising the scheduling of spinning reserve considering the cost of interruptions," *IEEE Proc. Gener., Trans. Distr.,* vol 153, no. 5, pp. 570-575, Sept. 2006.

[16] F. D. Galiana, F. Bouffard, J. M. Arroyo, and J. F. Restrepo, "Scheduling and pricing of coupled energy and primary, secondary, and tertiary Reserves," *Proc. IEEE*, vol. 93, no. 11, pp. 1970-1983, Nov. 2005.

[17] J. M. Arroyo and F. D. Galiana, "Energy and reserve pricing in security and network-constrained electricity markets," *IEEE Trans. Power Syst*., vol. 20, no. 2, pp. 634-643, May 2005.

[18] J. Wang, M. Shahidehpour, and Z. Li, "Security-constrained unit commitment with volatile wind power generation," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1319-1327, Aug. 2008.

[19] Y. Dvorkin, Y. Wang, H. Pandžić, D. Kirschen, "Comparison of scenario reduction techniques for the stochastic unit commitment", in *Proc. of IEEE PES General Meeting*, *Conference & Exposition 2014*., National Harbor, Maryland, July 2014, pp. 1-5

[20] Y. Dvorkin, H. Pandžić, M. Ortega-Vazquez, and D. S. Kirschen, "A hybrid stochastic/interval approach to transmission-constrained unit commitment," *IEEE Trans. Power Syst*., early access.

[21] S. Takriti, J. R. Birge, and E. Long, "A stochastic model for the unit commitment problem," *IEEE Trans. Power Syst.,* vol. 11, no. 3, pp. 1497-1508, Aug. 1996.

[22] L. Wu, M. Shahidehpour, and T. Li, "Stochastic security-constrained unit commitment," *IEEE Trans. Power Syst.,* vol. 22, no. 2, pp. 800-811, May 2007.

[23] R. Jiang, M. Zhang, G. Li, and Y. Guan, "Two-stage network constrained robust unit commitment problem," *Eur. J. Oper. Res*., vol. 234, no.3, pp. 751-762, May 2014.

[24] B. Zheng and L. Zhao, "Robust unit commitment problem with demand response and wind energy," in *Proc. IEEE PES General Meeting 2012*, San Diego, USA, July 2012, pp. 1-8.

[25] R. Jiang, J. Wang, and Y. Guan, "Robust unit commitment with wind power and pumped storage hydro," *IEEE Trans. Power Syst*., vol. 27, no. 2, pp. 800-810, May 2012.

[26] X. Sun and C. Fang, "Interval mixed-integer programming for daily unit commitment and dispatch incorporating wind power," in *Proc. Power System Technology (POWERCON) 2010*, Hangzhou, China, Oct. 2010, pp. 1-6

[27] C. Zhao and Y. Guan, "Unified stochastic and robust unit commitment," *IEEE Trans. Power Syst*., vol. 28, no. 3, pp. 3353-33761, Aug. 2013.

[28] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," in *Proc. of 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035, New Orleans, Louisiana, 2007.

[29] J. Dupacova, N. Growe-Kuska, and W. Romisch, "Scenario reduction in stochastic programming: An approach using probability metrics," *Math. Program*., Vol. 95, pp. 493-511, 2003.

[30] A. Papavasiliou and S. S. Oren, "Multiarea stochastic unit commitment for high wind penetration in a transmission constrained network," *Operation Research*, Vol. 61, No. 3, pp. 578-592, 2013.

[31] J. Zhao et all., "A comprehensive evaluation of robust unit commitment" 2014., [Online] Available at: https://e-reports-ext.llnl.gov/pdf/770804.pdf

[32] H. Pandžić, A. Conejo, I. Kuzle, "An EPEC Approach to the Yearly Maintenance Scheduling of Generating Units, " *IEEE Transactions on Power Systems*, Vol. 28, No. 2, pp. 922-930, 2013

Describing function analysis of systems with fuzzy compensated odd symmetric nonlinear actuators

O. Kuljača, and K. Horvat
Brodarski Institute, Zagreb, Croatia
ognjen.kuljaca@hrbi.hr, krunoslav.horvat@hrbi.hr

Abstract - Paper deals with fuzzy compensated odd symmetric nonlinear actuators. Most actuators are nonlinear sporting odd symmetric nonlinearities such as deadzone and saturation. One of ways of dealing with such actuators is to try to compensate for nonlinearities by fuzzy systems. Compensated actuators can improve behaviour of the system, but then arises the problem of stability analysis because compensated nonlinearity is now complex nonlinearity not described in common literature. One way of dealing with such problem is to perform stability analysis via describing function method. Paper will describe the method for compensating nonlinearities, recording describing function and performing stability analysis.

I. INTRODUCTION

Every control system contains at least one actuator. Virtually all actuators have nonlinearities. Depending on the control system these nonlinearities can be neglected or one has to deal with them in some way. Nonlinear control theory is rich with different methods of compensation or stability analysis of systems with nonlinear actuators ([9],[11],[12],[13],[14],[15] and many others.).

Many actuators have odd nonlinearity followed by linear filter type element. Examples are some electrical motors, hydraulic cylinders and solenoid actuators. There are different methods for dealing with such systems (introduction of ditter signal, different anti wind-up schemes etc,). This paper deals with fuzzy compensated odd symmetric nonlinear actuators. Such systems will be analyzed via describing function method.

II. DESCRIBING FUNCTION

Many actuators with odd nonlinearity can be represented by structure shown in Figure 1.

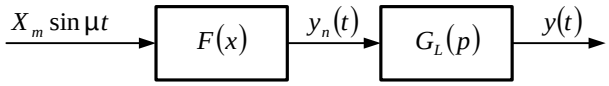


Figure 1. Actuator represented with decoupled nonlinear F(x) and linear parts $G_L(p)$, p=d/dt

Describing function is an equivalent gain of nonlinear element, defined by the harmonic linearization method of nonlinear static characteristic ([2], [9], [10], [11], [12] and many others). It is a known method of analysis and synthesis when nonlinear system can be decoupled into linear and nonlinear parts. If the linear part of the system has the characteristics of low-pass filter and if we apply periodical signal to the system, output signal will have the same base frequency as input signal with damped higher frequencies.

If the amplitudes of higher harmonics are relatively small when compared to the amplitude of the first harmonic, i.e. if filter hypothesis holds, output signal can be approximated by it's first harmonic. Mathematically, first harmonic of the output signal, for the sinusoidal input signal $X_m \sin \mu t$, can be expressed by Fourier expressions:

$$y_N(t) \approx Y_{P1} \sin \mu t + Y_{Q1} \cos \mu t \tag{1}$$

$$Y_{P1} = \frac{1}{\pi} \int_0^{2\pi} F(X_m \sin \mu t) \sin \mu t \, d(\mu t) \tag{2}$$

$$Y_{Q1} = \frac{1}{\pi} \int_0^{2\pi} F(X_m \sin \mu t) \cos \mu t \, d(\mu t) \tag{3}$$

where Y_{P1} and Y_{Q1} are first Fourier coefficients. Describing function is the ratio between first harmonic of the output signal and input signal in complex form:

$$G_N(X_m) = P(X_m) + jQ(X_m) = \frac{Y_{P1}}{X_m} + j\frac{Y_{Q1}}{X_m} \tag{4}$$

where $P(X_m)$ and $Q(X_m)$ are coefficients of the harmonic linearization ([2], [9], [10], [11], [12] and many others).

Determination of describing function boils down to the determination of integral expressions for the known static characteristic of the nonlinear part of the system.

There are many conventional nonlinearities for which static characteristics and describing functions are theoretically derived and given in analytical form.

The problem arises when the static characteristic of nonlinear system cannot be analytically expressed or integral expressions cannot be solved. In that case describing function can be determined by experiment or simulation ([3], [4], [5], [6], [7], [8]) or some method of numerical integration [10].

In this work we will use numerical integration in Simulink as given in [10]. The approach is very simple, yet effective. In essence, equations (2) and (3) are directly

applied in Simulink on nonlinear element outputs. More details on this approach to obtaining describing function are also given in [13] and [14].

III. FUZZY COMPENSATED ODD SYMMETRIC NONLINEAR ACTUATOR

In this section we describe simple compensation of nonlinearities via neural network. We deal with actuator with architecture shown in Figure 1. It is important that in practice $y_n(t)$ is measurable. This is often the case because many motor drives and hydraulic drives have nonlinearity in part of their electronic driver and this value can be measured. We then connect an in input with changeable amplitude and measure and save inputs and outputs. Once this is done we will use recorded data to train fuzzy system in such way that we try to achieve fuzzy system that will have opposite action to nonlinearity.

Let us illustrate the procedure on dead zone nonlinearity with Matlab simulation.

Schematic of data generation and recording is shown in Figure 2.

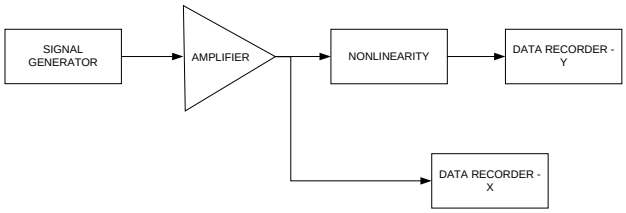


Figure 2. Data generation and recording

Datasets X and Y were generated using chirp signal generator with starting frequency 0.1 Hz, final frequency 20 Hz and simulation end time 1000s with fixed step size 0.001. Dead zone is static nonlinearity so frequency really does not have effect on it but we used chirp signal in order to generate a number of different data points. Deadzone had start of deadzone at -1 and end of deadzone at 1. Amplifier had gain 5.

Deadzone is shown in Figure 3.

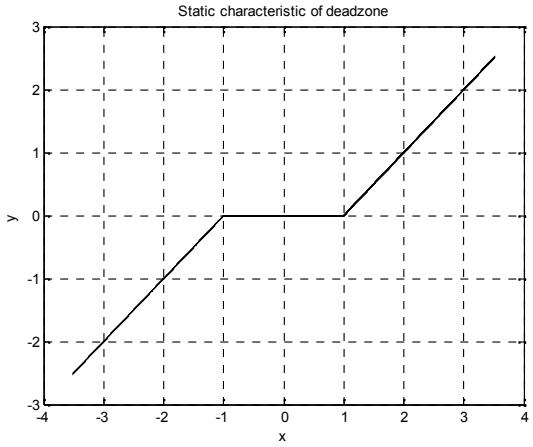


Figure 3. Deadzone

In principle any fuzzy single input single output system could be used. In this example, for compensation we used Takagi-Sugeno type fuzzy system and trained it using anfis procedure Matlab. We used max-prod inference with weights average deffuzification.

Training was performed for 20 epochs. Results are shown in Figure 4.

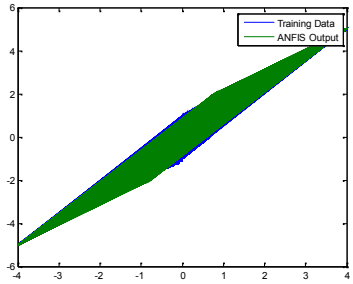


Figure 4. Training results

Input membership functions are shown in Figure 5. Output membership functions ar of type linear. The rule list is given as:

- 1. If (input1 is in1mf1) then (output is out1mf1) (1)
- 2. If (input1 is in1mf2) then (output is out1mf2) (1)
- 3. If (input1 is in1mf3) then (output is out1mf3) (1)
- 4. If (input1 is in1mf4) then (output is out1mf4) (1)
- 5. If (input1 is in1mf5) then (output is out1mf5) (1)

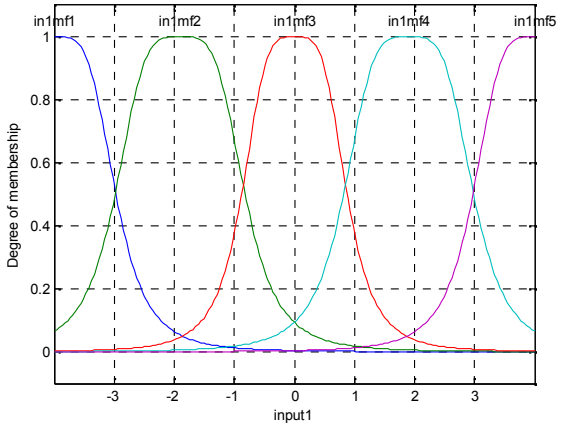


Figure 5. Input membership functions

Fuzzy system input-output (static) characteristic is shown in Figure 6.

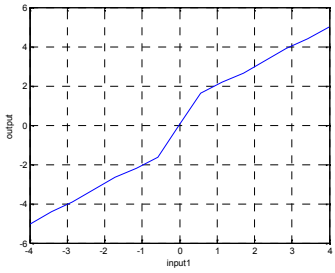


Figure 6. Fuzzy system static characteristic

Then, we can use fuzzy system to compensate for nonlinearity as shown in Figure 7 ([13], [14]),

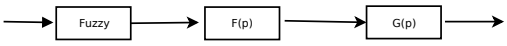


Figure 7. Fuzzy nonlinearity compensation

where fuzzy stands for neural network, F(p) is nonlinearity and G_L(p) is linear part of the system.

The idea is that fuzzy system compensates for some effects of nonlinearity since fuzzy system is trained to act as inversion of nonlinear function.

This procedure can be applied on any system with odd symmetric nonlinearity.

IV. DEADZONE EXAMPLE – SIMULATION RESULTS

Let us now see some simulation results for deadzone case. Simulink simulation scheme is shown in Figure 8.

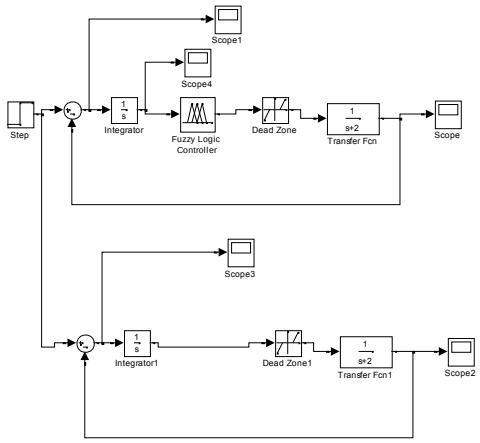


Figure 8. Simulink simulation scheme for dead zone case

Deadzone and fuzzy system are given in Section III. Responses for step input 1.7 are given in Figure 9.

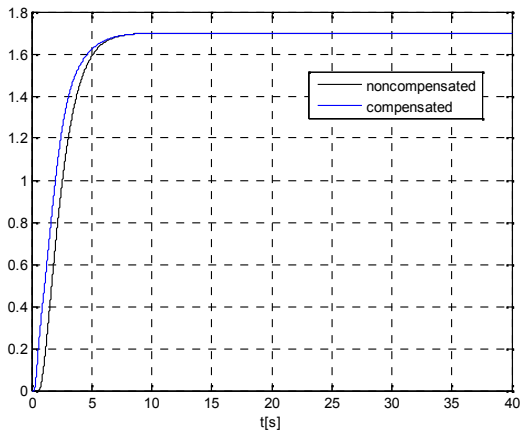


Figure 9. Responses for step input 1.7

Response of noncompensated system is depicted in black and response of compensated system is depicted in blue. It can be seen that compensated system reaches required value faster.

Let us now see what happens for smaller reference values, let's say with reference input of 0.1. Responses are shown in Figure 10.

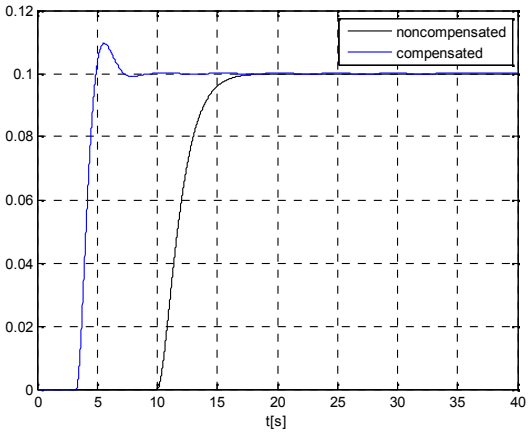


Figure 10. Responses for step input 0.1

Response of noncompensated system is depicted in black and response of compensated system is depicted in blue. It can be seen that compensated system reaches required value faster, although with some overshoot.

Let us also look at tracking case with sinusoidal input reference value. Simulink simulation scheme is shown in Figure 11.

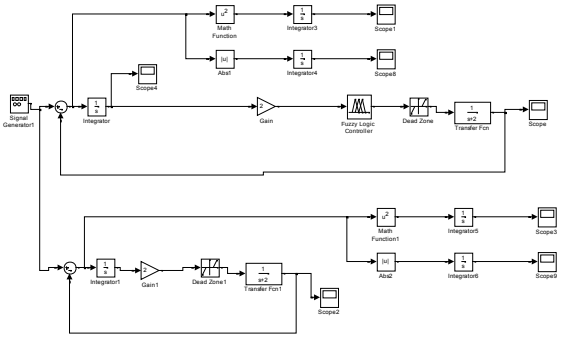


Figure 11. Simulink simulation scheme for dead zone case - sinusoidal input reference value

As it can be seen from Figure 11, in this case we used as a measure of quality performance indicators for each simulation, i.e Integral of Absolute Error (IAE)

$$J_{IAE} = \int_0^\infty |e(t)|dt. \quad (5)$$

and Integral of Squared Error (ISE)

$$J_{ISE} = \int_0^\infty e^2(t)dt. \quad (6)$$

Reference input was sinusoidal signal with amplitude of 1.7 and frequency of 0.01 Hz. Responses are shown in Figure 12.

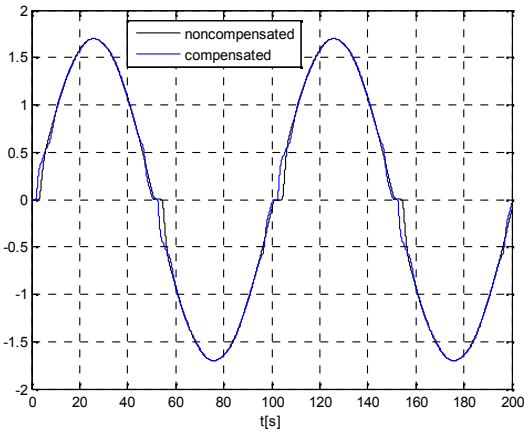


Figure 12. Responses for sinusoidal input with amplitude 1.7 and frequency 0.1 Hz

Response of noncompensated system is depicted in black and response of compensated system is depicted in blue. It appears that compensated system reaches required value faster. Let us check IAE and ISE (Table 1).

Table 1 Mean values of IAE and ISE with original parameters for sinusoidal input with amplitude 1.7 and frequency 0.1 Hz

	IAE	ISE
noncompensated	17.0532	2.6987
compensated	13.552	1.3281

Both integral parameters show smaller values for compensated case, thus it can be concluded that compensation indeed yields better performance.

Let us now see what happens for smaller reference values, with reference input of 0.1. Responses are shown in Figure 13.

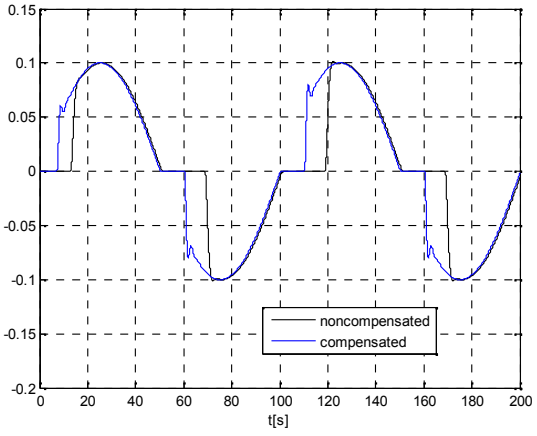


Figure 13. Responses for sinusoidal input with amplitude 0.1 and frequency 0.1 Hz

Response of noncompensated system is depicted in black and response of compensated system is depicted in blue. Again, t appears that compensated system reaches required value faster. Let us check IAE and ISE (Table 2).

Table 2 Mean values of IAE and ISE with original parameters for sinusoidal input with amplitude 0.1 and frequency 0.1 Hz

	IAE	ISE
noncompensated	4.3169	0.2531
compensated	1.4937	0.0511

Again, both integral parameters show smaller values for compensated case, thus it can be concluded that compensation indeed yields better performance.

It can be seen that responses are better in case of bigger input amplitude. This is expected situation since we deal with dead zone. If deadzone accounts for bigger percentage of input signal the response will be worse.

V. Stability Analysis of Fuzzy Compensated Odd Symmetric Nonlinear Actuator

It is clear that when we make compensation of nonlinearity with fuzzy system, in essence we have a new nonlinearity as shown in Figure 14.

Figure 14. Compensated nonlinearity

With compensated nonlinearity we can record describing function of compensated nonlinearity as described in [10], [13], [14] and [15]. In this example we recorded describing function of deadzone compensated with fuzzy system for parameters already given in this paper. Describing function is shown in Figure 15.

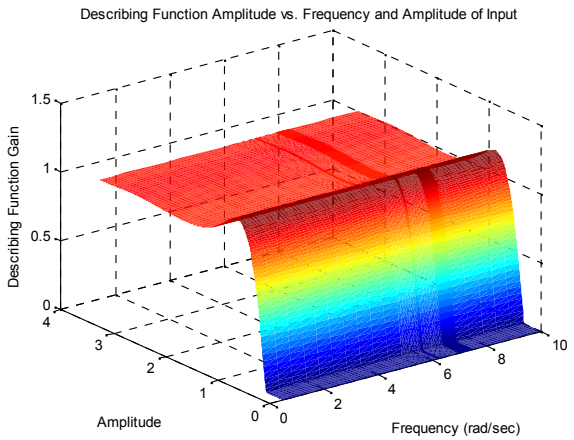


Figure 15. Describing function of compensated deadzone – $P(X_m)$

Because deadzone is symmetric nonlinearity there is no imaginary term in describing function. This is so for all symmetric nonlinearities. Thus, the same stability analysis procedure as given below can be used.

Characteristic equation of closed loop system with compensated nonlinearity as shown in Figure 8 can be written as

$$s^2 + Ks + G = 0 \tag{7}$$

where G is describing function of compensated nonlinearity and K is linear parameter system of actuator transfer function represented by

$$\frac{1}{s+K} \tag{8}$$

Once we have describing function we can use any stability criterion we use for linear systems. By Hurwitz criterion the system will be stable if

$$KG > 0 \tag{9}$$

Since K is always positive, system will be stable if G>0. In case of our system with dead zone that is always the case, thus the system is stable. The same analysis can be conducted with other types of odd symmetric nonlinearities. Also, actuator transfer function (8) can be of higher order and, as long as it is Hurwitz, this will not affect above described analysis.

VI. CONCLUSION

Paper gives the procedure for training fuzzy system to act as compensator for odd symmetric nonlinearities. It is shown how to record the describing function of such fuzzy system. Simulation examples are performed on case of dead zone and it was shown that compensation is effective. More analysis will be performed in future research for other nonlinearity types.

Finally, a simple procedure for analyzing stability of such systems is given.

REFERENCES

[1] Kreyszig, E. (1993). *Advanced Engineering Mathematics (7)*, John Wiley and Sons, New York, USA

[2] Novogranov, B.N. (1986). Determination of Frequency Domain Characteristics of Nonlinear Systems (in Russian), Masinostroenie, Moskow, Russia

[3] Kuljaca, O., Tesnjak, S., Vukić, Z. (1999). Describing Function of Mamdani Type Fuzzy Controller with Input Signals Derived From Single System Input and Singleton Ouput Membership Functions, *Proceedings of the 1999 IEEE Hong Kong Symposium on Robotics and Control*, Volume I, HKSRC '99, Hong Kong, 1999.

[4] Sijak, T., Kuljaca, O., Antonic, A., Kuljaca, Lj., (2005). Analytical Determination of Describing Function of Nonlinear Element with Fuzzy Logic, *Proceedings of the IEEE International Symposium on Industrial Electronics 2005*, ISBN 0-7803-8739-2, Dubrovnik, Croatia, June 2005

[5] Sijak, T., Kuljaca, O., Kuljaca, Lj. (2004). Describing function of generalized static characteristic of nonlinear element, *Proceedings of REDISCOVER 2004, Southeastern Europe, USA, Japan and European Community Workshop on Research and Education in Control and Signal Processing*, ISBN 953-184-077-6, Cavtat, Croatia, June 2004

[6] Sijak, T., Kuljaca, O., Kuljaca, Lj., (2007). Engineering Procedure for Analysis of Nonlinear Structure Consisting of Fuzzy Element and Typical Nonlinear Element, *Proceedings of the 15th Mediterranean Conference on Control and Automation*, Athens, Greece, July 2007

[7] Sijak, T., Kuljaca, O., Kuljaca, Lj., (2007). Computer Aided Harmonic Linearization of SISO Systems Using Linearly Approximated Static Characteristic. *Proceedings of EUROCON 2007 The International Conference on Computer as a Tool*. ISBN 1-4244-0813-X, Warsaw, Poland, September 2007.

[8] Sijak, T., Kuljaca, O., Tesnjak, S. (2001). Stability analysis of fuzzy control system using describing function method, *Proceedings of 9th Mediterranean Conference on Control and Automation*, ISBN 953-6037-35-1, Dubrovnik Croatia, June2001

[9] Slotine, J.J.E., Li, W. (1991). *Applied Nonlinear Control*, Prentice Hall, ISBN 0-13-040890-5, Engelwood Cliffs, New Jersey, USA

[10] Schwartz, C., Gran, R. (2001). Describing Function Analysis Using MATLAB and Simulink. *IEEE Control Systems Magazine*, Vol. 21, No. 4, (August 2001), pp. 19-26, ISSN 1066-033X

[11] Vukic, Z., Kuljaca, Lj., Donlagic, D., Tesnjak, S. (2003). *Nonlinear Control Systems*, Marcel Dekker, ISBN 0-8247-4112-9, New York, USA

[12] Netushil at al. *Theory of Automatic Control*, Mir, Moscow 1978

[13] Kuljaca, O., Horvat, K., Gadewadikar, J., "Describing function analysis of systems with NN compensated odd symmetric nonlinear actuators", *Proceedings of the 35th International Convention MIPRO 2012*, pp. 788-793, Opatija, Croatia, May 2012

[14] Kuljaca, O., Horvat, K., Gadewadikar, J., Tare, B., "Polynomial Compensation of Odd Symmetric Nonlinear Actuators via Neural Network Modelling and Neural Network Describing Function ", Engineering/Computing and Systems Research E-Conference CISSE 2012, International Conference on Industrial Electronics, Technology & Automation IETA 12, 2012

[15] Kuljaca, O., Horvat, K., Eskinja, Z., "Static Characteristic and Describing Function of Polynomial Compensated Nonlinearities", Proceedings of the 35th International Convention MIPRO 2014, pp. 1187-1192, Opatija, Croatia, May 2014

Simple Day-Ahead Bidding Algorithm for a System with Microgrids and a Distributor

Leonardo Jelenković, Toni Budrović
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
Email: leonardo.jelenkovic@fer.hr

***Abstract*—In this paper we present a heuristic distributed algorithm for day-ahead bidding between collection of microgrids that sell their surplus energy to a power grid controller – distributor, who can use that energy with its consumers. Each microgrid is modeled as a node with consumers, energy generator (from renewable sources) and energy storage capabilities. For each time period, distributor negotiate amounts and prices with each microgrid separately, in rounds. Although both sides are modeled, we focus our research on the microgrid's bidding algorithm for monetary gain from sold surplus energy. Presented algorithm is simple and produces good results from microgrid perspective. Due to model simplicity, produced results aren't always optimal, and convergence had to be additionally helped. In comparison to similar distributed bidding algorithms for systems with storage capacity presented algorithm is simplest and can be used for day-ahead but with presented adaptations also for more complex microgrids and distributor, real-time adjustments and asynchronous negotiations.**

I. Introduction

Using renewable energy sources is the only long term strategy that will ensure energy stability. Therefore, today we put a lot of effort to identify renewable energy sources and use them. Such researches are being promoted and widely supported by governments and organizations, as in [1]. In electric power systems, on global scale, power plants that use fossil fuels should be complemented or replaced with ones that use renewable energy. On a smaller scale, renewable energy could be collected with small power plants (e.g. wind turbines, solar panels) and used locally, in areas called microgrids, for its own consumers.

However, renewable energy production might be very variable even over single day. In some periods (hours) produced energy might be sufficient or insufficient for local consumers. Therefore, additional energy sources might be required. Such sources may include battery or other type of energy storage, diesel generator (or similar), or/and connection to a global power grid. Those sources can be used when renewable sources don't produce sufficient amount of energy for local microgrid. On the other side, when renewable sources produce more energy than its locally needed, surplus can be used to charge local storage or/and it can be sold through global power grid.

This work is supported by the research project MICROGRID, Grant no. I-4463-2011, Croatian Science Foundation.

Microgrids will not significantly influence global power grid (or smart grid), at least not in near future since their surplus energy is not significant when comparing with power grid needs. However, since they use renewable energy, microgrids might be supported by governments. In this paper we assume a model (Fig. 1) where a power grid controller – distributor (in rest of the paper), possibly under influence of government, should offer to buy some of microgrids' surplus energy (up to predefined amount for each period of day). That's where a trade market could be created – between a distributor and connected microgrids. To optimize usage of energy bought from microgrids, distributor defines different price ranges for different periods of day ahead (e.g. highest prices for peak demand periods) giving incentive to microgrids to adjust their selling strategy, to provide more energy in high-price periods. In this paper we focus on microgrid's control method that will maximize profit from sold surplus energy. Storage management influence microgrid's offers to distributor, and distributor's prices influence microgrid's storage management. This is where 'the game' is defined.

Similar problems were being addressed in literature. In [2] and [3] smart grid is simulated consisting of users (consumers) and suppliers (generators). Generator models include thermal and hydro generators. Thermal generators are modeled with: maximal power, minimum stable load, ramps and variable, no-load and startup costs. Hydro generators are modeled with maximal and minimal power outputs and daily energy reserves. Demand side was modeled as static, inelastic – does not changes with price. In [2] biding is performing iteratively, decrementing price in each iteration, while in [3] price was chosen at intersection of supply and demand curves. Convergence problems in [3] was addressed with heuristic approaches.

In [4] a system is modeled with (a simple) supply-side (distribution network) connected to central unit, which is connected to users (consumers). Users may be passive (just consumers) or active, that besides consumers also have energy generators and storage. Their demand side management consists of day-ahead bidding and real-time adjustments for variations of user needs and production. Their algorithm is distributed and iterative.

Methods from cited papers, and similar ones, with some effort might be adjusted even for our system model. However,

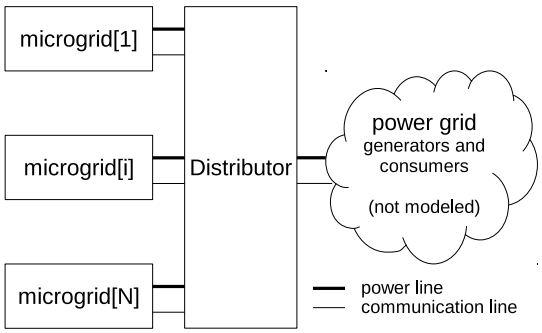


Fig. 1. The system model

we wanted to define and test a simpler models and highlight storage management in microgrids. Furthermore, our system model has some subtle difference. Price adjustment ("bidding process") is performed differently. Distributor proposes prices (one per each period) and sends offers to microgrids – each microgrid receive its own offers, one for each period. Microgrids evaluate offers individually, and send back replies to distributor. Each reply, from single microgrid and for single period, consists of accept or reject decision of appropriate offer, with units and prices distributor should use in next iteration of bidding process. For each period of day ahead distributor tries to buy up to previously calculated required amount of energy (which is fixed for whole bidding process). In bidding process, when microgrids (cumulatively) don't provide required amount, distributor might raise price for that period, if price isn't already on predefined limit. Similarly, when distributor detects that offered amount for particular period exceeds distributor demands, distributor might lower price for that period (if not already on predefined lower limit).

One of our main requirement for model was simplicity. We are trying to find a simple algorithm for a microgrid in previously described setup. Therefore, in our initial work on the subject, we present a simple heuristic algorithm that is ready to be 'copy-pasted' in control program. Simplicity enable usage of this algorithm even in automatically controlled small scale system consisting of small microgrids managed by micro-controllers (e.g. a house with solar panels, battery and consumers) connected together with one controlling node that acts as distributor and can sell energy to third party. We think that most difference from real systems (generators, storages, consumers) and presented model can be accounted locally on particular microgrid, adjusting only amounts, prices and capacity.

As we didn't yet put enough focus on distributor's side, our simulation system have problems with convergence and optimality. Convergence problem is for now (in used models) mitigated by slowing price changes. However, further work is required on distributor's side of algorithm if we want guarantee of convergence. Results from all experiments we used produced good results, but not always optimal from global perspective. However, global optimum is hard to reach in distributed algorithms. Furthermore, global optimum definition in described system is problematic – if we optimize for microgrids' gain, distributor will probably have to pay more than if optimization was toward distributor's gain.

Rest of the paper is structured as follows. In sections II and III a system model and negotiation algorithm are presented. Experimental results are presented in section IV, possible adaptation in section V and conclusion is given in section VI.

II. The system model

The system model consists of collection of microgrids and a distributor, as shown on Fig. 1. We do not model power grid with its generators and consumers. However, we assume they influence initial settings for distributor (demanded energy per periods, with price ranges).

A. Microgrid model

Every microgrid is defined with energy generators, local consumers and storage. Since we model renewable sources, energy production is based on weather conditions (e.g. wind or solar intensity) and can be (accurately) predicted for day ahead for each period (e.g. hour). Energy that can be produced in single period could be used for local microgrid consumers, for charging local energy storage (e.g. battery) and for selling it to distributor. First priority is to ensure enough energy for local consumers (cover it completely if possible). For that purpose microgrid can use storage: charge it when surplus is produced and discharge it when produced energy is insufficient. What energy is left can be sold to distributor. Storage can also be used for optimizing sale: store surplus energy when energy is 'cheap' and sell it later when its price is higher. In our model we didn't account storage manipulation costs and loses. Also, price of energy generation is not modeled directly. Instead, minimum selling price is defined for each period that might include those production costs.

Let $M = \{m_1, m_2, \ldots, m_i, \ldots, m_N\}$ represent system with N microgrids. Day (ahead) is divided into P periods (e.g. hours).

Parameters that define single microgrid m_i are:

- ST_i – storage capacity (kWh)
- $SC_{i,0}$ – storage initial charge (kWh)
- for each period of next day j:
  - constant (given as input):
    - $CE_{i,j}$ – local consumers need (kWh)
    - $PE_{i,j}$ – production (generator) capacity (kWh)
    - $MP_{i,j}$ – minimal selling price (kn/kWh)
  - calculated within algorithm in k-th iteration:
    - $SC^k_{i,j}$ – starting storage charge (kWh)
    - $RC^k_{i,j}$ – reserved storage charge (kWh)
    - $RS^k_{i,j}$ – reply status: $ACCEPT$ or $REJECT$
    - $RP^k_{i,j}$ – price for next iteration (kn/kWh)
    - $RU^k_{i,j}$ – units for next iteration (kWh)

Amount $RC^k_{i,j}$ represent storage charge that must be left in storage for next period. If $RC^k_{i,j} > SC^k_{i,j}$ than storage must be additionally charged in j-th period.

B. Distributor model

We assume that distributor has prediction for day-ahead demands from connected consumers and producing capabilities from connected generators, both from power grid side (excluding microgrids). Base on this data distributor can define appropriate demands toward microgrids (given per periods). Those demands with price ranges are inputs into our model. Distributor is defined with parameters:

- constant (given as input) for each period j:
  - RE_j – required energy (kWh)
  - LP_j – minimal (lowest) price (kn)
  - HP_j – maximal (highest) price (kn)
- calculated in k-th iteration for each m_i and period j:
  - $OU_{i,j}^k$ – offered amount (kWh)
  - $OP_{i,j}^k$ – offered price (kn)
  - $OS_{i,j}^k$ – status: $MORE$ or $ENOUGH$

Status is as a signal to microgrids in which periods demands are not met ($OS_{i,j}^k = MORE$) and in which periods offered amount exceeds demand ($OS_{i,j}^k = ENOUGH$).

C. Additional system constraints

For simplification and feasibility of result of each iteration, additional constraints are placed upon system model.

1) Fairness: In our model of distributor we defined that for each period, energy unit prices must be the same for each microgrid, i.e.:

$$OP_{1,j}^k = OP_{i,j}^k, \quad \forall i, j, k \tag{1}$$

Condition (1) doesn't affect microgrid's model. Used microgrid's model can be used in systems when previous condition isn't valid, i.e. where different microgrids gets different prices for same period.

2) Feasibility I: Amounts in offers for next iterations must not exceed units provided in replies, i.e.:

$$OU_{i,j}^{k+1} \leq RU_{i,j}^k, \quad \forall i, j, k \tag{2}$$

Therefore, initial offers in first iteration have:

$$OU_{i,j}^1 = 0, \quad OS_{i,j}^1 = MORE, \quad \forall i, j \tag{3}$$

Initial price must be from interval $[LP_j, HP_j]$. In our implementation we used middle value $(LP_j + HP_j)/2$.

3) Feasibility II: Cumulatively, offers for same period must not exceed required amount for that same period, i.e.;

$$\sum_{i=1}^{N} OU_{i,j}^k \leq RE_j, \quad \forall j, k \tag{4}$$

4) Feasibility III: The microgrid in its reply must not offer more energy that it can sell, i.e.:

$$RU_{i,j}^k \leq SC_{i,j}^k + PE_{i,j} - CE_{i,j} - RS_{i,j}^k, \quad \forall i, j \tag{5}$$

5) Feasibility IV: When status in particular reply is $ACCEPT$, amount in that reply must not be lower that amount in corresponding offer, and price must not be higher than one in same offer:

$$RP_{i,j}^k \leq OP_{i,j}^k, \quad \forall i, j \quad | \quad RS_{i,j}^k = ACCEPT \tag{6}$$

$$RU_{i,j}^k \geq OU_{i,j}^k, \quad \forall i, j \quad | \quad RS_{i,j}^k = ACCEPT \tag{7}$$

No restriction on amount and price is set when offer is rejected.

6) No buying from power grid: If local consumers can't be satisfied with local production and storage management we remodel microgrid with local consumers we can satisfy locally. Additionally energy must be bought, but we do not model this process. Algorithm 1 presents microgrid validation and remodel procedure. We assumed such procedure was performed on microgrids before negotiation process begun.

Algorithm 1 Discard excess consumers' demand in m_i

```
for j = 1 to P do
  if SC_{i,j}^k + PE_{i,j} < CE_{i,j} then {not enough energy?}
    CE_{i,j} = SC_{i,j}^k + PE_{i,j}          {remodel consumers}
  end if
  if j < periods then {calculate storage charge for j+1}
    SC_{i,j+1}^k = SC_{i,j}^k + PE_{i,j} - CE_{i,j}
    if SC_{i,j+1}^k > ST_i then {can't overcharge}
      SC_{i,j+1}^k = ST_i
    end if
  end if
end for
```

III. The Negotiation Algorithm

Negotiations shown on Fig. 2 are initiated by distributor, who creates offers and sends them to microgrids. Each microgrid gets its own offers, one for each period. After evaluating offers, microgrids send back replies, one for each period.

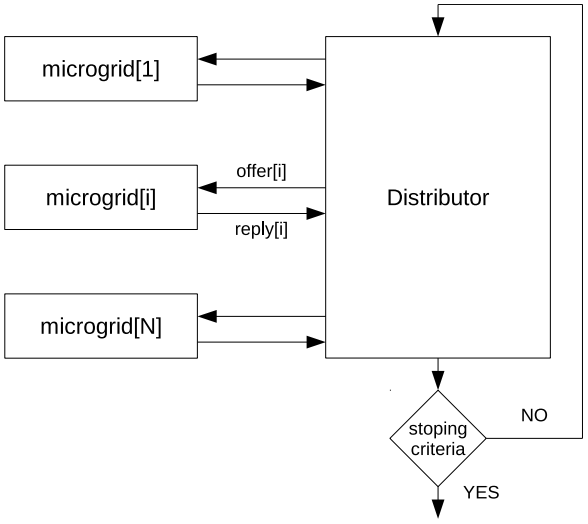


Fig. 2. Price negotiation process

A. The microgrid's algorithm

Microgrids evaluate offers separately, each microgrid looks at offers sent only to him. Main optimization criteria for a microgrid is maximization of profit from energy sold. However, the microgrid must first take into account its own local energy consumptions (per period). Units to sell are calculated iteratively: starting from period with the highest price per unit in offers. All surplus from that period and surplus from periods before that can be saved in storage and not already reserved can be included in offer for that period.

After any change in energy for sale, reservations on storage are first calculated, as presented in Algorithm 2.

Algorithm 2 $Update\_storage\_and\_reservations(i)$

for $j = 2$ **to** $periods$ **do** {calculate $SC^k_{i,j}$}
  $SC^k_{i,j} = SC^k_{i,j-1} + PE_{i,j-1} - CE_{i,j-1} - RU^k_{i,j-1}$
  if $SC^k_{i,j} > ST_i$ **then** {can't overcharge}
    $SC^k_{i,j} = ST_i$
  end if
end for
for $j = periods - 1$ **to** 1 **do** {calculate $RC^k_{i,j}$}
  $RC^k_{i,j} = CE_{i,j+1} + RC^k_{i,j+1} + RU^k_{i,j+1} - PE_{i,j+1}$
  if $RC^k_{i,j} < 0$ **then**
    $RC^k_{i,j} = 0$;
  end if
end for

First loop in Algorithm 2 calculate storage state after each period, accounting: initial storage state, production capability, local consumers and already allocated energy for selling. Second loop calculates amount of energy that must be left in storage for next period. This information is required in offer evaluation algorithm, Algorithm 3.

After optimization process (maximizing cumulative sell price), remaining units in periods are offered as well, because they might be requested in next iteration. Even if they are not requested, the microgrid can't use them (store for later) so its better to offer them anyway – distributor might ask for it in next iteration.

B. The distributor's algorithm

The distributor's behavior is simulated with simple model, presented in Algorithm 4. The main objective of algorithm is to get as much energy as can be obtained per period, up to defined value. Secondary objective is to minimize total cost per period. Used rules to achieve objectives were simple: when offered amount is below required amount – price is raised (for that period), and vice versa, when offered amount is equal or above required amount – price is lowered.

Optimization is stopped when no change is detected in successive iterations or when predefined maximum number of iterations is reached (for situations where algorithm don't converge), as shown in Algorithm 5.

Algorithm 3 Optimization algorithm for m_i

for $j = 1$ **to** P **do** {reset replies}
  $RS^k_{i,j} = REJECT$ {status}
  $RP^k_{i,j} = \max(MP_{i,j}, OP^k_{i,j})$ {price}
  $RU^k_{i,j} = 0$ {units}
  $sort[j] = j$
end for
for $l = 1$ **to** P **do** {optimize for period $sort[l]$}
  for $n = l + 1$ **to** P **do** {search for higher price}
    if $OP^k_{i,sort[l]} < OP^k_{i,sort[n]}$ **then**
      $\text{swap}(sort[l], sort[n])$ {swap values}
    end if
  end for
  $j = sort[l]$
  $Update\_storage\_and\_reservations(i)$
  if $OP^k_{i,j} \geq MP_{i,j}$ **then**
    $available = SC^k_{i,j} + PE_{i,j} - CE_{i,j} - RC^k_{i,j}$
    if $available \geq OU^k_{i,j}$ **then**
      $RS^k_{i,j} = ACCEPT$
      $RU^k_{i,j} = OU^k_{i,j}$
    end if
    if $OS^k_{i,j} = MORE$ **then**
      $RU^k_{i,j} = available$
    end if
  end if
end for
for $j = 1$ **to** P **do** {offer remaining power in periods}
  $Update\_storage\_and\_reservations(i)$
  $RU^k_{i,j} = SC^k_{i,j} + PE_{i,j} - CE_{i,j} - RC^k_{i,j}$
end for

IV. Experimental results

Proposed algorithm is evaluated on many examples. However, since single example requires large data set, only results from single example are presented in this paper. Microgrids and distributor parameters are just example, not chosen from real environment. In presented example ten microgrids are used. Each microgrid has storage capability for storing 50 kWh energy, but various predicted production capabilities during a day. In our simulation price change is limited to five discrete values defined by distributor separately for each period: lowest, low, normal (initial), higher and the highest, linearly distributed between $[LP_j, HP_j]$.

Cumulative input values for all ten microgrids and distributor are shown on Fig. 3 with: required energy from distributor (RE), production potential of microgrids (PE), local microgrid demands (LE) and initial price per period (IP). Initial price is set to roughly follow distributor's demands, but also taking into account microgrids' production capabilities, and to force

Algorithm 4 Optimization algorithm for distributor

for $j = 1$ **to** P **do** {for each period}
  $accepted = reserve = 0$
  for $i = 1$ **to** $microgrids$ **do** {sum offers}
    if $RS_{i,j}^k = ACCEPT$ **then**
      $accepted \mathrel{+}= OU_{i,j}^k$
      $OU_{i,j}^{k+1} = OU_{i,j}^k$ {offer to buy same amount again}
    else
      $OU_{i,j}^{k+1} = 0$
    end if
    if $RP_{i,j}^k \leq OP_{i,j}^k$ **then** {reply acceptable?}
      $reserve \mathrel{+}= RU_{i,j}^k$
    end if
  end for
  if $reserve < RE_j$ **then** {not enough? raise price}
    $new\_price = \min(OP_{1,j}^k + increment, HP_j)$
    $new\_flag = MORE$
  else {have enought, lower price}
    $new\_price = \max(OP_{1,j}^k - decrement, LP_j)$
    $new\_flag = ENOUGH$
  end if
  for $i = 1$ **to** $microgrids$ **do** {update price and flags}
    $OP_{i,j}^{k+1} = new\_price$
    $OS_{i,j}^{k+1} = new\_flag$
  end for
  $need = RE_j - accepted$
  $change =$ **true**
  while $need > 0$ **and** $change$ **do** {increase offers}
    $change =$ **false**
    for $i = 1$ **to** $microgrids$ **do**
      if $need > 0$ **and** $OU_{i,j}^{k+1} < RU_{i,j}^k$ **and**
      $(RP_{i,j}^k \leq OP_{i,j}^k$ **or** $RP_{i,j}^k \leq OP_{i,j}^{k+1})$ **then**
        $OU_{i,j}^{k+1} = OU_{i,j}^{k+1} + 1$
        $need = need - 1$
        $change =$ **true**
      end if
    end for
  end while
end for

Algorithm 5 Stopping criteria

$change = false$
for $i = 1$ **to** N **do**
  for $j = 1$ **to** P **do**
    if $RS_{i,j}^{k-1} \neq RS_{i,j}^k$ **or** $RP_{i,j}^{k-1} \neq RP_{i,j}^k$ **or**
    $RU_{i,j}^{k-1} \neq RU_{i,j}^k$ **or** $OU_{i,j}^{k-1} \neq OU_{i,j}^k$ **or**
    $OP_{i,j}^{k-1} \neq OP_{i,j}^k$ **or** $OS_{i,j}^{k-1} \neq OS_{i,j}^k$ **then**
      $change = true$
    end if
  end for
end for
if $k \geq k_{MAX}$ **or** $change = false$ **then**
  STOP
end if

them to save as much as they can for peak periods.

Presented algorithm produced results shown on Fig. 4 with: energy offered to distributor with acceptable price (RU), storage charge at period beginnings (SC) and final price difference from initial one (OP). We also put required energy (RE) on this figure to easier compare offers with demands.

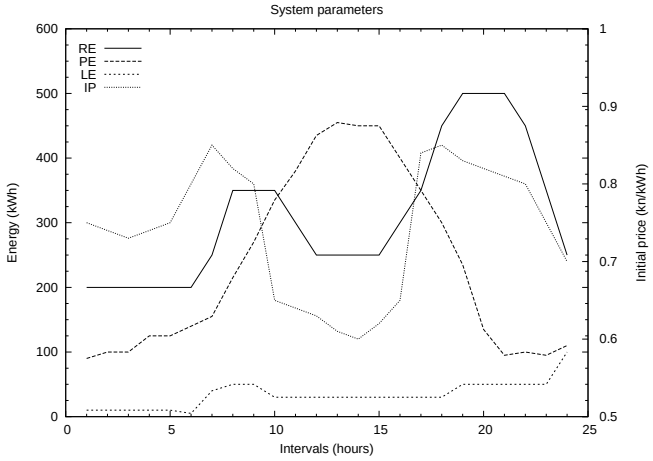


Fig. 3. Cumulative values per period for: required energy (RE), production potential (PE), local microgrid demands (LE) and initial price (IP)

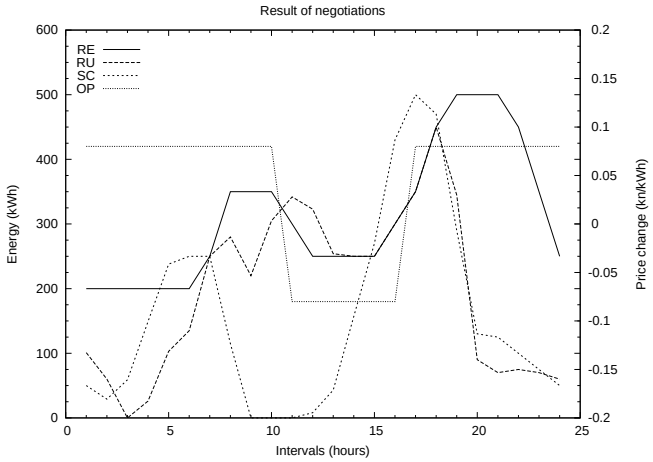


Fig. 4. Cumulative values per period for: required energy (RE), offered energy (RU), storage charge(SC), and price change from initial one (OP)

Analyzing presented results, we can extract several properties of proposed algorithm. Firstly, it behaves as expected: in periods where offered amount (RU) was not sufficient price reached the highest value, while in periods that offers exceed required amount (RE) price was lowered to the lowest possible value (periods 11 to 16). Periods with the highest price (7th and 18th) are covered as much as possible (completely in this example) using storage for difference between required and produced amount. Therefore, periods that precede ones with higher price may not be completely covered because storage was charged.

V. Possible improvements

Major problem with presented algorithm is in its stability. Single price change may trigger significant change in replies since microgrids sort periods by price and optimize periods in that order. We mitigated this problem by slowing price changes – enabling them only once in a few consecutive iterations, and by disabling of price lowering after predefined number of iterations. Those mitigations are not shown in Algorithm 4.

Fairness and cheating scenarios aren't analyzed. However, since single microgrid can't predict when negotiations will be completed, it should offer everything as soon as possible (as in proposed microgrid's algorithm). Otherwise, distributor might buy larger amounts from other microgrids (included behavior in proposed distributor's algorithm).

As already mentioned, optimality is not guaranteed with our algorithm. Within examples we used we found (not presented here) suboptimal solutions. We believe that better results might be achieved if distributor model (algorithm) is upgraded. However, even that might not guarantee optimal results for all cases.

Simple models for microgrids and distributor are used. In real application many parameters might be different and more complex. For example, minimal price per energy unit at which a microgrid is willing to sell energy might be calculated with a complex formula that take into account storage charging and discharging cost. Also, distributor may not offer all microgrids same price – price can be negotiated separately per microgrid. Used simplicity provides easier upgrade of particular model, distributor or microgrid. As long as constraints 2-6 from section II are met in all used algorithms in all nodes, every node might run different algorithm. That means that some microgrids might use one algorithm, other another, and so on. Same goes for distributor.

Presented algorithm is iterative and synchronous: distributor sends offer and collects all replies before evaluating them. However, algorithm could be adapted for asynchronous usage if constraint 1 is not enforced. In that case, communication between distributor and single microgrid could be achieved with three messages: 1) reply from microgrid, 2) offer from distributor and 3) a new reply from microgrid. All messages must conform to constraints 2-6. Microgrids replies, first and second one could be generated with same algorithm as presented. Distributor's algorithm must be adapted since price in offers for same period might not be the same to different microgrids (e.g. new_price could be replaced with $OP_{i,j}^{k+1}$).

What if prediction from any side proved to be incorrect? If we limit real-time adaptation only to energy redistribution from that moment forward, and ignore penalties imposed on node with wrong prediction that reflects on whole system (e.g. penalties could be handled as exception, separately), presented algorithm could be adapted even for real-time corrections. For example, algorithm will use current state of system (storage charge at that moment) as initial state, already negotiated prices as minimal prices (that can only increase), and only future periods in evaluations.

VI. Conclusion

In this paper we presented a simple algorithm for day-ahead negotiation between microgrids and a distributor. Separate algorithm is given for a microgrid and separate for a distributor. Its assumed that all entities in negotiations have accurate prediction on energy requirements and production, given per period.

Presented algorithm produces expected solutions – profit optimization for microgrids' surplus energy. Algorithms are simple, both microgrid's, which was primary objective, and distributor's, that was created for simulation purposed (and thus not fully optimized).

Stability problem was mitigated on distributor's side. When mitigating elements are activated stable solution is obtained within next ten to fifteen iterations. Therefore, at least twenty iterations should be put as upper limit, if one is required at all. In simpler examples, stability was reached within only four iterations.

Negotiation results weren't always optimal, even if Nash equilibrium was reached. Part of that problem lies with distributor's algorithm, which must be improved. However, optimal solution guarantee (as one presented in [4]) is generally hard to provide in distributed algorithms. Therefore, we are satisfied with a occasional 'only good' result which is provided by presented algorithm.

Simplicity of algorithm provides several advantages. Firstly, it enables algorithm implementation even on simplest hardware (microgrid's micro-controller). Secondly, adaptation for specific microgrid or distributor is simpler – not all nodes must run same algorithm, as elaborated in section V.

References

[1] E. Commission. (2010) Europe 2020. [Online]. Available: http://ec.europa.eu/europe2020/

[2] I. Otero-Novas, C. Meseguer, C. Batlle, and J. Alba, "A simulation model for a competitive generation market," *Power Systems, IEEE Transactions on*, vol. 15, no. 1, pp. 250–256, Feb 2000.

[3] J. Contreras, O. Candiles, J. de la Fuente, and T. Gomez, "Auction design in day-ahead electricity markets," *Power Systems, IEEE Transactions on*, vol. 16, no. 1, pp. 88–96, Feb 2001.

[4] I. Atzeni, L. Ordonez, G. Scutari, D. Palomar, and J. Fonollosa, "Noncooperative day-ahead bidding strategies for demand-side expected cost minimization with real-time adjustments: A gnep approach," *Signal Processing, IEEE Transactions on*, vol. 62, no. 9, pp. 2397–2412, May 2014.

MIA: A Multi-Purpose Intelligent Assistant

Mario Konecki *, Matjaž Gams ** and Nikola Ivković *
* Faculty of Organization and Informatics, Varaždin, Croatia
** Jožef Stefan Institute, Ljubljana, Slovenia
mario.konecki@foi.hr; matjaz.gams@ijs.si; nivkovic@foi.hr

Abstract - Business systems as well as corresponding websites are constantly growing in their complexity and it is harder and harder to use them and find desired functionality or information. Intelligent assistants can be used as a mean of enabling users to find information they seek and to get the answers they require in a way that is quicker and more suited for wider types of users compared to traditional approach because they are not required to know the system in detail but rather to communicate with appropriate intelligent assistant in a more natural language which is similar to talking with their colleagues. Intelligent assistants should be as easy to develop and maintain as possible because of the need to constantly improve them in order to get the intelligent assistants to provide the latest information. In this paper a multi-purpose intelligent assistant (MIA) is proposed and its architecture as well as the perception of various business systems about the usage of such assistant are presented and discussed.

Keywords **- business systems, information systems, websites, intelligent assistant, multi-purpose**

I. INTRODUCTION

Business systems are becoming more and more complex everyday and the same trend is present in the field of information systems that support these business systems. All kind of software is getting richer in features and thus more demanding from the user's point of view. Because of this exponential growth of software complexity a need for new user interfaces with more advanced technology becomes something that is more and more necessary in order for users to perform their software activities in a satisfactory way that is easy and quick enough to be usable [1]. This is particularly true if everyday activities are involved and that is the area where intelligent assistants are coming into focus.

There are a lot of aspects that intelligent assistants can aid in but the first and the foremost step for users is to get to know the system they want to use. The easiest way to learn about the system is to be able to ask something about the system in a way one would ask a colleague. This implies the usage of natural language to communicate with the system which enables users to have much easier and quicker way of getting information compared to browsing through the system's manual. Since getting the information about the system is one of the most common necessities of new users it would be beneficial to have a system that would enable simple creation of such assistants that would aid users in some particular areas of interest and that would be able to process natural language. In order to provide such concept a MIA (Multi-Purpose Intelligent Assistant) is proposed as an adequate solution. MIA provides its users with a suitable and accessible way of asking a variety of questions, in various forms, about the system of interest and it makes using the system that incorporates MIA as its intelligent assistant easier and more convenient to use. MIA is a multi-purpose intelligent assistant which means that it is not bound to just one particular area of use. MIA enables multiple configurations of its environment and creation of multiple instances that deal with different areas or systems of interest which enables quick creation of specific intelligent assistants.

Intelligent assistants are something that is gaining interest. Research results have shown that 81% of visitors would like to have an intelligent virtual assistant provided that it would improve their online user experience and 72% of visitors stated that they would welcome intelligent virtual assistants as a part of every website and that they would spend more time on a website that would incorporate such assistant [2]. These intelligent assistants are in fact intelligent agents [3] of a specific class and characteristics. Intelligent agents incorporate elements of artificial intelligence (AI) in order to provide users with a greater amount of helpful actions and with a more purposeful and more satisfactory experience. These concepts of AI enable intelligent assistants to aid their users in tasks that are difficult or time consuming for users and that require some form of intelligent behavior and reasoning. AI systems provide support for various tasks that include all from answering questions to helping in coordination or production, information gathering and management, large-scale planning and scheduling of shipments or autonomous robots and robotic assistants [4]. One of more known examples of working AI systems is Deep Blue [5].

In this paper MIA is proposed as a multi-purpose assistant that can be customized with specific rules and applied in various areas of interest as a quick and effective mean of enabling users of various systems to be able to get a quick and concise information about the system which promotes better understanding of the system in a shorter amount of time in a way that is close to new users since processing of natural language is incorporated in order to create an experience that is similar to the one that the users would have when talking with real people and agents of various helping services. In this way the path of learning about the system becomes more convenient and interesting for users which makes them consequently more willing to spend a greater amount of time in dealing with the system.

II. INTELLIGENT ASSISTANTS AND INTELLIGENT AGENTS

Intelligent assistants are based on intelligent agents with a specific set of features and characteristics. Intelligent agent in general is an entity that perceives its environment through sensors and acts upon that environment by using its effectors as can be seen in Fig. 1 [6]. Intelligent agents and intelligent assistants are a result of gradual development that includes five generations of AI systems: expert systems, joint cognitive systems, intelligent systems, intelligent assistant systems and the newest generation that includes context-based intelligent assistant systems [7]. There are many existing classification of intelligent agents. Franklin & Graesser have given a taxonomy of intelligent agents that is shown in Table I [8].

TABLE I. INTELLIGENT AGENTS TAXONOMY (FRANKLIN & GRAESSER)

Property	Other Names	Meaning
reactive	(sensing and acting)	responds in a timely fashion to changes in the environment
autonomous		exercises control over its own actions
goal-oriented	pro-active purposeful	does not simply act in response to the environment
temporally continuous		is a continuously running process
communicative	socially able	communicates with other agents, perhaps including people
learning	adaptive	changes its behavior based on its previous experience
mobile		able to transport itself from one machine to another
flexible		actions are not scripted
character		believable "personality" and emotional state

Nwana gives the following typology [9]:

- Collaborative agents
- Interface agents
- Mobile agents
- Information/Internet agents
- Reactive agents
- Hybrid agents
- Smart Agents

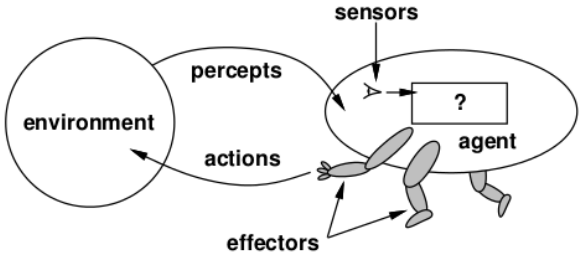


Figure 1. Intelligent agent functionality

Another classification of intelligent agents is given by Müller who distinguishes five categories of intelligent agents [10]:

- Reactive agents
- Deliberative agents
- Interacting agents
- Layered approaches
- Others (believable agents, softbots, Internet software agents)

Müller also proposes an intelligent agents taxonomy based on application areas according to the following two dimensions [10]:

- First dimension
  - Hardware agents
  - Software agents
- Second dimension
  - Autonomous agents
  - Multiagents
  - Assistant agents

Hardware agents are those agents that have some kind of physical form and interaction with physical environment by using sensors as input and effectors as a mean of output. These agents also use some kind of software that provides some form of logical control of these agents. Software agents are on the other hand purely software based and can be viewed as various computer programs that interact with environments which can be either of real or virtual nature. Autonomous agents refer to agents that can control their actions in greater or lesser amount and are focused on not just their own actions but also on their environment. Multiagents are agents that posses knowledge about other agents and have the ability to collaborate with them in order to achieve some goal. Assistant agents are agents that are mainly involved in interaction with another agent which can be and is in many cases a human agent who uses this kind of assistant agent. Based on given dimensions Müller proposes the following taxonomy of intelligent agents [10]: autonomous hardware agents, autonomous software agents, hardware assistant agents, software assistant agents, hardware multiagents and software multiagents.

Autonomous hardware agents are self-controlled agents that are usually involved in planning and implementing various activities as well as in controlling some physical environment by using their physical sensors and effectors. Autonomous software agents are program-based virtual agents that can control their own actions and as such influence their software environment. Hardware assistant agents are physical agents which are designed to help their users in some domain of physical environment. Software assistant agents are agents that are program-based and are designed to help their users in some domain of virtual environment. Hardware multiagents are physical agents that act in collaboration with other agents in physical environment in order to achieve some common

goal. Software multiagents are program-based agents that can communicate with other software agents in some virtual environment in order to achieve a mutual goal.

Intelligent agents can be developed by using a variety of approaches and are used in many different application areas [11]:

- Process Control
- Manufacturing
- Air Traffic Control
- Information Management
- Electronic Commerce
- Business Process Management
- Patient Monitoring
- Health Care
- Games
- Interactive Theater and Cinema

As mentioned before intelligent assistants are in fact a kind of intelligent agents. The term intelligent assistant system can be defined as an intelligent system that mediates interactions and aids its users in using some machine system as is shown in Fig 2. [12]. Intelligent assistants in many cases have some form of avatar that makes these assistants more distinctive and more interesting to use. Intelligent assistants' avatars are designed according to the area and purpose for which some particular intelligent assistant is being used. These avatars can be of either static or dynamic visual nature. This includes various images or animations that are used to represent the avatar. Research has shown that 59% of users have no preference when talking about the usage of dynamic or static avatars while 30% of users prefer dynamic and 11% of users prefer static avatars [2].

There are many examples of intelligent assistants in various areas of use. Some of these assistants are [13] [14] [15] [16] [17] [18] [19] [20]: PexA (an intelligent assistant for aiding in task and time management), WEBSELL (an intelligent assistant for aiding users in web sales), Triton (an intelligent assistant for aiding users in learning and using the air travel planning system), Office Assistant (an intelligent assistant that aids office visitors and helps in management of office's schedule), CAT (Composition Analysis Tool) (an intelligent assistant for aiding users in creation of complete and consistent workflows), TAC (The Architect's Collaborator) (an intelligent assistant for aiding architects in conceptual design), Patient Advocate (an intelligent assistant for patient-centered health care) and CAD (an intelligent assistant for aiding in interaction with computer-aided-design system).

When talking about intelligent assistants there is another important field of their usage and that is the field of education. Intelligent assistants can help students to learn but also the teachers to teach through various features such as: monitoring, error detection, answering questions, alerting teachers or generating feedback [21].

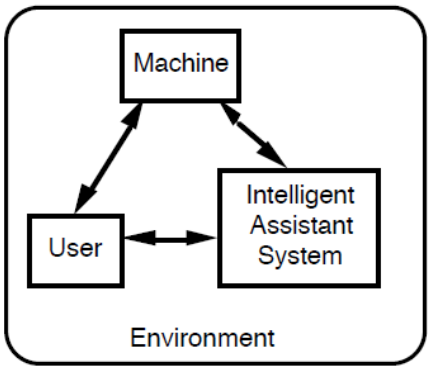


Figure 2. Intelligent assistant system

These kinds of assistants are also called intelligent educational systems and can be classified into three different classes based on the degree of autonomy they provide [21]: (1) Those that are fully autonomous and replace the teacher; (2) Those that assist the teacher, while remaining autonomous in their use: the student can practice without the teacher being present; (3) Those that assist the teacher and are not autonomous: the teacher must remain present during their use.

III. MIA ARCHITECTURE AND APPLICATION

MIA (Multi-Purpose Intelligent Assistant) has been designed as a multi-purpose assistant that can be applied in many different areas. MIA can be classified as an communicative (according to [8]), reactive (according to [9]) and reactive software assistant agent (according to [10]). MIA supports answering on various questions in a robust way that includes support for many different forms of essentially the same question. MIA also enables variations in answers to given questions which results in more natural and real experience. Users communicate with MIA by using a completely natural language just as they would communicate with any human agent. MIA is a multi-purpose assistant which means that it provides a well-defined and easy-to-use way of defining the rules and knowledge that MIA is supposed to have in order to be used in some particular areas of interest. MIA can be generally defined as:

$$\text{MIA: } q \rightarrow a$$

where q is a question and a is an answer. In the case of MIA there are many possible questions and their forms (Q) and also many possible answers and their forms (A). Every question also can have multiple variations of answers (answer can be omitted) and the same answer can be given to various forms of essentially the same question so the following applies:

$$q \in Q;\ a \in A;\ q \rightarrow A'^*;\ Q'^+ \rightarrow a$$

The former definition defines MIA question as $q \in Q$ and MIA answer as $a \in A$. The same question q can have zero or more forms of basically the same answer (A'^*) and one or more forms of basically the same question (Q'^+) can have the same answer a. MIA is developed as a web service which can be used alone or as a part of some existing web page or web application. MIA enables two important aspects that are required in order to customize

MIA for some particular purpose. The first element that needs to be customized is the avatar that represents the assistant. MIA support uploading of customized avatar design that is then shown in the intelligent assistant window as its representation. Since the users communicate with MIA by writing their questions, rules that define these questions and their answers must be produced before MIA can be usable for some particular purpose. To define these rules it is sufficient to use MIA settings where the rules can be written by using a very intuitive and simple interface which consists of a part where the user enters a set of keywords, which can be prioritized and paired with other keywords, and a part where one or more forms of the same answer can be entered. In order for questions to be robust MIA supports defining of various forms of the same question by identifying different keywords. It also enables developers to define which keywords are alternative and which have to be found together. In order to produce more convincing experience there is also a possibility of defining more answers for the same question that are randomly presented. Formal grammar for defining of question rules of MIA in EBNF form [22] is given below:

*rules = rule, (2 * eol, anotherrule) | /* nothing */ ;*

anotherrule = rule, anotherrule | / nothing */ ;*

word = {alphabetchar | digit}- ;

alphabetchar = 'A' | 'B' | 'C' | ... | 'X' | 'Y' | 'Z' | 'a' | 'b' | 'c' | ... | 'x' | 'y' | 'z' ;

digit = '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9' ;

eol = ['\r']'\n' ;

divider = ',' ;

ws = {' ' | '\t'}- ;

rule = part, (eol, moreparts) | (nothing *) ;*

part = word, {divider, ws, word} ;

moreparts = part, moreparts | (nothing *) ;*

As can be seen from the given grammar the MIA question rules are simply defined as a set of alternative keywords that are divided with the "," character. Additional keywords that have to be found together are entered as additional lines. MIA is web based and administrated through developed web interface in order to provide more convenient way of developing one's own assistant. MIA question rules enable MIA to be robust and insensitive to various forms of essentially the same question. For example, all following forms of the same question will result in the same answer:

- What is the weather in London?
- I want to know about the weather in London.
- I am interested in weather forecast of London.
- Please give me London's weather forecast.

In this way the user is able to communicate with MIA in a relaxed non-formal way, just as if it was some real person-to-person conversation. This promotes easier and more convenient way of getting the right information.

Figure 3. Customized MIA

The final result of MIA customization is a specialized intelligent assistant that is defined by its own avatar and a customized set of question-answer rules which make MIA applicable to some particular area of use. An example of MIA assistant is shown in Fig. 3. MIA also supports variations in the forms of answers which results in users getting a feeling that MIA is not predictable. This feature creates effect of more intelligent assistant comparing to just one form of answer. Another feature of MIA is that it supports opening of additional web resources in order to clarify given answers even further. An example of this feature is shown in Fig. 4. MIA can be used in many different areas but one of them that is particularly interesting are business systems that can benefit from easier way of getting help and the right information about the business system. In order to find out about the general attitude of business systems towards the usage of intelligent assistants a short research in the form of questionnaire was conducted on 22 business systems. All business systems where of small or medium size and were using various information systems as a part of their everyday business (finance, insurance, trading, web development, etc.). The results of conducted research are given in Table II.

Figure 4. MIA question and answer example with opening of additional web resource

TABLE II. BUSINESS SYSTEMS' ATTITUDE TOWARDS INTELLIGENT ASSISTANTS

Questionnaire item	Mean	Std. dev.
Intelligent assistants seem as an interesting technology.	4,14	0,68
We think that intelligent assistant would help our customers to find information quicker and in easier manner.	3,87	0,92
We think that intelligent assistant would help our employees to learn about new information systems quicker and easier.	3,53	1,19
We would use intelligent assistant as a part of our business system support	3,64	0,83
We would recommend intelligent assistants to our business colleagues.	3,77	1,34

As can be seen in Table II, the conducted research among business system has shown that they perceive intelligent assistants as an interesting technology that they would recommend to others and that they find it useful for both customers and employees. However, research has also shown that although business systems have shown positive attitude towards usage of intelligent assistants their answers weren't as positive as expected and this can be explained with the fact that intelligent assistants represents a new technology that is still rather unknown to their potential users so one step in the right direction would be to make this technology more familiar to its potential users.

IV. CONCLUSION

Business systems and their information systems are getting more and more complex. Increasing complexity of software asks for new technologies and new user interfaces that will make using this kind of software quicker and easier. Intelligent assistants provide their users with many aiding features that can make usage of various systems more convenient. One of possible applications of intelligent assistants is in helping users to learn about the system and to get the desired information about the system. In this paper MIA (Multi-Purpose Intelligent Assistant) is proposed as a generic intelligent assistant that can be applied in many different areas. MIA provides its users with possibilities of customization in both graphical and logical sense. MIA users are able to create their own MIA avatar and customize MIA behavior by defining a set of question-answer rules that will make MIA applicable in the desired area of application. In this way MIA provides an easy and effective way of creating one's own intelligent assistant for some particular purpose. In order to find out about the attitude of business systems towards intelligent assistants a short research has been conducted which has shown that business system find this technology interesting but need additional information in order to understand its full potential. Adding new features to MIA and a more detailed research about its effectiveness will be a part of future research.

REFERENCES

[1] D. Guzzoni, A. Cheyer, and C. Baur, "Active, a platform for building intelligent software", In Computational Intelligence pp. 126-131, 2006.

[2] Artificial Solutions , The top traits of intelligent Virtual Assistants, available at: http://www.artificial-solutions.com/wp-content/uploads/protected/wp_top-traits-of-intelligent-virtual-assistants.pdf, accessed: 5.2.2015.

[3] M. Wooldridge, and N. R. Jennings, "Intelligent agents: theory and practice", The Knowledge Engineering Review, vol. 10, no. 2, 1995, pp. 115-152.

[4] G. Ferguson, and J. F. Allen, "TRIPS: An Integrated Intelligent Problem-Solving Assistant", In Proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 567-572, Madison, Wisconsin, 1998.

[5] S. Hamilton, and L. Garber, "Deep Blue's hardware-software synergy", IEEE Computer, vol. 30, no. 10, 1997, pp. 29-35.

[6] S. Russell, and P. Norvig, Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs, NJ, 1995.

[7] P. Brézillon, "From Expert Systems to Context-based Intelligent Assistant Systems: A testimony." The Knowledge Engineering Review, vol. 26, no. 1, 2011, pp. 19-24.

[8] S. Franklin, and A. Graesser, "Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents", In Intelligent Agents 3: Agent Theories, Architectures, and Languages, pp. 21-35, Springer-Verlag, 1997.

[9] H. S. Nwana, "Software Agents: An Overview", Knowledge Engineering Review, vol. 11, no. 3, 1996, pp. 205-244.

[10] J. P. Müller, "The Right Agent (Architecture) to Do the Right Thing", In Intelligent Agents V: Agents Theories, Architectures, and Languages, pp. 211-225, Springer-Verlag, 1999.

[11] N. R. Jennings, and M. Wooldridge, "Applications of Intelligent Agents", In Agent technology, pp. 3-28, Springer-Verlag, 1998.

[12] G. Boy, and T. R. Gruber, "Intelligent Assistant Systems: Support for Integrated Human-Machine Systems", In Proceedings of 1990 AAAI Spring Symposium on Knowledge-Based Human-Computer Communication, Knowledge Systems Laboratory, Computer Science Department, Stanford University, pp. 7-9, 1990.

[13] K. Myers, P. Berry, J. Blythe, K. Conley, M. Gervasio, et al., "An Intelligent Personal Assistant for Task and Time Management", AI Magazine, vol. 28, no. 2, 2007, pp. 47-61.

[14] P. Cunningham, R. Bergmann, S. Schmitt, R. Traphöner, S. Breen, et al., "WEBSELL: Intelligent sales assistants for the World Wide Web", KI - Künstliche Intelligenz 1, vol. 15, no. 1, 2001, pp. 28-32.

[15] J. R. Davies, A. S. Gertner, N. Lesh, C. Rich, C. L. Sidner, et al., "Incorporating Tutorial Strategies into an Intelligent Assistant", In Proceedings of the 6th International Conference on Intelligent User Interfaces, pp. 53-56, ACM, 2001.

[16] H. Yan, and T. Selker, "Context-Aware Office Assistant", In Proceedings of the 5th International Conference on Intelligent User Interfaces, ACM, pp. 276-279, 2000.

[17] J. Kim, M. Spraragen, and Y. Gil, "An Intelligent Assistant for Interactive Workflow Composition", In Proceedings of the 9th International Conference on Intelligent User Interfaces, pp. 125-131, ACM, 2004.

[18] K. Koile, "An Intelligent Assistant For Conceptual Design: Informed Search Using A Mapping Of Abstract Qualities To Physical Form", In Proceedings of Design Computing and Cognition '04, pp. 3-22, 2004.

[19] S. Miksch, K. Cheng, and B. Hayes-Roth, "An Intelligent Assistant for Patient Health Care", In Proceedings of the First International Conference on Autonomous Agents, pp. 458-465, ACM, 1997.

[20] O. St-Cyr, Y. Lespérance, and W. Stuerzlinger, "An Intelligent Assistant for Computer-Aided Design", In Smart Graphics: AAAI Spring Symposium, pp. 24-28, 2000.

[21] K. Yacef, "Intelligent Teaching Assistant Systems", In International Conference on Computers in Education, pp. 136-140), Auckland, New Zealand, IEEE, 2002.

[22] R. S. Scowen, "Extended BNF - A generic base standard", Technical report, ISO/IEC 14977, available at: http://www.cs.kent.edu/~durand/CS43101Fall2004/resources/iso-ebnf-paper.pdf, accessed: 5.2.2015.

GUIDL IA: An Intelligent Assistant for Aiding Visually Impaired in Using GUIDL

Mario Konecki *, Robert Kudelić * and Hristijan Gjoreski **
* Faculty of Organization and Informatics, Varaždin, Croatia
** Jožef Stefan Institute, Ljubljana, Slovenia
mario.konecki@foi.hr; robert.kudelic@foi.hr; hristijan.gjoreski@ijs.si

Abstract - GUIDL (Graphical User Interface Description Language) system has been proposed as an effective aiding technology that enables inclusion of visually impaired computer users into activities of graphical user interface design. GUIDL system implements many aiding concepts that make creation of graphical user interfaces suitable for visually impaired. In order to make GUIDL system even more accessible to visually impaired a GUIDL IA (Intelligent Assistant) is proposed as a mean of providing a more suitable and simpler way of getting help about various parts of GUIDL system. GUIDL IA enables visually impaired to ask questions in the form of natural language compared to browsing through instruction manual which makes getting the right information easier and quicker. GUIDL IA also enables visually impaired to hear the answer through integrated text-to-speech synthesizer that enables more natural pronunciation of given answers. In this paper GUIDL IA is presented and discussed along with the results about its effectives.

Keywords **- GUIDL, visually impaired, graphical user interfaces, intelligent assistant**

I. INTRODUCTION

Programming is hard and challenging for sighted students to learn because of its abstract nature and a specific way of thinking that it requires [1; 2; 3, 4]. The challenge is much more prominent when it comes to visually impaired students [5]. Visually impaired students experience difficulties in trying to imagine various programming structures and concepts. Nevertheless, visually impaired have shown interest in programming since the very beginning of computer industry [6; 7] and their interest has not disappeared [8]. Along with this interest various means of enabling visually impaired to use computers were developed. Most commonly used aiding tools for visually impaired were based on text-to-speech synthesizers that enabled visually impaired computer users to work with computer programs because all programs' interfaces were based on text. Some of the most known text-to-speech synthesizers are [7]:

- JAWS
- HAL Screen Reader
- COBRA
- Window Eyes
- Easy Web Browsing

These aiding tools also enabled visually impaired to participate in professional programming activities. The adoption of GUIs (Graphical User Interfaces) has brought a huge revolution to the world of computers and made computers much more accessible to a larger number of users. Unfortunately this progress has made things much more difficult for visually impaired. Existing aiding technology that was more than enough to support programming activities of visually impaired has become insufficient to support all new graphical concepts and to describe the graphical screens' context. Along with these problems additional problems have emerged regarding tools for programming and development which have become based on graphical controls and point-and-click actions that have become necessary in order to create graphical user interfaces and to define many other settings. Writing program code still wasn't so problematic since it was still based on text although some concepts and fragmentation of code that came along with new object-oriented programming languages and environments have made programming more complex.

Visually impaired programmers have become unable to participate as equal team members in all activities of software development especially in activities of graphical user interfaces development because textual definition of graphical user interfaces has become practically impossible and although some efforts in the area of aiding visually impaired to learn programming and perceive graphical elements have been made [9; 10; 11; 12; 13] none of them gave any long-term solution for creation of graphical user interfaces.

In order to enable visually impaired programmers to participate as equals in all activities of overall software development, including design of graphical user interfaces a GUIDL (Graphical User Interface Description Language) system with GUIDL language as its core part was developed as an aiding technology that is designed to overcome the obstacle of being unable to create graphical user interfaces in an adequate way for visually impaired [6; 7; 14]. GUIDL system is used as an aiding technology because it does not replace programming development environments. It is only used in a part of development that is dealing with graphical user interfaces creation which is done by using the GUIDL system upon which the rest of the development process is again performed in original development environment. In order to be usable GUIDL system had to meet the following requirements [7]:

- Easiness of usage
- Intuitive, simple and easy to understand syntax
- Independence of descriptive language
- Extensibility of descriptive language

GUIDL system has been designed to make creation of graphical user interfaces possible in a suitable manner for visually impaired programmers and to include them into all activities of programming development as equals. It is as such based on social model of aiding technology [15; 16]. Its GUIDL core language is also designed to be very simple and easy to use since easiness is one of the most important aspects of any aiding technology. In order to make learning about GUIDL system and its usage even easier an intelligent assistant is proposed that enables its users to ask a variety of questions about GUIDL system and to get a quick and clear answer opposed to going through manual page by page. In this way the usage of GUIDL becomes even more efficient and suitable for all visually impaired programmers. In this paper proposed intelligent assistant is described and a research about perception of visually impaired programmers regarding such solution is presented and discussed. This paper also gives some ideas for further research in this area.

II. GUIDL SYSTEM

GUIDL system as an aiding technology has been developed in order to enable visually impaired programmers to work in actual up-to-date technologies rather than having an isolated aiding ecosystem. GUIDL system with its GUIDL language provides easy and understandable way of creating graphical user interfaces that can be translated into desired programming language development environment format by the GUIDL system's mediators which in the end leaves visually impaired programmers with interfaces that can be included into development environments in which some actual projects are being developed. In this way visually impaired are actually included into real projects rather than just having a way to show their ideas in isolated environment without any actual contribution to the real projects. Conceptual model of GUIDL system functioning is shown in Fig. 1 [7; 14].

GUIDL system has been developed to be simple to use with simple syntax of GUIDL language which is designed to be easy and intuitive to use. In order to provide an aiding mean in its full sense GUIDL system includes a number of aiding concepts which are aimed to make creation of graphical user interfaces simple and intuitive for visually impaired. These aiding concepts include [14]:

- Predefined gradual sizes of forms
- Predefined gradual sizes of graphical elements
- Predefined width/height attribute values
- Division of forms into quadrants
- Possibility to position graphical elements into one of form quadrants

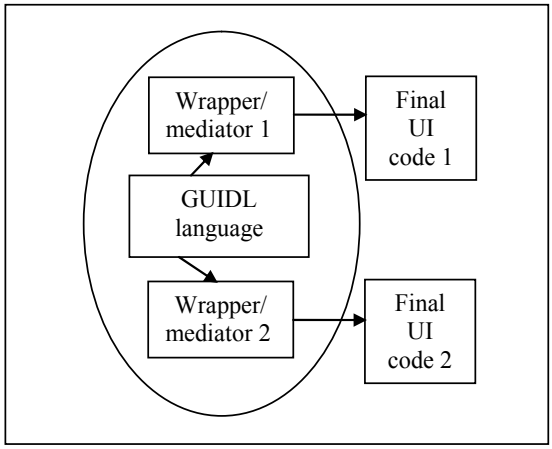


Figure 1. GUIDL system's conceptual model

- Possibility to define the position offset of forms
- Possibility to define the position offset of graphical elements
- Detection of problems with position of graphical elements (graphical element out of form boundaries)
- Automatic correction of problems with form dimension and position (form out of screen boundaries)

The aim of these aiding concepts is to catch those errors that occur on the logical level and are not of syntactical nature. These kind of errors can occur because it is very hard not to make them without seeing the overall screen context. By using these concepts visually impaired are able to create valid graphical user interfaces that do not have overlapping concepts, objects out of screen boundaries, or too small forms. Very important aspect of GUIDL system that is necessary in order for visually impaired to be able to organize objects on the screen is the possibility to divide the form into quadrants in which a certain object can then be placed. This makes organizing of form's elements much easier which in the end results in better and more polished interface. These features combined with simple and accessible GUIDL language make this aiding technology usable and suitable for visually impaired. A part of GUIDL language grammar in EBNF form is shown below [14]:

- project = projectcode, controlname, form;
- projectcode = 'Project ' | 'project ';
- form = formcode, controlname, formattributes, [controldeclarations], formend;
- formcode = 'Frm ' | 'frm ';
- formend = ('End' | 'end'), [eol];
- controlname = qoute, word, qoute, eol;
- word = alphabeticcharacter, {alphabeticcharacter | digit};
- formattributes = frmcommonattributes, windowstateattribute, {colorattribute};

- frmcommonattributes = textattribute, frmrestcommonattributes;
- frmrestcommonattributes = frmsizeattribute, locationattribute;
- frmsizeattribute = sizecode, (frmsize | frmwidth, ws ,frmheight), eol;
- locationattribute = locationcode, xposition, ws, yposition, eol;
- frmsize = 'frmsize1' | 'frmsize2' | 'frmsize3'
- locationcode = 'Location = ' | 'Location=' | 'location = ' | 'location=';
- xposition = 'left' | 'center' | 'right';
- yposition = 'top' | 'middle' | 'bottom';

Although GUIDL system has been designed to be simple to use some effort is needed to learn its rules and concepts. In order to provide visually impaired with something more than just textual or Braille manual an intelligent assistant is proposed as an aiding solution for helping visually impaired to learn and use GUIDL system.

III. GUIDL INTELLIGENT ASSISTANT

GUIDL IA (Intelligent Assistant) is proposed as a suitable aiding mean that is aimed at making GUIDL system easier to learn and use. GUIDL IA is designed to appear in a small part of the screen and to provide an easy way for visually impaired programmers to ask questions about GUIDL system in a form of natural language that they use in everyday conversation. In order to make GUIDL IA simple to use its interface consists of just two fields, one for input of questions and one read-only field for displaying the answers. In this way it is very easy to change focus from one field to another by using the keyboard. The interface of GUIDL IA is shown in Fig. 2. Another feature that is implemented in GUIDL IA is a possibility of opening additional materials in order to clarify or give more details about some aspect of GUIDL system. This feature is shown in Fig. 3. Using GUIDL IA is quite simple. It is sufficient to just enter any GUIDL system related question by typing it and hit the enter key to get the textual answer.

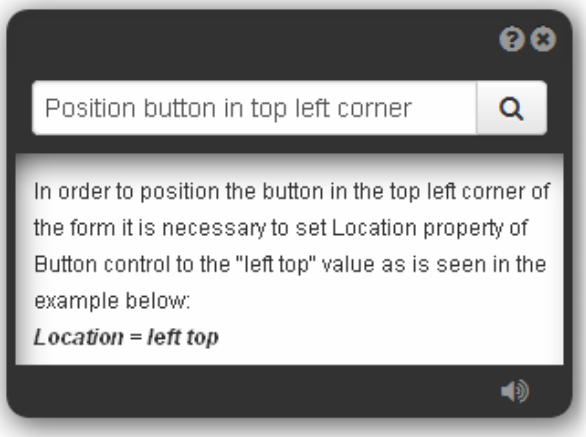


Figure 2. GUIDL IA interface

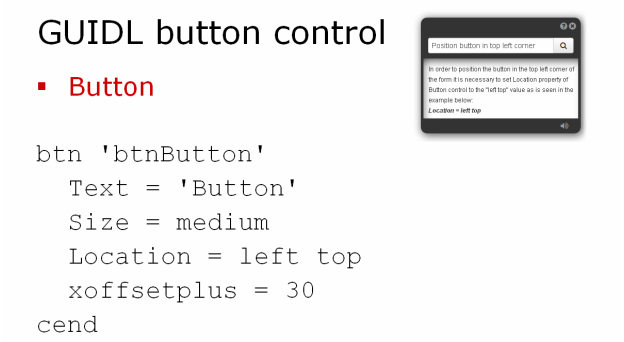


Figure 3. GUIDL IA answer and additional web resource

Since read-only field of GUIDL IA that displays the answers is sometimes not appropriate for a larger amount of information it is sometimes also convenient to open a web page or other web resource with additional information about the asked question and the given answer. In this way visually impaired are presented with more detailed information which can include a larger portion of example GUIDL code and in that way make it easier to understand certain elements of GUIDL system. GUIDL IA incorporates certain aspects of intelligent and expert systems that enable visually impaired programmers to use this assistant in a natural way just as they would communicate with their peers. This includes an extensive tolerance to various forms of questions. For example, all following forms of the same question are accepted and understood by GUIDL IA:

- Tell me how to create buttons?
- What is button and how it is used?
- I am not able to create a button.

Aside from extensive tolerance to different forms of questions GUIDL IA also supports variations of answers so that the same question can have many forms of answer that are randomly displayed which in the end results in somewhat more interesting form for GUIDL IA users. Another aspect that is built into GUIDL IA is a possibility to hear the answer in audio format. Converting text to speech is something that visually impaired computer users have been using for years and although this feature is supported by existing aiding technology which is mostly based on various text-to-speech synthesizers having its own variation of text-to-speech support leaves space for further customization and improvement of this feature and making it even more suitable for GUIDL IA. It is however still possible to use existing text-to-speech tools to convert textual GUIDL IA answers to audio format. Making the usage of GUIDL system even more simple with support of intelligent assistant has a goal to make GUIDL system even more effective and efficient in its main task of enabling visually impaired programmers to create graphical user interfaces.

IV. GUIDL IA EFFECTIVENESS

GUIDL IA has been included into GUIDL system in order to test its level of usefulness in making GUIDL system easier to use. The research was conducted among 24 visually impaired programmers who were given a

GUIDL system with GUIDL IA and were asked to try to do several tasks and report whether GUIDL IA has made using GUIDL system easier compared to using GUIDL system's standard documentation. Research about the experience of participants with GUIDL IA was conducted by using a questionnaire with several questions about their attitude towards provided intelligent assistant and its way of functioning. All answers to given questions were based upon the Likert scale (5 - I strongly agree, 1 - I strongly disagree). Research results are shown in Table I.

TABLE I. RESULTS OF CONDUCTED RESEARCH

Questionnaire item	Mean	Std. dev.
GUIDL IA helps me in using GUIDL system.	4,01	0,84
GUIDL IA makes getting help about GUIDL system quicker.	4,23	0,75
GUIDL IA makes using GUIDL system simpler.	3,89	1,12
GUIDL IA is an interesting aiding tool.	4,38	0,69
I find the feature of being able to ask questions in natural language to be effective way of getting help about GUIDL system.	4,11	0,93
I would recommend this kind of assistant to my peers.	4,04	0,72

The research results have shown that visually impaired programmers find GUIDL IA to be simple and useful tool that helps GUIDL system users in an interesting way that is quick and effective.

V. CONCLUSION

To learn how to program is very challenging for visually impaired who have nevertheless been included in computer and programming activities since the very beginning of computer usage. For years developed aiding technology that was designed to enable visually impaired to use computers was sufficient to support programming activities of visually impaired since all computer programs' interfaces were based on text. With the occurrence of graphical user interfaces in computer industry visually impaired programmers have been left in difficult position because of graphical nature of software development environments, usage of point-and-click actions to set various settings, more complex object-oriented concepts and mostly because of usage of mouse to design graphical interfaces. Visually impaired programmers that were an equal team members found themselves unable to create graphical user interfaces in a suitable manner. In order to include visually impaired into all parts of overall software development a GUIDL system has been developed as an aiding solution which does not replace actual development environments but serves as an aiding technology in a part of software development that includes creation of graphical user interfaces. Although GUIDL system has been made simple it still requires some effort to learn how to use it. To make this process even simpler than it is an intelligent assistant GUIDL IA is proposed that enables its users to ask questions in natural language and to get an instant textual answer with a possibility of opening additional web resources to clarify given answers even further. In order to conclude about the effectiveness of GUIDL IA a research has been conducted in which visually impaired programmers where given GUIDL system with GUIDL IA for use on several tasks and were then asked to report about their experience. Research results have shown that GUIDL IA is simple, interesting and useful aiding solution for making usage of GUIDL system easier and quicker. Further improvements of GUIDL IA and research on a larger scale with additional objective measurements will be a part of future research.

REFERENCES

[1] A. Gomes, and A. J. Mendes, "An environment to improve programming education", In Proceedings of the 2007 International Conference on Computer Systems and Technologies, pp. 88:1-6, ACM, New York, USA, 2007.

[2] T. Jenkins, "On the difficulty of learning to program", In Proceedings of the 3rd Annual Conference of the LTSN Centre for Information and Computer Sciences, pp. 53-58, 2002.

[3] W. Peng, "Practice and experience in the application of problem-based learning in computer programming course", In Proceedings of the International Conference on Educational and Information Technology (ICEIT), pp. 1:170-172, IEEE, 2010.

[4] A. Robins, J. Rountree, and N. Rountree, "Learning and teaching programming: A Review and Discussion", Journal of Computer Science Education, vol. 13, no. 2, 2003, pp. 137-172.

[5] M. Konecki, "Aiding visually impaired programming novices in their programming education", In Proceedings of IAC-EIaT 2014, pp. 238-244. Czech Institute of Academic Education z.s., Vestec, Czech Republic, 2014.

[6] M. Konecki, A. Lovrenčić, and R. Kudelić, "Making Programming Accessible to the Blinds", In Proceedings of the 34th MIPRO International Convention on Computers in Technical Systems, pp. 180-184, 2011.

[7] M. Konecki, "A New Approach Towards Visual Programming for the Blinds", In Proceedings of the 35th MIPRO International Convention on Computers in Technical Systems, pp. 1076-1081, 2012.

[8] S. Alexander, "Blind Programmers Face An Uncertain Future", ComputerWorld, vol. 32, no. 44, 1998, pp. 86-87.

[9] J. Sanchez, and F. Aguayo, "Blind learners programming through audio", ACM CHI '05 Extended Abstracts on Human Factors in Computing Systems, ACM Press, pp. 1769-1772, 2005.

[10] A. C. Smith, J. M. Francioni, and S. D. Matzek, "A Java programming tool for students with visual disabilities", In Proceedings of the Fourth International ACM Conference on Assistive Technologies, ACM Press, pp. 142-148, 2000.

[11] J. Alty, and D. Rigas, "Communicating Graphical Information to Blind Users Using Music: The Role of Context", In Proceedings of the SIGCHI Conference on Human factors in Computing Systems, ACM Press, pp. 574–81, 1998.

[12] Y. Bellik, and D. Burger, "Multimodal interfaces: New solutions to the problem of computer accessibility for the blind", In CHI '94 Conference Companion on Human factors in Computing Systems, ACM Press, pp. 24–28, 1994.

[13] M. Rotard, C. Taras, and T. Ertl, "Tactile web browsing for blind people", Multimedia Tools Appl., Springer, vol. 37, no. 1, 2008, pp. 53–69.

[14] M. Konecki, "GUIDL as an Aiding Technology in Programming Education of Visually Impaired", Journal of Computers, vol. 9, no. 12, 2014, pp. 2816-2821.

[15] L. Damodaran, "User involvement in the systems design process - a practical guide for users", Behaviour and Information Technology, vol. 15, no. 6, 1996, pp. 363–377.

[16] J. Rowley, "Towards a methodology for the design of multimedia public access interfaces", Journal of Information Science, vol. 24, no. 3, 1998, pp. 155–166.

Anomaly Detection in Thermal Power Plant using Probabilistic Neural Network

A. Hajdarevic*, I. Dzananovic *, L. Banjanovic-Mehmedovic**, F. Mehmedovic***
* Thermal Power Plant "Tuzla", 75000 Tuzla, Bosnia and Herzegovina
** Faculty of Electrical Engineering, University of Tuzla, P.O.Box 75000 Tuzla, Bosnia and Herzegovina
*** ABB Ltd. Representative Office in Bosnia and Herzegovina, P.O.Box 75000, Tuzla, Bosnia and Herzegovina
hajdarevicamel@gmail.com, i.dzananovic@elektroprivreda.ba, lejla.mehmedovic@untz.ba,
fahrudin.mehmedovic@hr.abb.com

Abstract - Anomalies are integral part of every system's behavior and sometimes cannot be avoided. Therefore it is very important to timely detect such anomalies in real-world running power plant system. Artificial neural networks are one of anomaly detection techniques. This paper gives a type of neural network (probabilistic) to solve the problem of anomaly detection in selected sections of thermal power plant. Selected sections are steam superheaters and steam drum. Inputs for neural networks are some of the most important process variables of these sections. It is noteworthy that all of the inputs are observable in the real system installed in thermal power plant, some of which represent normal behavior and some anomalies. In addition to the implementation of this network for anomaly detection, the effect of key parameter change on anomaly detection results is also shown. Results confirm that probabilistic neural network is excellent solution for anomaly detection problem, especially in real-time industrial applications.

I. INTRODUCTION

With around 40% share in global electricity production, thermal power plants are still significant power generation facilities, although their influence is reduced in favor of the renewable energy plants. Adequate control of thermal power plant units is necessary for the system to function properly, since improper functioning of the system may result in electricity supply interruption and financial losses for electricity producers.

One of the most important systems involved in thermal power plant operation is the plant's boiler. Pulverized coal combustion occurs in this system. Its combustion raises the temperature which helps feedwater conversion into steam with needed parameters. Such steam is driven further to the turbine system propelling it, which is essential for electricity production in thermal power plant. From the above, it is clear that the malfunction of any part in boiler system affects the operation of thermal power plant. Outage of some part or the whole thermal power plant unit can often be avoided with timely anomaly detection.

Many of those techniques are being researched and one of those are neural networks based anomaly detection techniques. If the neural network is created to detect anomalies in the system, reaction time on the anomaly could be significantly reduced. This would provide greater system stability and therefore fewer losses.

This paper presents one of the possible solutions to the anomaly detection problem in boiler system using probabilistic neural network. Given the complexity of the system, because of necessary simplification, only some characteristic sections and some characteristic process variables were selected. All of the input data that are used are coming from the system for monitoring and control present in Thermal Power Plant "Tuzla", Bosnia and Herzegovina.

II. SELECTED SECTIONS OF THERMAL POWER PLANT

A thermal power plant, as a large and complex system, consists of multiple smaller systems that work together and ensure continuous electricity generation. This paper is focused on one of the main systems of every thermal power plant and that is the boiler.

The boiler is one of the most important systems in a thermal power plant because of its role in electricity generation [1, 2]. The boiler represents the entire system that participates in the conversion of water into steam. A single division of the boiler system into individual sections does not exist, because many sections are connected and there is often no clear distinction between them. In this paper, the division is made according to available process images with certain corrections concerning data availability.

A. Water-Steam System

This system is primarily engaged in converting water into steam and consists of multiple subsystems with separate functions. The system consists of multiple pipes and vessels. Heat exchange between different media occurs in this system in order to achieve optimal steam parameters. The steam is driven further to the turbine propelling it, which is essential for electricity generation. Given that the quality and parameters of steam directly affect the electricity generation, anomaly detection in this system is of great importance for the plant. The most important measurements of process variables related to steam are temperature, pressure and steam flow which directly affects the current power output.

B. Feedwater System

One of the important subsystems of the boiler is the system for its feedwater supply, because without the feedwater supply there is no steam generation. After the raw water is treated at the chemical water treatment plant, the water is stored in the feedwater tank. The water is

distributed further from the tank into the feedwater pipe system using feedwater pumps. The pumps maintain the specified feedwater flow which is determined by the required power output. The purpose of this system is distribution of feedwater to the most important subsystem of the water-steam system – the steam drum.

C. Steam Superheaters

The steam generated in the steam drum is distributed through this system. After distribution of steam through this system, it is called the superheated steam. Steam superheaters system consists of pipes mounted in the boiler that distribute the steam to the turbine. This system has a major role in removing moisture from the steam, which improves its quality.

The temperature of steam generated in the drum still does not match the temperature needed in the process. The steam generated in the drum still contains a certain percentage of moisture. Such steam should not be distributed to the turbine due to the possibility of condensation on the blades. In every thermal power plant, there is a tendency to produce 100% quality steam. Because of that, the steam is heated in this system using flue gases. Increasing the steam temperature results in removal of moisture from the steam and that is the primary goal of this system. The steam temperature is increased by around 160 °C compared to steam temperature in the drum. The goal of the system is to maintain the temperature around 535 °C, which is optimal for boiler analyzed in this paper.

D. Steam Drum

The most important subsystem of the water-steam system is the steam drum, which is a large tank with the task of steam extraction from a water-steam mixture stored in the drum. Given the importance of the drum for electricity generation, it is logical that timely anomaly detection is required in this system. The most important process variable in the drum is the water level. In addition to that, steam pressure, conductivity and pH value are also measured. A rough block diagram is shown in "Fig. 1". It shows important boiler sections and selected sections of the boiler for anomaly detection (marked in red) and some connections between the sections.

III. Analysis of Existing Anomaly Detection Techniques

Conventional anomaly detection techniques have been used for a long time, but with the development of computer technology modern anomaly detection techniques can be developed.

Anomaly detection techniques division can be done in multiple ways [3, 4]. Based on the means of acquiring knowledge anomaly detection techniques can operate in one of the following three modes: supervised, semi-supervised and unsupervised. Techniques that operate in unsupervised mode detect anomalies by analyzing an unlabeled data set and making an implicit assumption that majority of the data set represents normal behavior. Anomalous data is the data that least fits the characteristics of the data set. Techniques that operate in semi-supervised mode build a model for the class corresponding to normal behavior. Then the model is used to identify anomalies in the test data. Techniques that operate in supervised mode require two data sets, one labeled as normal behavior and the other labeled as anomaly. Then it is needed to train the classifier that determines if the test data belong to the class labeled as normal behavior.

In addition to this division, there is also another division that is used in this paper, proposed in [3]. It is division based on anomaly type and classifies anomaly detection techniques in one of the following three groups: point anomaly detection techniques, contextual anomaly detection techniques and collective anomaly detection techniques. If an individual data instance from a data set can be considered as anomalous with respect to the rest of data, then such instance can be classified as a point anomaly. The focus of this paper are point anomaly detection techniques.

A. Classification based Anomaly Detection Techniques

Classification is used to learn a model from a set of labeled data instances (training data set). After that, the model is used to classify a test instance into one of the classes. These techniques operate under the assumption that there is enough data in the training data set, so the classifier can learn how to classify test data.

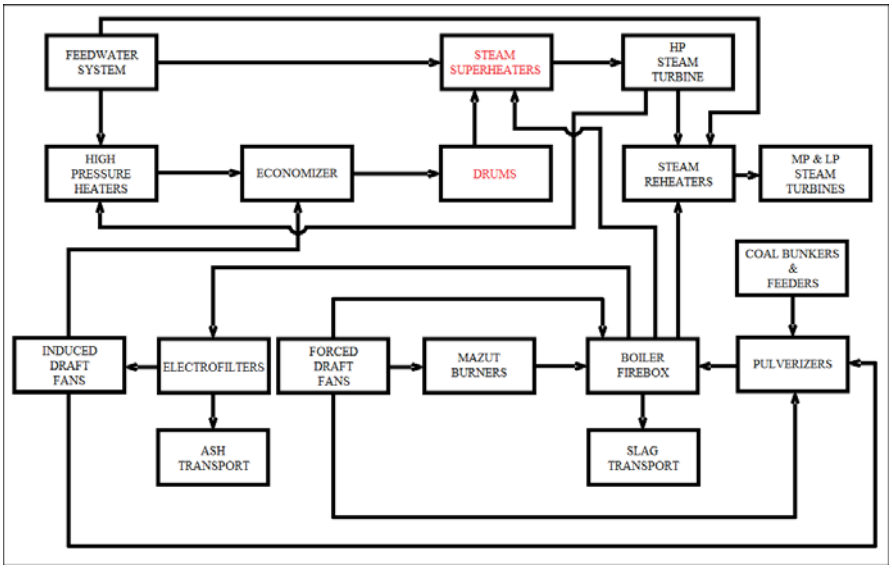


Figure 1. Block diagram of boiler sections (some turbine sections included)

Classification based anomaly detection techniques can be categorized into two following groups: one-class and multi-class anomaly detection techniques. One-class classification based anomaly detection techniques assume that all training instances have only one class label. A test instance different from the labeled ones are declared as anomalous. Multi-class classification based anomaly detection techniques assume that the training data contains labeled instances belonging to multiple normal classes. A test instance is declared as anomalous if it does not belong to any of the normal classes [5].

Commonly used classification based anomaly detection techniques are: rule based techniques, neural networks based techniques and support vector machines (SVM) based techniques.

B. Neural networks based anomaly detection techniques

Neural networks are used for anomaly detection and they represent anomaly detection techniques that use artificial intelligence. They can perform as one-class as well as multi-class classifiers. Neural networks are trained with normal instances and anomalous instances. After that the network should be able to determine anomalous instances in a test data set.

Many of the published papers are dealing with applications of neural networks in various systems. Papers [6] and [7] are dealing with the application of artificial intelligence in anomaly detection. Many techniques that use neural networks are represented in those papers. The application of neural networks in power systems is considered in [8]. In particular, the application of neural networks for anomaly detection in power plants is considered in papers [9] and [10]. Neural networks in combination with fuzzy logic for anomaly detection are considered in [11].

Multiple types of neural networks could be used for anomaly detections: MLP neural networks, recurrent neural networks (RNN), probabilistic neural networks, etc.

IV. PROBABILISTIC ARTIFICIAL NEURAL NETWORK

Probabilistic neural networks (PNN) belong to the stochastic neural networks group. PNN represent an implementation of a statistical algorithm called kernel discriminant analysis (KDA) [12, 13]. The probabilistic neural net is based on the theory of Bayesian classification and the estimation of probability density function (PDF). Because of ease of training and a sound statistical foundation in Bayesian estimation theory, PNN has become an effective tool for solving many classification problems. A PNN consists of several sub-networks, each of which is a Parzen window pdf estimator for each of the classes. The development of the probabilistic neural network relies on Parzen windows classifiers. The Parzen windows method is a non-parametric procedure that synthesizes an estimate of a probability density function (pdf) by superposition of a number of windows, replicas of a function (often the Gaussian). The Parzen windows classifier takes a classification decision after calculating the probability density function of each class using the given training examples. The multicategory classifier decision is expressed as follows:

$$p_k f_k > p_j f_j \tag{1}$$

where p_k is the prior probability of occurrence of examples from class k and f_k is the estimated pdf of class k.

The advantages of this network type are: rapid training process (faster than BP based networks), guaranteed convergence to the optimal classification with increasing data set and the ability to change the number of learning inputs with minimal or no additional training. The disadvantages are: large memory requirements, slow execution and the representative data set requirement. Unlike many other neural networks, training of PNN is done in a single iteration.

The PNN architecture consists of four layers: input layer, pattern layer, summation layer and decision layer, presented in "Fig 2". This PNN structure recognizes c classes. The first layer shows the input pattern with n features. The number of nodes in the pattern layer is equal to the number of training instances. The number of nodes in the summation layer is equal to the number of classes in the training instances. The input layer is fully connected to the pattern layer. The input layer doesn't perform any computation and simply distributes the input to the neurons in the pattern layer, which is semi-connected to the summation layer. The summation units sum the inputs from the pattern units that correspond to the category from which the training pattern was selected.

The PNN works by creating a set of multivariate probability densities that are derived from the training vectors presented to the network. The input instance with unknown category is propagated to the pattern layer. Once each node in the pattern layer receives the input, the output of the node will be computed:

$$\pi_i^c = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}} \tag{2}$$

where n is the number of features of the input instance x, σ is the smoothing parameter and x_{ij} is a training instance corresponding to category c.

The summation layer neurons compute the maximum likelihood of pattern x being classified into c by summarizing and averaging the output of all neurons that belong to the same class:

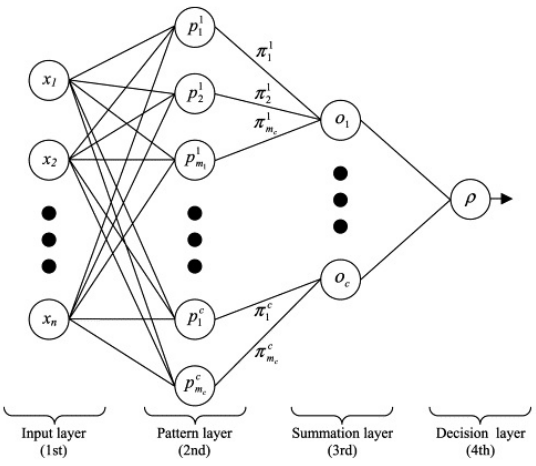


Figure 2. An example of probabilistic artificial neural network

$$p_i(x) = \frac{1}{(2\pi)^{n/2}\sigma^n}\frac{1}{N_i}\sum_{i=1}^{N_i} e^{-\frac{(x-x_{ij})^T(x-x_{ij})}{2\sigma^2}} \quad (3)$$

where N_i denotes the total number of samples in class c. If the *a priori* probabilities for each class are the same, and the losses associated with making an incorrect decision for each class are the same, the decision layer unit classifies the pattern x in accordance with the Bayes' decision rule based on the output of all the summation layer neurons:

$$C(x) = argmax\{p_i(x)\}, i = 1,2,\dots,c \quad (4)$$

where C(x) denotes the estimated class of the pattern x and m is the total number of classes in the training samples. If the *a priori* probabilities for each class are not the same and the losses associated with making an incorrect decision for each class are different, the output of all the summation layer neurons will be

$$C(x) = argmax\{p_i(x)cost_i(x)apro_i(x)\}, \quad i = 1,2,\dots,c \quad (5)$$

where $cost_i(x)$ is the cost associated with misclassifying the input vector and $apro_i(x)$ is the prior probability of occurrence of patterns in class c.

V. Evaluation of Acquired Knowledge

After obtaining the actual classification results, some way of quality assessment is needed for a certain type of neural network. The purpose of the assessment is to determine which type of neural network fits best for some classification problem. Therefore the evaluation of acquired knowledge is an important part of neural network application. During this process, it is needed to determine which verification and validation measures will be used. Different types of these measures exist through specific application [14] and some of them are applied in this paper.

A. Result Verification

There are several different measures used for result verification. Majority of the measures are adapted to one-class classification. This corresponds to the aim of this paper, because neural networks used for anomaly detection in this paper are trained as one-class classifiers. Table I provides an overview of possible outcomes of one-class classification problem solution. There are four possible outcomes of anomaly detection. True positive (TP) and true negative (TN) outcomes represent a correct classification, while false negative (FN) and false positive (FP) outcomes represent an incorrect one. Both types of incorrect classifications represent a hazard. False positives can cause an action which is not needed, but false negatives are actually more dangerous because an anomaly would be ignored.

TABLE I. Possible putcomes of one-class classification problem solution

Outcome	**Actual**		
Desired	***Behavior***	***Normal***	***Anomalous***
	Normal	True positive	False positive
	Anomalous	False negative	True negative

Based on these counts, the following perfomance metrics are calculated (all expresed as a percentage):

- Accuracy (ACC):

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \quad (6)$$

- Sensitivity or true positive rate (TPR) or recall:

$$TPR = \frac{TP}{TP + TN} \quad (7)$$

- Specificity or true negative rate (TNR):

$$TNR = \frac{TN}{TP + TN} \quad (8)$$

- Precision (PR) or positive predictive value (PPV):

$$PR = \frac{TP}{TP + FP} \quad (9)$$

- Negative predictive value (NPR):

$$NPR = \frac{TN}{TN + FN} \quad (10)$$

- F1 score:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (11)$$

B. Result Validation

Result validation techniques are mainly linked to neural networks if an input data set is not large enough. In addition to that, they are used to test a neural network with some random sample from the data set. If the data set is large enough, which does not guarantee the neural network will be trained properly. This is due to a possibility of overfitting the network with a large number of similar instances. If that occurs, the neural network could treat some noise as an anomaly. In order to prevent this, some kind of result validation should be performed. Result validation techniques can also show if the neural network parameters are good enough. Commonly used result validation techniques are: cross-validation, regularization, early stopping etc. Result validation techniques are not exclusively used with neural networks. They are also used with other anomaly detection techniques and with value prediction techniques.

The most popular result validation technique is cross-validation [15]. The idea is to separate the available data into a training data set (usually 80% to 90% of the data) and a test data set (the remaining 10% to 20% of the data). The simplest form of cross-validation randomly separates the available data into a single training set and a single test set. It is a risky approach, because an unlucky split could lead to an ineffective neural network. A better approach would be to repeat the previous procedure multiple times, but this approach is also risky because there is a chance that some instances could be used only for training and never for testing, or vise versa. Therefore, a technique called k-fold cross-validation was developed.

The idea behind k-fold cross-validation is to divide all the available data into *k* roughly equal-sized groups. In each iteration of k-fold cross-validation, *k-1* groups are used for training and the remaining one is used for testing. The k-fold cross-validation iterates through a number of folds. After the first iteration, the next group is used for testing, and the remaining data are used for training. This procedure repeats until all of the groups are used for testing once.

The main advantage of k-fold cross-validation compared to other techniques is utilization of all the available data in the process. The main disadvantage is the potential for different validation results due to stochastic process of group forming at the beginning of the validation process. The k-fold cross-validation can give different results each time it is performed. This can be avoided by repeating the process multiple times and using the mean validation result.

VI. RESULTS AND DISCUSSION

The results are presented as a set of different verification measures which show the effect of important parameters change. Given that the k-fold cross-validation technique is mentioned in the paper, in addition to the test results, 10-fold cross-validation results are also presented.

Scores of an ideal classification would all be the same (1), except TPR and TNR (0.5 – their sum is always 1). Given that the initial synaptic weights are chosen randomly, it is possible to obtain different results if a neural network is trained and tested multiple times. Therefore, all of the obtained results are obtained from average confusion matrices (average of multiple training and testing cycles).

A. Data Used as Neural Network Input Data

Neural networks were applied to data from two sections of the boiler, the steam superheaters and the steam drums. These sections consist of two separate systems each, which produce steam of adequate quality together. Process variables used in this paper represent both systems (I and II). The process variables from the steam superheaters are: superheated steam temperature (TS), superheated steam flow (FS) and superheated steam cooling water flow (CWF). The variables from the steam drums are: drum level (DL), drum pressure (DP) and feedwater flow (FWF). All of the input data come from an actual system for monitoring and control present in unit 4 of Thermal Power Plant "Tuzla" with the sampling period of 1 s.

There are 962 instances that represent normal behavior and the same amount representing anomalous behavior in the input data set. That makes a total of 1924 instances for each section. Given that, the 10-fold cross-validation is not necessary because the data set is not large enough, but to test the neural networks with more random data sets. The data set is divided into a training data set (70% of the data), a validation data set (15% of the data) and a test data set (15% of the data).

B. Results of Probabilistic Neural Network Application

Due to the nature of this neural networks type, there are not many parameters that can be changed. Number of neurons depends on the number of instances used in input data set. Also, the learning process is completed in only one epoch, so the only parameter that can be changed is the spread coefficient. The initial value of this parameter is 1 and the changes are performed by decreasing and increasing the value. The minimum value is 0.01, and the maximum is 10. The best results can be obtained if the coefficient value is between 0.5 and 5 for both section. Results of PNN application for anomaly detection in the steam superheaters are given in the Table II. From the results, PNN gives excellent results if the spread coefficient is properly selected (spread = 2). Results of PNN application for anomaly detection in the steam drums are given in the Table III. The best results can be obtained if the value is between 0.5 and 5, but the steam drums results are a bit better if other values are used. The inputs (6 values for both systems) from Steam Superheats and PNN classification outputs are shown in "Fig 3". The inputs (6 values for both systems) from Steam Drums and PNN classification outputs are shown in "Fig 4". The results confirm that PNN is excellent solution for anomaly detection problem.

TABLE II. RESULTS OF PNN (STEAM SUPERHEATERS)

Validation results					
Measure	**Spread = 0.01**	**Spread = 0.1**	**Spread = 0.5**	**Spread = 2**	**Spread = 10**
ACC	0.5418	0.9214	0.9995	0.9994	0.8925
TPR	0.9227	0.5425	0.4999	0.4998	0.5601
TNR	0.0773	0.4575	0.5001	0.5002	0.4399
PR	0.5218	0.8643	0.9997	0.9997	0.8231
NPR	0.9977	0.9998	0.9992	0.9991	0.9998
F1	0.6857	0.9271	0.9995	0.9994	0.9029
Testing results					
Measure	**Spread = 0.01**	**Spread = 0.1**	**Spread = 0.5**	**Spread = 2**	**Spread = 10**
ACC	0.6501	0.9982	0.9993	0.9994	0.8990
TPR	0.7690	0.5009	0.4997	0.4997	0.5562
TNR	0.2310	0.4991	0.5003	0.5003	0.4438
PR	0.5883	0.9965	0.9999	0.9999	0.8320
NPR	0.9999	0.9999	0.9987	0.9989	0.9999
F1	0.7408	0.9982	0.9993	0.9994	0.9083

TABLE III. RESULTS OF PNN (STEAM DRUMS)

Validation results					
Measure	**Spread = 0.01**	**Spread = 0.1**	**Spread = 0.5**	**Spread = 2**	**Spread = 10**
ACC	0.5566	0.8932	0.9961	0.9974	0.9877
TPR	0.8982	0.5597	0.5004	0.4997	0.5036
TNR	0.1018	0.4403	0.4996	0.5003	0.4964
PR	0.5300	0.8241	0.9954	0.9979	0.9809
NPR	0.9983	0.9998	0.9968	0.9969	0.9946
F1	0.6928	0.9035	0.9961	0.9974	0.9877
Testing results					
Measure	**Spread = 0.01**	**Spread = 0.1**	**Spread = 0.5**	**Spread = 2**	**Spread = 10**
ACC	0.6822	0.9987	0.9999	0.9999	0.9916
TPR	0.7329	0.5006	0.4997	0.4999	0.5041
TNR	0.2671	0.4994	0.5003	0.5001	0.4959
PR	0.6114	0.9975	0.9999	0.9999	0.9838
NPR	0.9999	0.9999	0.9999	0.9999	0.9998
F1	0.7589	0.9987	0.9999	0.9999	0.9917

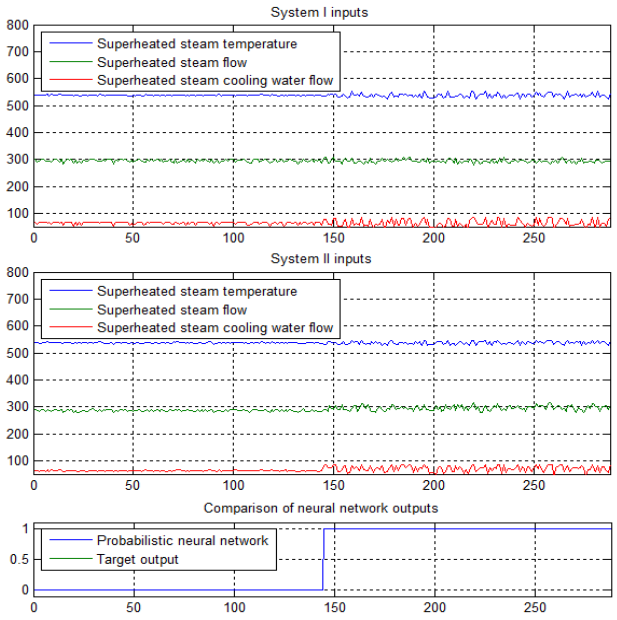


Figure 3. Anomaly detection using probabilistic neural network for input data from steam superheaters (spread=2)

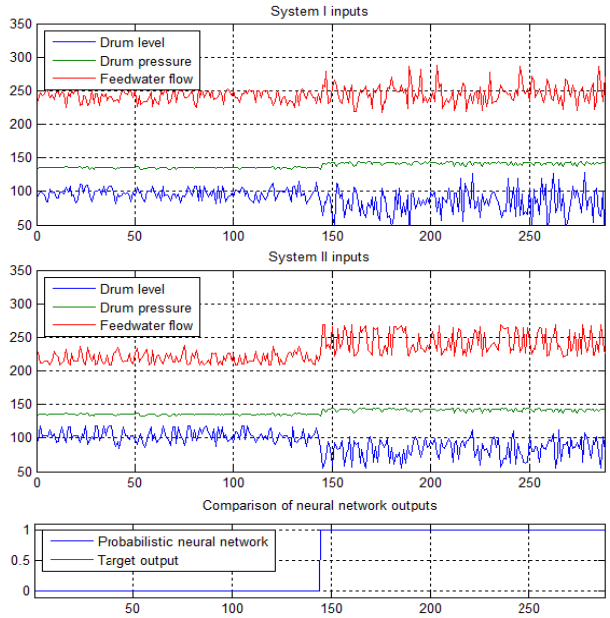


Figure 4. Anomaly detection using probabilistic neural network for input data from steam drums (spread=2)

VII. CONLUSION

In this paper the probabilistic neural network (PNN) was tested with data from real-time thermal power system (TPS) to explore possibilities of anomaly detection.

An anomaly in any part of the boiler affects the operation of thermal power plant unit. Malfunctions can be prevented by timely anomaly detection. Given the system complexity, for the purpose of necessary simplification, only some of the characteristic boiler sections and process variables from those sections were selected for further analysis. All of the input data come from an actual system for monitoring and control present in Thermal Power Plant "Tuzla".

It was shown that the PNN provide excellent results for anomaly detection and with the used data. As for the significance of the results for power generation process control in a thermal power plant, it could be great. Due to the fact that the data are divided so that the very beginning of anomalous state is represented in anomalous data instances, neural networks could be used as a tool for early anomaly detection. Given that, an anomaly reaction time could be significantly reduced. That would provide greater system stability and therefore fewer losses.

Only some data from large sections of the boiler were included in the analysis provided in this paper. In an actual application of neural networks for anomaly detection, data from many other (smaller) sections should be included. In addition to the problem of anomaly detection, future research could be extended to predictive control in industrial systems using neural networks. In any case, in the area of modern advanced automation, neural networks and other forms of artificial intelligence are labeled as the future of complex systems.

REFERENCES

[1] J.B. Kitto, S.C. Stultz, "Steam: It's Generation and Use", The Babcock & Wilcox Company, Edition 41, 2005.

[2] G.F. Gilman, "Boiler Control Systems Engineering", International Society of Automation (ISA), Second Edition, 2010.

[3] V. Chandola, A. Banerjee, V. Kumar, "Anomaly Detection: A Survey", Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA, 2007.

[4] V.J. Hodge, J. Austin, "A Survey of Outlier Detection Methodologies", Dept. of Computer Science, University of York, York, UK, 2004.

[5] K. Hempstalk, E. Frank, "Discriminating Against New Classes: One-Class versus Multi-Class Classification", AI 2008: Advances in Artificial Intelligence Lecture Notes in Computer Science, Vol. 5360, 2008.

[6] T. Ahmed, B. Oreshkin, M. Coates, "Machine Learning Approaches to Network Anomaly Detection", 2nd USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques (SysML), Cambridge, MA, USA, 2007.

[7] S.X. Wu, W. Banzhaf, "The Use of Computional Intelligence in Intrusion Detection Systems: A Review", Applied Soft Computing, Vol. 10, Issue 1, 2010.

[8] M.T. Haque, A.M. Kashtiban, "Application of Neural Networks in Power Systems: A Review", International Journal of Electrical, Electronic Science and Engineering, Vol. 1, No. 6, 2007.

[9] Z. Liu, W. Gao, "Wind Power Plant Prediction by using Neural Networks", IEEE Energy Conversion Conference and Exposition, Raleigh, NC, USA, 2012.

[10] A. Kusiak, Z. Song, "Sensor Fault Detection in Power Plants", Journal of Energy Engineering ASCE, 2009.

[11] F. Amiri, C. Lucas, N. Yazdani, "Anomaly Detection using Neuro Fuzzy System", World Academy of Science, Engineering and Technology Vol. 3, 2009.

[12] F. Ancona, A.M. Colla, S. Rovetta, R. Zunino, "Implementing Probabilistic Neural Networks", Neural Computing and Applications 7, Springer-Verlag London Ltd., 1998.

[13] S. N. Sivanadam, S. Sumathi, S. N. Deepa, "Special Networks" in Introduction to Neural Networks using MATLAB 6.0, New Delhi: Tata McGraw-Hill, pp. 311-314, 2006.

[14] M. Suljic, L. Banjanovic-Mehmedovic, I. Dzananovic: "Determination of coal quality using Artificial Intelligence Algorithms", Journal of Scientific and Industrial Research (JSIR), Vol. 72 (06), 2013.

[15] J. McCaffrey, "Understanding and Using K-Fold Cross Validation for Neural Networks", 2013.

Neutral Grounding in Middle Voltage Power Systems: Simulation model

S. Ravlić*, A. Marušić **, J. Havelka **
* Ravel d.o.o., Zagreb, Hrvatska
** Faculty of Electrical Engineering and Computing, Zagreb, Hrvatska
sonja.ravlic@ravel.hr

Abstract - In middle voltage networks in Croatia there are four neutral earthing concepts: isolated neutral point, low-impedance grounded networks, resonant grounded networks and networks with partial compensation. The paper will generally discuss modelling and simulation software developed for analysis of four grounding concepts mentioned earlier. Model was made using Matlab. Real time simulations were performed to test the behavior of middle voltage networks with different grounding concepts in case when single phase earth faults occur. The manner in which neutral point is treated has a fundamental influence on the value of the current and on overvoltages. The paper will also discuss influences that cable network connected to middle voltage side of network transformer and the moment when a fault occurs have on the value of the current and on overvoltages. Based on the simulation results, the paper presents the possibilities and limitations of various neutral earthing concepts in medium voltage systems.

I. INTRODUCTION

Single phase to earth faults are the most frequent failures in medium and high voltage systems [1]. The values of the currents and of the temporary overvoltages with power frequency in case of a single pole earth fault depend on the neutral point treatment, earth capacitances of the overhead lines or cables connected to the network, voltage level, the value of fault resistance and on the distance between the supply busbars and the fault location [2]. When considering single phase earth faults in medium voltage networks two important aspects need to be taken into account: value of the short circuit current (and consequently touch voltages) and overvoltages. In order to describe the behavior of the network during single phase earth faults simulation model made using MATLAB SimPowerSystem software [3]. Several simulation models were made for same network: isolated neutral point, low-impedance grounded networks, resonant grounded networks and networks with partial compensation. The model described in this paper observes substation during fault and the influence of various neutral earthing methods combined with the influence of the length of middle voltage power lines and the moment when a fault occurs. Every neutral grounding concept has its own advantages and disadvantages.

A real time simulation model like this can be powerful analysis tool to validate which neutral grounding concept to choose for a specific substation regarding network connected to the substation, safety issues, network importance, earth system requirements and costumer interruptions [2].

II. SIMULATION MODEL ELEMENTS

The aim of this paper is to determine the characteristics of single phase earth faults in middle voltage networks in Croatia. All simulation models are consisted of: three-phase 110 kV source, network transformer 110/10(20) kV, middle voltage cables and overhead lines, 10(20)/0.4 kV transformers, series load and single pole fault with earth. MATLAB SimPowerSystems provides component libraries and analysis tools for modeling and simulating electrical power systems [3]. Figure 1 shows a network model in Matlab SymPowerSystem. Simulations of earth fault currents and overvoltages were conducted in 20 kV network with a different grounding concept, different middle voltage cable length and different moment when a fault occurs. All network elements are in detail described below.

The three-phase source is connected in Y with a grounded neutral. This block generates a three-phase sinusoidal 110 kV voltage, 50 Hz with no time variation. Phase angle of phase A is zero degrees.

Network transformer was modelled according to the real network transformer. Network transformer is 110/10(20)/10.5 kV, 40 MVA with winding connection YNynD5. In order to provide real parameters of network transformer equivalent circuit was calculated. Figure 2 gives equivalent circuit of three phase network transformer [4]. This part of network dimensioning was especially complex and important in order to provide real network state.

All cables are represented as distributed parameter lines. In order to dimension middle voltage cable parameters from Elka's catalogue [5] were used. Line to earth capacitances of middle voltage cables are one of the most important factors that need to be taken into account during dimensioning the network model. Two cases were observed in this paper: cable length 1 km and cable length 40 km. The aim of the simulation was to observe the impact that cable length have on overvoltages and on the value of the current.

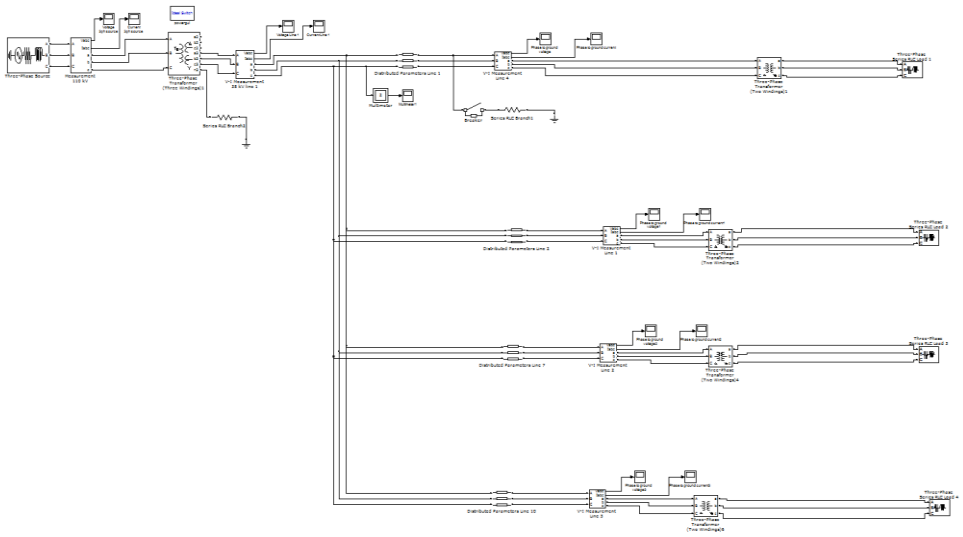


Fig. 1. The network model in Matlab

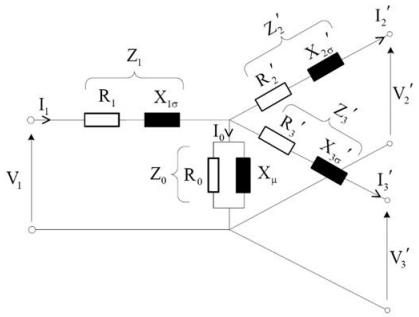


Fig. 2. Equivalent circuit of three phase network transformer

Single pole fault with earth is represented as one-phase circuit breaker connected to ground. In series with circuit breaker is resistance (characterizes the ground resistance). Two cases of single pole fault with earth were observed in order to see what influence the moment when the fault occurs has on overvoltages and current value. Table 1 gives moments when a fault occurs and phase shift of faulted phase. The moment when a fault occurs is represented by t (ms) and φ (°) represents a phase shift of faulted phase. It is expected that the highest values of the overvoltages will occur in cases when the phase shift of faulted phase is $\varphi = 90$ °.

TABLE 1. SINGLE POLE FAULT WITH EARTH

	t (ms)	***φ (°)***
case 1	10	0
case 2	5	90

Neutral grounding resistor was dimensioned according to the real neutral grounding resistors in Croatia (generally 40 Ω for the 20 kV network). The purpose of a neutral grounding resistor is to limit ground fault currents to safe levels so that all the electrical equipment in a power system is protected. In this case ground fault current is limited to 300 A. In case of low ohmic grounded neutral point it is connected between neutral point of middle voltage side of network transformer and the earth.

In order to be able to determine inductance L of Peterson coil (in case of a resonant grounded neutral point and partial compensation) it is necessary to know the three phase grounding capacitance of middle voltage cables:

$$\omega \cdot L = \frac{1}{3 \cdot \omega \cdot C_0} \tag{1}$$

In (1) C_0 represents line to earth capacitances of middle voltage cables, ω represents angular frequency. For the suppression of the arc the Peterson coil should be well tuned within limits [6]. If the value of the Peterson coil is correctly tuned the current of the possible earth fault can be reduced [7]. In this case current will be reduced to zero. In case of resonant grounded neutral point Peterson coil is connected between neutral point of middle voltage side of network transformer and the earth. In case of a partial compensation resistor is connected parallel to low-ohmic resistor between the neutral point of middle voltage side of network transformer and the earth [1].

III. MEASUREMENTS

Number of measurement blocks were used to measure three-phase values of voltages and currents in a circuit. When connected in series with three-phase elements, the three-phase V-I measurement block returns the three phase-to-ground or phase-to-phase peak voltages and currents. The block can output the voltages and currents per unit or in volts and amperes. It measures phase to phase voltages as well as a phase to ground. Comparison of the values before and after the single pole fault with earth was made in order to understand the influence it has to network.

IV. SIMULATION EXPECTATIONS

Length of the middle voltage power lines, neutral earthing method should determine the earth fault current and overvoltages [2]. The moment when fault with earth occurs should influence overvoltages significantly. Highest overvoltages for all four neutral grounding concepts and both cable lengths are expected to appear if a fault occurs at moment t = 5 ms (φ = 90°).

It is expected that currents in the case of the isolated neutral point increase significantly with the increase of the length of middle voltage cables. Internal overvoltages during earth faults are expected to be the highest in the case of the isolated neutral point (compared to all other neutral earthing methods).

In case of the low-ohmic neutral point large number of long cables should increase the value of the earth fault current. Earth fault currents are expected to be the highest in the case of low-ohmic neutral point (current increases especially with the increase of underground cables). Overvoltages in this case are expected to be the lowest.

In case of resonant grounded neutral point the inductive reactance of the Peterson coil is set near the value of the capacitive reactance of the whole network. The capacitive and inductive currents are opposite and resulting current is expected to be very small. In the simulation, it is expected that result current is set to zero. Increase of the length of the underground cables should not influence the value of the current. Overvoltages in case of resonant grounded neutral point are expected to be very high (almost as high as in the case of an isolated neutral point).

Increase of the length of the middle voltage cables should not influence the value of the current in case of the partial compensation, too. Current is supposed to depend only on the value of the low-ohmic resistor. Overvoltages in case of partial compensated networks should be similar as in the case of low-ohmic neutral point.

During a single phase earth fault the voltage of the faulted phase in all simulations is expected to collapse while the voltage of the healthy phases is expected to rise. Figure 2 shows the voltage of the faulted phase for all four ways of transformer neutral point grounding treatment.

Phase to phase voltage of all three phases for all four ways of transformer neutral point grounding treatment should remain unaffected regardless the cable length and moment when a fault occurs. Figure 3 shows the phase to phase voltage for isolated neutral point in case when cable length is 1 km, and the moment when a fault occurs is 10 ms.

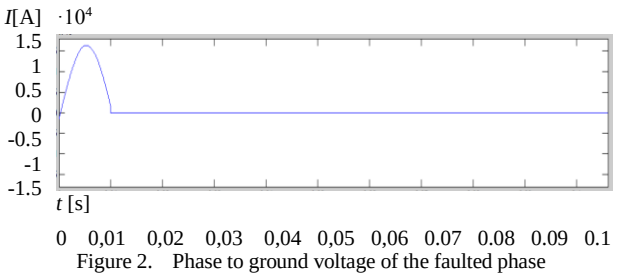


Figure 2. Phase to ground voltage of the faulted phase

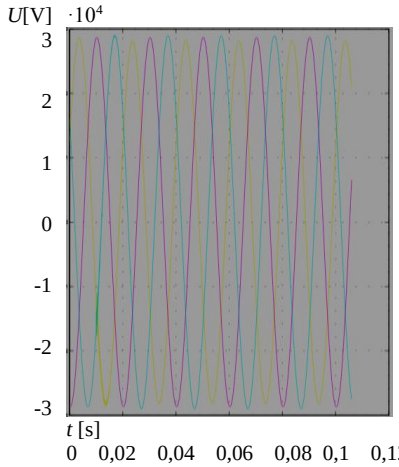


Figure 3. Phase to phase voltages before and after fault in case of isolated neutral point

V. SIMULATION RESULTS

A. Isolated neutral point

Results for two cable lengths and two moments when a fault occurs in networks with isolated neutral point are given in table 2. Figure 4 and figure 5 show the increase of current due to long cables. In case of long cables (40 km) current increases 42.8 times, as expected. It is evident that cable length has little impact on the value of overvoltages. Figure 6 and figure 7 show influence that moment when a fault occurs have on overvoltages. As expected, overvoltages are the highest in case if a fault occurs at 5 ms. The moment when a fault occurs has no influence on the value of fault current.

TABLE 2. SIMULATION RESULTS IN CASE OF ISOLATED NEUTRAL POINT

l (km)	***t* (ms)**	***Fault current***	***Overvoltage***	***Max. overvoltage***
1	10	4.78	1.733	1.823
1	5	4.78	1.735	3.316
40	10	204.67	1.754	1.818
40	5	204.67	1.756	2.275

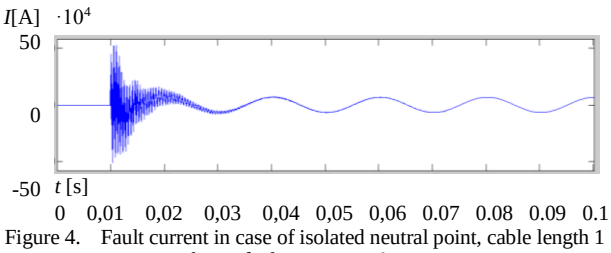


Figure 4. Fault current in case of isolated neutral point, cable length 1 km, a fault occurs at 10 ms

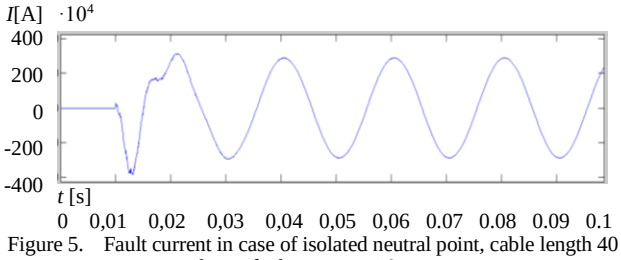


Figure 5. Fault current in case of isolated neutral point, cable length 40 km, a fault occurs at 10 ms

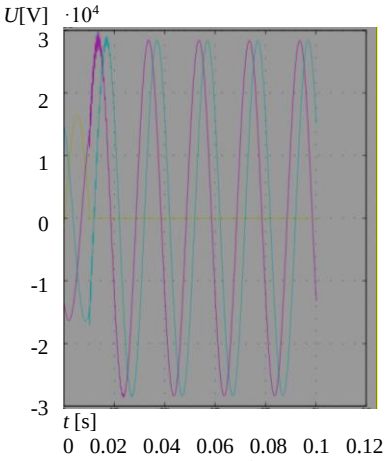


Figure 6. Overvoltages in case of isolated neutral point, cable length 1 km, a fault occurs at 10 ms

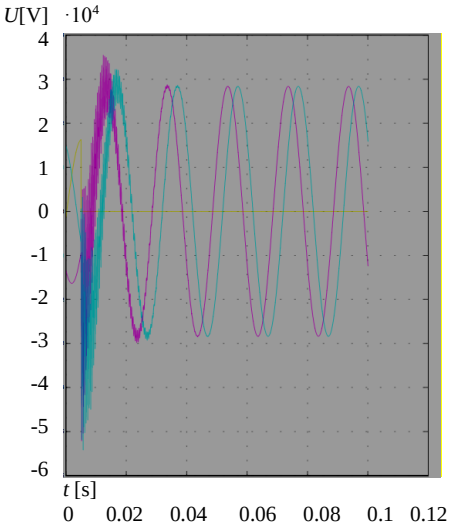


Figure 7. Overvoltages in case of isolated neutral point, cable length 1 km, a fault occurs at 5 ms

B. *Low-ohmic grounded neutral point*

Table 3 gives results for two cable lengths and two moments when a fault occurs in case of low-ohmic grounded neutral point. Figure 8 and figure 9 show increase of current due to long cables. In case of long cables (40 km) current increases, as expected. Overvoltages are the highest in case if a fault occurs at 5 ms. Figure 10 and figure 11 show influence that moment when a fault occurs have on overvoltages in case of resonant grounded neutral point.

TABLE 3. SIMULATION RESULTS IN CASE OF LOW-OHMIC GROUNDED NEUTRAL POINT

l (km)	***t* (ms)**	***Fault current***	***Overvoltage***	***Max. overvoltage***
1	10	283.27	1.697	1.747
1	5	283.27	1.697	2.216
40	10	314.31	1.659	1.761
40	5	314.31	1.611	2.068

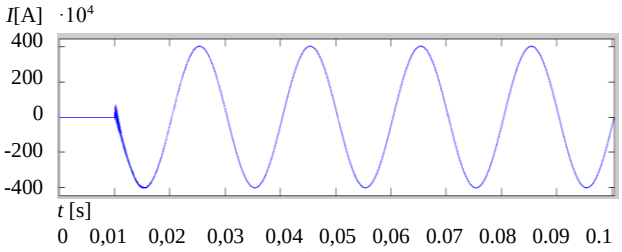


Figure 8. Fault current in case of low-ohmic grounded neutral point, cable length 1 km, a fault occurs at 10 ms

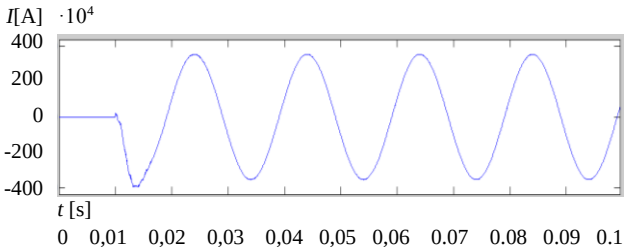


Figure 9. Fault current in case of low-ohmic grounded neutral point, cable length 40 km, a fault occurs at 10 ms

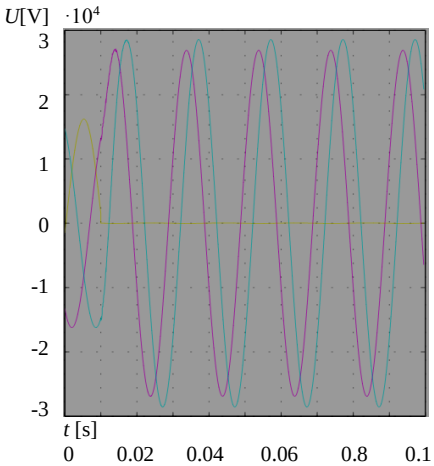


Figure 10. Overvoltages in case of low-ohmic grounded neutral point, cable length 40 km, a fault occurs at 10 ms

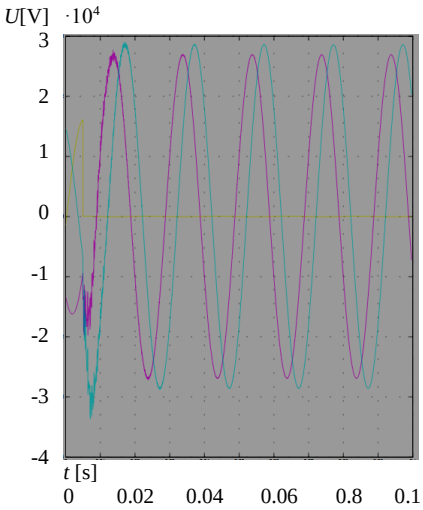


Figure 11. Overvoltages in case of low-ohmic grounded neutral point, cable length 40 km, a fault occurs at 5 ms

C. *Resonant grounded neutral point*

Results for two cable lengths and two moments when a fault occurs in networks with resonant grounded neutral point are given in table 4. Figure 12 and figure 13 show

that current does not increase due to long cables, as expected. In case of resonant grounded neutral point overvoltages are expected to be almost as high as in the case of isolated neutral point. Figures 14 and 15 show overvoltages in case of resonant grounded neutral point for a 40 km cable length.

TABLE 4. SIMULATION RESULTS IN CASE OF RESONANT GROUNDED NEUTRAL POINT

l (km)	***t* (ms)**	***Fault current***	***Overvoltage***	***Max. overvoltage***
1	10	0	1.734	2.140
1	5	0	1.746	2.049
40	10	0	1.719	1.770
40	5	0	1.731	2.213

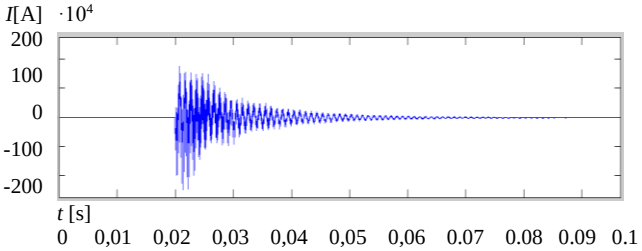


Figure 12. Fault current in case of resonant grounded neutral point, cable length 1 km, a fault occurs at 10 ms

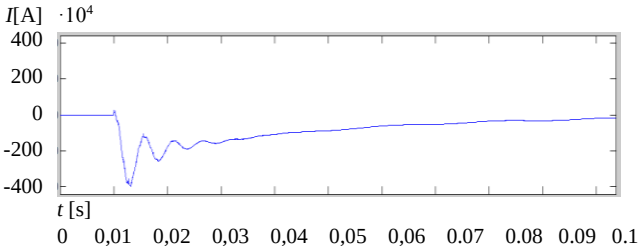


Figure 13. Fault current in case of resonant grounded neutral point, cable length 40 km, a fault occurs at 10 ms

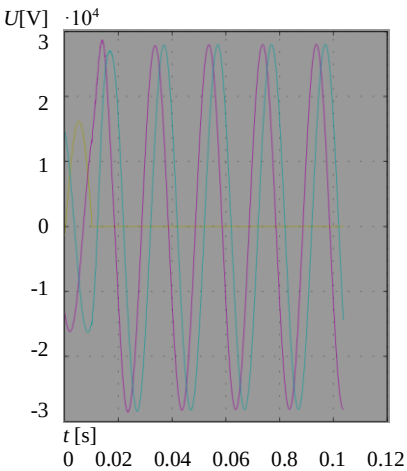


Figure 14. Overvoltages in case of resonant grounded neutral point, cable length 40 km, a fault occurs at 10 ms

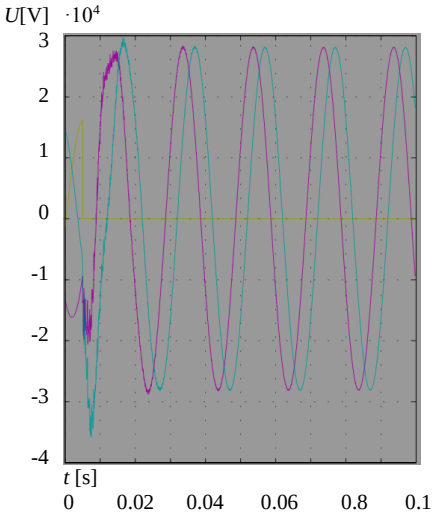


Figure 15. Overvoltages in case of resonant grounded neutral point, cable length 40 km, a fault occurs at 5 ms

D. Partial compensation

Results for two cable lengths and two moments when a fault occurs in networks with partial compensation are given in table 5. Figure 16 and figure 17 shows that the current does not increase due to long cables. Figure 18 and figure 19 show influence that moment when a fault occurs have on overvoltages in case of partial compensation.

TABLE 5. SIMULATION RESULTS IN CASE OF PARTIAL COMPENSATION

l (km)	***t* (ms)**	***Fault current***	***Overvoltage***	***Max. overvoltage***
1	10	282.49	1.697	1.745
1	5	282.94	1.692	2.214
40	10	244.66	1.649	1.686
40	5	244.63	1.655	2.065

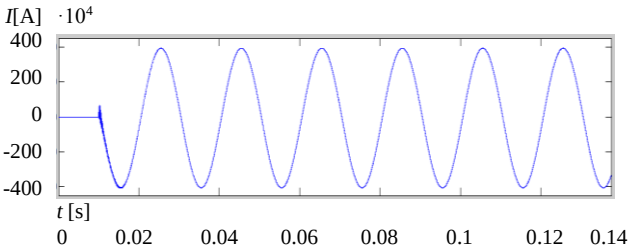


Figure 16. Fault current in case of partial compensation, cable length 1 km, a fault occurs at 10 ms

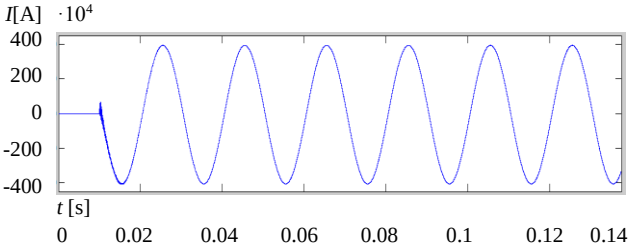


Figure 17. Fault current in case of partial compensation, cable length 40 km, a fault occurs at 10 ms

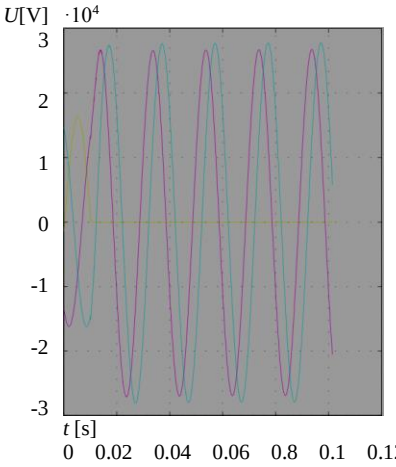


Figure 18. Overvoltages in case of partial compensation, cable length 40 km, a fault occurs at 10 ms

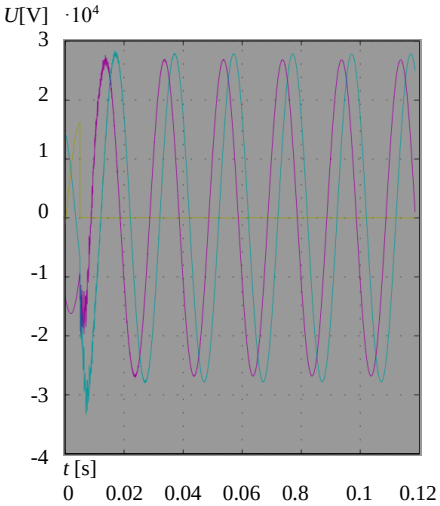


Figure 19. Overvoltages in case of partial compensation, cable length 40 km, a fault occurs at 5 ms

VI. ANALYSIS OF RESULTS

As expected, during a single phase earth fault the voltage of the faulted phase collapsed while the voltage of the healthy phases increased. The voltage of the faulted phase for all four transformer neutral point earthing methods behaved as expected. Phase to phase voltage of all three phases for all four methods of transformer neutral point grounding treatment remained unaffected.

Highest overvoltages were expected to occur in case of isolated neutral point and in case of resonant grounded neutral point. In case of low-ohmic grounded neutral point and partial compensation overvoltages were expected to be similar. Results show matching of expected and simulated results.

It has been showed that overvoltages strongly depend on the neutral point method and on the moment when the fault occurs. Maximum calculated overvoltage factor (transient overvoltages during an earth fault) in isolated network was 3.316.

Highest single pole fault current was expected to occur in case of networks with low-ohmic grounded neutral point. Calculated and simulated results show match with expected values. Fault current in case of networks with low-ohmic grounded neutral point increases from 283.27A to 314.31due to larger number of long cables. In case of isolated neutral point single pole fault current increases from 4.78 A to 204.67 A. In case of resonant grounded neutral point single pole fault current remains the same (total compensation), as expected. Increase of the length of the cables does not influence the value of the current in case of the partial compensation, too. Current depends only on the value of the low-ohmic resistor.

The aim of the study was to show influence various neutral earthing methods, different length and type of the middle voltage power lines and different moments when a fault occurs have on overvoltages and fault current magnitudes. For this particular substation observed in paper resonant grounded neutral point is recommended. In that case danger of too high potential (touch voltage) on earth of substation is reduced to a minimum (fault current is zero). When considering single phase to earth faults in medium voltage networks handling safety issues, especially touch voltages, is the most important aspect.

VII. CONCLUSION

The aim of this paper was to observe influence several neutral grounding concepts, different middle voltage cable lengths and various moments when a fault occurs have on the value of fault current and overvoltages. As expected, the neutral grounding concept determines the behavior of a power system during a single phase to earth fault. With the increase of cable lengths current increases, too, except in case when Peterson coil is connected in order to compensate capacitive component of fault current. In case when Petersen coil is connected between neutral point and earth cable length doesn't influence the value of fault current. Cable length doesn't influence overvoltages as much it influences the value of the current. The moment when a fault occurs has great influence on overvoltages. If a fault occurs at 5 ms overvoltages increase significantly.

Listed neutral earthing concepts should be considered for a specific substation. A decision which concept to choose should be adopted based on simulation results for values of the short circuit currents and overvoltages. The model described in this paper can be used in order to present what particular kind of neutral earthing method to use in specific substation.

REFERENCE

[1] V. Komen and R. Ćućić, "Fault resistances in medium voltage distribution networks" Annals of DAAAM & Proceedings, DAAAM International Vienna, 2007

[2] "Application Guidelines - Overvoltage Protection: Dimensioning, testing and application of metal oxide surge arresters in medium voltage systems", ABB,; 4th revised and expanded edition, Wettingen; 2009

[3] "Mathworks, Accelerating the pace of engineering and science", http://www.mathworks.com/products/simpower/, 26.11.2014.

[4] R. Goić, D. Jukus, I. Penović, "Electric power distribution", Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia, 2008

[5] Elka, "Medium voltage power cables for rated voltage up to 36 kV", http://elka.hr/en/category/energetski-srednjenaponski-kabeli-za-napone-do-36-kv/, 26.11.2014.

[6] G. Druml, A. Kugi and B. Parr, “Control of Peterson coils”, XI. International Symposium on Theoretical Electrical Engineering, Liz, Austria, 2001.

[7] B. Dolnik and J. Kurmisky, “ Contribution to earth fault current compensation in middle voltage distribution networks”, Przglad Electrotechniczny (Electrical Review), 0033-2097, R. 87, 2/2011

[8] I. G. Kuliš, A. Marušić and G. Leci, “Protection relay software models in interaction with power system simulators”, MIPRO, Opatija, 2012.

[9] I. G. Kuliš, A. Marušić and S. Žutobradić, “Insufficiency of Wattmetric Protection in Resonant Grounded Networks”, Proceedings of the 8th International Conference on Developments in Power System Protection (DPSP), Amsterdam, 2004., pp. 486-489.

[10] Y. Chun-wen, W. Ming-hui, J. Litingvand L. Ling-ling, “MATLAB Power System Simulation of Resonant Grounding Fault”, Energy Procedia, 17, 2012., pp. 1761-1768.

[11] S. Hänninen, “Single phase earth faults in high impedance grounded networks, Characteristics, indication and location”, disertation for the degree of doctor of technology, Helsinki University of Technology, Espoo, Finland, 2001.

[12] O. Siirto, M. Loukkalahti, M. Hyvarinen, P. Heine and M. Lehtonen, “Neutral point treatment and earth fault suppression”, IEEE Electric Power Quality and Supply ReliabilityConference, Tartu, Estonia, 2012

Utjecaj momenta tereta na brzinu vrtnje sinkronog elektromotornog pogona

Marko Pavetić, Toni Bjažić
Tehničko veleučilište u Zagrebu
Konavoska 2, 10000 Zagreb

***Sažetak*—U ovom radu opisan je sustav dvaju mehanički spregnutih elektromotornih pogona. Primarni elektromotorni pogon je sinkroni s permanentnim magnetima, zasnovan na Siemensovom Sinamics S120 modularnom sustavu, a sekundarni pogon je asinkroni, također zasnovan na Sinamics S120 modularnom sustavu. Na primarnom pogonu projektirani su regulatori momenta i brzine vrtnje, a sekundarni pogon je konfiguriran za promjenu momenta tereta, odnosno za emulaciju opterećenja primarnog elektromotornog pogona. Dana je usporedba odziva na promjenu referentne veličine i momenta tereta za regulator brzine vrtnje sinkronog pogona projektiran prema tehničkom i simetričnom optimumu. Budući da je rad prvenstveno edukativnog karaktera, dane su i smjernice za emulaciju rada radnih mehanizama.**

I. Uvod

Za učenje i proučavanje rada elektromotornih pogona najčešće se koriste laboratorijske makete manjih snaga. Pri tome se opterećenje elektromotornog pogona vrši drugim elektromotornim pogonom bliskih karakteristika. Mehanička sprega vratila dvaju pogona izvodi se krutim ili elastičnim spojkama. Prednost elastičnih spojki je jednostavnije prenošenje maketa na mjesto demostracije, budući da se pogoni mogu lakše odvojiti i pojedinačno prenijeti na potrebno mjesto.

U ovom radu je opisan sustav elektromotornih pogona, koji su mehanički spregnuti zupčastim remenim prijenosom. Prikazano je ponašanje sinkronog elektromotornog pogona pod utjecajem momenta tereta te je direktan nastavak na završni rad [1], izrađen na Tehničkom veleučilištu u Zagrebu, te objavljen kao članak u [2]. U tom je završnom radu opisana optimizacija sinkronog elektromotornog pogona regulatorima struje i brzine vrtnje, u neopterećenom stanju, kojom se postiže brži odziv pogona na promjenu referentne brzine vrtnje, uz ostvareno manje nadvišenje odziva, nego u slučaju automatski određenih parametara od strane proizvođača. U ovom radu prikazano je ponašanje istog sinkronog pogona, ali uz promjenjivi moment tereta koji se izvodi asinkronim motorom spojenim preko zupčastog remenog prijenosa.

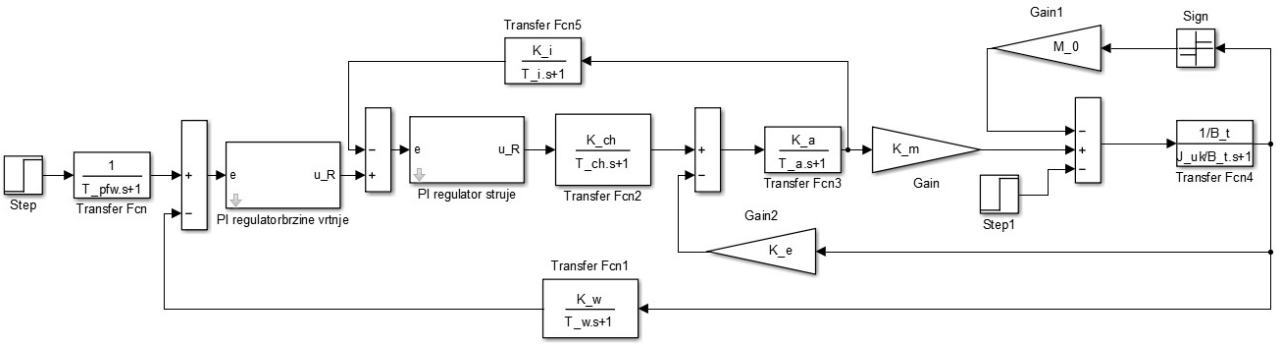
Važan dio pri proučavanju rada elektromotornih pogona predstavlja sposobnost predviđanja određenih pojava, uz poznavanje parametara pogona. Zbog toga su u ovom radu pri iznošenju eksperimentalnih rezultata mjerenja dani i simulacijski rezultati, dobiveni pomoću programskog paketa *Matlab - Simulink*, te su komentirana pojedina odstupanja. Time se bitno povećava kvaliteta edukacije iz područja elektromotornih pogona.

U drugom poglavlju opisan je sustav elektromotornih pogona te je dan matematički model s parametrima za *Simulink*. U trećem poglavlju dani su parametri regulatora struje (momenta) dobiveni pomoću tehničkog optimuma te parametri regulatora brzine vrtnje dobiveni pomoću tehničkog i simetričnog optimuma. U četvrtom poglavlju dan je opis asinkronog pogona kao i postupak podešenja istog za regulaciju momenta. Korištena je alternativna metoda regulacije momenta iz razloga što je direktna onemogućena zaštitom postavljenom od strane proizvođača. U petom poglavlju dani su odzivi veličina sinkronog elektromotornog pogona na promjenu momenta tereta uzrokovanu asinkronim pogonom. U šestom poglavlju dana su zaključna razmatranja i planirani budući rad.

II. Opis sustava i matematički model za *Simulink*

Elektromotorni pogon na kojem su vršena testiranja prikazan je na slici 1. Zasnovan je na Siemensovom Sinamics S120 modularnom sustavu [3] koji se upravlja pomoću programskog paketa *Starter*. U lijevom ormaru na slici 1 prikazan je sinkroni elektromotorni pogon. Sastoji se od sinkronog motora s permanentnim magnetima nazivne snage 0,31 kW i nazivne brzine vrtnje 3000 o/min, energetske jedinice (SINAMICS) nazivne snage 0,37 kW te upravljačke jedinice (SINAMICS S120) s DRIVE CLiQ sučeljem za prihvat signala s rezolvera za mjerenje pozicije motora. U desnom ormaru na slici 1 prikazan je asinkroni elektromotorni pogon. Sastoji se od četveropolnog kaveznog asinkronog motora nazivne snage 0,18 kW, energetske jedinice (SINAMICS) nazivne snage 0,37 kW te upravljačke jedinice (SINAMICS S120). Pozi-

Slika 1. Sinkroni i asinkroni pogon spojeni zupčastim remenom.



Slika 2. Simulink model elektromotornog pogona.

cija motora mjeri se inkrementalnim enkoderom. Sinkroni pogon je konfiguriran i optimalno parametriran postupkom opisanim u poglavlju 3. Asinkroni pogon je konfiguriran na temelju tvorničkih postavki, bez optimizacije.

Simulink model sinkronog elektromotornog pogona s anti-windup PI regulatorima momenta i brzine vrtnje prikazan je na slici 2. Model je ekvivalentan klasičnom istosmjernom pogonu s nezavisnom i konstantnom uzbudom, uz drugačija značenja pojedinih parametara pogona [1], [2]. U tablici I dani su parametri dobiveni iz *Startera* i kataloga pogona. Proračunati i procijenjeni parametri dani su u tablici II.

Pri modeliranju pogona osnovna je pretpostavka kruta mehanička sprega dvaju pogona. Zbog toga se spajanjem vratila asinkronog motora na sinkroni preko krutog zupčastog remena promijenio ukupni moment inercije koji sinkroni motor vidi. Ukupni moment inercije dobiven je zbrajanjem momenta inercije sinkronog motora i momenta inercije asinkronog motora pomnoženog s kvadratom omjera remenica obaju motora [4], te je konačni iznos $J_u = 4 \cdot 10^{-4}$ kgm^2.

Iznos modeliranog Coulombovog trenja određen je mjerenjem q-komponente struje sinkronog pogona u stacionarnom stanju $I_{q,ss}$, uz mehanički vezan asinkroni pogon s isključenim pretvaračem. Dobivena vrijednost je:

Tablica I
PARAMETRI POGONA TEMELJENI NA KATALOGU I PARAMETRIMA DOBIVENIM U *Starteru*.

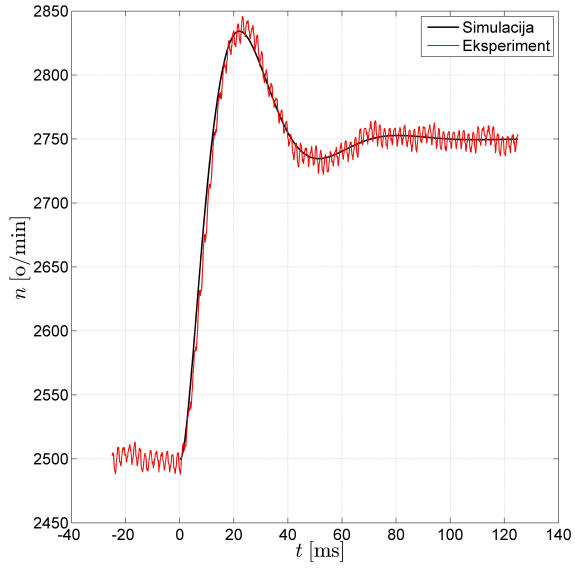
Parametar	Iznos
P_n	310 W
n_n	3000 o/min
M_n	1 Nm
I_{an}	1,6 A
n_{pp}	3
η	85 %
R_s	5,53135 Ω
R_{ch}	0,135 Ω
L_d	19,56624 mH
L_q	16,89075 mH
f_{sw}	4 kHz
$U_{d,ref}$	0 V
$U_{q,ref}$	145 V
E_n	135 V
U_{an}	153 V

Tablica II
PARAMETRI POGONA TEMELJENI NA PRORAČUNU I PROCJENI.

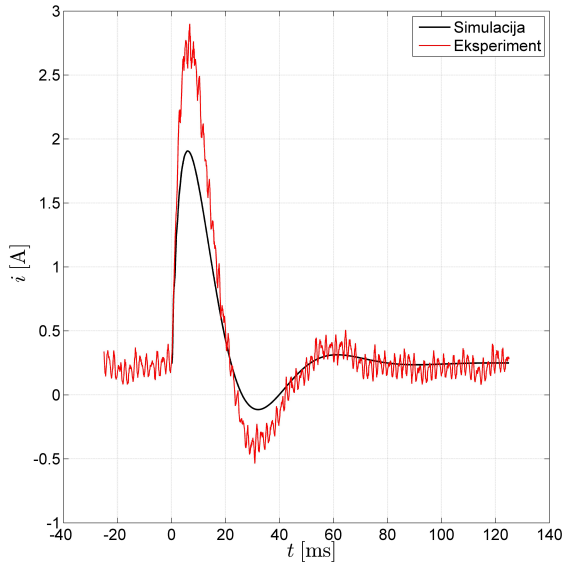
Parametar	Iznos
J_{am}	$1{,}45 \cdot 10^{-3}$ kgm^2
M_0	0,1474 Nm
B_t	$6{,}484 \cdot 10^{-5}$ Nms/rad
K_a	0,1765 A/V
T_a	3 ms
K_{ch}	3,08
T_{ch}	31,25 μs
K_i	1,00
T_i	46,576 μs
K_w	1
T_w	1 ms
K_e	0,4927 Vs/rad

$$M_0 = K_m \cdot I_{q,ss} = 0{,}1474 \text{ Nm}. \quad (1)$$

Usporedni rezultati eksperimenta i simulacije odziva brzine vrtnje na promjenu referentne brzine za 250 o/min dani su na slici 3, a struje na slici 4.



Slika 3. Usporedba simulacijskog i eksperimentalnog odziva brzine vrtnje na promjenu referentne brzine od 250 o/min.



Slika 4. Usporedba simulacijskog i eksperimentalnog odziva q-komponente struje na promjenu referentne brzine od 250 o/min.

Simulacija brzine vrtnje približno se poklapa sa rezultatom eksperimentalnog testiranja.

Ako usporedimo simulaciju i eksperimentalna mjerenja u slučaju struje, vidljiva je velika razlika u maksimalnom iznosu struje, gotovo 1 A. Razlika je takva zbog toga što je kod simulacije u obzir uzeto samo Coulombovo trenje. Bolji rezultati mogli bi se dobiti drukčijim modelom trenja (npr. Stribeckovo trenje). U obzir nije uzeta ni elastičnost remena koja također doprinosi ovoj razlici.

III. Parametri regulatora momenta i brzine vrtnje

Rezultati proračuna parametara regulatora dani su u ovom poglavlju, a potpuni proračuni novih parametara regulatora dani su u [4]. Korišteni su tehnički optimum (TO) za projektiranje regulatora struje, te tehnički i simetrični (SO), za projektiranje regulatora brzine vrtnje [5]. Parametri PI regulatora momenta dobiveni tehničkim optimumom uz zadano nadvišenje 10% su:

$$K_{Rit} = 27{,}9, \quad T_{Iit} = 3 \text{ ms}. \tag{2}$$

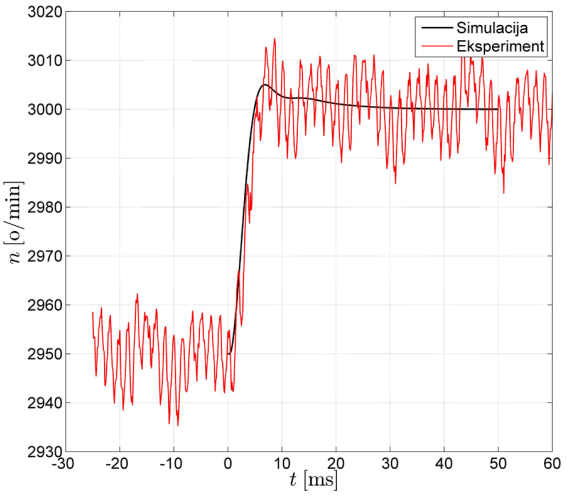
Parametri PI regulatora brzine vrtnje dobiveni simetričnim optimumom uz zadano nadvišenje 25%, korekciju na 40%, te uz PT_1 prefiltar da se postigne 10%, su [4], [5]:

$$K_{R\omega s} = 0{,}4, \quad T_{I\omega s} = 9 \text{ ms}, \quad T_{pf\omega} = 2{,}72 \text{ ms}. \tag{3}$$

Parametri PI regulatora brzine vrtnje dobiveni tehničkim optimumom uz zadano nadvišenje 40% su [4], [5]:

$$K_{R\omega t} = 1{,}2, \quad T_{I\omega t} = 6{,}2 \text{ s}. \tag{4}$$

Proračunati parametri regulatora uneseni su u *Starter* i izvršeni su eksperimenti. Usporedni rezultati simulacije i eksperimenta u slučaju korištenja simetričnog optimuma s parametrima regulatora (3) prikazani su slikom 5 za brzinu vrtnje, te slikom 6 za struju. Testiranja su odrađena na



Slika 5. Usporedba simulacijskih i eksperimentalnih rezultata u primjeni regulatora brzine vrtnje projektiranog pomoću simetričnog optimuma i uključenim prefiltrom brzine vrtnje za postizanje 10% nadvišenja.

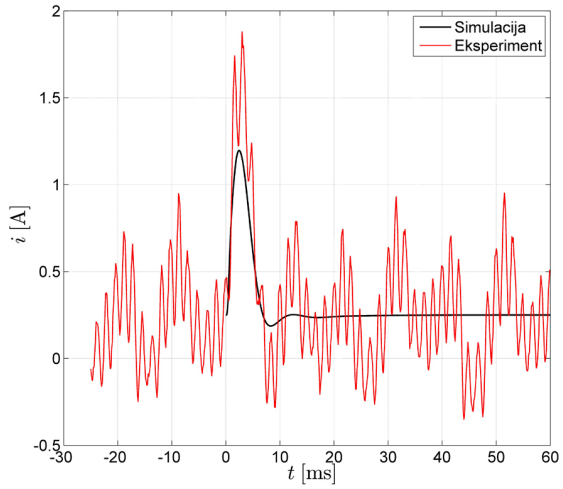
Tablica III
Pokazatelji kvalitete prijelazne pojave dobiveni iz simulacijskih rezultata prikazanih na slici 5.

Parametar	iznos
$K_{R\omega s}$	0,4
$T_{I\omega s}$[ms]	9
t_m[ms]	6,85
t_p[ms]	30
σ_m[%]	10
$T_{pf\omega}$[ms]	2,72

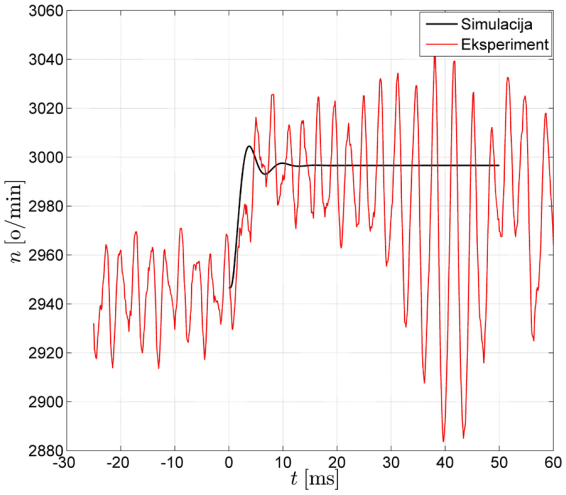
način da je referentna brzina promijenjena za 50 o/min. Zadana promjena je relativno mala kako bi regulator u svim testiranim uvjetima ostao u linearnom režimu rada (više rezultata dano je u [4]). Dodatno je potrebno napomenuti da je pogon vektorski upravljan, u d, q koordinatnom sustavu. Struja i_d regulacijom se održava na nuli, a mjerena struja je i_q i ona predstavlja mjeru elektromagnetskog momenta [1], [2]. Na slikama je struja i_q označena samo oznakom i.

Odzivi sustava korištenjem regulatora projektiranog pomoću tehničkog optimuma prikazani su na slici 7. Rezultati odziva simulacije dani su u tablici IV. U praksi nije bilo moguće primijeniti te parametre. Kao što je vidljivo na slici 7, zbog proklizavanja remena izražene su oscilacije. Potrebno je smanjiti pojačanje regulatora dok se ne dobije ono sa kojim su oscilacije prihvatljive. Korigirani koeficijent pojačanja iznosi $K_{R\omega t}$ = 0,4, a prefiltar brzine vrtnje je isključen ($T_{pf\omega}$ = 0 s). Odzivi simulacije i eksperimenta, brzine vrtnje i struje, dani su na slikama 8 i 9, s rezultatima u tablici V.

Odabrani koeficijent pojačanja omogućuje rad sustava bez izraženih oscilacija, no pojačanje je sniženo toliko da prefiltar brzine vrtnje nije potreban jer nadvišenje je manje od 10%.



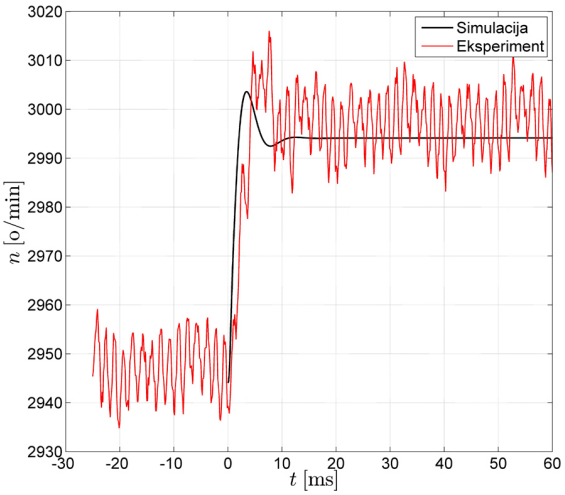
Slika 6. Usporedba simulacijskih i eksperimentalnih odziva struje i_q kod primjene regulatora brzine vrtnje projektiranog pomoću simetričnog optimuma i uključenim prefiltrom brzine vrtnje za postizanje 10% nadvišenja.



Slika 7. Usporedba simulacijskih i eksperimentalnih rezultata u primjeni regulatora brzine vrtnje projektiranog pomoću tehničkog optimuma za nadvišenje 40% i uključenim prefiltrom brzine vrtnje za spuštanje nadvišenja na 10%.

Tablica IV
POKAZATELJI KVALITETE PRIJELAZNE POJAVE DOBIVENI IZ SIMULACIJSKIH REZULTATA PRIKAZANIH NA SLICI 7.

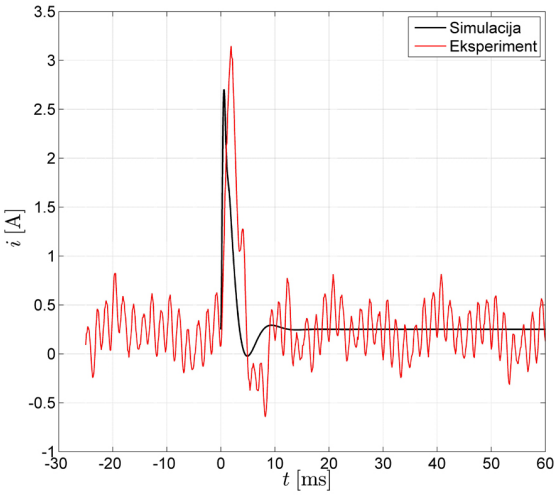
Parametar	iznos
$K_{R\omega t}$	0,7
$T_{I\omega t}$[s]	6,234
t_m[ms]	3,75
t_p[ms]	13
σ_m[%]	10
$T_{pf\omega}$[ms]	1,58



Slika 8. Usporedba simulacijskih i eksperimentalnih rezultata u primjeni regulatora brzine vrtnje projektiranog pomoću tehničkog optimuma sa pojačanjem odabranim tako da remen ne proklizava.

Tablica V
POKAZATELJI KVALITETE PRIJELAZNE POJAVE DOBIVENI IZ SIMULACIJSKIH REZULTATA PRIKAZANIH NA SLICI 8.

Parametar	iznos
$K_{R\omega t}$	0,4
$T_{I\omega t}$[s]	6,234
t_m[ms]	3,75
t_p[s]	$\approx 5 \cdot T_{I\omega t}$
σ_m[%]	6
$T_{pf\omega}$[ms]	0



Slika 9. Usporedba simulacijskih i eksperimentalnih odziva struje i_q kod primjene regulatora brzine vrtnje projektiranog pomoću tehničkog optimuma sa pojačanjem odabranim tako da remen ne proklizava.

IV. PRILAGODBA ASINKRONOG POGONA ZA REGULACIJU MOMENTA TERETA

Asinkroni pogon je Siemensov proizvod, sličan sinkronom pogonu na kojem se vrše testiranja. Sastoji se od upravljačke jedinice CU320-2PN i pretvarača PM340 [1], [2]. Upravljani elektromotor u tom pogonu je trofazni asinkroni, model 1LA9063-4KA11 (tablica VI).

Asinkroni pogon se pušta u rad po uputama danim u [6]. Dodatne postavke uključuju isključivanje rampe u grani referentne struje, čime se omogućuje gotovo skokovita promjena momenta, te ograničenje momenta (slika 10).

Kao što je vidljivo na slici 10, imamo ikonu pod nazivom *Torque limits*, i u njoj se nalaze polja za promjenu iznosa momenta. Moment se u polju *Upper Limit* ograničava na iznos 0, tako da ne može djelovati u smjeru momenta sinkornog motora, a u polju *Lower limit* na iznos -1 Nm. Iznos u polju *Lower limit* je ključna postavka. Kada bi prijenosni omjer između remenica pogona bio 1:1 tada bi iznos u polju *Lower limit* bio ekvivalentan nominalnom iznosu momenta sinkronog pogona kojem stvaramo teret. No, omjer remenica je 1:1,36. To znači da kada asinkronim pogonom postignemo maksimalni moment (zadan u polju *Lower limit*), u našem slučaju 1 Nm, sinkroni pogon vidi (terećen je sa) otprilike 0,73 Nm. Isto tako, kada bi sinkronim motorom htjeli teretiti asinkroni motor 1 Nm tada bi na sinkronom trebalo postići 0,73 Nm.

Testiranje sinkronog motora pod utjecajem momenta tereta uzrokovanog sa asinkronim motorom vršeno je kod brzine vrtnje 1200 o/min. Uključivanjem sinkronog motora na 1200 o/min, uključeni i podešeni (s gore opisanim postavkama) asinkroni motor se, s obzirom na prijenosni omjer, vrti otprilike 880 o/min i u tom slučaju sinkroni pogon je opterećen momentom tereta iznosa 0,73 Nm.

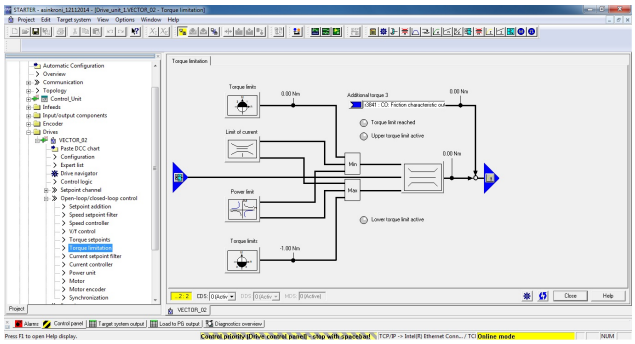
Sinkroni pogon opterećen je s 0,73 Nm iz razloga što asinkroni pogon nije ubrzan, tj. njegova zadana brzina (u *Starteru*) je 0 i on pokušava zadržati tu brzinu (odupire se), a kada podignemo brzinu asinkronog motora preko 880 o/min, npr. 1500 o/min, sinkroni motor je opterećen momentom tereta iznosa 0,22 Nm, što predstavlja iznos momenta trenja i ventilacije (modelirano Coulombovo trenje). Dakle, asinkroni motor je povećanjem zadane (referentne) brzine smanjio moment tereta koji sinkroni motor vidi, jer je pri zadanoj brzini 1500 o/min moment asinkronog pogona ograničen na nulu pa tako sinkroni pogon ne osjeća ništa osim momenta trenja i ventilacije.

Postavljanjem referentne brzine asinkronog pogona na iznos 0, uzrokuje se gotovo skokovito povećanje momenta tereta (udar momenta) sinkronog pogona. To uzrokuje pad brzine vrtnje sinkronog pogona i povećanje struje, čiji su valni oblici snimljeni i analizirani u sljedećem poglavlju.

U postavkama prikazanim na slici 10 moguće je podešavanje za emulaciju različitih radnih mehanizama (ventilatora, pumpi, itd.) posredstvom DCC-a (Drive Control Chart). DCC je grafička konfiguracija i proširenje funkcionalnosti uređaja korištenjem kontrolnih, aritmetičkih i logičkih blokova [7].

Tablica VI
SPECIFIKACIJE TROFAZNOG ASINKORNOG MOTORA 1LA9063-4KA11.

Parametar	Iznos
P_n[s]	0.18 kW
n_n[ms]	1395 o/min
M_n[s]	1,2 Nm
n_{pp}	4
U_n	230V trokut /400V zvijezda



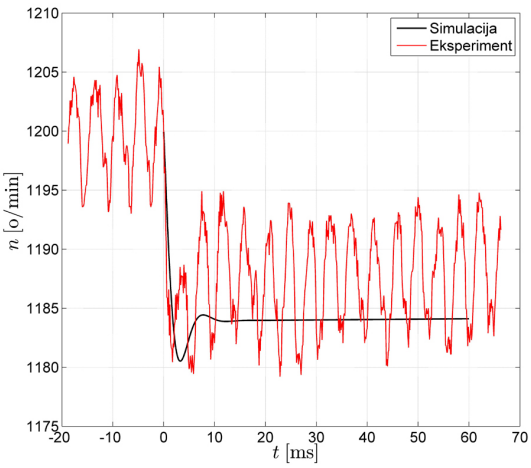
Slika 10. Postavke momenta za asinkroni pogon u *Starteru*.

V. USPOREDBA REZULTATA S MOMENTOM TERETA

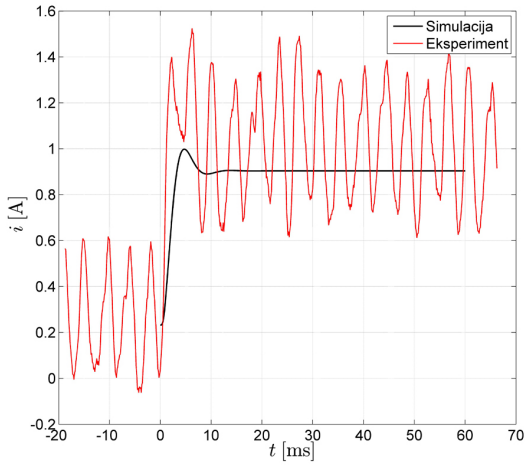
Na slikama 11 – 14 dani su simulacijski i eksperimentalni odzivi brzine vrtnje i struje na promjenu momenta tereta, za tehnički i simetrični optimum.

Vidljive su izražene oscilacije u odzivu eksperimenta i kod brzine vrtnje i kod struje. Amplituda oscilacija u prosjeku iznosi 13 o/min *peak-to-peak*. Javlja se iz razloga što stabilnost pogona (motora) u odnosu na podlogu nije u potpunosti zadovoljena. Također, utjecaj elastičnosti remena i trenje na obje remenice pridonosi većem iznosu oscilacija. Isto tako, čvršći spoj motora sa kućištem te robusnija izvedba kućišta, smanjili bi utjecaj vibracija na odziv sustava.

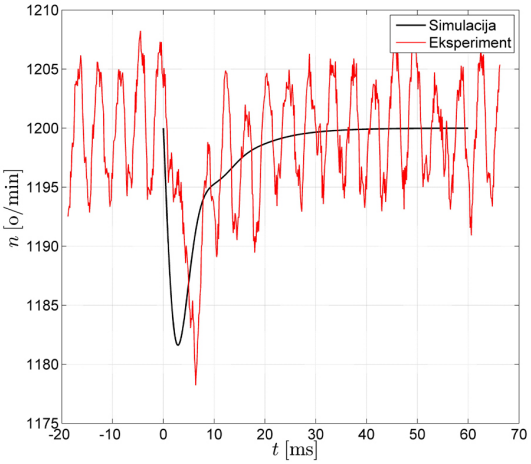
Ono što je najbitnije uočiti na slikama 11 – 14 jest to da je simetrični optimum za projektiranje regulatora brzine vrtnje neosporno bolji izbor od tehničkog optimuma.



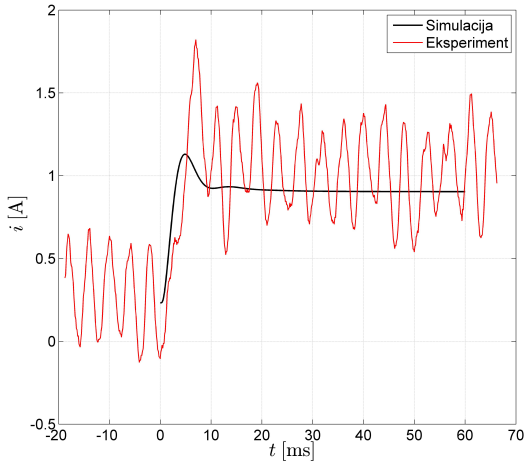
Slika 11. Usporedba simulacijskih i eksperimentalnih odziva brzine vrtnje uz primjenu tehničkog optimuma s parametrima (4).



Slika 12. Usporedba simulacijskih i eksperimentalnih odziva struje uz primjenu tehničkog optimuma s parametrima (4).



Slika 13. Usporedba simulacijskih i eksperimentalnih odziva brzine vrtnje uz primjenu tehničkog optimuma s parametrima (3).



Slika 14. Usporedba simulacijskih i eksperimentalnih odziva struje uz primjenu tehničkog optimuma s parametrima (3).

Odzivi na promjenu referentne brzine su po kvaliteti sumjerljivi, ali glavni razlog izbora simetričnog optimuma u praksi je duljina trajanja kompenzacije poremećaja (momenta tereta). Kod tehničkog optimuma vrijeme kompenzacije poremećaja iznosi otprilike 30 sekundi (cca. 5 integralnih vremenskih konstanti regulatora), a kod simetričnog optimuma vrijeme kompenzacije iznosi otprilike 45 milisekundi. Iznosi maksimalnog propada brzine vrtnje pri djelovanju momenta tereta su također sumjerljivi u oba slučaja, zbog jednakog iznosa pojačanja regulatora.

VI. Zaključak

Terećenje motora drugim motorom zahtjeva poznavanje načina rada upravljačkih jedinica oba pogona, jer se u njima određuju ključna podešenja. Siemensova upravljačka jedinica S120 i pripadajući programski paket *Starter* pružaju mnogo mogućnosti podešavanja sustava, ali ne omogućuju direktnu regulaciju momenta. Stoga je u ovom radu opisano indirektno rješenje za regulaciju momenta.

Detaljnim eksperimentiranjem potvrđena je važnost dobre mehaničke sprege remena i remenice, te stabilnosti pogona u odnosu na podlogu na kojoj se nalaze. Oscilacije u odzivu brzine vrtnje od prosječno 13 o/min *peak-to-peak* su neizbježne zbog elastičnosti remena, no mogu biti i puno veće ako se ne zadovolji uvjet mehaničke stabilnosti pogona. Oscilacije u odzivu struje su također neizbježne zbog samog načina doziranja iste – sklapanjem u pretvaraču, ali u ovom postavu sustava dominantan utjecaj na jako veliku oscilatornost struje u stacionarnom stanju, a time i brzinu vrtnje, ima mehanička sprega dvaju pogona remenom preko kojeg se prenose vibracije.

Usporedbom rezultata učinaka oba optimuma za projektiranje regulatora brzine jasno se može vidjeti koji je od navedenih superirorniji po kvaliteti prijelazne pojave sustava za koji su projektirani. Simetrični optimum proračunom gotovo da ne odstupa od predviđenog, a tehnički optimum ima veća odstupanja i uzrokuje forsiranje sustava. Također, ukupno vrijeme kompenzacije poremećaja je predugačko te je time tehnički optimum neprihvatljiv za brzinu vrtnje elektromotornih pogona.

Budući rad na ovoj temi mogao bi biti promjena granica momenta u *Starteru*, kako bi se mogla ostvariti emulacija rada različitih radnih mehanizama.

Literatura

[1] M. Boršić, *Upravljanje brzinom vrtnje sinkronog motora s permanentnim magnetima*. Tehničko veleučilište u Zagrebu, 2014.

[2] M. Boršić and T. Bjažić, "Upravljanje brzinom vrtnje sinkronog motora s permanentnim magnetima," *Polytechnic and Design*, vol. 3, no. 1, pp. 52–65, 2015.

[3] Siemens, "Sinamics S120/150 user manual," http://www.automation.siemens.com/doconweb/pdf/SINUMERIK_SINAMICS_02_2012_E/LH1.pdf?HTTPS=REDIR&p=1", 13.11.2014.

[4] M. Pavetić, *Utjecaj momenta tereta na brzinu vrtnje sinkronog elektromotornog pogona*. Tehničko veleučilište u Zagrebu, 2014.

[5] T. Bjažić and P. Crnošija, *Osnove automatike (1. dio: Analiza i sinteza kontinuiranih sustava - teorija i primjena)*. Zagreb: Element, 2011.

[6] M. Majcenić, *Rad s asinkronim pogonom*. Tehničko veleučilište u Zagrebu, 2014.

[7] Siemens, "Siemens forum," http://www.automation.siemens.com/WW/forum/guests/PostShow.aspx?PageIndex=1&PostID=591863&Language=en#top", 13.11.2014.

Funkcije naprednih mjernih sustava za električnu energiju

Zdravko Lipošćak
HEP-Operator distribucijskog sustava d.o.o., Zagreb, Hrvatska
zdravko.liposcak @hep.hr

Funkcije naprednih mjernih sustava za električnu energiju utvrđuje operator distribucijskog sustava te ih dostavlja regulatornoj Agenciji (HERA) koja na temelju dostavljenih podataka provodi analizu troška i dobiti te pribavlja mišljenje predstavnika tijela za zaštitu potrošača za uvođenje naprednih mjernih uređaja za krajnje kupce. Nadležni ministar na temelju analize utvrđuje odlukom plan i program mjera za uvođenje naprednih mjernih uređaja za krajnje kupce, a provođenje odluke nadzire Agencija. Prilikom korištenja naprednih mjernih uređaja nužno je prikupljati, obrađivati i koristiti mjerne podatke u skladu s propisima o zaštititi osobnih podataka. Od sustava naprednog mjerenja potrošnje električne energije očekuje se veliki doprinos u ostvarenju ciljeva 20/20/20 Europske Unije do kraja 2020. godine; smanjenje potrošnja energije za 20%, smanjenje emisije stakleničkih plinova za 20% i povećanje udjela energije iz obnovljivih izvora za 20%.

U radu je prezentiran problema donošenja odluke o potrebnim funkcijama naprednog mjernog sustava, koji je zbog velikih očekivanja i nužnosti zadovoljenja potreba različitih subjekata, vrlo složen, važan i osjetljiv početni korak u realizaciji cijelog projekta.

I. UVOD

Direktive europske unije 2009/72/CE i 2009/73/CE od 13. srpnja 2009. godine u području naprednog mjerenja zahtijevaju da država članica osigura implementaciju naprednih mjernih sustava koji će pridonijeti aktivnom sudjelovanju kupaca na tržištu električne energije. Isto tako potrebno je provesti ekonomsku analizu dugoročnih troškova i dobiti te pripremiti vremenski plan izgradnje naprednog mjernog sustava u periodu od najviše deset godina. Ukoliko ekonomska analiza pokaže isplativost ugradnje naprednih mjernih sustava za određene kategorije kupaca ili područja primjene, najmanje 80% obračunskih mjernih mjesta je potrebno opremiti naprednim brojilima do 2020. godine.

Zahtjevi europskih direktiva prema izgradnji naprednih mjernih sustava ugrađeni su u hrvatsko zakonodavstvo kroz Zakon o energiji , NN 120/2012 i Zakon o tržištu električne energije, NN 22/2013.

Zakonom o energiji definirano je da operator mreže/sustava utvrđuje tehničke zahtjeve i troškove uvođenja naprednih mjernih uređaja i sustava za njihovo umrežavanje te ih dostavlja Agenciji (HERA). Na temelju podataka Agencija provodi analizu troška i dobiti te pribavlja mišljenje predstavnika tijela za zaštitu potrošača za uvođenje naprednih mjernih uređaja za krajnje kupce. Ministar na temelju analize utvrđuje odlukom plan i program mjera za uvođenje naprednih mjernih uređaja za krajnje kupce, a provođenje odluke nadzire Agencija. Prilikom korištenja naprednih mjernih uređaja nužno je prikupljati, obrađivati i koristiti mjerne podatke u skladu s propisima o zaštititi osobnih podataka.

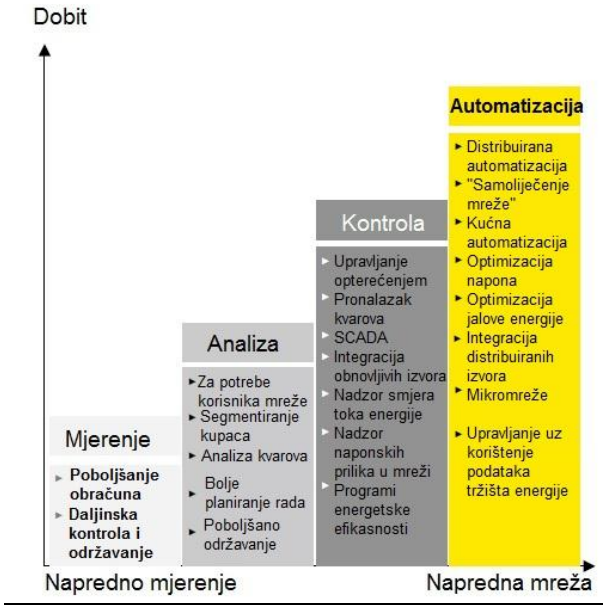
Od sustava naprednog mjerenja potrošnje električne energije očekuje se veliki doprinos u ostvarenju ciljeva 20/20/20 Europske Unije do kraja 2020. godine; smanjenje potrošnja energije za 20%, smanjenje emisije stakleničkih plinova za 20% i povećanje udjela energije iz obnovljivih izvora za 20%.

Zbog velikih očekivanja i nužnosti zadovoljenja potreba različitih subjekata, vrlo je složen i osjetljiv početni korak utvrđivanja tehničkih karakteristika naprednih mjernih uređaja i sustava za njihovo umrežavanje za potrebe hrvatskog tržišta električne energije. Posebno se preporučuje da sustav naprednih mjernih uređaja podržava barem minimalni skup funkcija predložen u Preporuci Europske Komisije 2012/148/EU, a koje su u skladu s aktivnostima normiranja u ovom području. To je potrebno za osiguravanje tehničke i komercijalne interoperabilnosti u naprednom mjerenju, osiguranje privatnosti i sigurnosti podataka te omogućavanje kreiranja i razvoja novih energetskih usluga.

II. NAPREDNI MJERNI SUSTAV

Iako je obveza ugradnje naprednih mjernih sustava za mjerenje potrošnje i proizvodnje električne energije prenesena u hrvatsko zakonodavstvo, niti na razini EU niti na razini našeg zakonodavstva nije definirano što je to napredno mjerenje. Ovaj nedostatak Europska komisija pokušala je ublažiti donošenjem Preporuke 2012/148/EU u kojoj su navedene minimalne funkcionalnosti naprednog mjernog sustava, a konačna odluka o funkcionalnostima prepuštena je državama članicama.

Napredni mjerni sustav osim kroz skup minimalnih funkcija koje mora podržavati, definira se i kao elektronički sustav koji pri mjerenju potrošnje energije osigurava više dodatnih informacija u odnosu na tradicionalna brojila električne energije te ima mogućnost razmjene podatke korištenjem različitih oblika elektronske komunikacije. Napredni mjerni sustav je jedna od najvažnijih karika napredne energetske mreže.



Slika 1. Razvoj od naprednog mjerenja do napredne distribucijske mreže i moguće dobiti novih sustava

Važnost odluke o opsegu funkcija naprednog mjernog sustava je u tome što postoji velika povezanost između željenih funkcija sustava i troškova izgradnje. Dostupni podaci iz do sada provedenih studija troškova i dobiti [1] ukazuju da bi napredni sustav mjerenja mogao stajati prosječno 200 EUR do 250 EUR po potrošaču. Trošak po mjernom mjestu kreće se od oko 100 EUR (77 EUR na Malti, 94 EUR u Italiji) do 766 EUR u Češkoj Republici.

Očekuje se da će napredni sustavi mjerenja donijeti ukupnu korist od 160 EUR po korisniku za plin i 309 EUR za električnu energiju uz pretpostavljenu uštedu energije od 3 %. Procjene uštede energije kreću se u rasponu od 0 % u Češkoj Republici do 5 % u Grčkoj i na Malti. Od zemalja koje su završile s uvođenjem, Finska i Švedska navele su uštede energije u rasponu od 1 % do 3 %.

Za opremanje svih obračunskih mjernih mjesta HEP-Operatora distribucijskog sustava, uz procjenu troškova opremanja i izgradnje sustava od 200 EUR po obračunskom mjernom mjestu, ukupna investicija za 2,3 milijuna kupaca iznosila bi oko 460 mil. EUR.

Prema Preporuci Europske komisije 2012/148/EU napredni mjerni sustav mora imati slijedeći minimalni set funkcionalnosti:

a) Za potrebe kupca, osigurati očitanja potrošnje izravno za kupca i bilo koje treću stranu određenu od strane kupca, standardno sučelje koje će omogućiti rješenja upravljanja energijom u 'stvarnom vremenu', kao što su kućne automatizacije, opremljena sa standardiziranim sučeljem koje omogućuje vizualan prikaz podataka o potrošnji.
b) Očitavati mjerne podatke dovoljno često kako bi se omogućilo njihovo korištenje za postizanje ušteda energije. Osigurati očitanja kupcu najmanje u periodima od 15 minuta te pohranu podataka za daljnje analize (da kupac može očitati povijesne podatke).
c) Za operatora mjerenja, omogućiti daljinsko očitavanja mjernih i kontrolnih podataka,
d) osigurati dvosmjernu komunikaciju između naprednog mjernog sustava i vanjske mreže za održavanje i kontrolu mjernog sustava,
e) očitanje treba vršiti dovoljno često kako bi informacije bilo moguće koristiti za planiranje rada mreže.
f) Za opskrbljivače, podržavanje naprednih tarifnih sustava, korištenje time-of-use registara, daljinska kontrola tarifa, prosljeđivanje informacija o naprednim tarifnim sustavima direktno do kupca.
g) Omogućavanje daljinskog uključenja i isključenja te ograničavanje vršnog opterećenja.
h) Za sigurnost i zaštitu podataka, osigurati sigurnu podatkovnu komunikaciju između brojila i operatora mjerenja, ali i na komunikaciju brojila prema kupcu. Za komunikaciju unutar prostora kupca osim zaštite mjernih podataka potrebno je voditi računa i o zaštiti privatnosti.
i) Sprječavanje i otkrivanje prijevara i neovlaštenog korištenja energije.
j) Za distribuiranu proizvodnju, osigurati dvosmjerno mjerenje radne i jalove energije.

III. FUNKCIONALNOSTI NAPREDNIH MJERNIH UREĐAJA U ZEMLJAMA EUROPSKE UNIJE

Napredni sustavi mjerenja trebaju:

- biti opremljeni namjenskim funkcijama u skladu s normizacijom kako je predloženo u Preporuci Komisije 2012/148/EU radi osiguranja tehničke i komercijalne interoperabilnosti, ili osiguranja mogućnosti naknadnog dodavanja funkcionalnosti,
- jamčiti privatnost i sigurnost podataka,
- omogućavati razvoj usluga za odgovor na potražnju i drugih energetskih usluga i
- podržavati tržišta maloprodaje kojima se potrošačima i energetskom sustavu osigurava potpuna korist.

U osam država članica koje nastavljaju s masovnim uvođenjem naprednih brojila za električnu energiju do 2020., funkcije su u potpunosti kako se preporučuje u Preporuci 2012/148/EU.

Funkcija čija realizacija predstavlja najveći izazov, odnosi se na učestalost kojom se može ažurirati podatke o potrošnji i učiniti ih dostupnima potrošačima i trećim osobama u njihovo ime. Ova će funkcija podržavati izravne povratne informacije o troškovima za potrošače, omogućiti potrošačima donošenje informiranih odluka o načinu potrošnje i olakšati razvoj novih maloprodajnih usluga i proizvoda. Sedam država članica koje nastavljaju s masovnim uvođenjem naprednog mjerenja do 2020. i tri koje ih ne namjeravaju uvesti ne zadovoljavaju tu

funkciju. Ako sustav naprednog mjerenja nema tu funkciju, države članice trebale bi osigurati da se ta funkcija može dodati naknadno ili da se ta funkcija može ponuditi na drugi način.

TABLICA I. PODRŽAVANJE MINIMALNOG SETA FUNKCIONALNOSTI NAPREDNOG MJERNOG SUSTAVA U ZEMLJAMA EU KOJE SU DONIJELE ODLUKU O MASOVNOJ UGRADNJI NAPREDNIH BROJILA

	Napredni sustav sadrži minimalnu funkcionalnost									
Zemlja	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**
Austrija	+	+	+	+	+	+	+	+	+	+
Danska	+	djel.	+	+	+	+	+	+	+	+
Estonija	+	djel.	+	+	+	+	+	+	+	+
Finska	+	djel.	+	+	+	+	+	+	+	+
Francuska	+	+	+	+	+	+	+	+	+	+
Grčka	+	+	+	+	+	+	+	+	+	+
Irska	+	+	+	+	+	+	+	+	+	+
Italija	+	djel.	+	+	+	+	+	+	+	+
Luksemburg	-	-	+	+	-	-	+	+	-	-
Malta	djel.	+	+	+	+	djel.	+	djel.	+	+
Nizozemska	+	+	+	+	+	+	+	+	+	+
Poljska	+	+	+	+	+	+	+	+	+	+
Rumunjska	+	+	+	+	+	+	+	+	+	+
Španjolska	+	-	+	+	+	+	+	+	+	+
Švedska	+	djel.	+	+	+	+	djel.	djel.	+	+
Velika Britanija	+	+	+	+	+	+	+	+	+	+
+	sadrži funkcionalnost									
-	ne sadrži funkcionalnost									
djel.	djelomično sadrži funkcionalnost									

Čini se da ne postoji izravna veza između raspona zajedničkih minimalnih funkcija koje se razmatraju za uvođenje naprednih sustava mjerenja i njihovog ukupnog troška. Drugim riječima, odabir manjeg broja funkcija iz skupa zajedničkih minimalnih funkcija nužno ne podrazumijeva manje skupe sustave. Štoviše, razlike u podacima o „trošku po mjernom mjestu" u državama članicama pokazuju da na ukupno ulaganje daleko više utječu drugi parametri, uključujući:

- početne uvjete,
- lokalne troškove rada,
- zemljopisna konfiguraciju,
- dodatne značajke osim minimalnog skupa funkcija i
- ukupne scenarije, diskontne stope i razdoblja procjene koji se uzimaju u obzir u pojedinačnim analizama troškova i koristi.

To snažno govori u prilog tome da se od samog početka potrebno u potpunosti pridržavati zajedničkih minimalnih funkcija. Ako se u analizi troškova i koristi države članice ne podupire taj pristup, preporučuje se da se barem omogući nadograđivanje sustava koji će se uvesti kako bi podržavali „pametne" usluge i proizvode u budućnosti. Odabir nedovoljno optimalnog, nefleksibilnog sustava koji se ne može ažurirati u konačnici će uzrokovati više troškove ako, na primjer, ubrzo nakon ugradnje budu potrebne značajne izmjene ili čak potpuna zamjena kao odgovor na zahtjeve tržišta i potrošača.

Trenutačno je samo nekoliko država članica utvrdilo smjernice o funkcionalnim zahtjevima sustava naprednog mjerenja. Ostale su prepustile analizu mogućnosti odgovornima za uvođenje, u većini slučajeva operatorima distribucijskog sustava, bez davanja jasnih poticaja ili zahtjeva kakva obilježja trebaju imati funkcije koje koriste i potrošačima.

TABLICA II. PLANIRANI VREMENSKI OKVIRI UGRADNJE NAPREDNOG MJERENJA U DRŽAVAMA ČLANICAMA EU

Država	Broj brojila [milijun]	Početna godina	Završna godina
Italija	36,7	2001	2011
Finska	3,3	2009	2013
Malta	0,26	2009	2014
Španjolska	27,7	2011	2018
Austrija	5,7	2012	2019
Poljska	16,5	2012	2020
Velika Britanija	31,99	2012	2020
Estonija	0,709	2013	2017
Rumunjska	9	2013	2020
Grčka	7	2014	2020
Francuska	35	2014	2020
Nizozemska	7,6	2014	2020
Danska	3,28	2014	2020
Luksemburg	0,26	2015	2018
Irska	2,2	2016	2019

IV. KUPCI ELEKTRIČNE ENERGIJE – KRAJNJI KORISNICI SUSTAVA

Potreban je intenzivan komunikacijski napor kako bi se pomoglo kupcima da shvate svoja prava, koristi uvođenja naprednih brojila i sudjelovanja u programima upravljane potrošnje [2]. Kupci moraju biti informirani o funkcijama brojila i o tome koji će se podaci prikupljati i u koju će se svrhu te podatke koristiti.

TABLICA III. PROSJEČNI OSNOVNI REZULTATI POZITIVNIH EKONOMSKIH ANALIZA MASOVNOG UVOĐENJA NAPREDNIH MJERNIH SUSTAVA

	Područje vrijednosti	Prosjek (na osnovu pozitivnih CBA)
Smanjenje cijene energije	od 3 do 10 %	5,7 % ± 1,8 %
Životni vijek opreme	od 8 do 20 godina	15 % ± 4 godine
Ušteda energije	od 0 do 5 %	3 % ± 1,3 %
Smanjenje vršnog opterećenja	od 1 do 13%	-
Trošak po mjernom mjestu	od 450 kn do 5.800 kn	1.640 kn ± 990 kn
Dobit po mjernom mjestu	od 140 kn do 5.800 kn	2.300 kn ± 1.200 kn
Dobit za korisnika mreže	od 20 do 80%	-

Loša informiranost kupaca generira različite strahove od nove tehnologije koji mogu uzrokovati i blokadu provođenja projekta:

Informacije -Imaju li potrošači potrebne informacije kako bi donijeli ispravne odluke o potrošnji energije?

Aktivni angažman potrošača zahtijeva slijedeće informacije:

- usporedbu s povijesnim mjernim podacima i podacima drugih potrošača,
- troškove i način korištenja energije ovisne o pojedinom uređaju i vremenu korištenja (tarifi),
- ostvarene troškove po satu i procjenu budućih troškova,
- alarme o neobičnoj potrošnji i trenutne informacije o rezultatima poduzetih aktivnosti,
- novi tipovi podataka koji mogu biti raspoloživi (kvaliteta napona i opskrbe energijom),
- kako dobiti pomoć i detaljnije informacije.

Odabir - Da bi se omogućilo kupcu da dobije najbolju ponudu, potrebno mu je:

- jasan cilj i pouzdane informacije o dostupnim ponudama,
- pristup i mogućnost korištenja neovisnih alata za usporedbu,
- jednostavniji i brži proces promjene opskrbljivača.

Privatnost - Zabrinutost za privatnost može dovesti do gubitka samopouzdanje i angažmana kupca, a uzrok zabrinutosti može biti:

- priče u medijima o prijetnjama hakera, cyber terorizmu,
- korištenje detaljnih podataka o potrošnji kupca bez njegovog dopuštenja.

Troškovi - Kupci će platiti troškove naprednih mjernih uređaja, ali ostvarene uštede su male i neki kupci neće imati koristi;

- prosječni trošak naprednih brojila očekuje se u granicama od € 200 do € 250,
- očekivane uštede energije od 5-6% su nedovoljno dokazane,
- upitna je održivost početnih ušteda ostvarenih u nekim pilot projektima,
- teško je postići neke koristi ukoliko nisu osigurane potrebne funkcionalnosti cijelog sustava.

Kako bi se spriječio najveći broj potrošačkih strahova, osim dobrog informiranja potrošača, najveći izazov u uvođenju naprednog mjerenja biti će osmišljavanje sveobuhvatnog modela upravljanja prikupljanim podacima. Model upravljanja podacima mora omogućiti:

- razmjenu podataka između svih sudionika,
- pohranu ogromnih količina podataka,
- sigurnost,
- iskoristivost podataka za različite analize,
- osiguranje koristi za sve sudionike i promicanje čvrste uloge krajnjih korisnika mreže,
- podržavanje ostvarenje europskih 20/20/20 ciljeva.

V. UVOĐENJE NAPREDNOG MJERENJA U REPUBLICI HRVATSKOJ

Kao što je opisano u uvodnom poglavlju (I.) zahtjevi europskih direktiva prema izgradnji naprednih mjernih sustava ugrađeni su u hrvatsko zakonodavstvo kroz Zakon o energiji i Zakon o tržištu električne energije. Prijedlogom novih Općih uvjeta za opskrbu električnom energijom iz prosinca 2014. godine, Hrvatski regulator tržišta energije (HERA) opisao je i model uvođenja elektroničkih brojila s mogućnošću daljinskog očitavanja, bez zahtijevanja ostalih naprednih funkcija. Ovaj model primijenio bi se ukoliko se ne bi donijela odluka o uvođenju naprednog mjerenja.

Prema trenutnom prijedlogu teksta novih Općih uvjeta rok za opremanje svih 2,3 milijuna obračunskih mjernih mjesta u Republici Hrvatskoj [3] bio bi određen rok od 15 godina, što znači da bi godišnje trebalo ugraditi nešto više od 150.000 brojila. Uz prosječni trošak za ugradnju naprednih brojila od 223 eura, godišnje bi bilo potrebno investirati oko 34 milijun eura ili 263 milijuna kuna.

Na razini pojedinačnog kupca kategorije kućanstvo, dobit od uvođenja naprednog mjerenja ovisi o korištenom tarifnom modelu i količini potrošnje energije. Uz očekivano smanjenje cijene energije od 5,7% i moguće uštede energije od 3%, za kupce s jednom tarifom (plavi model) i prosječnom godišnjom potrošnjom od 3.000 kWh, može se očekivati godišnja ušteda od 66,20 kn.

Daljinsko očitavanje svih brojila omogućilo bi izdavanje mjesečnih računa za potrošenu električnu energiju na osnovu stvarnih očitanja.

VI. ZAKLJUČAK

Direktive europske unije 2009/72/CE i 2009/73/CE od 13. srpnja 2009. godine u području naprednog mjerenja zahtijevaju da država članica osigura implementaciju naprednih mjernih sustava koji će pridonijeti aktivnom sudjelovanju kupaca na tržištu električne energije. Zahtjevi europskih direktiva prema izgradnji naprednih mjernih sustava ugrađeni su u hrvatsko zakonodavstvo kroz Zakon o energiji , NN 120/2012 i Zakon o tržištu električne energije, NN 22/2013.

Oko dvije trećine zemalja članica Europske unije odlučilo se za uvođenje naprednog mjerenja do 2020. godine. Ovim projektom ugraditi će se oko 240 milijuna naprednih brojila, a ukupno ulaganje procjenjuje se na oko 45 milijardi eura. Kod pripreme masovne ugradnje brojila vrlo je bitno razmisliti o funkcijama takvog sustava, a pomoć u tome je zasigurno Preporuka komisije 2012/248/EU.

Napredna brojila svojim funkcijama trebala bi doprinijeti razvoju novih tarifnih modela i omogućiti izravnije sudjelovanje krajnjih korisnika na tržištu električne energije.

Za operatora distribucijskog sustava daljinsko očitavanja mjernih i kontrolnih podataka, omogućiti će brži postupak obračuna potrošnje energije, izračun

gubitaka energije po dijelovima mreže, priključenje većeg broja obnovljivih izvora, brže otkrivanje i otklanjanje kvarova te detaljnije informacije za planiranje rada i razvoja mreže.

Fazi određivanje funkcija naprednog mjernog sustava potrebno je dati posebnu pozornost te temeljiti odluke na do sada provedenim analizama u zemljama članicama Europske unije, kako bi se uz što manja ulaganja, osigurala što veća korist za krajnje kupce, ali i za operatora sustava i ostale subjekte na tržištu električne energije.

LITERATURA

[1] European Commission, *Cost-benefit analyses & state of play of smart metering deployment in the EU-27*, Brussels, 2014.

[2] Manuel Sanchez, Overview of the benchmarking report on Smart metering roll-out, Brussels, 2013.

[3] HEP–Operator distribucijskog sustava d.o.o., Godišnje izvješće za 2013. godinu

Određivanje mjerne nesigurnosti 3D laserskog skenera

Mr Nikolina Mijić,dipl.ing.geod.*,
* EKO EURO TIM d.o.o / Tehničko odjeljenje, Banja Luka, Bosna i Hercegovina
nikolinamijic7@gmail.com

Sažetak - 3D laserski skeneri nude korisnicima nove mogućnosti za konformniji i efektivniji rad, ali u stručnoj javnosti još uvijek ne postoji konsenzus u vezi ustaljene procedure za testiranje tačnosti i kalibraciju 3D laserskih skenera. Proces određivanja mjerne nesigurnosti laserskog skenera realizovan je u Metrološkoj laboratoriji Građevinskog fakulteta u Beogradu. Test polje za metrološku kontrolu formirano je pomoću 46 crno - bijelih signalnih markica. Obrada rezultata mjerenja izvršena je u softverskom paketu MATLAB. Mjerna nesigurnost je određena na osnovu dužine, horizontalnog i vertikalnog ugla lasera.

I. UVOD

Ubrzan tehnološki razvoj u sferi elektroničkog mjerenja dužina rezultovao je upotrebom impulsne metode koja omogućava bezreflektorno mjerenje dužina. Dalji razvoj u tom smjeru rezultovao je pojavom sasvim novih instrumenata koji se zovu laserski skeneri. Od pojave laserskih skenera prošlo je skoro već desetak godina, međutim zbog svojih mjernih karakteristika nezadrživo tehnološki napreduju, cijena im pada i postaju ekonomski sve prihvatljiviji. Primjena terestričkog laserskog skenera može biti zaista široka, počevši od mjerenja građevinskih objekata, mjerenje deformacija na branama, topografski premjer, industrijski premjer, arheološka mjerenja, snimanje fasada, mjerenje kulturnih objekata pa sve do savremenog 3D pogonskog katastra [2]. Za određivanje tačnosti 3D laserskih sistema u laboratorijskim uslovima vrši se testiranje rezultata mjerenja i kalibracija pomoću test polja, i to se vrši upoređenjem sa etalonskim vrijednostima.

U cilju metrološkog obezbjeđenja laserskog skenera potrebno je kritički procjeniti performanse instrumenta, korištenjem nezavisnih procedura, kao što je upoređenje rezultata skeniranja sa test poljem kontrolnih tačaka [3].

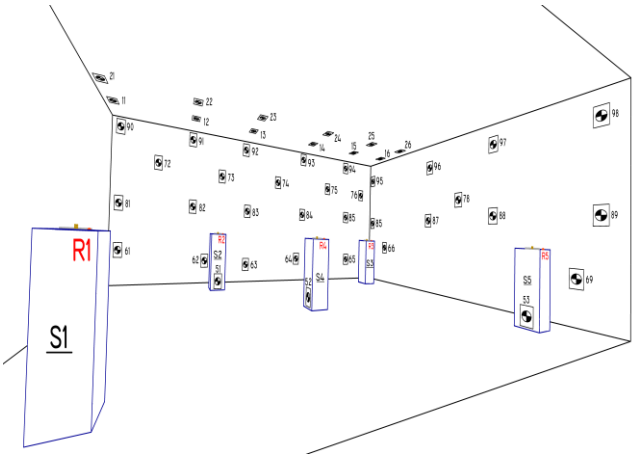
U radu će biti opisan postupak određivanja mjerne nesigurnosti 3D laserskog skenera. Terestrički laserski skener sve se više primjenjuje u praksi, ali ne postoji međunarodni konsenzus oko njihovog etaloniranja i određivanja mjerne nesigurnosti. Predložene su različite metode za postupak etaloniranja terestričkog laserskog skenera, ali se jedna izdvaja kao najvjerovatnija osnova budućeg standarda, a to je korištenje većeg broja kontrolnih tačaka [9].

Eksperimentalna mjerenja su izvršena u Metrološkoj laboratoriji, Građevinskog fakulteta u Beogradu. Pri terestričkom laserskom skeniranju razvijen je jedinstven pristup za etaloniranje terestričkih laserskih skenera, koji je primjenjen na instrument serije Leica Scan Station P20 [3].

II. ETALONIRANJE 3D LASERSKOG SKENERA

Eksperimentalni dio ovog rada odnosi se na etaloniranje 3D laserskog skenera u laboratorijskim uslovima. Model skenera koji je ispitivan je Leica ScanStation P20 [6]. Test polje za metrološku kontrolu laserskog skenera realizovano je u Metrološkoj laboratoriji. Test polje je formirano pomoću 46 crno-bijelih signalnih markica (Slika 1). Izgled test polja odnosno poligona za određivanje pravouglih koordinata markica poligona u koordinatnom sistemu laboratorije. Markice su izrađene na papiru dobrog kvaliteta, sa takvim osobinama da omogućavaju refleksiju prilikom mjerenja totalnom stanicom i da ih prepoznaje skener prilikom mjerenja [3]. Markice su dobro raspoređene u laboratoriji i nalaze se na zidovima i plafonu laboratorije, što je značajno za ocjenu preciznosti mjerenja horizontalnog i vertikalnog ugla. U laboratoriji se takođe nalazi 5 stubova sa opremom za prisilno centrisanje geodetskih mjerila. Provera mjerila u cjelini kao i kontrola rada pojedinih njegovih djelova vrši se neposredno i obuhvata spoljašni pregled i proceduru mjerenja.

Spoljašnji pregled, kojim se mjerilo postavlja na stub uz obezbjeđenu potpunu stabilnost postolja mjerila obuhvata grubu provjeru.



Slika 1. Poligon za određivanje pravouglih koordinata markica.

Mjerilo se izlaže laboratorijskim uslovima u vremenu od najmanje 30 minuta, postavlja se na stub uz obezbjeđenu potpunu stabilnost postolja mjerila. Gruba provjera mjerila obuhvata mehaničku provjeru instrumenta, i ogleda se u:

- Rad položajnih zavrtnjeva postolja instrumenta.
- Ispravnost ekrana (Display) i tastature (Keyboard) mjerila.
- Ispravnost kontakata izvora napajanja i konekcije (USB, Bluetooth, WLAN).
- Podešavanja položaja rotirajućeg ogledala (vidnog polja).

Određivanje mjerne nesigurnosti mjerila, nije moguće izvršiti u slučaju da se prilikom grube provjere samog instrumenta utvrdi neispravnost kontakata izvora napajanja [9].

III. ODREĐIVANJE PRAVOUGLIH KOORDINATA OZNAČENIH MARKICAMA

Svaka tačka u skeniranom oblaku tačaka je definisana sfernim (r, φ, θ) ili pravouglim koordinatama (x, y, z) u koordinatnom sistemu skenera i ima vrijednost inteziteta lasera intensity. Veza između rezultata merenja (r, φ, θ) i koordinata (x, y, z) je oblika:

$$\begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix} = \begin{bmatrix} r_j \cos\theta_j \sin\varphi_j \\ r_j \sin\theta_j \cos\varphi_j \\ r_j \sin\varphi_j \end{bmatrix} \quad (1)$$

gdje su r_j, φ_j, θ_j mjerene veličine - dužina, horizontalni pravac i vertikalni ugao za j-tu tačku u oblaku tačaka. Pravougle koordinate tačke u koordinatnom sistemu skenera postavljenog na stubu br. 1 su (x_j, y_j, z_j), označene indeksom i=1 u vektoru koordinata X_i.

A. *Određivanje spoljnih parametara orijentacije skenera*

Georeferenciranje je transformacija oblaka tačaka iz koordinatnog sistema skenera u spoljni (geodetski) koordinatni sistem, u koordinatni sistem laboratorije.

U georeferenciranju je zajedničko to što markice moraju imati poznate koordinate u oba sistema. Veza između dva sistema se opisuje, pomoću 6 transformacionih parametara. Faktor razmere je nebitan u transformaciji [2]. Ovi parametri se nazivaju spoljni orijentacioni parametri (EOPs). U cilju jedinstvenog određivanja 6 EOPs parametara potrebno je znati makar 6 koordinata u oba sistema, tako da tačke nikako ne bi smjele da budu na istoj liniji [5]. U poligonu se koriste 3 ili više markica sa poznatim 3D koordinatama. Ove tačke se zovu kontrolne tačke. Njihove koordinate su određene totalnom stanicom [7]. Geometrija vizurnih markica je takva da su one raspoređene na različitim rastojanjima i ne leže na istoj liniji [3].

Vektor translacije je oblika:

$$\Delta X_{ie} = [\Delta X \ \Delta Y \ \Delta Z]^T \quad (2)$$

R_{ie} je matrica rotacije između koordinatnih sistema, od skenera do koordinatnog sistema laboratorije. Rotaciona matrica predstavlja rotaciju između različitih koordinatnih sistema. Matrica je funkcionalna veza uglova rotacije ω, ϕ, κ, oko x, y i z koordinatne ose. Matrica rotacije se računa kao:

$$\mathbf{R}_1(\omega) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\omega & \sin\omega \\ 0 & -\sin\omega & \cos\omega \end{bmatrix}, \quad (3)$$

$$\mathbf{R}_2(\phi) = \begin{bmatrix} \cos\phi & 0 & -\sin\phi \\ 0 & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{bmatrix}, \quad (4)$$

$$\mathbf{R}_3(\kappa) = \begin{bmatrix} \cos\kappa & \sin\kappa & 0 \\ -\sin\kappa & \cos\kappa & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

gdje su $R_1(\omega)$, $R_2(\phi)$ i $R_3(\kappa)$ matrice rotacije oko x, y i z koordinatnih osa.

U georeferenciranju transformacija između skenera i spoljnog koordinatnog sistema se opisuje kao:

$$X_e = X_{scanner} + R_{ie}(\kappa) X_i \quad (6)$$

gdje su: $X_{scanner}$ koordinate centra koordinatnog sistema skenera u spoljnom koordinatnom sistemu, $R_{ie}(\kappa)$ je matrica rotacije između dva koordinatna sistema. Matrica rotacije je funkcija samo ugla za orijentaciju (azimut κ). Praktično su uglovi ω i ϕ jednaki nuli. Matrica rotacije se računa po izrazu $R_3(\kappa)$.

Parametri transformacije koordinatnog sistema skenera u koordinatni sistem laboratorije se određuju rješavanjem jednačina (6), metodom najmanjih kvadrata [9]. Parametri transformacije se računaju na osnovu tačaka na zidovima laboratorije (ukupno 46 tačaka).

B. *Određivanje pravouglih koordinata markica poligona u koordinatnom sistemu laboratorije*

Matematički model je razvijen na principima self-calibration. Self-calibration se izvodi u laboratorijskim uslovima unutar stabilizovanog poligona. Poligon čine označene tačke na zidovima, plafonu i stubovima laboratorije. U matematičkom modelu uslovnih jednačina učestvuje vektor mjerenih veličina i vektor nepoznatih parametara. Transformacija koordinatnog sistema skenera u spoljni koordinatni sistem može biti implicitno izražena na sledeći način:

$$f = (L, X) = 0 \quad (7)$$

gdje su: L - vektor izravnatih mjerenih veličina

X - vektor ocjenjenih nepoznaih parametara.

$$\mathbf{f}_j^{(i)}\left(\hat{\mathbf{L}}_j^{(i)},\hat{\mathbf{X}}_j^{(i)}\right)=\Delta\mathbf{X}^{(i)}+\mathbf{R}\left(\omega^{(i)},\phi^{(i)},\kappa^{(i)}\right)\mathbf{x}_{scan\ j}^{\ (i)}-\mathbf{X}_j^e=0,\quad i=1,2,...,p;\ j=1,2,...,m, \tag{8}$$

gdje je: p ukupan broj stanica skeniranja (p = 1), m je ukupan broj tačaka unutar poligona za kalibraciju (ukupno 40 tačaka), $\Delta X^{(i)}$ - vektor translacionih parametara za i-tu stanicu skeniranja:

$$\mathbf{\Delta X}^{(i)}=\begin{bmatrix}\Delta X^{(i)} & \Delta Y^{(i)} & \Delta Z^{(i)}\end{bmatrix}^{\mathrm{T}},\quad i=1,2,...,p. \tag{9}$$

$R(\omega^{(i)}, \phi^{(i)}, \kappa^{(i)})$ - matrica rotacije između dva koordinatna sistema za i-tu stanicu skeniranja:

$$\mathbf{R}\left(\omega^{(i)},\phi^{(i)},\kappa^{(i)},\right)=\mathbf{R}^{(i)}=\begin{bmatrix}r_{11}^{(i)} & r_{12}^{(i)} & r_{13}^{(i)}\\ r_{21}^{(i)} & r_{22}^{(i)} & r_{23}^{(i)}\\ r_{31}^{(i)} & r_{32}^{(i)} & r_{33}^{(i)}\end{bmatrix}=\mathbf{R}_3\left(\kappa^{(i)}\right)\mathbf{R}_2\left(\phi^{(i)}\right)\mathbf{R}_1\left(\omega^{(i)}\right)=$$

$$=\begin{bmatrix}\cos\kappa^{(i)} & \sin\kappa^{(i)} & 0\\ -\sin\kappa^{(i)} & \cos\kappa^{(i)} & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos\phi^{(i)} & 0 & -\sin\phi^{(i)}\\ 0 & 1 & 0\\ \sin\phi^{(i)} & 0 & \cos\phi^{(i)}\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\omega^{(i)} & \sin\omega^{(i)}\\ 0 & -\sin\omega^{(i)} & \cos\omega^{(i)}\end{bmatrix},\ i=1,2,...,p \tag{10}$$

$X_{scan\ j}^{\ (i)}$ - je vektor koordinata za j-tu skeniranu tačku na poligonu sa i-te stanice skeniranja u koordinatnom sistemu skenera (CS) korigovan za uticaj parametara kalibracije skenera (koji se još nazivaju i dodatni parametri):

$$\mathbf{x}_{scan\ j}^{\ (i)}=\begin{bmatrix}x_j^{(i)}\\ y_j^{(i)}\\ z_j^{(i)}\end{bmatrix}=\begin{bmatrix}\left(r_j^{(i)}-\Delta r\right)\cos\left(\varphi_j^{(i)}-\Delta\varphi\right)\cos\left(\theta_j^{(i)}-\Delta\theta\right)\\ \left(r_j^{(i)}-\Delta r\right)\sin\left(\varphi_j^{(i)}-\Delta\varphi\right)\cos\left(\theta_j^{(i)}-\Delta\theta\right)\\ \left(r_j^{(i)}-\Delta r\right)\sin\left(\theta_j^{(i)}-\Delta\theta\right)\end{bmatrix},\quad i=1,2,...,p;\ j=1,2,...,m \tag{11}$$

gdje je: p = 1 i m = 40, a r, φ i θ označavaju dužinu, horizontalni pravac i vertikalni ugao, koji su sračunati iz približnih koordinata skeniranih tačaka u skeneru CS (x, y, z) kao:

$$r_j^{(i)}=\sqrt{\left(x_j^{(i)}\right)^2+\left(y_j^{(i)}\right)^2+\left(z_j^{(i)}\right)^2}$$

$$\varphi_j^{(i)}=\arctan\left(\frac{y_j^{(i)}}{x_j^{(i)}}\right),$$

$$\theta_j^{(i)}=\arctan\left(\frac{z_j^{(i)}}{\sqrt{\left(x_j^{(i)}\right)^2+\left(y_j^{(i)}\right)^2}}\right) \tag{12}$$

IV. POSTUPAK OPAŽANJA I ODREĐIVANJE MJERNE NESIGURNOSTI 3D LASERSKOG SKENERA

Mjerilo, u ovom slučaju laserski skener serije Leica ScanStation P20, se postavi i fiksira na stub broj 1, djelujući na položajne zavrtnje približno se horizontira dovođenjem mjehura centrične libele da vrhuni. Instrument izvrši samohorizontiranje odmah nakon uključivanja. Postupak se ponovi i prilikom postavljanja instrumena na stub broj 4, jer su i sa ovog stuba vršena opažanja [9].

Opažanje je izvršeno u zatvorenom poligonu, gdje su markice označene kao mete. Poligon za testiranje terestričkog laserskog skenera čine signalne markice na zidovima, plafonu i stubovima laboratorije. Skeniranje je obavljeno sa jednim skenom sa stuba broj 1. Sve signalne markice su vidljive sa stuba broj 1, ali instrument serije Leica ScanStation P20 nije mogao da skenira signalne markice koje se nalaze na plafonu [3]. Idući od markice broj 61 u smjeru kretanja kazaljke na satu izvršeno je skeniranje svih ostalih markica. Postupak je ponovljen i za mjerenja sa stuba broj 4 [9].

A. *Određivanje mjerne nesigurnosti mjerenja dužina*

Postupak određivanja mjerne nesigurnosti mjerenja dužina je vršeno identično za oba stuba. Mjerna nesigurnost mjerenja dužina sračunata je za 32 signalne markice, a potom su vršene kombinacije mjerenja tako što su izbacivana pojedina mjerenja u cilju upoređenja rezultata mjerenja i određivanja broja tačaka sa kojima treba da se vrši etaloniranje. Mjerna nesigurnost mjerenja dužina je sračunata prvo za 32 signalne markice a potom je smanjen broj mjerenja na 20, 15, 10, 5 i 4. Tako je urađeno za oba stuba sa kojih je vršeno opažanje [8].

B. *Određivanje mjerne nesigurnosti mjerenja horizontalnih pravaca*

Postupak određivanja mjerne nesigurnosti mjerenja horizontalnih pravaca je vršen identično za oba stuba. Mjerna nesigurnost mjerenja horizontalnih pravaca sračunata je za 32 signalne markice, a potom su vršene kombinacije mjerenja tako što su izbacivana pojedina mjerenja u cilju upoređenja rezultata mjerenja i određivanja broja tačaka sa kojima treba da se vrši etaloniranje. Mjerna nesigurnost mjerenja horizontalnih pravaca sračunata je prvo za 32 signalne markice a potom je smanjen broj mjerenja na 20, 15, 10, 5 i 4. Tako je urađeno za oba stuba sa kojih je vršeno opažanje [9].

C. *Određivanje mjerne nesigurnosti mjerenja vertikalnih uglova*

Mjerna nesigurnost mjerenja vertikalnih uglova je sračunata na osnovu izraza (12). Postupak određivanja mjerne nesigurnosti mjerenja vertikalnih uglova je vršen identično za oba stuba. Mjerna nesigurnost mjerenja vertikalnih uglova sračunata je za 32 signalne markice, a potom su vršene kombinacije mjerenja tako što su izbacivana pojedina mjerenja u cilju upoređenja rezultata mjerenja i određivanja broja tačaka sa kojima treba da se vrši etaloniranje [8]. Nesigurnost mjerenja vertikalnih uglova sračunata je prvo za 32 signalne markice a potom je smanjen broj mjerenja na 20, 15, 10, 5 i 4. Tako je urađeno za oba stuba sa kojih je vršeno opažanje [9].

V. EKSPERIMENTALNA ISTRAŽIVANJA

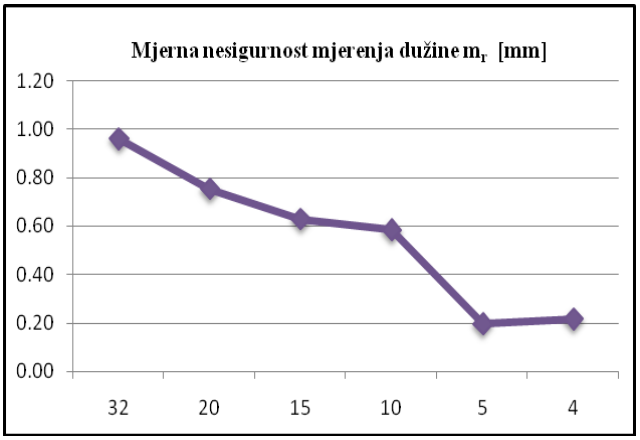
Unos i obrada podataka mjerenja izvršena je u softverskom paketu Microsoft Office Excel a provjera rezultata mjerenja je vršena u softverskom paketu MATLAB, gdje je posebno kreiran zapis za

transformaciju koordinata i određivanje vrijednosti mjerne nesigurnosti. Određene su vrijednosti mjerne nesigurnosti za mjerenje dužina, horizontalnih pravaca i nesigurnosti mjerenja vertikalnih uglova. Mjerenja su izvršena sa stubova broj 1 i 4. Vršene su kombinacije mjerenja sa ovih stubova u cilju određivanja broja tačaka sa kojima treba da se vrši etaloniranje [9].

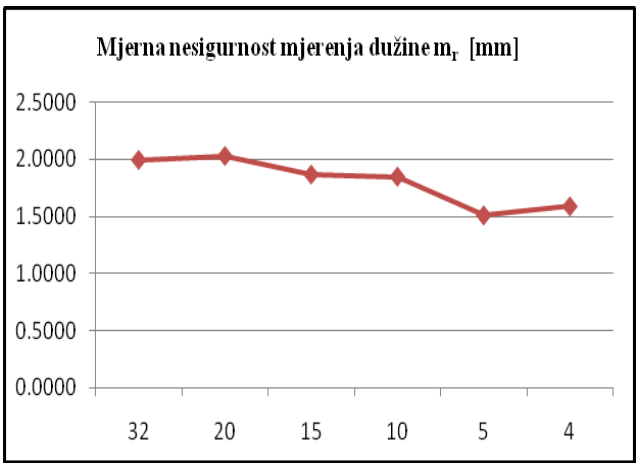
TABLICA I. VRIJEDNOSTI MJERNE NESIGURNOSTI MJERENJA DUŽINA SA STUBA 1 I 4.

Mjerenje sa stuba broj 1	
Br.markica	m_r [mm]
32	0.96
20	0.76
15	0.63
10	0.59
5	0.20
4	0.22

Mjerenje sa stuba broj 4	
Br.markica	m_r [mm]
32	2.00
20	2.03
15	1.87
10	1.85
5	1.51
4	1.59



Slika 2. Zavisnost vrijednosti mjerne nesigurnosti sa promjenom broja mjerenja na stubu broj 1.



Slika 3. Zavisnost vrijednosti mjerne nesigurnosti sa promjenom broja mjerenja na stubu broj 4.

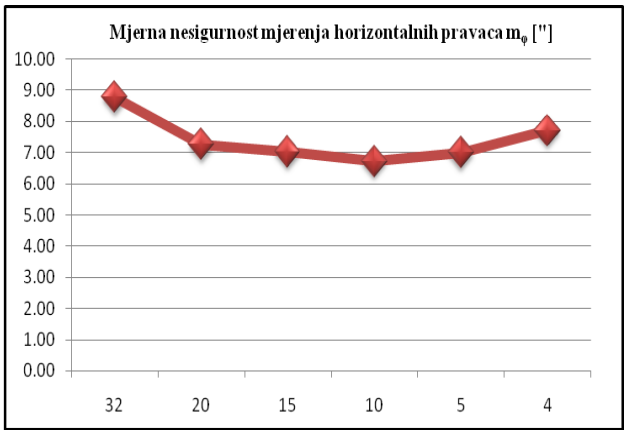
Deklarisana tačnost mjerne nesigurnosti mjerenja dužina je 2 mm [6]. Sve vrijednosti koje su sračunate su u okviru ove tačnosti [9].

Tačnost određivanja mjerne nesigurnosti mjerenja horizontalnih pravaca je 8" [6]. Sve vrijednosti koje su sračunate sa stuba 1 i 4 su u okviru deklarisane tačnosti. Vrijednosti mjerne nesigurnosti horizontalnih pravaca su prikazani u Tablici II.

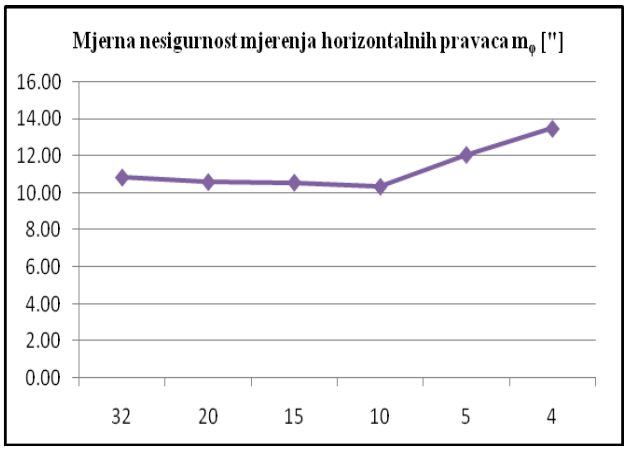
TABLICA II. VRIJEDNOSTI MJERNE NESIGURNOSTI MJERENJA HORIZONTALNIH PRAVACA SA STUBA 1 I 4.

Mjerenje sa stuba broj 1	
Br.markica	m_φ ["]
32	8.78
20	7.26
15	7.02
10	6.72
5	6.99
4	7.70

Mjerenje sa stuba broj 4	
Br.markica	m_φ ["]
32	10.84
20	10.59
15	10.56
10	10.34
5	12.07
4	13.48



Slika 4. Zavisnost vrijednosti mjerne nesigurnosti sa promjenom broja mjerenja na stubu broj 1.



Slika 5. Zavisnost vrijednosti mjerne nesigurnosti sa promjenom broja mjerenja na stubu broj 4.

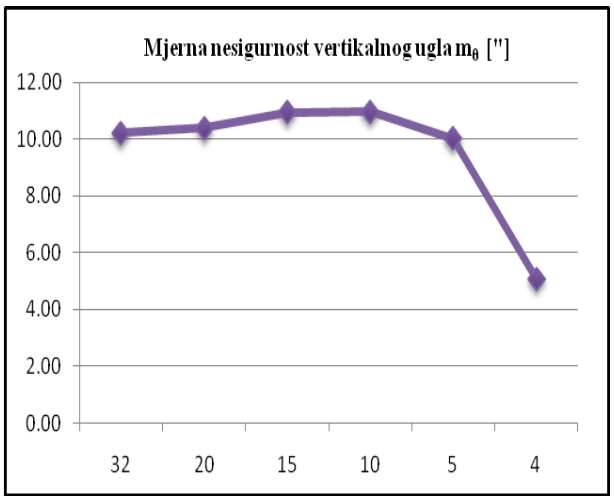
Deklarisana tačnost mjerenja vertikalnih uglova je 12". Tačnost mjerenja vertikalnih uglova prema uputstvu

za etaloniranje je 12" [6]. Sva mjerenja koja su dobijena eksperimentalnim putem su u okviru deklarisane tačnosti, prema uputstvu za etaloniranje [9].

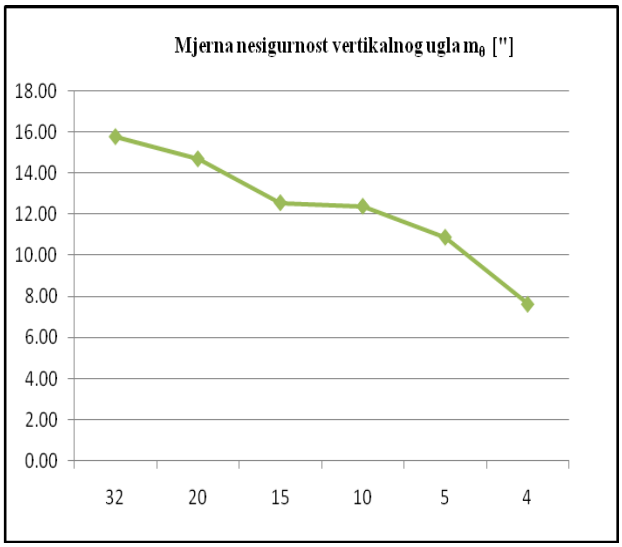
TABLICA III. VRIJEDNOSTI MJERNE NESIGURNOSTI MJERENJA VERTIKALNIH UGLOVA SA STUBA 1 I 4.

Mjerenje sa stuba broj 1	
Br.markica	m_0 ["]
32	10.20
20	10.38
15	10.94
10	10.96
5	10.01
4	5.08

Mjerenje sa stuba broj 4	
Br.markica	m_0 ["]
32	15.79
20	14.71
15	12.58
10	12.40
5	10.90
4	7.64



Slika 6. Zavisnost vrijednosti mjerne nesigurnosti sa promjenom broja mjerenja na stubu broj 1.



Slika 7. Zavisnost vrijednosti mjerne nesigurnosti sa promjenom broja mjerenja na stubu broj 4.

Prilikom etaloniranja terestričkog laserskog skenera, snimljeno je 32 signalne markice, na osnovu kojih su sračunate vrijednosti za mjernu nesigurnost mjerenja dužine, horizontalnih pravaca i vertikalnog ugla. Ove vrijednosti su upoređivane sa manjim brojem mjerenja. Vršene su kombinacije sa vrijednostima koje su dobijene sa 20, 15, 10, 5 i 4 signalne markice. Vrijednosti koje su sračunate su u okviru deklarisanih tačnosti mjerenja [6]. Kombinacije mjerenja su vršene u cilju da se odredi broj tačaka, sa kojima bi bilo najbolje vršiti etaloniranje terestričkog laserskog skenera. Prilikom iscrtavanja dijagrama koji pokazuje zavisnost vrijednosti mjerne nesigurnosti mjerenja, vrijednosti koje su dobijene se lome na 10 tačaka, rastu ili opadaju. U slučaju računanja mjerne nesigurnosti mjerenja dužine (Slika 2 i 3) i vertikalnog ugla sa stuba 1 i 4 (Slika 6 i 7), vrijednosti opadaju, dok vrijednosti mjerne nesigurnosti mjerenja horizontalnih pravaca sa stuba 1 i 4 (Slika 4 i 5), rastu sa smanjivanjem broja mjerenja. Na osnovu obrade rezultata mjerenja i kombinacije pojedinih mjerenja, prijedlog je da se etaloniranje terestričkog laserskog skenera vrši sa 10 tačaka.

VI. ZAKLJUČAK

Na tržištu se pojavljuje sve veći broj proizvođača terestričkih laserskih skenera, mnogi od njih na različite načine izražavaju karakteristike svojih instrumenata, koristeći različite kriterijume kao što su: tačnost, preciznost i rezolucija. Pored određivanja ukupne tačnosti rezultata mjerenja laserskim skenerom drugi važan aspekt osiguranja kvaliteta 3D informacija je geometrijska kalibracija mjerila [3].

U oblasti kalibracije laserskih skenera koristi se analogija sa fotogrametrijskim metodama, gdje se kalibracijom smatra određivanje spoljnih parametara orjentacije i sistematskih grešaka svih snimaka u datom projektu. Kalibracija terestričkih laserskih skenera je zahtjevna procedura zbog velikog broja različitih komponenata iz kojih se sastoji instrument. Etaloniranje komponenata terestričkog laserskog skenera bazira se na poznatom modelu konstrukcije instrumenta i njegovih komponenata, a i takođe u dobrom poznavanju instrumentalnih grešaka [1].

Svaka instrumentalna greška istražuje se pojedinačno pomoću specifičnih eksperimentalnih mjerenja, podešavanju instrumenta i registrovanju uticajnih veličina na proces mjerenja terestričkim laserskim skenerom [2]. Za postizanje milimetarske tačnosti neophodno je testiranje terestričkog laserskog skenera i u laboratoriji i u terenskim uslovima, jer postoje velike razlike koje utiču na rezultate mjerenja [3].

Pri terestričkom laserskom skeniranju razvijen je jedinstven pristup za etaloniranje terestričkih laserskih skenera, koji je primjenjen na instrument serije Leica ScanStation P20. Proces etaloniranja je izvršen za jedan dan. Etaloniranje terestričkog laserskog skenera je proces koji osim poznavanja instrumenta i rada na njemu zahtjeva znanja i iz drugih oblasti geodezije koja su morala biti primjenjena prilikom obrade podataka mjerenja [2].

Na osnovu mjerenja, sračunate su vrijednosti za mjernu nesigurnost mjerenja dužina, horizontalnih pravaca i vertikalnog ugla. Nakon što su sračunate

vrijednosti za 32 signalne markice, postupak je ponovljen i za manji broj mjerenja, tj. sračunate su vrijednosti za 20, 15, 10, 5 i 4 signalne markice [9].

Ove vrijednosti su sračunate za mjerenja sa stuba broj 1 i 4. Cilj računanja vrijednosti mjerne nesigurnosti za manji broj mjerenja bio je da se odredi sa koliko tačaka bi se trebalo vršiti etaloniranje terestričkog laserskog skenera.

Sve vrijednosti koje su dobijene kombinacijom mjerenja su u okviru deklarisane tačnosti [6]. Prijedlog je da se etaloniranje vrši sa 10 tačaka, a to je zaključeno na osnovu dijagrama jer se na toj vrijednosti kod oba stuba i za sve vrijednosti lomi dijagram. Rezultati mjerenja su sljedivi do nacionalnog etalona jedinice dužine i nacionalnog etalona jedinice ugla Švajcarske (METAS) preko laboratorije broj 02-025 [5].

LITERATURA

[1] Yuriy Reshetyuk: Self-calibration and direct georeferencing in terrestrial laser scanning, doktorska disertacija, Royal Institute of Technology, Stockholm, 2009.

[2] Jasmina D. Nedeljković Ostojić: Analiza metode premera građevinskih objekata 3D laserskim skenerima, doktorska disertacija, Banja Luka, 2010.

[3] Marko M. Pejić: Tačnost modeliranja objekata tehnologijom terestričkog laserskog skeniranja, doktorska disertacija, Beograd, 2013.

[4] J. G. Lahoz, D. G. Aguilera, J. Finat, J. Martinez, J. Fernandez and J. San Jose, "Terestrial laser scanning metric control assessment of metric accuracy for cultural heritage modeling,", Land and Cartography Engineering Department, Univ. of Salamanca, Spain.

[5] Fabio Remondino, "From point cloud to surface the modeling and visualization problem," Institute of Geodesyv and Photogrammetry, ETH Zurich, Switzerland, 2014.

[6] Leica ScanStation P20, User Manual Version 1.1, English, Leica Geosystems, Switzerland, 2013.

[7] W. Boehler, M. Bordas Vincent, A. Marbs, "Institute for spatial information and surveying Germany Technology," FH Meinz, Holzstrasse 36, 55116 Meinz, 2013.

[8] Rudolf Staiger, "Terestrial laser scanning, technology, systems and applications," 2nd Fig Conference, Marocco, 2003.

[9] Jelena Gučević, Siniša Delčev, "Etaloniranje terestričkih laserskih skenera - radno upustvo", Metrološka laboratorija za etaloniranje merila ugla i dužine, Beograd, 2014.

[10] S. Schuhmacher, J. Böhm,"Georeferencing of terrestrial laser scanner data for applications in architectural modelling," In: Procedings of the ISPRS Working Group V/4 Workshop 3DARCH "Virtual Reconstruction and Visualization of Complex Architectures", Mestre-Venice, Italy, August 22 – 24, 2005.

[11] Gordon SJ, Lichti DD, "Terrestrial laser scanners with a narrow field of view: the effect on 3D resection solutions." Surv Rev 37(292): 448 – 468, 2004.

[12] Derek D. Lichti, "Error modelling, calibration and analysis of an AM–CW terrestrial laser scanner system," ISPRS Journal of Photogrammetry and Remote Sensing, Volume 61, Issue 5, January 2007.

[13] S. Delčev, J. Gučević, V. Ogrizović, "A-priori accuracy estimation of terrestrial laser scanners calibration," XIth International scientific conference Coordinate Measuring Technique, 2014.

[14] R. Mrkić, Geodetska metrologija, Naučna knjiga, Građevinski fakultet, Beograd 1991.

[15] N. Činklović, Analiza i prethodna ocena tačnosti metoda preciznih geodetskih merenja, Građevinski fakultet, Beograd, 1976.

[16] ISO 17123-1:2002: Optics and optical instruments - Field procedures for testing geodetic and surveying instruments - Part 1: Theory.

[17] VIM: Međunarodni rečnik osnovnih i opštih termina u metrologiji, 1996.

[18] EA-4/02, Expression of the Uncertainty of Measurement in Calibration, April 1997.

Analiza operativnih karakteristika javnog gradskog prijevoza temeljem prikupljanja podataka pokretnim osjetilima

Damir Budimir*, Davor Brčić **, Niko Jelušić*
* Fakultet prometnih znanosti/Samostalne katedre, Vukelićeva 4, Zagreb, Hrvatska
** Fakultet prometnih znanosti/Zavod za gradski promet, Vukelićeva 4, Zagreb, Hrvatska
damir.budimir@fpz.hr

Sažetak

Prikupljanje podataka za analizu operativnih karakteristika površinskog *javnog gradskog prijevoza* (JGP) danas se može izvesti na više načina uz pomoć različitih tehnologija. Cilj prikupljanja podataka s pokretnim osjetilima je jeftina i učinkovita analiza operativnih karakteristika JGP-a. Upravo u razvoju JGP-a traže se rješenja prometne problematike većih gradova.

U radu se prikazuje koncept agilnog prikupljanja podataka o prometnim tokovima JGP-a, te transformacija prikupljanih podataka dobivenih pokretnim osjetilima u definirani logički model prostornih podataka koji je ujedno i sredstvo za mjerenje svojstava prikupljanih podataka. Logički model prostornih podataka razvijen je temeljem *linearnih referentnih sustava* (LRS). LRS-ima se podaci dovode u isti prostorni okvir, a složenost obrade podataka je smanjena za dvije dimenzije. Ostvaren je koncept za računalnu analizu prikupljanih podataka i testiran na sveučilišnom projektu[1].

Analiza se temelji na predviđenim i prepoznatim karakterističnim pravilima ponašanja prikupljanih podataka. Primjena pojedinog pravila vrši se na osnovu teorije prometnih tokova i *mjera težine svojstava* (MTS) dobivenih u odnosu na postavljene LRS-e kao i prema tome donosi li pravilo neku novu informaciju/podatak koji doprinosi analizi ili potvrđuje neku pretpostavku za procjenu operativnih karakteristika.

Aplikativni doprinos ogleda se u izradi fizičkog modela podataka ostvarenog LRS-ima koji pruža podlogu za prometne analize i izračun operativnih karakteristika JGP-a.

I. UVOD

Da bi se stanje u površinskom javnom gradskom prijevozu moglo evaluirati potreban je niz pokazatelja. Posebno su bitni mjerljivi pokazatelji koji služe i za evaluaciju primijenjenih mjera u cilju unaprjeđenja javnog gradskog prijevoza. Niz evaluacijskih aktivnosti povezan je sa smanjenjem vremena potrebnog da vozilo JGP-a prijeđe udaljenost između dvije stanice, što je često povezano sa smanjenjem vremena koje vozilo provede na raskrižjima da bi se u konačnici povećale komercijalne brzine. U tom smislu rade se *logističke analize* utvrđivanja vjerojatnosti izbora prijevoznog sredstva u ovisnosti o poželjnosti njegove uporabe i očituje se pokazateljima koji se ogledaju kroz:

- trajanje putovanja nastalo korištenjem različitih prijevoznih sredstva,
- troškove putovanja za dva prijevozna sredstva s obzirom na prosječne troškove,
- broj dijelova putovanja (pješačenje, čekanje i vožnja pojedinim prijevoznim sredstvom),
- vrijeme utrošeno na vožnju i čekanje prijevoznog sredstva.

Operativne karakteristike JGP-a bitno su obilježene prometnim tokovima drugih načina odvijanja prometa (pješački, automobilski, itd.). Stoga stoji tvrdnja da je JGP kao podsustav ukupnog prometnog sustava izložen, poput mnogih dinamičkih sustava, slučajnim (*heurističkim*) prostorno-vremenskim utjecajima okoline, da uvelike ovisi o prostornim konstrukcijskim elementima dionica i pojedinih odsječaka prometnica, te da na njega utječe niz faktora koji nisu u neposrednoj vezi s tim elementima [1] [2] [3].

Pod pojmom *prostorno-vremenski pokazatelj* najčešće se podrazumijeva vrijeme provedeno na stajalištu, vrijeme izlaska/ulaska putnika, vrijeme vožnje na promatranim dionicama, vrijeme putovanja, zadržavanja na raskrižju, vrijeme prolaska kroz raskrižje te operativna brzina vozila. Navedeni pokazatelji često služe i za evaluaciju primijenjenih mjera u cilju unaprjeđenja javnog gradskog prijevoza [4]. Vrijednosti pokazatelja se koriste i kao ulazne veličine za razne ITS (*Inteligentne Transportne Sustave*) aplikacije [5] [6]. Primjerice, u [7] se tvrdi da se "unutarnja" kvaliteta usluga javnog prijevoza može ocjenjivati i pratiti njeni ciljevi izvedbe temeljem *prostorno-vremenskih pokazatelja*, ali procjena "istinske" kvalitete oslanja se na percepciju performansi sustava od strane korisnika usluge.

Prikupljanje podataka o operativnim karakteristikama JGP-a se vrši na više različitih načina, primjerice: ručnim

[1] Istraživanje, koncept obrade i rezultati istraživanja sačinjeni su u sklopu projekta „*Razvoj planova održive urbane mobilnosti*“ koji je odobren i proveden u sklopu projekta kratkoročne financijske potpore istraživanju Sveučilišta u Zagrebu, temeljem namjenskog višegodišnjeg institucijskog financiranja znanstvene djelatnosti na sveučilištima.

mjerenjem vremena, anketiranjem, podacima iz naplatnog sustava kao i drugim, više ili manje, preciznim podacima i metodama. U znanstvenoj literaturi prepoznat je veliki potencijal prikupljanja podataka *pokretnim osjetilima*. Osnovni takav sustav temelji se na više podsustava različitih tehnologija. Primjenom vozila opremljenih opremom za satelitsko pozicioniranje i komunikaciju te sustava koji bilježe putanje kretanja imamo tzv. „sonda-vozila“ ili *pokretna osjetila* (eng. *Probe vehicle ili Floating Car Data*, skraćeno - FCD) [8] [9] [10] [11].

Razvijeno je niz modela koji se temelje na vremenskim serijama tako prikupljanih podataka i predviđaju vrijednost varijable na temelju povijesnih podataka. Pri tom je problem u nepredvidivosti prometnog toka, najčešće iz razloga jer je podvrgnut prometnim opterećenjima (gužvama), koja nisu ravnomjerna u vremenu, već imaju svoju prostornu raspodjelu i vremenska kolebanja. Razlozi za to su brojni, a najčešće se mogu opisati ograničenjima nametnutim ponašanjem čovjeka, prometnom potražnjom, prometnom regulativom, fizičkim značajkama prometne infrastrukture i sl.

Kako modeliranje vremenskih pokazatelja u stvarnom vremenu i kratkoročno predviđanje zahtijeva da je proces *stacionaran* (eng. *stationarity*), što u slučaju prometnog toka nije zadovoljeno, najčešće se uvodi pojam *sezonalnosti* (eng. *seasonality*) i umjesto teorije o prometnom toku koristi se statistika. Rezultati su pokazali u određenim slučajevima veoma visoku točnost, uz nisko korištenje računalnih resursa, dok su u nekim drugim slučajevima bili neupotrebljivi. Postoji dosta modela koji se koriste za predviđanje i procjenu stanja [12] [13] [14]. Provedena je i njihova verifikacija na njemačkim autocestama [15] i urbanim sredinama gdje su pokazale realistične rezultate s makroskopske razine, *TABLICA I*. Najčešći problem koji je uočen bio je vezan *za nedovoljnu količinu podataka od pokretnih osjetila* kao i *utjecaj specifičnih događaja* na prikupljane podatke.

Za razliku od drugih metoda, pokazatelji operativnih karakteristika JGP-a računati su u odnosu na LRS-ove koji svojom strukturom (*crte i kružni lukovi*) i pridodanim atributima predstavljaju prometne tokove i mogu kao sredstvo za računalnu obradu podataka omogućiti komparaciju podataka na *virtualnoj lokaciji*. Učinak takvog pristupa primjene LRS-a veoma dobro se pokazao na križanjima na kojima dolazi do sukobljavanja većeg broja prometnih tokova. Obrada podataka ostvarena je transformacijom podatkovnih točki u LRS-ove što je u suštini *različito* od linearnog referenciranja iako dosta slično. Linearno referenciranje omogućuje nalaženje točaka i raspona uzduž linearnih sredstava dok transformacijom točki u LRS one ostaju na istom mjestu u prostoru ali je omogućeno mjerenje njihovih svojstava u odnosu na LRS.

U prostornoj bazi podataka LRS-ovi predstavljaju zasebnu tematsku cjelinu. U analizi su *sredstvo* za komparaciju i temelj logičkog modela podataka. Sa stanovišta teorije prometnih tokova pridruživanjem atributa LRS-ovima određena su *očekivana svojstva* prometnog toka na lokaciji. Procesom *transformacije prikupljanih podataka* prema LRS-u ostvareno je prepoznavanje *karakterističnih promjena* u podatcima u odnosu na očekivanje na toj lokaciji (LRS-u). Prepoznavanje *karakterističnih događaja* koji se vezuju za lokaciju ostvareno je kroz *grupiranje podatkovnih točki prema svojstvima* u odnosu na LRS, njegovu geometriju i atribute koji su karakteristični za lokaciju. Time se u analizi smanjila računalna složenost obrade i kompariranja podataka za dvije dimenzije. Stoga da bi proces *grupiranja* mogao ostvariti svoju funkciju nužno je osigurati i dobro organiziranu infrastrukturu prostornih podataka (IPP; eng. *Spatial Data Infrastructure* - SDI) kao širi pojam prostorne baze podataka.

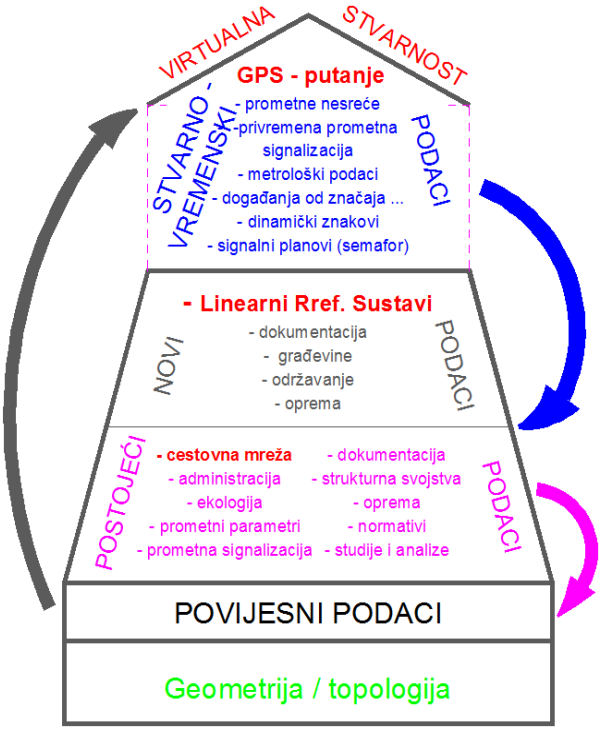
TABLICA I. PREGLED ČESTO PRIMJENJIVANIH METODA PREDVIĐANJA PROMETNOG TOKA [9]

Tip metode	Metoda	***Prostorna karakteristika***			***Vrsta prometne mreže***		***Formativno vrjednovanje rezultata od 1 do 5***	
		Točka	Ruta	Prometna mreža	Brze ceste	Urbana sredina	Jednostavnost izračuna	**Točnos**
INTUITIVNA	Povijesno uprosječivanje	**Koristi se!**	**Koristi se!**	**Koristi se!**	**Koristi se!**		**5**	1
	Grupiranje		**Koristi se!**		**Koristi se!**		**5**	2
PARAMETARSKA	Makroskopski prometni modeli			**Koristi se!**	**Koristi se!**	**Koristi se!**	2	3
	Mezoskopski prometni modeli			**Koristi se!**	**Koristi se!**	**Koristi se!**	2	3
	Mikroskopski prometni modeli			**Koristi se!**	**Koristi se!**	**Koristi se!**	2	3
	Linearna regresija	**Koristi se!**	**Koristi se!**	**Koristi se!**	**Koristi se!**	**Koristi se!**	**4**	**4**
	Kalmanov filtar		**Koristi se!**		**Koristi se!**	**Koristi se!**	3	**4**
NEPARAMETARSKA	Bayesove mreže	**Koristi se!**	**Koristi se!**			**Koristi se!**	3	**4**
	Metoda particioniranja (k-Najbližih susjeda)		**Koristi se!**			**Koristi se!**	**4**	2
	Neizrazita logika		**Koristi se!**	**Koristi se!**		**Koristi se!**	**5**	2
	Neuronske mreže		**Koristi se!**		**Koristi se!**	**Koristi se!**	2	3
	Lokalna regresija s težinskim faktorima	**Koristi se!**	**Koristi se!**			**Koristi se!**	**4**	**4**

Ovakvim pristupom višestruko se olakšalo grupiranje podataka i povećala iskoristivost prikupljanih prostorno-vremenskih podataka. Olakšalo se kompariranje s drugim infrastrukturama u smislu jednostavnije uporabe, što stvara temelj za ostale aktivnosti i usluge. Na *slici 1.* prikazan je širi opseg i načelna organizacija podataka u modelu IPP-a koja zadovoljava potrebe analize operativnih karakteristika JGP-a. Uspješna implementacija pojedinih logičkih dijelova IPP-a podrazumijeva standardizaciju procedura koje omogućuju integraciju svih podataka u *tematski sloj* (eng. *layer*) koji sadrži LRS-ove.

Istraživanja koja su provedena pokazala su značajan napredak u kvaliteti i iskoristivosti prikupljanih podataka (GPS-putanja) uvođenjem *logičkog modela podataka* i prikladnih procedura za linearno referenciranje. Dobar primjer takve prakse je integracija prostornih podataka u europske infrastrukture prostornih podataka kao što je INSPIRE (eng. *Intermodal referencing system for traffic-related data*) [16]. Istraživanja u okviru projekta OKSTRA (Njem. *Objektkatalog für das Straßen- und Verkehrswesen*) ukazuju na važnost linearnog referenciranja za opisivanje objekata u blizini prometne infrastrukture kao i u procesu spajanja stvarno-vremenskih podataka o prometnim tokovima s ciljem da se postignu što točnije percepcije uvjeta u mreži [17].

U eksperimentalnom dijelu procjena pokazatelja odnosno operativnih karakteristika javnog gradskog prijevoza ostvarena je temeljem prikupljanja podataka GPS-zapisivačem (eng. *Logger*). Za potrebe istraživanja prostorna baza podataka u GIS-u [2] nadopunjena je tematskim slojem *linearnih referentnih sustava* i isprogramirane su odgovarajuće funkcije. To je omogućilo realizaciju sustava primanja i obrade podataka od pokretnih osjetila na način koji je omogućio analizu pokazatelja JGP-a. Na *slici 2.* prikazan je segment tramvajske prometne mreže s postavljenim LRS-ovima.



Slika 1. Opseg i organizacija prometnih podataka u modelu IPP.

[2] Geografski informacijski sustav koji je razvijen pod nazivom „OptaGis“ na Fakultetu prometnih znanosti za potrebe istraživanja

II. OBRADA PODATAKA POKETNIH OSJETILA

A. Formaliziranje procesa mjerenja svojstava prikupljanih podataka

Postupak izračuna kvantitativnih pokazatelja JGP-a korištenjem podataka globalnog pozicioniranja direktno je vezan za infrastrukturu prostornih podataka. Polazi se od teze da su karakteristične vrijednosti o položaju i vremenu sa *svojstvima površina* na kojima se odvija promet dovoljne za pouzdanu prostorno-vremensku procjenu stanja prometnog toka vozila JGP-a.

Svojstva površina opisana su *atributima* u *logičkom modelu podataka* koji je ostvaren LRS-ima. Nakon definiranja LRS-a ostvarene su procedure grupiranja prema istima. Samo grupiranje promatrano je kroz izračun udaljenosti te geometrijske usporedbe vektoriziranih ulaznih podataka koje su u radu označene kao *mjere težine svojstva*, a dobijaju se mjerenjem kao primjerice na s*lici 3.* Na slici je prikazano mjerenje udaljenosti u odnosu na prometni tok na osnovu kojeg se može sa sigurnošću utvrditi da se radi o stajanju unatoč greški u pozicioniranju.

Mjerama težine svojstva (MTS) definirani su entiteti tj. prikupljani podaci ili podatkovne točke kroz niz vrijednosti. Formalni zapis podatkovnih točki je najčešće vektor u kojem svaka dimenzija, odnosno vrijednost prikazuje neko svojstvo. Obzirom na kompleksnost vektor sadrži diskretne i kontinuirane vrijednosti, a odabir svojstava je u direktnoj vezi s površinom na kojoj se odvija prometni tok.

Pored jednostavnog prikaza entiteta kroz geometrijske značajke važno je svojstvo vremena koje karakterizira svaku podatkovnu točku. U tom prikazu vrijeme je *različnica* iz cijelog skupa entiteta i čini tj. definira novo svojstvo koje je ključno za logički model podataka. Stoga, prikazom grupiranja prema svojstvima *različnica*

Slika 2. Grafički prikaz dijela tramvajske mreže s LRS-ovima (*projekt: „Razvoj planova održive urbane mobilnosti“*

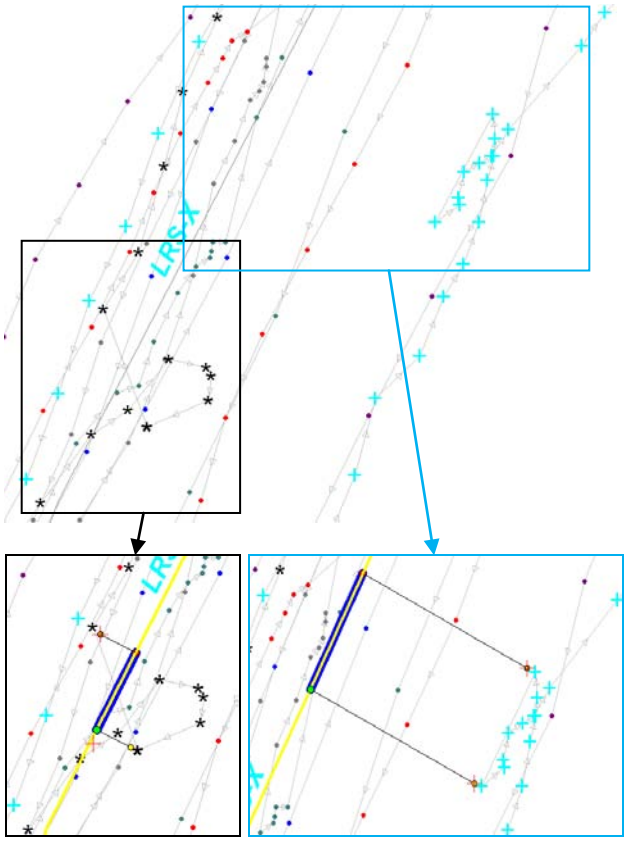
možemo dodatno upotpuniti zaključivanje na temelju prostornih značajki.

Postupak formaliziranja procesa mjerenja svojstava prikupljanih podataka na lokaciji, u daljnjem tekstu *pojavnica*, može se sažeti u dva zadatka:

- promatranje grupe točaka prikupljenih pokretnim osjetilima na mikro razini i određivanje njihovih MTS-a u odnosu na LRS,
- grupiranje na osnovu izračunatih svojstava i izračun novog svojstva koje se vezuje za lokaciju.

Analiza je pokazala da čistoća pojavljivanja neke *pojavnice* predstavlja *„događaj"* koji je važan za analizu operativnih karakteristika kroz *različnice*. Pokazalo se da je *bilježenje učestalosti pojavnice* kao dodatne mjere težine svojstva podatkovnih točaka dobar način za određenje relevantnosti *pojavnice*.

Evidentno je da su najčešće *pojavnice* upravo funkcijski vezane za događaje na prometnoj mreži ali kao izoliran događaj koji se ne može direktno i sadržajno uvijek uzeti u razmatranje. Bez uporabe funkcije cilja u puno slučajeva one nisu relevantne za pojedine pokazatelje upravo zato što im je *razlikovna vrijednost* u direktnoj sprezi s drugim utjecajnim čimbenicima, primjerice regulacijom prometa. Iz tog razloga potrebno je kao MTS koristiti one mjere koje uspijevaju prikazati koliko je neko svojstvo posebno za neki *„događaj"*, odnosno koliko ono uspijeva razlikovati taj *„događaj"* od drugih *„događaja"* ili očekivanih vrijednosti.



Slika 3. Primjer određivanja geometrijskih mjera težine

B. *Primjena linearnih referentnih sustava*

Glavni zadatak LRS-a je ostvarenje logičkog modela podataka koji omogućuje jednostavniju interpretaciju prikupljanih podataka i time olakšava mjerenja, analizu i izračun prometnih značajki. Na *slici 4.* grafički su prikazani LRS-ovi definirani na nekim od križanja koji se često pojavljuju u gradskim sredinama.

Uloga i značaj LRS-ova može se sagledati kroz učinke procedura mjerenja svojstava, a može se opisati kroz aktivnosti:

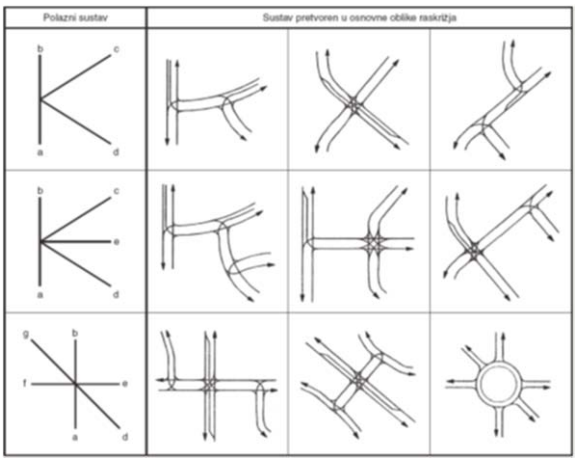
- segmentacije podataka (eng. *data segmentation*)
- pronalaženje prvog „karakterističnog" podatka u segmentu (eng. *first data detection*)
- pronalaženje veza između tih segmenata (eng. *story link detection*)
- grupiranje podataka prema svojstvima (eng. *cluster detection*)
- komparacija s referentnim vrijednostima uvažavajući geometriju prometnih tokova
- definiranje *„događaja"* u kontekstu prikupljanja podataka

Temeljne odrednice razvoja procedura za prihvat prikupljanih podataka su direktno vezane za strukturu i atribute definiranih LRS-a.

C. *Funkcije udaljenosti i sličnosti u procesu integracije podataka*

Funkcije sličnosti obzirom na prikupljane podatke mogu se podijeliti u funkcije iz područja *teorije skupova*, područja *geometrije* kao i područja *teorije informacija*. Kako se radi o prepoznavanju sličnosti među podacima koji su sadržani u *podatkovnim točkama*, koje su povezane i promatraju se kroz vrijeme, razvijen je veliki broj različitih funkcija. Zadatak tih funkcija je omogućavanje izračuna sličnosti/različitosti među prikupljanim podacima i očekivanim vrijednostima na temelju MTS-a s tim da je domena za grupiranje definirana LRS-ovima.

U procesu grupiranja podataka i zadacima



Slika 4. Tipični primjeri definiranja LRS-a na križanjima

prepoznavanja *pojavnica* (*„događaja“*) potrebno je primijeniti funkcije tj. mjerenja koje omogućuju izračun vjerojatnosti da podaci koji se obrađuju pripadaju točno određenoj grupi. *Pojavnice* ukazuju na karakteristične osobine prikupljanih podataka i u direktnoj su vezi sa *očekivanim svojstvom* prometnog toka na lokaciji. Takve mjere koje se rade na osnovu određenih kriterija zvat će se *mjerama sličnosti*. Jedna od mjera sličnosti koja je primijenjena je geometrijska mjera izračuna udaljenosti podataka od LRS-a.

Za izračun geometrijske udaljenosti u svrhu komparacije podataka korištene su *Manhattan* i *Euklidova* udaljenost. Njihovom primjenom moguće je izvršiti mjerenje geometrijskih udaljenosti između podatkovnih točaka kao i objekata u vektorskom prostoru. Ako su podatkovne točke $\boldsymbol{i}$ i $\boldsymbol{j}$ prikazane kao p-dimenzionalni vektori (1):

$$\boldsymbol{i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \quad \text{i} \quad \boldsymbol{j} = (x_{j1}, x_{j2}, \ldots, x_{jp}). \tag{1}$$

Manhattan udaljenost se računa (2)*:*

$$Udaljenost_{manhatten}(i,j) = \sum_{m=1}^{p} w_m |x_{im} - y_{im}| \quad . \tag{2}$$

Dok se *Euklidova* udaljenost računa (3):

$$Udaljenost_{euklid}(i,j) = \sqrt{\sum_{m=1}^{p} w_m (x_{im} - y_{im})^2} \quad , \tag{3}$$

gdje je $\boldsymbol{w_m}$ težina $\boldsymbol{m}$-te varijable (veća težina dodjeljuje se važnijim varijablama).

Pored izračuna udaljenosti podataka u odnosu na LRS kao mjera sličnosti korištena je *„kosinus“* mjera koja je popularizirana u pretraživanju informacija. Glavna joj je prednost prema *Manhattan* i *Euklidovoj* udaljenosti činjenica da olakšava grupiranje povezanih podataka (putanja-trajektorija) u odnosu na postavljene LRS-ove i otpornija je na stršeće podatke. Analitički je identična skalarnom produktu dvaju normaliziranih vektora. U primjeni s GPS podacima veoma je korisna jer svaka podatkovna točka ima i smjer kretanja. Opći obrazac za računanje *„kosinusne“* mjere dviju točaka je (4):

$$Slicnost_{kosinus}(i,j) = \frac{\boldsymbol{i} \times \boldsymbol{j}}{|i||j|} = \frac{\sum_{m=1}^{N} i_m \cdot j_m}{\sqrt{\sum_{m=1}^{N} i_m^{\,2}} \cdot \sqrt{\sum_{m=1}^{N} j_m^{\,2}}} \quad . \tag{4}$$

Nakon geometrijskih mjera i grupiranja podataka primjenom LRS-a često su isti podaci pridruženi u više grupa, pa je potrebno usporediti skupove grupiranih podataka kako bi se donijela konačna odluka o pripadnosti podatka. U tom smislu koristi se *Jaccard* koeficijent koji korijen ima u teoriji skupova te predstavlja mjeru sličnosti dvaju skupova. Formalna definicija te mjere može se napisati kako slijedi (5):

$$Slicnost_{jaccard}(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \quad . \tag{5}$$

U ovom izrazu presjek skupova podrazumijeva sumu vrijednosti manjih dimenzija, dok se unija smatra sumom većih vrijednosti u pojedinim grupama.

D. Postupak grupiranja prikupljanih podataka

Grupiranje prikupljanih podataka je postupak razdjeljivanja podataka u grupe (*klastere*) koje sadrže podatke prema nekoj pripadnosti, tako da su slični podaci (*slični po nekom svojstvu*) svrstani u istu grupu, a različiti podaci u različite grupe. Svrha grupiranja je nalaženje “*prirodnih*” grupa u skupu prikupljanih podataka. Grupiranje podataka kao metoda primjenjuje se u različitim stručnim i znanstvenim područjima. Pod pojmom grupiranje podataka (eng. Clustering) u metodi koja se u ovom radu istražuje podrazumijeva se proces klasifikacije podataka pokretnih osjetila (eng. *data points*) temeljem mjera težine svojstva (eng. *feature vectors*) u grupe (eng. *Clusters*) koje su definirane LRS-ovima.

Sljedeća uočena prednost ovakvog tretiranja podataka od pokretnih osjetila je u segmentu *spremanja prikupljanih podataka* čime oni postaju *povijesni podaci* koji su važni za analizu operativnih karakteristika JGP-a. Cilj je da su primjenjivi i korisni bez potrebe za ponovnim složenim računalnim zahtjevima obrade, a smanjuje se količina prikupljanih podataka. U narednim istraživanjima taj dio će se detaljnije istražiti.

E. Višekriterijska analiza prikupljanih podataka

Višekriterijska analiza ostvarena je grupiranjem svojstava na lokaciji prema težinskim vrijednostima *pojavnica* i *različnica* tj. prema udaljenosti i vremenu gdje je rezultat dobiven temeljem uvjetne vjerojatnosti. Uvjetna vjerojatnost definirana je na osnovu teoretskih postavki koje se temelje na poznavanju ponašanja prometnih tokova u realnim uvjetima (*koji vladaju na lokaciji*). Definiranje kriterija se temelji na predviđenim i prepoznatim karakterističnim pravilima ponašanja prikupljanih podataka koja asociraju na događaje iz stvarnog svijeta.

Pored važnosti pojedinog kriterija neophodno je ustanoviti i njegovu pouzdanost. Stoga je za neke kriterije dodatno definirano svojstvo *podrška pravila* (eng. *Support rules*) ($A \Rightarrow B$) = $P(A \cup B)$. Ako je podrška pravila $\boldsymbol{p\%}$, to znači da je tom vjerojatnošću ($\boldsymbol{p\%}$) ono ostvareno. Ukazuje na pouzdanost prepoznavanja događaja iz prometnog toka na temelju analize.

Na ovaj način analizom i primjenom kriterija koji su specifični za lokaciju, uvažavajući teoriju dinamike prometni tokova, ostvarena je analiza operativnih karakteristika JGP-a.

III. ZAKLJUČAK

Po svojoj naravi *infrastruktura prostornih podataka* predstavlja fenomen koji još uvijek nije u potpunosti istražen, posebice kada se radi o prometnim tokovima. Njena implementacija, poboljšanje i razvoj predstavljaju izrazito kompleksan i dinamičan proces koji se konstantno mijenja u skladu s promjenama u njenom okruženju. Teza od koje kreće ovaj rad je postojanje infrastrukture prostornih podataka u određenom obliku i na određenoj razini.

Uz pojam infrastruktura prostornih podataka u literaturi se koristi i pojam geoinformacijska infrastruktura (GII; eng. *Geoinformation infrastructure*) koja nastoji udovoljiti sve prisutnijem zahtjevu za stalnu interakciju vozila i/ili pješaka s infrastrukturom i centralnim mjestom za upravljanje u svrhu podizanja kvalitete usluge, optimizacije prometa, te podizanja razine pouzdanosti („*putovanje – vrijeme – pouzdanost*“; eng. „*Journey Time Reliability*“). Kako se u pravilu radi o analizi prometnih tokova i njima svojstvenim *događajima* istraživanje je potvrdilo nužnost primjene LRS-a. Njome je ostvaren znatan napredak u iskoristivosti podataka dobivenih pokretnim osjetilima, a ujedno je smanjena složenost analize prometnih tokova za dvije dimenzije.

Proces određivanja *svojstava* i *grupiranja* podatkovnih *točaka* korelira s pitanjem kako prepoznati „*događaj*“? Ostvaruje se prema nekom od *kvantitativnih svojstava* dobivenih na temelju LRS-a koje nazivamo *mjera težine svojstva*, a ukazuju na važnost tog svojstva za prometni tok i lokaciju kojoj pripadaju podatkovne točke. Stoga je nužno definirati i odrediti kriterije za određivanje MTS-a kao mjere relevantnosti nekog „*događaja*“ ili grupe točaka. Dvije su mogućnosti za to: izračunom (*teorija prometnih tokova*) i *analizom povijesnih podataka*, vodeći računa o zakonitostima kretanja vozila u urbanim sredinama ne *zanemarujući* opće uvjete odvijanja prometa (primjerice, *prometna regulacija*), složenost prometnog toka (primjerice, *broj kolničkih traka*) i vremensku neravnomjernost prometne potražnje.

Radnje i procesi koji su potrebni za realizaciju metode prepoznavanja *karakterističnih stanja prometnog toka* na temelju prikupljanih podataka pokretnim osjetilima su:

- definiranje logičkog modela prostornih podataka (*atributi i oblik LRS-a*),
- proces transformiranja i grupiranja podataka prema LRS-ima (metodama mjerenja: *Manhattan, Euklidova udaljenost i kosinusna metoda*),
- utvrđivanje svojstava i referentnih vrijednosti (*varijabli*) na osnovu općih uvjeta odvijanja prometa i/ili „*povijesnih*“ podataka neovisno za svaku lokaciju,
- definiranje kriterija za *prepoznavanje karakterističnih stanja* prometnog toka temeljem svojstava prikupljanih podataka.

U procesu obrade prikupljanih podataka, transformacijom prikupljanih podataka u LRS, omogućeno je mjerenje *svojstava* i grupiranje na osnovu tih *svojstava* što je omogućilo prepoznavanje *„specifičnih događaja“* na unaprijed određenim lokacijama (LRS-ovima), a rezultiralo je *pokazateljima* operativnih karakteristika JGP-a.

Njihova primjena je višestruka, pa mogu poslužiti i za evaluaciju mjera unaprjeđenja gradskog javnog prijevoza kao i za druge ITS aplikacije koje ovise o dinamičkim podacima prometnih tokova. Imajući na umu sam proces prikupljanja podataka može se konstatirati da je prikupljanje podataka pokretnim osjetilima jednostavno i jeftino, rezultati obrade prikupljanih podataka su prihvatljivi, a mogućnosti primjene velike.

LITERATURA

[1] Mrvelj, Š., Matulin, M., Jelušić, N.: Public transport priority system: Impact on quality of service, Proceedings of the 21st Central European Conference on Information and Intelligent Systems, Varaždin, 22-24 September, 2010, pp. 457-462

[2] Maerivoet, S., De Moor, B.: Traffic Flow Theory, Katholieke Universiteit Leuven, Belgium, 2008.

[3] NCHRP, FHWA, TCRP, TRB: Highway Capacity Manual, National Research Council,Washington DC, 2010.

[4] J. Esser, M. Schreckenberg, "Microscopic online simulation of urban traffic", International Journal of Modern Physics C, Vol. 8, No. 5, 1997. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?-doi=10.1.1.40.6931&rep=rep1&type=pdf

[5] Matulin, M., Mrvelj, Š., Gold, H.: Evaluating the public transport priority system - Identifying background data impact, Proceedings of the 17th ITS World Congress, Busan, Korea, 25-29 October, 2010, 10p.

[6] Jelušić, N., Anžek, M., Ivanković, B.: Information Source Quality in Intelligent Transport Systems, Promet – Traffic & Transportation, Scientific Journal on Traffic and Transportation Research, Volume 22, No.2, pp. 125-131, 2010.

[7] Tiedong, W., Tingjian F., Jianghong, H., Jian W.: Traffic Monitoring Using Floating Car Data in Hefei, Proceeding of IPTC10 International Symposium on Intelligence Information Processing and Trusted Computing, Huanggang, 2010.

[8] Messelodi, S., Modena, C.M., Zanin, M., De Natale, F.G.B., Granelli F., Betterle, E., Guarise A.: Intelligent extended floating car data collection, Expert Systems with Applications, ELSEVIER, Vol. 36, Issue 3, 2009, pp. 4213-4227

[9] Barceló, J., Kuwahara, M.: Traffic Data Collection and its Standardization, Chapter 11, Sheider, T., Kuwahara, M.: Extended Floating Car Data in Co-operative Traffic Management, Springer, July 2010, pp. 161-170

[10] Díaz, J.J.V., Llorca D.F., González, A.B.R., Mínguez , R.Q., Llamazares, Á.L., Sotelo, M. Á.: Extended Floating Car Data System: Experimental Results and Application for a Hybrid Route Level of Service, Institute of Electrical and Electronics Engineers, Transactions on Intelligent Transportation Systems, Vol. 13, No. 1, March 2012, pp. 25-35

[11] Ehmke, J.F, Meisel, S., Mattfeld, D.C.: Floating car based travel times for city logistics, Transportation Research Part C, Volume 21, Issue 1, April 2012, pp. 338-352

[12] S. Maerivoet, B. De Moor, "Traffic Flow Theory", 2005. [Online]. Available: http://arxiv.org/abs/physics/0507126

[13] C.P.IJ. van Hinsbergen, J.W.C. van Lint, F.M. Sanders, "Short Term Traffic Prediction Models", 2007.

[14] Yang Zhang, Yuncai Liu, "Comparison of Parametric and Nonparametric Techniques for Non-peak Traffic Forecasting", 2009., World Academy of Science, Engineering and Technology vol. 51. Available: http://www.akademik.unsri.ac.id/download-/journal/files/waset/v51-40.pdf

[15] Vandenberghe, W., Vanhauwaert, E., Verbrugge, S., Moerman, I., Demeester, P.: Feasibility of expanding traffic monitoring systems with floating car data technology, Intelligent Transport Systems-IET Research Journals, Vol. 6, Issue 4, United Kingdom, 2012, pp. 347–354

[16] Kochs, A., Weidner, B., Marquardt, T.: Auswirkungen der EU-Geodaten-Richtlinie INSPIRE auf das Straßen- und Verkehrswesen, Straßenverkehrstechnik, Nr. 6, 2010, pp. 349-353.

[17] Izvještaji s projekata: E-FRAME (http://www.frame-online.net/; http://www.aktiv-online.org/index.html, 13.11.2013.), OKSTRA (http://www.okstra.de, 7.02.2014.)

Optimizacija procesa naplate cestarine na Autocesti Rijeka-Zagreb

Josip Ostojić, Dr.Sc.* i Elmin Kalajdžić**
*HAC-ONC d.o.o., Zagreb, Hrvatska
josip.ostojic@hac-onc.hr
**Enel-Split d.o.o., Zagreb, Hrvatska
nino@enel.hr

Sažetak - Autocesta Rijeka-Zagreb je tijekom 2013. pokrenula nekoliko projekata iz domene unapređenja poslovanja i uvođenja novih poslovnih i organizacijskih rješenja. Obzirom na nužnu potrebu unaprjeđenja poslovnih procesa, prepoznata je poslovna potreba optimizacije poslovnih procesa naplate cestarine.

Poslovni proces naplate cestarine na Autocesti Rijeka-Zagreb, kao jedan od ključnih poslovnih procesa uključuje sve relevantne aktivnosti od trenutka evidencije ulaska vozila na autocestu pa sve do evidentiranja ove transakcije u financijskom modulu SAP ERP sustava.

Cilj projekta „Optimizacija poslovnog procesa naplate cestarine na Autocesti Rijeka-Zagreb“, je bio osigurati transparentnost poslovnih procesa uzevši u obzir cjelokupnu postojeću infrastrukturu tvrtke i sve postojeće uvjete plaćanja cestarine. Jedan od prioriteta projekta je bio i izrada novog, pouzdanog sučelja koji vrši prijenos podataka iz Sustava naplate cestarine u SAP ERP sustav.

Kao potpora unapređenju poslovnih procesa odabrana je platforma ARIS, njemačke firme Software AG. Tijekom analize poslovnih procesa ciljano su se uočavale slabe točke koje su se tretirale kao prigode za poboljšanja, te su na kraju projekta dokumentirano predstavljene zajedno sa prijedlozima na koji način poboljšati i unaprijediti poslovne procese.

I. UVOD

Temeljem odluke Vlade Republike Hrvatske, osnovano je dioničko društvo Autocesta Rijeka-Zagreb d.d. (ARZ) za građenje i gospodarenje autocestom, koje je započelo s radom u ožujku 1998. godine.

Društvu je dodijeljena koncesija na vremenski rok od 28 godina. Prihodi Društva su od naplate cestarine i naplate korištenja pratećih i uslužnih objekata na trasi autoceste.

Gospodarska svrha Društva je zatvaranje financijske konstrukcije, građenje, gospodarenje i održavanje autoceste Rijeka - Zagreb, te njenih cestovnih i pratećih objekata na cestovnom zemljištu u sklopu ostvarivanja prava iz koncesije koje se Društvu dodjeljuje na temelju odluke Vlade.

U cilju unapređenja poslovanja, tijekom 2013-te godine Društvo je iniciralo više projekata. Cilj ovih projekata je podizanje učinkovitosti i kvalitete poslovanja i poslovnih procesa, uvođenje novih tehnologija, usklađivanje sa obvezujućim propisima, normama i standardima, te uvođenje organizacijskih promjena kao kontinuirane poslovne aktivnosti.

Proces naplate cestarine kao temeljni proces za realizaciju prihoda Društva je imao određenih problema. Ovaj poslovni proces uključuje sve povezane aktivnosti od trenutka ulaska vozila na autocestu do trenutka okončanja transakcije u financijskom modulu SAP ERP (Enterprise Resource Planning) sustava.[3][4]

Rezultati projekta „Optimizacija poslovnih procesa naplate cestarine na Autocesti Rijeka-Zagreb“ morali su omogućiti učinkovitu transformaciju poslovnih procesa i zadovoljiti slijedeće zahtjeve:

- Dokumentiranje mape procesa ARZ i dokumentiranje detaljnog tijeka svih poslovnih procesa ključnih za naplatu cestarine uključujući sve resurse nužne za izvršavanje poslovnog procesa
- Dokumentiranje svih podataka nužnih za učinkovitu naplatu cestarine
- Dokumentiranje svih IT (Informatička tehnologija) sustava uključenih u proces naplate cestarine i njihove integracije
- Analiza dokumentiranih procesa, podataka i IT sustava, definiranje slabih točaka i izrada prijedloga optimizacije procesa
- Razvoj aplikacije za integraciju koja vrši prijenos podataka iz Sustava naplate cestarine u SAP ERP sustav po načelima definiranim prijedlogom optimizacije procesa.

U radu će se izložiti analiza postojećih procesa sustava naplate cestarine, softverske infrastrukture koja podržava taj proces, naglasiti slabe točke uočene tijekom analize i dati prijedlozi za njihovo otklanjanje.

II. ARIS – PLATFORMA ZA ANALIZU POSLOVNIH PROCESA

Poslovanje u suvremenim uvjetima se nalazi pod stalnim pritiskom velike konkurencije i striktne regulative. Kako bi postala i ostala konkurentna, poduzeća sve više obraćaju pozornost na svoj poslovni model i svoje poslovne procese. Uspjeh poduzeća ovisi o sposobnosti efikasnog upravljanja vlastitim poslovnim procesima.

Kako bi shvaćanje kompleksnih poslovnih procesa bilo učinkovito, razvijeni su posebni softverski alati za

modeliranje poslovnih procesa. Za razliku od običnih alata za grafičko prikazivanje, aplikacije za analizu poslovnih procesa kreiraju repozitorij modela i njihovih sastavnih elemenata kao bazu podataka u kojoj se za svaki element vežu odgovarajući atributi korisni za provođenje različitih analiza i simulacija.

Repozitorij modela poslovnih procesa tako postaje baza znanja koja, osim grafičkog prikaza poslovnih procesa, sadržava i sve međusobne veze i odnose elemenata u modelima. Time su omogućeni pogledi na procese iz drugih perspektiva: organizacijske (npr. koje poslove obavlja ovo radno mjesto ili odjel), informatičke (npr. koje poslovne procese pokriva ova aplikacija), podatkovne (npr. za koje poslovne procese je bitan ovaj dokument ili podatak), i slično…

Na tržištu postoji veliki broj aplikacija koje podržavaju analizu i izradu modela poslovnih procesa. Gartner u svojim izvješćima iz kategorije „BPA tools“ navodi ključne karakteristike koje treba imati kvalitetan alat za analizu poslovnih procesa i po tim kriterijima donosi preporuke za odabir0.

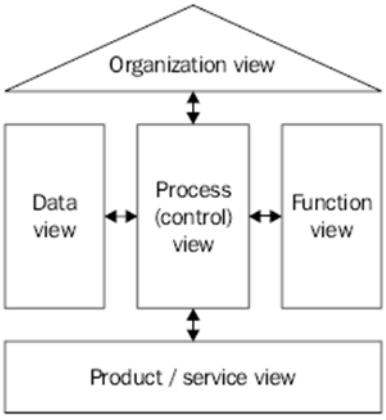
Lideri na polju alata za analizu poslovnih procesa (Slika 2) su IBM (WebSphere Business Modeler i BlueWorks), Metastorm (ProVision BPA), Mega (Mega Modeling Suite) i Software AG (ARIS).

Prije nekoliko godina (2007) Autocesta Rijeka-Zagreb je odabrala ARIS za dizajn i analizu poslovnih procesa, pa je ova platforma korištena i tijekom ovog projekta.

Na odabir ARIS-a kao standarda u ARZu utjecala je činjenica da ARIS u sebi sadrži metodološki okvir za modeliranje poduzeća (enterprise modelling). Taj okvir podrazumijeva skup metoda za analizu poslovnih procesa, podataka, IT sustava i drugih elemenata ključnih za razumijevanje i upravljanje poduzećem.

ARIS je nastao na osnovama akademskog istraživanja Prof. Dr. August-Wilhelma Scheera tijekom devedesetih. Ima vrlo raširenu industrijsku primjenu, osobito u Europi. Praktički ne postoji niti jedna velika europska firma koja ga ne koristi u svom poslovanju.[2]

Osnova ARIS metodologije je „ARIS kuća“ (Slika 1)



Slika 1: ARIS kuća

Source: Gartner (December 2011)

Slika 2: ARIS u Gartnerovom izvješću za BPA Tools0

koja predstavlja različite perspektive/poglede na poduzeće:

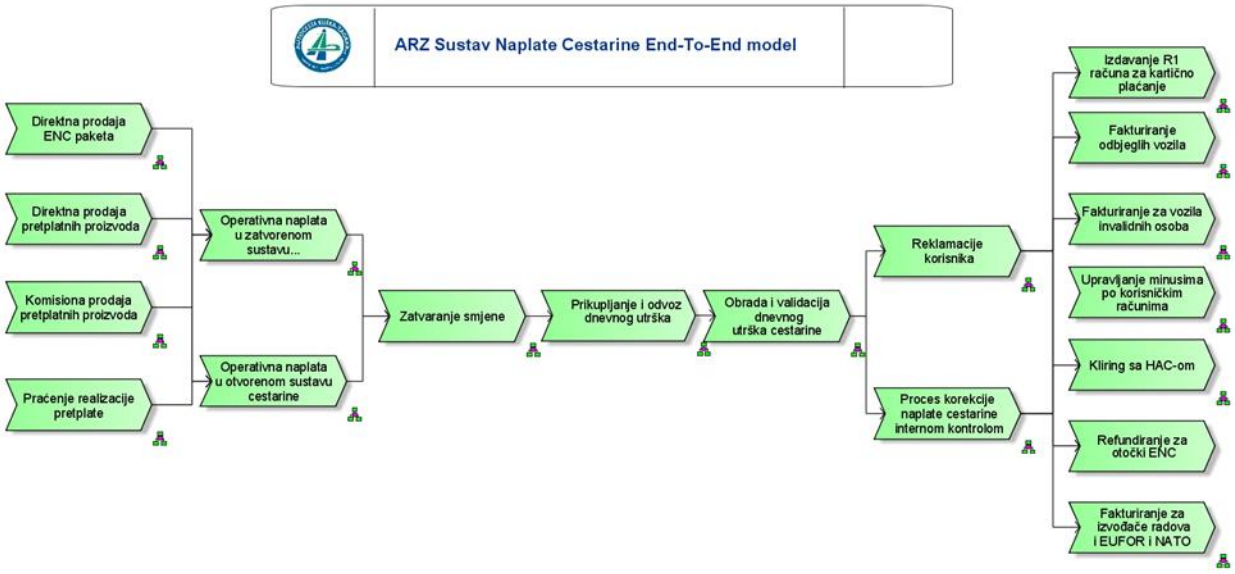
- Organizacijski pogled
- Podatkovni pogled
- Procesni/kontrolni pogled
- Funkcijski pogled
- Proizvodno-servisni pogled

Vrlo intuitivan, grafički jasan i metodološki u potpunosti zaokružen koncept ističe ARIS kao jedan od omiljenih i rado korištenih alata.

III. AS-IS ANALIZA POSLOVNIH PROCESA NAPLATE CESTARINE

Pod pojmom „Procesi naplate cestarine“ objedinili smo sve procese od prodaje pretplatnih proizvoda, preko operativne naplate do završnih knjiženja u jedinstven end-to-end proces.

Analitičkim pristupom, dekompozicijom procesa na elementarne funkcije [5][6], napravljeni su modeli koji grafički prikazuju stvarne (AS-IS) odnose izvršitelja, radnji/postupaka, te dokumenata, odnosno IT aplikacija. Tijekom radionica sa neposrednim izvršiteljima i rukovoditeljima, te uvidom u ISO (International Organization for Standardization) standard dokumentaciju, osim samog modeliranja, raspravljalo se i o problemima u radu, te mogućim načinim rješavanja tih problema. Na taj način se prikupljalo i dokumentiralo „procesno znanje“ poduzeća. Prezentirani ARIS modeli su odraz tog znanja i njihovim objavljivanjem na intranetu poduzeća pruža se prilika svim djelatnicima poduzeća da to znanje usvoje, ali i da interveniraju (kroz definiranu proceduru) kao procesni inovatori i tako malim, ali svakodnevnim koracima potiču procesnu kulturu, kao i kulturu inovacije.



Slika 3: ARZ Sustav naplate cestarine End-To-End VAC (Value-Added Chain) model

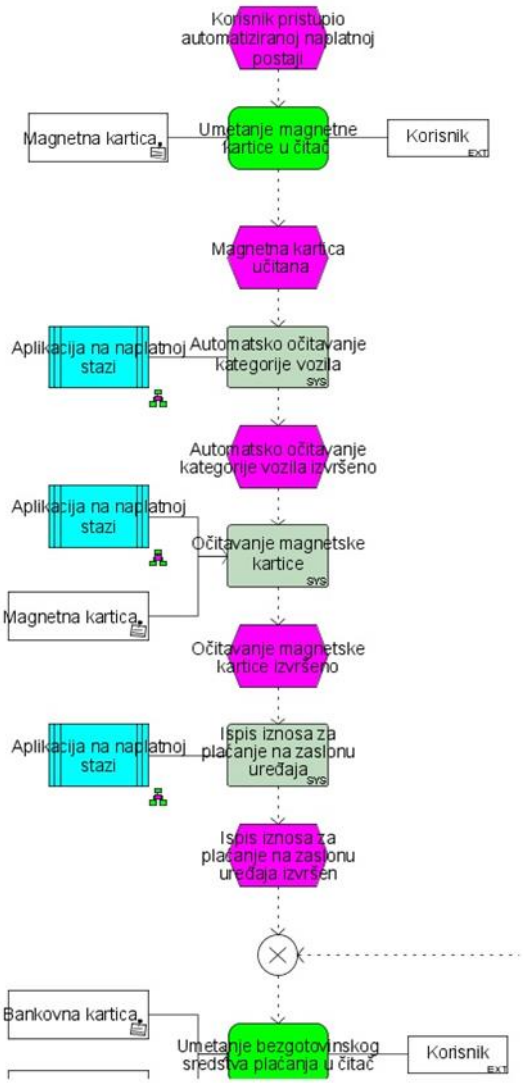
Tijekom analize prvo su prepoznate osnovne grupe procesa u procesu naplate cestarine i izrađeni su procesni modeli niske razine detalja. Zatim se analiza nastavila modeliranjem procesa sve veće i veće razine detalja uvođenjem i drugih ovisnih elemenata koji prate procesne korake. Na kraju se, sintetičkim pristupom integrirajući modele, kreirao tzv. „End-To-End“ model (Slika 3), kojim je cijeli proces naplate cestarine zaokružen u preglednu cjelinu. Uz svaki element modela u ARIS-u vidimo simbol u donjem desnom kutu koji nam „otvara“ modele veće razine detalja.

Početni modeli osnovnih procesa niske razine detalja, kao i „End-To-End“ model se prikazuju VAC (Value-Added Chain) metodom, čiji je prvenstveni cilj prikazati slijed dodavanja vrijednosti u procesu.

Modeli veće razine detalja se uobičajeno prikazuju EPC (Event-driven Process Chain) metodom, koja prikazuje slijed procesa na način da postoji uzrok (event) za neki proces, a kao rezultat tog procesa nastaje novi uzrok za slijedeći proces. Na taj način se uzroci i procesi vežu u lančani slijed (Slika 4).

Osim uzroka i procesa, na tim modelima se prikazuje i cijeli niz povezanih grafičkih simbola kao što su izvršitelj, dokument, IT sustav, rizik, KPI (Key Performance Indicator), strateški cilj, itd. Odabir koji će se elementi uključiti se definira dokumentom „Konvencije modeliranja“ koji se usvoji na početku projekta. Naravno da količina odabranih elemenata i razina detalja modela ovisi o predviđenom trajanju projekta, potrebama korisnika i budžetom projekta, pa je na početku potrebno imati konsenzus oko očekivanih rezultata projekta.

VAC i EPC modeli se, u cilju povećavanja prikaza razine detalja slažu u slojeve na način da svaki slijedeći sloj prikazuje više detalja. Uobičajeno je da se prve dvije ili tri razine prikazuju VAC modelima, a najveća razina detalja EPC modelima. Nije nužno da svi procesi imaju isti broj slojeva.



Slika 4: Prikaz dijela EPC modela

IV. ANALIZA SOFTVERSKE INFRASTRUKTURE KOJA PODRŽAVA PROCESE NAPLATE CESTARINE

Analiza procesa nužno otvara pitanje „Koji informacijski sustavu podržavaju pojedine procese?“. Odgovor na to pitanje nam pokazuje adekvatnost i pokrivenost procesa softverskim aplikacijama i informatičkom infrastrukturom, te ukazuje na potencijalne praznine (engl. gap) gdje informacijski sustav nedovoljno podržava poslovne procese.

Tijekom analitičkih radionica otvaraju se i pitanja upotrebljivosti informacijskih sustava, lakoće rada, pouzdanosti i ostalih svakodnevnih problema u radu. Prikupljena znanja se koriste kao podloga za poboljšanje poslovnih procesa, aplikacija, informatičke infrastrukture, organizacije i ostalih elemenata o kojima ovisi uspješno poslovanje.

Aplikacije koje podržavaju Sustav naplate cestarine u ARZ-u su:

- Aplikacija na naplatnoj stazi NJN-TST
- Aplikacija na naplatnoj stazi QFREE
- Aplikacija na naplatnoj stazi GEA
- Aplikacija na naplatnoj postaji NJN
- Aplikacija u centrali ARZ-a WARZ
- ERP sustav SAP

Podaci kroz ove sustave putuju u oba smjera. U smjeru od naplatne kućice do ERP SAP-a i obrnuto (Slika 5). To je ključno za naplatu, jer su u SAP-u tekuća salda svih pretplatničkih sredstava plaćanja. Brzo i pouzdano propagiranje ovih informacija kroz sustav sprječava malverzacije i osigurava točnu naplatu.

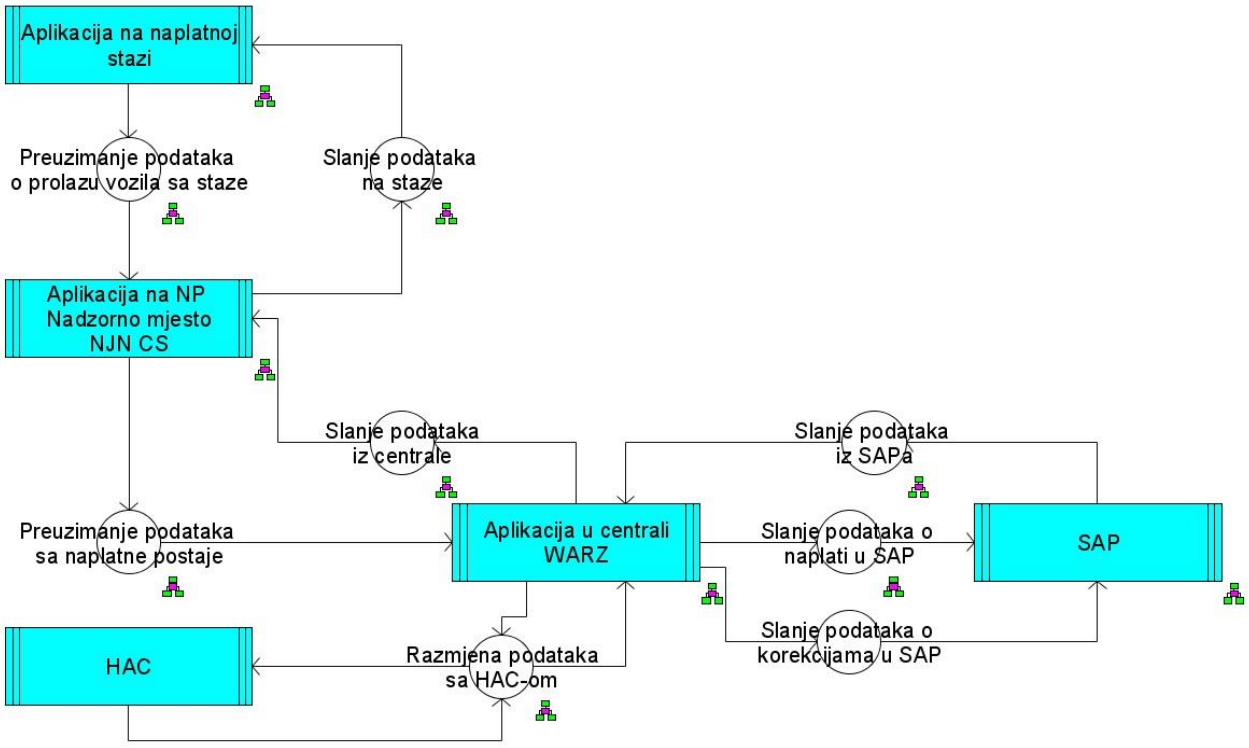
Obzirom na postojanje više različitih sustava raznih proizvođača, integracija tih sustava predstavljala je stručni izazov izvođačima. Izvedeno rješenje sa raznim sučeljima koja djeluju u oba smjera, izvedena point-to-point načinom integracije (klasični način integracije aplikacija prema kojoj se svaka aplikacija pojedinačno povezuje sa svim aplikacijama s kojima treba komunicirati) usložnjavaju razumijevanje, nadzor i održavanje, pogotovu iz razloga što je svako sučelje izvedeno različitim načinom i programirano različitim alatima. Neka sučelja su izvedena koristeći razmjenu datoteka, neka direktnim prebacivanjima SQL (Structured Qurey Language) upitima iz baze u bazu.

Naravno da bi se upotrebom nekog servera specijaliziranog za integracije (Microsoft BizTalk Servera, SAP NetWeaver Process Integration Servera, Software AG Integration Servera ili IBM Message Broker Servera,…[7]) povećala stabilnost, olakšao nadzor i održavanje, ali bi se time višestruko povećala inicijalna cijena investicije, pa bi nužno bilo tu činjenicu razmotriti uzimajući u obzir širi kontekst investicija u ARZ-u.

V. SLABE TOČKE I PRILIKE ZA UNAPRJEĐENJE

Tijekom analize poslovnih procesa i pripadajućih organizacijskih i infrastrukturnih elemenata posebna pažnja se pridaje pronalaženju tzv. „slabih točaka“ (engl. weak points) koje u ovom kontekstu treba promatrati kao prilike za unaprjeđenje.

Te se prilike klasificiraju kao „Quick Win“, odnosno prilike koje se mogu relativno brzo implementirati bez većih troškova, te kao „To be“, odnosno prilike za koje je



Slika 5: Model tijeka podataka između aplikacija

potreban veći organizacijski i/ili troškovni angažman.

Neke od slabih točaka su strateške prirode, a neke su vrlo detaljne i razumljive samo direktnim operativcima koji se s njima svakodnevno susreću. Upravo takav široki spektar pokazuje kako se prilike za poboljšanja mogu pronaći u svim segmentima poduzeća i kako je potreban sustavan pristup na svim razinama kako bi se iskoristio taj inovacijski potencijal.

Na završnim radionicama sa odgovornim osobama u ARZ-u ti problemi su analizirani. Rezultat analize su prijedlozi rješenja koji su klasificirani i prikazani u tablicama. Ovako formirane tablice predstavljaju smjernice za daljnja djelovanja.

Od ukupno 26 uočenih prilika za poboljšanja ovdje možemo navesti par primjera (TABLICA I,TABLICA II).

TABLICA I. PRIMJERI STRATEŠKE RAZINE

Opis problema	Prijedlog unaprjeđenja
Neadekvatna međusektorska komunikacija. Iako je rješavanje operativnih problema na razini jedne organizacijske cjeline dosta efikasno, operativni problemi za čije rješavanje je potrebna suradnja između dvije ili više organizacijskih cjelina dugo se rješavaju.	Poboljšati komunikaciju među organizacijskim cjelinama prilikom rješavanja otvorenih pitanja. Komunikaciju je potrebno i formalizirati i proceduralno opisati, te za svako otvoreno pitanje definirati nositelje i rokove za rješenje. Uspostaviti listu otvorenih pitanja i po njoj pratiti uspješnost rješavanja.
Veliki značaj, složenost i distribuiranost informacijskog sustava nalaže bolje svakodnevno praćenje operativnog funkcioniranja informacijskog sustava.	Potrebno je poboljšati svakodnevni nadzor nad funkcioniranjem informacijskog sustava u cilju bržeg uočavanja hardverskih kvarova, softverskih pogreški, ispravnosti prijenosa podataka i slično. U tom cilju predlaže se izrada nadzornih konzola, posebnih automatskih softverskih procedura koje prate i kontroliraju tijek podataka i alarmiraju neslaganja,…

TABLICA II. PRIMJERI TAKTIČKE RAZINE

Opis problema	Prijedlog unaprjeđenja
Iste informacije se mogu dobiti iz različitih aplikacija i sa različitim ulaznim parametrima, pa često dolazi do nesuglasica u tumačenju podataka. Česta je upotreba ad-hoc formiranih Excel tablica u kojima su često podaci stavljeni u pogrešan kontekst.	Potrebno je definirati „jednu verziju istine“ na način da se odrede mjesta u sustavu (izvještaji) koja su mjerodavna za tumačenje. Preporuka je formirati listu izvješća iz sustava sa točnom definicijom za koje poslovne potrebe se koristi i po mogućnošću sa uputom kako se ispravno koristi. Ostala izvješća time se proglašavaju kao operativna i nemjerodavna za tumačenje konkretnih poslovnih rezultata.
Način rada sa kunskim i deviznim polozima otežava kontrolu izvoda banke u računovodstvu. Računovodstvo prilikom knjiženja pologa dnevnog utrška ne može provjeriti da li su podaci ispravni ili je došlo do pogreške, jer ne mogu jednostavno zbrojiti kunski polog i devizni (u kunama) i dobiti zaduženje za naplatnu postaju.	Za ovaj problem postoji više prijedloga rješenja od drugačijeg pakiranja deviznih pologa, pa do izmjene načina knjiženja u SAP-u ili izrade kontrolnih izvještaja. Izmjena pakiranja donosi dodatne troškove, dok izmjene knjiženja usložnjavaju rad knjigovodstva. Najbolja varijanta bi bila izrada kontrolnih izvještaja, za što je informatika već predložila rješenje opisano u posebnom dokumentu.
Neprepoznate bankovne kartice od strane softvera na naplatnoj stazi ne dolaze ispravno u SAP, što stvara razlike u knjiženjima u odnosu na doznaku banke koja je karticu prepoznala i po njoj doznačila sredstva.	Dogovorom između informatike i financija problem je riješen otvaranjem posebnog konta za sve neprepoznate kartice na koje će se knjižiti promet.

Jedna od bitnih slabih točaka, uočenih i prije nego je projekt počeo, je problematičan rad sučelja koje prebacuje podatke iz aplikacije WARZ u SAP.

Uočena nestabilnost u radu, s vremena na vrijeme dosta usporen rad i sporadično pojavljivanje netočnosti u podacima, potakle su odgovorna osobe iz Sektora za informatiku ARZ-a da se u okviru ovog projekta razmotri redizajn tog sučelja.

Višestrukim pokušajima prije ovog projekta da se u postojećem sučelju pronađu pogreške nisu urodila plodom, pa je izvedba novog sučelja u ovom slučaju bila optimalno rješenje.

Obzirom na raspoloživi budžet projekta, odlučeno je da se novo sučelje izvede sličnom metodologijom point-to-point integracije. Korišteni su SQL upiti kojima su se iz Oracle baze WARZ aplikacije podaci prebacivali direktno u Oracle bazu SAP ERP sustava u transportne tzv. Z-tablice, a onda su se vremenski upravljanim procedurama pokretali posebno razvijeni BAPI-ji (Business Application Programming Interface) koji su iz tih tablica knjižili stavke u financijskim tablicama SAP ERP sustava.

Na taj način se dobilo zadovoljavajuće ubrzanje prijenosa podataka, podaci su bili točni (bez pogreški u zaokruživanju kod izračuna rabata), a cijelo sučelje je donekle i pojednostavljeno.

VI. ZAKLJUČAK

Ovim projektom izgrađena je važna baza podataka poslovnih procesa u Sustavu naplate cestarine. Uz grafički dokumentirane poslovne procese, povezani su svi izvršitelji i aplikacije koji prate te procesne korake.

Sve dokumentirane informacije o poslovnim procesima, podacima i IT sustavima su pohranjene u ARIS repozitoriju poslovnih procesa, ali i u HTML formatu kako bi im se moglo pristupiti koristeći Internet preglednik.

Time je postignuta transparentnost kao preduvjet za upravljanje i permanentno usavršavanje poslovnih procesa na svim razinama odgovornosti, kao i usavršavanje informatičke infrastrukture koja podržava poslovanje.

Analizom su uočene određene slabe točke, te se njihovim identificiranjem stvorila podloga za unapređenja.

Izmjenom SAP sučelja prema centralnoj WARZ aplikaciji otklonjene su smetnje koje su predstavljale problem u svakodnevnom funkcioniranju informacijskog sustava.

Projektom su postavljene osnove za uvođenje procesne kulture u poslovanju poduzeća. Takva korporativna kultura u svakom, pa i najmanjem procesnom koraku traži način kako bolje poslovati i svakodnevno razmatra bolje varijante rada.

LITERATURA

[1] Norton, D. (Analyst); Magic Quadrant for Business Process Analysis Tools, 2011, Gartner, Licensed for distribution by Software AG.

[2] ARIS Business Process Analysis, Software AG, http://www.softwareag.com/corporate/products/aris_alfabet/bpa/overview/

[3] Jakupovic, A.; Pavlic, M.; Fertalj, K.; Analysis and classification of ERP producers by Business Operations, 2009, CIT. Journal of computing and information technology, 17(3), 239-258.

[4] Candrlic, S.,;Pavlic, M.; Poscic, P; Model of a system for team software development, 2006, In Information Technology Interfaces, 28th International Conference on (pp. 117-122), IEEE.

[5] Pavlić, M;Razvoj informacijskih sustava: projektiranje, praktična iskustva i metodologija,1996, Znak.

[6] Pavlić, M.; Jakupović, A.;Čandrlić, S.; Modeliranje procesa, 2014.

[7] What is Integration Server?, SearchCIO TechTarget Internet portal, http://searchcio.techtarget.com/definition/integration-server

Vođenje elektroenergetskih mikromreža na osnovi sinkroniziranih mjerenja

S. Skok *, M. Filipović ** i V. Kirinčić *
* Tehnički fakultet Sveučilišta u Rijeci, Zavod za elektroenergetiku, Rijeka, Hrvatska
** HEP-Operator distribucijskog sustava d.o.o., Gospić, Hrvatska
srdjan.skok@riteh.hr, milan.filipovic@hep.hr, vedran.kirincic@riteh.hr

Sažetak - Područje elektroenergetike nezaustavljivo se sve više usmjerava prema tematici naprednih elektroenergetskih mreža (engl. Smart Grid) i mikromreža (engl. Microgrids). Naime, operatori sustava suočeni su sa sve većim izazovima koje nameću ekonomski pokazatelji, odnosno kako sa što manjim ulaganjima u elektroenergetsku mrežu postići što veću "zaradu". Ovaj je rad pokušaj pomoći u vođenju elektroenergetskoga sustava, odnosno rješenja vezanog uz podjelu distribucijskoga elektroenergetskog sustava na mikromreže s integriranim obnovljivim izvorima energije. Rad se posebno odnosi na mogućnost odvajanja mikromreže u otočni pogon te vođenje i nadzor mikromreže u stvarnom vremenu na osnovi već primijenjene tehnologije sinkroniziranih mjerenja fazora napona i struja. Cilj je rada otvoriti pitanja vezana uz primjenu mikromreža, kao i uz moguće izmjene postojećih Mrežnih pravila, te u konačnici odrediti prioritete razvoja u elektroenergetici.

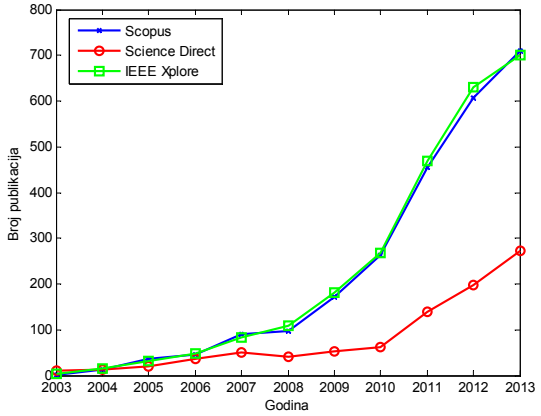
I. UVOD

Teme istraživanja u elektroenergetici mijenjaju se tijekom godina u skladu s primjenom novih tehnoloških rješenja, primjenom novih matematičkih modela, adaptivnih tehničkih rješenja ili drastičnijih primjera kao što su promjene topologije i koncepcije mreže uzrokovane integriranjem distribuiranih izvora (obično tzv. obnovljivi izvori energije) ili pak energetskim paketima i "novim" energetskim strategijama razvoja elektroenergetskoga sustava (EES). Gledajući vremenski, tematika istraživanja mijenja se gotovo s jednakom vremenskom konstantom (5-7 godina). Naime, u svjetskim okvirima, određeno područje postaje zanimljivo prilikom dolaska kapitala i potencijalnih ulaganja u istraživanje te implementacijom tehničkih rješenja koja su rezultat spomenutih istraživanja.

Pregledom stanja na području mikromreža (elektroenergetskih, naravno) u RH može se uočiti da trenutačno ima nekoliko projekata koje sufinancira država te postoje udruge i klasteri koji se bave tim područjem, ali do sada nema konkretne primjene [1, 2, 3, 4]. Posebice je indikativno da trenutačne projekte ne sufinanciraju subjekti iz privrede, osim pojedinih vezanih uz mikromreže s integriranim vjetroelektranama. No sadašnja, konzervativna, Mrežna pravila EES-a [5] definiraju da vjetroelektrane ne mogu raditi u otočnom pogonu, što uvelike smanjuje njihovo korištenje kao pomoć operatoru u vođenju sustava te uspostavu mikromreže s vjetroelektranom.

Svjetska istraživanja u području mikromreža mogu se okarakterizirati kao eksponencijalno rastuća i prema broju objavljenih članaka i prema broju projekata te održanih i predstojećih konferencija. Slika 1 i tablica 1 prikazuju broj članaka vezanih uz područje mikromreža prema tri relevantne baze – Scopus, Science Direct i IEEE Xplore.

U nastavku će se dati pregled mikromreža i stanje razvoja u svijetu i RH.



Slika 1. Godišnji broj publikacija za upit "Microgrids and Power System" [6, 7, 8]

TABLICA I. GODIŠNJI BROJ PUBLIKACIJA ZA UPIT "MICROGRIDS AND POWER SYSTEM" [6, 7, 8]

Godina/Baza	Scopus	Science Direct	IEEE Xplore
2003.	2	10	4
2004.	13	13	16
2005.	36	20	32
2006.	46	37	48
2007.	89	50	83
2008.	98	41	108
2009.	173	52	181
2010.	262	61	267
2011.	454	139	469
2012.	608	198	630
2013.	709	273	701

II. NAPREDNE ELEKTROENERGETSKE MREŽE – MIKROMREŽE

A. Napredne elektroenergetske mreže – Smart Grid

Današnja elektroenergetska mreža obuhvaća više od 50 vrsta dalekovoda u uporabi, a još više u razvojnoj fazi. Nadzor, zaštita i vođenje prijenosne mreže zasnovane na sinkroniziranim mjerenjima (engl. Wide Area Monitoring, Protection and Control – WAMPAC) više nije apstrakcija, nego se primjenjuje u mnogim elektroprivredama [9]. Razvijeni su sustavi koji služe zaštiti cjelovitosti prijenosnoga EES-a, uključujući stabilnost, sprečavanje poremećaja i minimiziranje njihovih posljedica, jedinstvenoga naziva SIPS (engl. System Integrity Protection Scheme) ili SPS (engl. Special Protection Systems) [10].

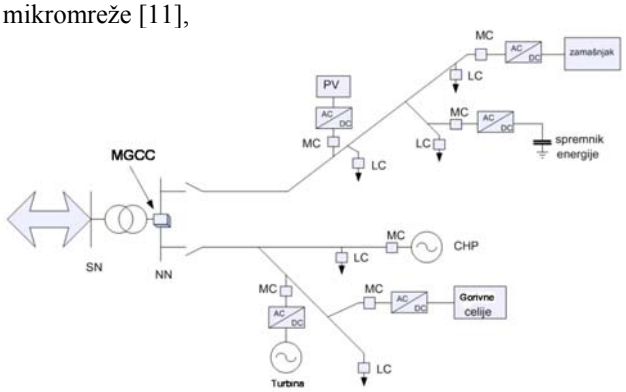
Napredne elektroenergetske mreže odnosno Smart Grid nova su generacija elektroenergetskih mreža, koje koriste nove dostupne tehnologije radi poboljšanja pouzdanosti i ekonomičnosti.

Napredne prijenosne mreže razvijaju se usporedno s tehnološkim inovacijama, u prijenosu i distribuciji i u IT i telekomunikacijskim inovacijama. Napredne mreže pružaju dvosmjerni tok energije u stvarnom vremenu, povezivanje svih sudionika u lancu proizvođača i korisnika električne energije – omogućuju bolju komunikaciju između proizvođača i operatora prijenosnoga sustava, bolju koordinaciju između distributera električne energije i krajnjih potrošača.

B. Mikromreže

Napredne mikromreže (engl. Smart Microgrids ili µGrid) moderni su, mali centralizirani elektroenergetski sustavi koji sadrže sve sastavnice kao i veliki elektroenergetski sustavi (proizvodnja – uobičajeno obnovljivi izvori električne energije; sustavi za skladištenje energije; prijenos – od izvora do potrošača; potrošači). Osnovno je svojstvo da se mikromreža može voditi kao i veliki EES te da može raditi otočno ili vezano za veliki EES. Također je bitno naglasiti da mikromreže imaju vlastite izvore proizvodnje električne energije – distribuirane izvore električne energije (uobičajeno obnovljivi izvori energije). Kao i veliki EES, napredne mikromreže proizvode, prenose, skladište i distribuiraju električnu energiju i reguliraju tok električne energije prema potrošačima, ali na lokalnoj razini. Napredne mikromreže imaju točno određene granice, jedinstveno su upravljiv entitet i mogu raditi vezano na EES kao i u otočnom pogonu. Napredne mikromreže čine sastavnice – blokove napredne elektroenergetske mreže. Naime, s razine cjelokupnoga EES-a upravlja se lokalno ograničenim naprednim mikromrežama.

Određivanje optimalne veličine, vrste i lokacije mikromreža u razdjelnim mrežama važno je za korisnika mikromreže (upravljanje proizvodnjom i dobavom energije poslovni je oportunizam) i operatora razdjelne mreže (predviđanje opterećenja i planiranje investicijskih ulaganja, pomoć pri održavanju konstantnoga napona i frekvencije). Metode korisne i upotrebljive u tržišnim uvjetima moraju se temeljiti na višekriterijskim metodama optimiranja i teoriji odlučivanja te moraju uključivati neizvjesnosti i odgovarajuće vrednovati moguće rizike. Slika 2 prikazuje tipičnu topologiju mikromreže [11],



Slika 2. Tipična topologija mikromreže

gdje je:

MGCC – centralni kontroler i nadzor mikromreže (u ovom se radu predlaže da to bude Local Area Network, odnosno nadzor i vođenje zasnovano na sinkroniziranim mjerenjima fazora napona i struje, što će biti opisano poslije) (engl. MicroGrid Central Controller). MGCC trebao bi imati tehničke i ekonomske funkcije da bi potpuno optimirao proizvodnju iz mikromreže.

MC – kontroler izvora u mikromreži (engl. Microsource Controller)

LC – kontroler potrošnje u mikromreži (engl. Load Controller).

III. NADZOR, ZAŠTITA I VOĐENJE MIKROMREŽE U STVARNOM VREMENU

A. Motivacija

Mrežna pravila EES-a [5] za sve pogonske probleme koje donosi utjecaj distribuiranih izvora u EES-u imaju rješenje u bezuvjetnom odvajanju od mreže. Danas su sve glasnija razmišljanja o potrebi promjene Mrežnih pravila [12] na strani opstojnosti paralelnoga pogona s mrežom te omogućivanju otočnoga pogona određenih vrsta elektrana s dijelom mreže.

Glavni izazovi identificirani u sklopu razmišljanja o promjeni Mrežnih pravila vezanih uz predmetnu problematiku su:

- zahtjevi prema elektrani s gledišta statičke i dinamičke potpore stabilnosti napona u točki priključenja
- mogućnosti prolaska pojedinih vrsta elektrana kroz stanje kvara u mreži i održanje elektrane u paralelnom pogonu (N-1 kriterij za distribuirane izvore energije)
- utvrđivanje primjerenih uvjeta paralelnoga pogona elektrane i mreže te uvjeta za odvajanje iz paralelnoga pogona i prelazak u otočni pogon
- otočni pogon elektrane s dijelom distribucijske mreže.

Sve to dovodi do zaključka o potrebi promjene Mrežnih pravila i donošenja novih Tehničkih uvjeta za priključenje distribuiranih izvora na distribucijsku mrežu.

Kao posljedica spomenutih glavnih razloga za stvaranje novih uvjeta koji vode k paralelnom pogonu distribuiranih izvora s distribucijskom mrežom te otočnoga pogona potrebno je definirati i sljedeće:

- izrada odgovarajućih matematičkih modela, algoritama i simulatora kao alata za pomoć u vođenju distribucijskoga EES-a
- korištenje distribuiranih izvora radi smanjenja gubitaka u distribucijskom sustavu
- održavanje kvalitete napona
- uređenje postupaka resinkronizacije distribuiranih izvora na mrežu nakon odvajanja
- procedure i uvjeti automatskoga uključenja izvora na mrežu.

B. Moguće rješenje

Problem koji se javlja ako se Mrežna pravila budu mijenjala kako je opisano jest kako ostvariti nadzor, zaštitu i vođenje distribucijske mreže. Rješenje u prijenosnom sustavu je poznato i zasniva se na uređajima za sinkronizirano mjerenje fazora (PMU uređaji – engl. Phasor Measurement Unit). Tako razvijen sustav naziva se WAMPAC odnosno Nadzor, zaštita i vođenje prijenosne mreže na osnovi sinkroniziranih mjerenja [9-10], [13-16].

Brojne su studije pokazale da tradicionalna zaštita može otkriti i izolirati kvarove u mikromreži u različitim režimima rada, pogotovo kada su u otočnom pogonu. Razlog je u značajnim razlikama struje kvara i struje opterećenja kao i u smjeru struje. Pokazalo se da se mikromreža može odgovarajuće zaštititi korištenjem naprednih digitalnih releja razmještenih na svako čvorište unutar mikromreže [17]. To se postiže uzorkovanjem napona i struje valnih oblika po stopi uzorkovanja od 1 kHz ili više te prijenosom izmjerenih vrijednosti drugim relejima u sustavu. Time se zapravo gradi shema diferencijalne zaštite koja će otkriti kvar bez obzira na način rada mikromreže. Taj je postupak veoma učinkovit u pogledu zaštite, ali je preskup.

Rješenje se može naći u preslikanoj shemi s prijenosnoga sustava, primjerice nazvanoj LAMPAC (engl. Local Area Monitoring, Protection and Control) ili Nadzor, zaštita i vođenje mikromreže. Predloženom bi se shemom u mikromreži koristili PMU uređaji i centralni kontroler. PMU uređaji mjerene vrijednosti fazora napona i struja u 20 ms razlučivosti prenose centralnom kontroleru, koji provođenjem određenih algoritama može primjerice odvojiti, mikromrežu u otočni rad. U nastavku će biti opisano predloženo moguće rješenje.

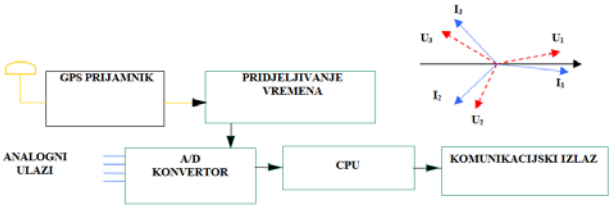
C. Local Area Monitoring, Protection and Control – LAMPAC

PMU uređaji

Uređaji za sinkronizirano mjerenje fazora tehnološki su napredak u korištenju sistemskoga nadzora i čine njegovu osnovu. Ti su uređaji objedinili klasične uređaje sekundarnih sustava i potrebno novo funkcijsko svojstvo za prikupljanje podataka u stvarnom vremenu. Tek tako oblikovani omogućili su kvalitetnu realizaciju sistemskoga nadzora.

Razvoj tih ili sličnih uređaja započeo je ranih 1980-ih, uglavnom u Sjedinjenim Američkim Državama (SAD). Početkom devedesetih objavljena je i prva norma američkoga strukovnog udruženja IEEE (eng. Institute of Electrical and Electronics Engineers). Norma IEEE 1344 o sinkrofazorima doživjela je više promjena, a najnovija verzija nosi naziv IEEE 1344-1995.

Razvojem uređaja za sinkronizirano mjerenje fazora izrađena je nova norma, koja je globalno prihvaćena (norma nosi oznaku C.37-118-2005).



Slika 3. Osnovna blok-shema uređaja za sinkronizirano mjerenje fazora

Kao i svaki drugi uređaj sekundarnih sustava, uređaji za sinkronizirano mjerenje fazora pretvaraju analogni signal u digitalni, uz određeno filtriranje i brzinu uzorkovanja. Ključna prednost uređaja za sinkronizirano mjerenje fazora pridjeljivanje je vremenske oznake za svaki uzorak. Tek se nakon toga informacija o struji, naponu i frekvenciji šalje prema daljinskim centrima vođenja. Glavne karakteristike uređaja za sinkronizirano mjerenje fazora su:

- brzina uzorkovanja od 10 do 50 uzoraka u sekundi
- uzorkovanje napona i struja
- pridjeljivanje točnog vremena uzorku, s točnošću od 10-6 (1 μs)
- vrijeme obrade unutar sinkronizirane mjerne jedinice do nekoliko milisekundi
- prijenos direktne komponente, napona, struje i frekvencije
- lokalna mjerenja i lokalne funkcije
  - funkcija podnaponske zaštite
  - funkcija nadnaponske zaštite
  - funkcija podfrekventne zaštite
  - funkcija nadfrekventne zaštite
  - funkcija nadstrujne zaštite
  - funkcije lokalnoga upravljanja i signalizacije
- brzine komunikacijskih ulaza/izlaza 10, 100 ili 200 MB
- lokalni prikaz mjerenja na uređaju.

Centralni sustav prikupljanja podataka – centralni kontroler

Centralni kontroler sastoji se od strojnoga dijela (hardware - HW) i programske podrške (software - SW), koja je najčešće smještena u sklopu opreme i podrške za nadzor, zaštitu i vođenje mikromreža u regionalnim, nacionalnim i područnim centrima upravljanja.

Hardver se u principu uvijek sastoji od konfiguracije server + klijenti. Moguće su dvije osnovne konfiguracije:

I. jedan server + klijenti

II. jedan glavni server + više podservera + klijenti.

Server može objedinjavati više funkcija:

- prihvat podataka iz mreže PMU uređaja
- obrada podataka

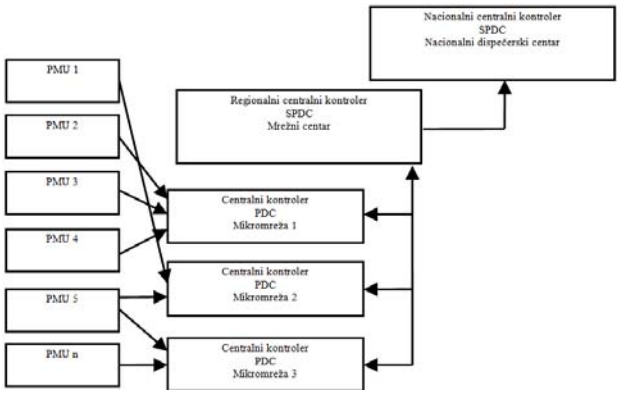
- ostvarenje funkcija nadzora, zaštite i vođenja
- arhiviranje podataka
- sučelje prema operatorima.

Osnovna razlika između konfiguracija I i II u načinu je prikupljanja podataka. Konfiguracija I ima samo jednu razinu na kojoj se prikupljaju podaci (centralni kontroler, engl. Phasor Data Concentrator – PDC), a konfiguracija II ima dvije ili više razina. Server koji prikuplja podatke na višoj razini naziva se i Super Phasor Data Concentrator (SPDC).

LAMPAC

Objedinjena ugradnja opisanih PMU uređaja na točno određena čvorišta u mikromrežama (proračuni tokova snaga i struja kratkoga spoja za različita pogonska stanja, kao i dinamički odziv distribuiranih izvora) povezana s opisanim centralnim kontrolerom, na kojem se instalira programska podrška koja pomaže operatoru sustava u nadzoru, zaštiti i vođenju, predstavlja LAMPAC sustav.

Na slici 4 prikazana je blok-shema LAMPAC sustava.



Slika 4. Blok-shema LAMPAC-a

Ako u mikromreži postoji relativno mali broj PMU uređaja, moguće ih je kvalitetno i sigurno povezati samo jednim centralnim kontrolerom. Kada se taj broj poveća, odnosno povezuje se više područja, potrebno je migrirati u pravcu više centralnih kontrolera, koji su na kraju objedinjeni s jednim regionalnim centralnim kontrolerom. Obrada podataka unutar centralnoga kontrolera ne usporava bitno protok i brzinu slanja – kašnjenje je u tom slučaju nekoliko stotina milisekundi.

Centralni kontroleri prikupljaju i obrađuju podatke iz PMU uređaja, iz cijele mikromreže u stvarnom vremenu, čime se dobivaju trenutačne snimke stanja mikromreže. Tako se ostvaruje stvarna slika stanja unutar mikromreže. Vrijeme za ostvarenje uvida u stanje mikromreže i osvježavanje podataka manje je od 200 ms. Ekstrakcijom, obradom i uporabom ekspertnih algoritama (programska podrška) omogućuje se i pravodobno informiranje i upozoravanje operatera u dispečerskim centrima. Bitna razlika u odnosu na klasične SCADA sustave je u brzini prikupljanja podataka. Klasični SCADA sustavi kasnili su s prikazom stanja unutar EES-a do nekoliko minuta i uz takvu brzinu nisu mogli pružiti sliku stanja mikromreže u stvarnom vremenu.

D. Primjeri nadzora, zaštite i vođenja mikromreža u svijetu

U svijetu se nadzor, zaštita i vođenje mikromreža zasnovano na sinkroniziranim mjerenjima fazora sve više naglašava kao rješenje te se ispituje simulacijama i pilot-projektima.

Kao primjer se može istaknuti dokument Office of Electricity Delivery and Energy Reliability – Smart Grid R&D Program, Department of Energy, USA [18], koji ističe sljedeće: "Potrebno je razviti normirani sustav koji obuhvaća hardver i softver te podržava komunikacijske protokole i Cybersecurity norme da bi omogućili tzv. plug and play distribuiranih izvora; razviti sustav na osnovi sinkroniziranih mjerenja (PMU uređaja) za proračun estimacije stanja, vođenje sustava, odnosno sustav koji bi omogućio sinkronizaciju i rekonekciju mikromreže."

IV. HRVATSKO RJEŠENJE

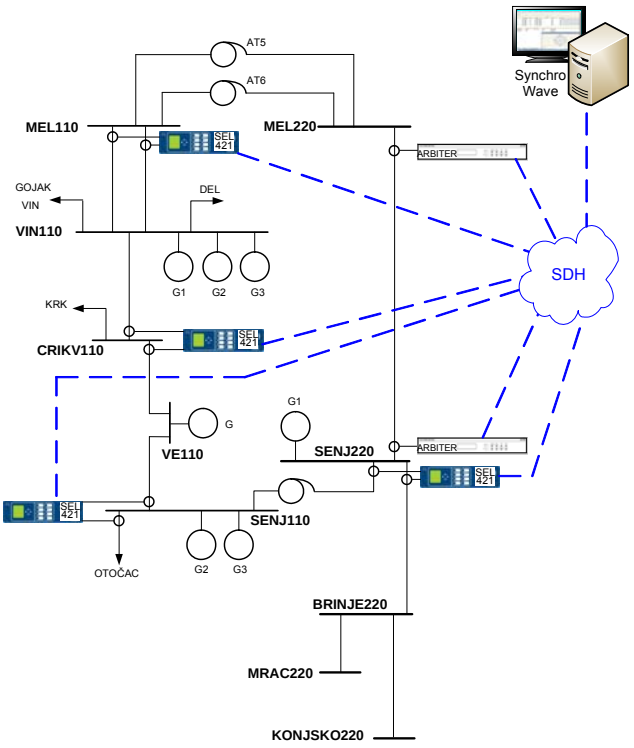
Tijekom rada na dva znanstvena projekta [19, 20] razvijen je sustav nadzora, zaštite i vođenja prijenosnoga sustava s integriranim obnovljivim izvorima energije.

Razvijeni sustav imao je smjernice da u svakom trenutku treba održavati ravnotežu između proizvodnje i potrošnje električne energije te osiguravati sigurnost i stabilnost EES-a (i to uz minimalne troškove).

Ciljano, odnosno ostvareno stanje nadzora, zaštite i vođenja prijenosnoga EES-a rješenje je koje se odnosi na tri temeljna cilja: očuvanje stabilnosti i integriteta EES-a u cjelini (odnosno SIPS), omogućivanje regulacije proizvodnje odnosno evakuacije viška snage iz dijelova EES-a u svakom trenutku te mogućnost napajanja potrošača, odnosno dobavu manjka energije. Potrebno je naglasiti da ni u kojem slučaju sustav nije zamjena za dispečera nego pomoć pri odlučivanju.

Razvijeni sustav podrazumijeva zajedničko automatsko djelovanje, zasnovano na funkcijama upravljanja, zaštite i mjerenja, radi očuvanja zadanog stanja dijela prijenosne mreže. Zahtjevi na brzinu djelovanja su ekstremni te je potrebno djelovati u stvarnom vremenu, uz veoma sigurne i pouzdane telekomunikacijske putove. Razvijeni sustav, kompletna oprema s PMU jedinicama implementirana je u Prijenosnom području (PrP) Rijeka, gdje se nalazi vjetroelektrana Vrataruša te više konvencionalnih izvora električne energije. Na slici 5 prikazana je shema nadzora, zaštite i vođenja PrP Rijeka s naznačenim PMU uređajima.

Na principima razvijenoga sustava za prijenosnu mrežu pristupilo se modificiranju za potrebe nadzora, zaštite i vođenja mikromreže, uvažavajući postojeća Mrežna pravila te zahtjeve operatora sustava i vlasnika distribuiranih izvora električne energije.

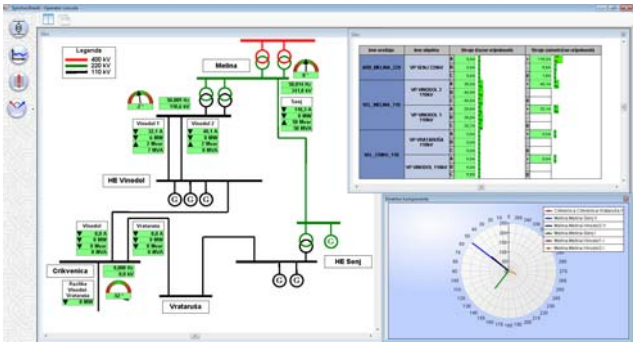


Slika 5. Shema nadzora, zaštite i vođenja PrP Rijeka s naznačenim PMU uređajima

Osnovne naznake arhitekture sustava su sljedeće:

- sustav je zasnovan na PMU uređajima opisanima u poglavlju 3.3., slika 4
- sustav je razvijen na Microsoft .NET 4.0 platformi te podržava Microsoft Windows Server 2008/2012 operacijski sustav za poslužiteljske komponente i Microsoft Windows 7/8 za korisničke aplikacije
- grafičko sučelje za operatore sustava omogućuje prikaz izmjerenih, uvezenih, izračunatih i arhiviranih vrijednosti te alarma i događaja kroz sljedeće vizualne module: prikaz vektorskih dijagrama, topološki prikaz promatrane elektroenergetske mreže (transformatorske stanice i vodovi) s pridijeljenim izmjerenim i izračunatim vrijednostima na elektroenergetskom objektu (napon, struja, frekvencija, radna snaga, jalova snaga i prividna snaga), prikaz podataka iz arhive
- alarmiranje i lista događaja – sustav upozorenja prati i signalizira promjene u radu EES-a. Pragovi za upozorenja i alarme mogu se podešavati po svim parametrima za sve granice za stanje EES-a i za sve proračune
- implementirani su proračuni: izračun radne snage (P), jalove snage (Q) i prividne snage (S), izračun razlike kutova napona sabirnica u EES-u, izračun temperature voda, izračun oscilacija frekvencije, radne snage i napona
- sustav podržava protokol IEEE C37.118 za prikupljanje i izvoz podataka
- sustav koristi Microsoft SQL Server 2008 ili višu verziju RDBMS-a za spremanje konfiguracije svih komponenti sustava i modela mreže
- u bazu podataka sprema se model mreže koji se koristi za proračune koji zahtijevaju poznavanje topologije mreže
- konfiguracijska baza i baza modela mreže podržava verzioniranje, tj. mogućnost spremanja više verzija konfiguracija i modela da bi se održala konzistentnost arhive podataka uslijed promjena konfiguracije i promjene topologije mreže
- arhiva izmjerenih i izračunatih veličina sprema se u raspodijeljenu bazu (bazu koja može biti raspodijeljena na više fizičkih servera), koju korisnik vidi kao jednu bazu, čime se postiže redundancija i segmentiranje podataka po serverima. Baza nema ograničenja vezana uz količinu i vremenski period čuvanja podataka.
- korisniku je omogućeno da bilo koji set podataka izveze u CSV format
- omogućeno je parcijalno arhiviranje na događaj u EES-u.

Na slici 6 prikazano je grafičko sučelje razvijenoga programskog paketa (prikaz je za prijenosni sustav i trenutačno se prilagođava za potrebe mikromreža).

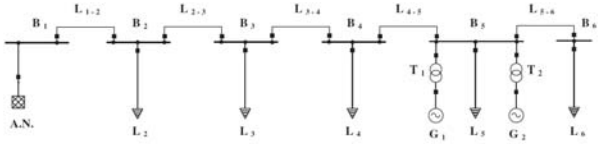


Slika 6. Prikaz razvijenoga programskog sučelja za nadzor, zaštitu i vođenje

Analiza primjenjivosti LAMPAC sustava provedena je na primjeru mikromreže Sveučilišnoga kampusa u Rijeci, za koji je predviđena kogeneracija (Combined Heat and Power – CHP) kao distribuirani izvor [21]. CHP je dizajniran da bi osigurao toplinsku energiju za Sveučilišni kampus, bolnicu na Sušaku i okolne privatne potrošače te omogućio proizvodnju električne energije.

CHP projekt obuhvaća potrebne prilagodbe u distribucijskom sustavu, ali daje LAMPAC rješenje koje omogućuje otočni pogon toga dijela distribucijske mreže, odnosno mikromreže s neprekinutom opskrbom kritičnih potrošača (npr. bolnica) električnom i toplinskom energijom. Takav je slučaj razmatran u slučaju ozbiljnog poremećaja u distribucijskom sustavu.

Na slici 7 prikazana je mikromreža Sveučilišnoga kampusa u Rijeci,



Slika 7. Mikromreža Sveučilišnoga kampusa u Rijeci

gdje je:

Bx – sabirnica

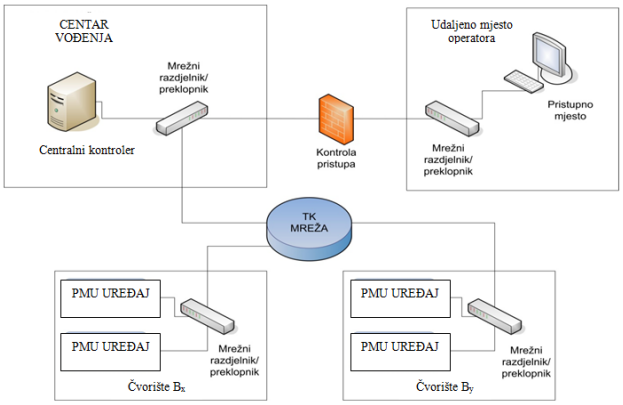
A.N. – mjesto priključka na distribucijsku mrežu

Lx-x – kabeli

Lx – tereti

Gx – generator za proizvodnju električne energije snage 1063 kW.

Na slici 8 prikazano je idejno rješenje LAMPAC sustava.



Slika 8. Idejno rješenje LAMPAC sustava za potrebe mikromreže Sveučilišnoga kampusa

V. ZAKLJUČAK

Značajan i rastući trend integriranja distribuiranih izvora u elektroenergetsku mrežu nameće operatorima sustava izazove u smislu nadzora, zaštite i vođenja distribucijske mreže. Također je potrebno preispitati postojeća Mrežna pravila te pravilnike o priključku distribuiranih izvora radi korištenja pozitivnog djelovanja distribuiranih izvora i pomoći u vođenju sustava. Da bi se to ostvarilo, potrebno je promijeniti pristup nadzoru, zaštiti i vođenju sustava prema stvarnom vremenu i voditi se smjernicama vođenja dijelova distribucijskih sustava – mikromreža paralelno s distribucijskom mrežom i u otočnom pogonu. U radu je prikazano jedno od mogućih rješenja nadzora, zaštite i vođenja mikromreža, zasnovano na sinkroniziranim mjerenjima, nazvano LAMPAC (engl. Local Area Monitoring Protection and Control). Predstavljeno rješenje podrazumijeva raspodjelu distribucijskoga sustava na više mikromreža te postavljanje PMU uređaja u određena čvorišta svake mikromreže, što se određuje provedbom proračuna tokova snaga i struja kratkoga spoja za karakteristična pogonska stanja. Mikromreže se nadziru i vode pomoću centralnoga kontrolera, koji se nalazi u centru vođenja. To rješenje obuhvaća već razvijenu programsku podršku koja ima svoju pozadinu u prijenosnom sustavu. Programska podrška vezana uz vođenje prijenosnoga sustava s integriranim obnovljivim izvorima osnova je za izradu LAMPAC sustava. Za potpuno gotov LAMPAC sustav potrebno je izraditi algoritme koji su specifični za svaki pojedini distribucijski sustav, a izraditi ih treba u suradnji s uputama operatora sustava.

Prikazano rješenje nalazi svoju financijsku i tehničku opravdanost na primjeru primjene LAMPAC sustava u mikromreži Sveučilišnoga kampusa u Rijeci. Daljnji razvoj treba svakako usmjeriti prema sličnim sustavima u stvarnom vremenu, primijenjenim na mikromreže kao temelj budućega vođenja distribucijskoga sustava. Važno bi bilo i ostvariti suradnju s operatorima sustava te izraditi pilot-projekt predloženoga LAMPAC sustava.

LITERATURA

[1] Udruga za napredne energetske mreže - Smart Grid Hrvatska. http://www.smartgrid.hr (pristupljeno 5. lipnja 2014.).

[2] <http://www.koncar-institut.hr/istrazivanje/projekti/HRZZ> (pristup 5. lipnja 2014.)

[3] <http://zprojekti.mzos.hr/public/c-prikaz_det.asp?psid=2-02&offset=20&ID=2335> (pristup 5. lipnja 2014.)

[4] <http://www.enhems-buildings.fer.hr/_news/41937/Panel_F_FER_2013.pdf> (pristup 5. lipnja 2014.)

[5] Mrežna pravila elektroenergetskog sustava, Narodne novine, 36/2006.

[6] IEEE Xplore. Online: <http://ieeexplore.ieee.org> (pristup 5. lipnja 2014.)

[7] Science Direct. Online: <http://www.sciencedirect.com> (pristup 5. lipnja 2014.)

[8] Scopus. Online: <http://www.scopus.com> (pristup 5. lipnja 2014.)

[9] S. Skok, V. Kirinčić, Projekt ugradnje PMU uređaja u cilju automatskog vođenja dijela prijenosnog EES-a PrP Rijeka, Tehnički fakultet, Sveučilište u Rijeci, 2011.

[10] Idejni projekt automatskog vođenja dijela Prijenosnog područja Rijeka, elaborat, Tehnički fakultet, Sveučilište u Rijeci, 2011.

[11] F. Katiraei, R. Iravani, N. Hatziargyriou, A. Dimeas, "Microgids Management," IEEE Power & Energy, vol. 6, no. 3, pp. 54-65, svibanj/lipanj 2008.

[12] D. Karavidović, "Distribucijska mreža i distribuirani izvori zajedno u paralelnom i otočnom pogonu," CIRED Hrvatski ogranak, 2. (8.) savjetovanje, Umag, svibanj 2010.

[13] V. Terzija, G. Valverde, D. Cai, P. Regulski, V. Madani, J.Fitch, S. Skok, M. Begovic, A. Phadke, "Wide Area Monitoring, Protection and Control of Future Electric Power Networks," Proceedings of the IEEE, vol. 99, no. 1, kolovoz 2010.

[14] S. Skok, I. Šturlić, R. Matica,"Multipurpose Open System Architecture Model of Wide Area Monitoring," Bukurešt, Rumunjska, PowerTech, lipanj/srpanj 2009.

[15] D. Novosel, K. Vu, V. Centeno, S. Skok, M. Begović, "Benefits of Synchronized-Measurement Technology for Power-Grid Applications," HICSS, Havaji, SAD, siječanj 2007.

[16] S. Skok, V. Kirinčić, R. Rubeša, K. Frlan, Z. Zbunjak,"Dinamička analiza pogona dijela elektroenergetskog sustava Prijenosnog područja Rijeka obzirom na rad VE Vrataruša," 9. simpozij o sustavu vođenja EES-a, Zadar, Hrvatska, studeni 2010.

[17] S. S. Venkata, E. Sortomme,"Using advanced measurement systems for microgrid protection," IEEE PES Innovative Smart Grid Technologies, Washington, DC, USA, siječanj 2012.

[18] DOE (Department of Energy, United States of America) – Microgrid Workshop, Chicago, Illinois, USA, srpanj 2012.

[19] Projekt "Inteligentni sustavi u prijenosnoj elektroenergetskoj mreži", lipanj 2010. – lipanj 2013.

[20] Projekt "Nadzor, zaštita i vođenje širokog područja elektroenergetskog sustava u okruženju dereguliranog i liberaliziranog tržišta električne energije", rujan 2006. –rujan 2009.

[21] D. Radulovic, S. Skok, V. Kirincic, "Cogeneration – investment dilemma, " Energy, vol. 48, no. 1, pp. 177-187, prosinac 2012.

CIS

International Conference on

INTELLIGENT SYSTEMS

Steering Committee

Chairs: Slobodan Ribarić, University of Zagreb, Croatia
Andrea Budin, Ericsson Nikola Tesla, Zagreb, Croatia

PAPERS

SPECIAL SESSION ON BIOMETRICS, SOFT-BIOMETRICS AND DE-IDENTIFICATION

Binary Pattern Flavored Feature Extractors for Facial Expression Recognition: An Overview

Rasmus Lyngby Kristensen*, Zheng-Hua Tan†, Zhanyu Ma‡ and Jun Guo‡
*Section of Image Analysis and Computer Graphics
DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
Email: raly@dtu.dk
†Signal and Information Processing section (SIP)
Department of Electronic Systems, Aalborg University, Aalborg, Denmark
Email: zt@es.aau.dk
‡Pattern Recognition and Intelligent System Lab.
Beijing University of Posts and Telecommunications, Beijing, China
Email: mazhanyu@bupt.edu.cn

***Abstract*—This paper conducts a survey of modern binary pattern flavored feature extractors applied to the *Facial Expression Recognition* (FER) problem. In total, 26 different feature extractors are included, of which six are selected for in depth description. In addition, the paper unifies important FER terminology, describes open challenges, and provides recommendations to scientific evaluation of FER systems. Lastly, it studies the facial expression recognition accuracy and blur invariance of the Local Frequency Descriptor. The paper seeks to bring together disjointed studies, and the main contribution is to provide a solid overview for future research.**

I. INTRODUCTION

When a scientific field sees a burst of published articles it often results in a clutter where an overview of state-of-the-art is lost. Today, this is what has happend to the field of *Facial Expression Recognition* (FER). Many of the recent articles focus on feature extraction and description. Especially extractors that are somewhat related to *Local Binary Patterns* (LBP) are popular. We term this kind of extractors: *Binary Pattern Flavored* feature extractors. The main contribution of this paper is to provide an overview of how this feature extractor type has been applied to FER. From our literature survey it is evident, that a substantial amount of researchers still base their FER systems on the relatively outdated LBP feature. Often it is not explicitly stated why LBP was chosen, but better performance and general robustness could have been achieved if a different and newer feature had been used. Thus, a clear overview of the current state-of-the-art will allow for the construction of better systems as well as extending them.

In order to provide a thorough overview, we make the following contributions:

- Unification of the FER terminology
- Elucidation of unresolved challenges in feature extraction for FER
- Considerations for choosing a feature extractor
- Overview of binary pattern flavored features currently used for FER
- Recommendations for good evaluation practice of feature extractors for FER

The remainder of the paper is divided according to the list above. Section II defines the FER terminology, Section III elucidates unresolved challenges, Section IV defines feature extractor considerations, Section V provides an overview of current binary pattern flavored features, Section VI provides evaluation recommendations, Section VII performs an example study of how the *Local Frequency Descriptor* (LFD) feature extractor performs for FER, and Section VIII concludes the paper.

II. TERMINOLOGY

By looking through the literature, one can see that different researchers use a slightly different FER terminology. To encourage consistency in the future, we seek to define a set of pivot terms based on the general FER surveys [1], [2], [3], and [4].

Spontaneous expressions are based on a true, underlying felt emotion. They are typically what FER systems aim at recognizing as they are the ones carrying true emotion information.

Posed expressions are acted expressions. According to [5], posed and spontaneous expressions look differently. Spontaneous expressions are hard to obtain, which is why most facial expression databases contain posed expressions.

Micro expressions are expressions which never reach full apex because the executing person suppresses the expressions before they are fully formed. Micro expressions can be very informative in some applications, such as clinical psychology.

Basic facial expressions were defined by [6]. They are six expressions which has been proved universal and are as follows: Happiness, Sadness, Anger, Surprise, Disgust and Fear. Sometimes Neutral is included as well. The basic expressions have been extensively used in FER research, but [7] and [8] argue that they are inadequate.

Entity based FER recognizes expressions by considering the entire face as one entity. This is by far the most popular

and most researched form of FER [9].

Action Unit based FER recognizes muscle contractions and relaxations in the face and use them to recognize the expression. These activations of face muscles are known as *Action Units* (AU) as described by the *Facial Action Coding System* (FACS) [10].

Permanent facial features are features being permanent in the face, such as the eyes and mouth [11].

Transient facial features are volatile face features, meaning that they occur and disappear as the face moves, such as wrinkles and furrows [11].

Geometric features are image features which describe some form of physical parameters of the face, such as positions, shapes, sizes, or similar. Usually, a set of easily recognizable salient points are defined on the face and their positions and mutual distances are used to describe the face.

Appearance features describe image properties based directly on pixel intensities. Often, they are formed by filtering images by a bank of image filters and use their outputs as features. Binary pattern flavored features fall in this category.

Global feature descriptors describe the features over the entire face. Thus, a very small amount of feature location information is kept.

Local features descriptors contain information about the spatial position of features. Many ways of retaining locality information exist, but it is often done by dividing faces into smaller regions, each with its own descriptor.

Engineered features arefeatures which have been explicitly designed and therefore convey some form of meaningful information about the objects in the image.

Learned features are features which are learned from a dataset of observations by a learning algorithm. As a result, the features do usually not convey any meaningful information.

Binary pattern flavored features describe objects in an image by binary patterns extracted from small local image patches. The standard example is LBP. Some authors use the term ”local features” but obviously this term is too easily confused with the term *local feature descriptors*, as defined above.

III. RESEARCH CHALLENGES

In 2011, the first *Facial Expression Recognition and Analysis Challenge* (FERA2011) [12] was held as a way to uncover the then current status of the FER field. A meta analysis of the submitted papers was published in 2012 [9]. Based on this meta analysis, and our thorough literature survey, there seems to be a recent overweight of papers that focus on entity based FER from high-resolution faces viewed front on. The meta analysis concluded that this problem can be considered largely solved for the person-specific case.

We believe that the continued substantial amount of research going into this problem is due to missing awareness of the state-of-the-art. Clarifying the current general challenges of the FER field also results in clarification of the feature specific challenges, as they must be correlated.

Be aware that research has already been done on some of the challenges in the following list, but that more is needed. Also note that most papers only tackle few challenges at a time. Thus, there is unexplored territory to conquer for researchers who are interested in creating multi-invariant systems.

Blurred faces can result from disturbances in the atmosphere or out-of-focus images. *Local Phase Quantization*, developed by [13] for face recognition in blurred images and applied to FER by [14] and [15], is particularly interesting. Note that neither [14] nor [15] investigates FER in blurred images. They use LPQ solely for its high descriptiveness.

Illumination variations results from changes in direction or intensity of light used to illuminate faces. This is often seen in face images obtained in a real-world setting. A substantial amount of research has been done on intensity invariance, and it is a natural property of most binary pattern flavored features. In our survey, we only encountered research on direction invariance as a subparts of larger studies, such as [16].

Transformation of face position and orientation happens if a person is allowed to move freely, as would be the case in a real-world setting. Quite a lot of research has gone into rotation and scaling invariance, such as rotation and scale invariant LBP [17]. Translations aligned with the image plane are often handled by the segmentation algorithm. Changes to face orientation have been studied, such as [18], but it remains an open problem.

Face occlusion happens if there is not a direct line of sight between the camera and the face. Occlusion is often due to mustaches, glasses or hair. It has seen a relatively large amount of research, such as the comparative study done by [4]. However, the topic still remains active and unsolved.

Variations in faces between different people make facial expressions person dependent. [19] proposed to use the difference between the expressive face and the neutral face to eliminate the person specific contributions. However, this requires the setting to be person specific, thus with known identities.

Multiple faces in one image could be a problem. We are aware that this challenge has been studied for face recognition, but so far we have not encountered any such research related to FER.

Variations in facial expression intensity happen because humans can form both subtle and explicit versions of facial expressions. A smile might be small and subtle, it can be large and explicit, or somewhere in between. So far, we have not encountered any research which addresses this challenge.

Recognition of AUs has been done, but as noted in the meta analysis of FERA2011 there is room for a lot more. Reliable recognition of AUs is still a challenge.

Processing time is an important factor in instant feedback systems, which is usually the case for social robots [56] and the like. Some researchers provide notes about the processing time of their systems, but often just as a minor detail. A comparative study would be worthwhile.

TABLE I. FEATURE ABBREVIATIONS

Abbr.	Phrase
LBP	Local Binary Patterns [26]
$LBP^{riu2}_{R,P}$	Rotation Invariant LBP [27]
FLBP	Fuzzy LBP [28]
PLBP	Pyramid of LBP [29]
MBP	Median Binary Patterns [30]
LTP	Local Transitional Patterns [31]
MTP	Median Ternary Patterns [32]
VLBP	Volume LBP [33]
LBP-TOP	LBP from Three Orthogonal Planes [33]
SLMBP	Spatiotemporal Local Monogenic Binary Patterns [34]
LGBP	Local Gabor Binary Patterns [35]
LGBP-TOP	LGBP from Three Orthogonal Planes [36]
LGDP	Local Gabor Directional Patterns [21]
RGJ	Radial encoded Gabor Jets [37]
LGTP	Local Gabor Transitional Patterns [38]
LDP	Local Directional Patterns [39]
LDPv	LDP variance [40]
GDP	Gradient Directional Patterns [41]
WLD	Weber Local Descriptor [42]
LPQ	Local Phase Quantization [14]
RI-LPQ	Rotation Invariant LPQ [43]
VLPQ	Volume LPQ [44]
LPQ-TOP	LPQ from Three Orthogonal Planes [15]
LCVBP	Local Color Vector Binary Patterns [45]
TPCF	Tensor Perceptual Color Framework [16]
CLBP	Curvature Local Binary Patterns [46]
PRICoLBP	Pairwise Rotation Invariant Co-occurrence Local Binary Pattern [47]
Ferns	[48], [49]
LEP	Local Energy Patterns [50]
LFD	Local Frequency Descriptor [51]

IV. FEATURE EXTRACTOR CONSIDERATIONS

A set of considerations for selecting a feature extractor for FER has to be made. A flowchart of the choices can be seen in Fig. 1. Five binary primary choices, all based on ones research focus, has to be made, some with sub-choices. Note that two opposite choices can be combined for potential higher recognition rates as reported by [20].

Most researchers tend to make these choices implicit without arguments. We believe that making choices explicit with solid argumentation will lead to better understanding and therefore better science. Therefore, we encourage future researchers to provide argumentation for the choices they make.

V. BINARY FLAVORED FEATURES

This section provides an overview of modern binary pattern flavored features which have previously been applied to the FER problem. Fig. 2 shows the relationships between these features, including four new, promising features. These four features were selected because they have provided promising results for other applications. We strived to make the list of previously used features as comprehensive as possible, but please note that some could be missing by accident.

From the figure it is evident, that there is a large overweight of feature extractors which use still images containing 2D intensity data as input. In fact, only a single extractor relies on 3D depth data. Only a handful of extractors relies on image sequences. The rest use still images. Based on our survey, this tendency seems to be a good indicator for the status of the FER field in general. We hope that this survey will help to change that. The used abbreviations are explained in Table I.

We encourage readers to use this overview to select a proper feature extractor for their own research, and to minimize the development of redundant features. In 2011, [18] proposed to use *Local Gabor Binary Patterns* (LGBP) for makeing a face-orientation invariant FER system. Later the same year, a system using LGBP for AU based FER won the AU recognition sub-challenge of FERA2011 [35]. As recent as 2014, a new paper on entity based FER using LGBP was published by [52]. The authors seems to be unaware of the previous use of LGBP, as they make no notes or citations to the two articles listed above. Thus, redundant work could have been avoided by a clear overview of the state-of-the-art.

Below we have selected six of the most interesting recently used feature extractors for a more detailed description. We hope that the descriptions will help readers choose a good, modern feature extractor, and thereby enable them to develop highly accurate and robust FER systems for practical use. Note that the novel *Local Frequency Descriptor* (LFD) will be studied in Section VII. As proposed by [53], all of the described features use local descriptors formed by dividing the input images into a set of non-overlapping blocks, each with its own descriptor. The descriptors are then concatenated into one, large descriptor, thus keeping some locality information. This is informally known as the *grid division approach*. Also note that all the referenced methods use entity based FER.

A. Features from static image data

Pyramid of Local Binary Patterns (PLBP) was proposed in 2013 by [29]. PLBP is interesting due to their high level of descriptiveness and low complexity compared to other pyramid methods. The key idea is to extract $LBP^{u2}_{R,P}$ features from a given rectangular region of the image. That region is then sub-divided into four smaller regions, each of which can be further sub-divided and so on. LBP features are extracted from all regions, and their resulting descriptors are concatenated into one. The authors show that two subdivisions are enough, thus creating a total of five histograms for each region. They use the region covering the mouth and the region covering the eyes. PLBP requires less memory and provides faster computation than previous pyramid features. Further, PLBP provides a stronger discriminative ability than previous non-pyramid methods. However, it has not been compared against LPQ or similar modern features.

Median Ternary Patterns (MTP) was proposed in 2013 by [32]. MTP is interesting due to their robustness and high level of descriptiveness. MTP combines the strengths of MBP and LTP to form a highly illumination and random noise invariant feature. The idea is to quantize each pixel into a ternary pattern. For a given pixel, the median of the intensities of its eight directly neighboring pixels is calculated. The neighbors are quantized into three states depending on their relation to the median. The ternary pattern is split into two binary patterns and their two corresponding descriptors are concatenated together. MTP has a higher recognition rate and better illumination and noise robustness than pixel intensity *Principal Component Analysis* (PCA), LBP, MBP, and LTP. No comparison against LPQ or similar has been done.

Local Color Vector Binary Patterns (LCVBP) was proposed in 2013 by [45]. LCVBP is interesting because they

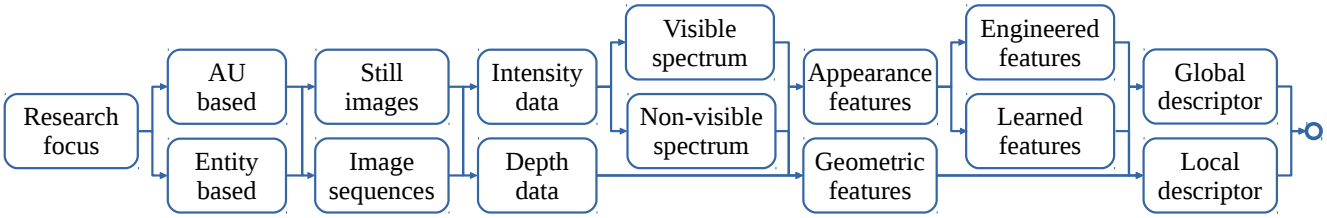


Fig. 1. Flowchart of considerations for selecting a feature extractor. Making these choices should always be based on what is best in relation to the research focus. Opposite choices can be combined, e.g. a system could use both intensity data and depth data to achieve a higher descriptiveness.

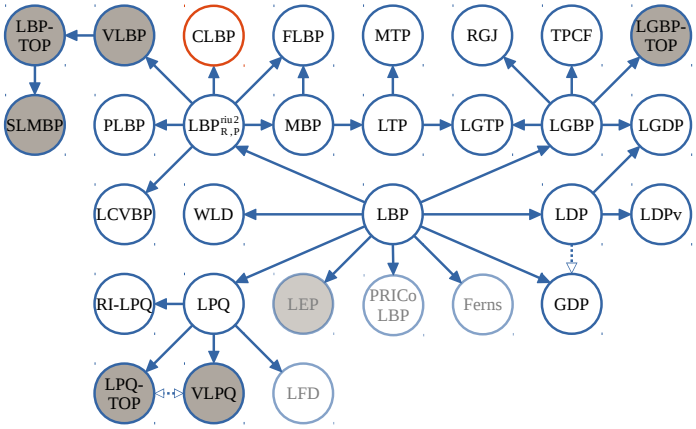


Fig. 2. Overview of 26 binary flavored feature extractors which has been applied to FER, along with four new features. Each node is a feature. Coloring codes: White node = still image data, Gray node = image sequence data, Blue rim = 2D intensity image data, Orange rim: 3D depth data. Solid arrows indicate a direct extension of a previous method. Dotted arrows indicate an inspirational link to a previous method. Transparent nodes indicate new features. The abbreviations are defined in Table I.

are among the few binary flavored features which incorporates color information. The key idea is to treat the color of each pixel as a three dimensional vector of which the magnitude can be calculated by $r = \sqrt{R^2 + G^2 + B^2}$. Binary patterns are created from the magnitude response similarly to the process used by $\text{LBP}_{R,P}$. In addition to these patterns, also magnitude difference patterns, color directional patterns, and color directional arc patterns are extracted. The descriptors from the four pattern types are concatenated to form the final descriptor. LCVBP provides a higher recognition accuracy than a large set of previous features, including $\text{LBP}^{u2}_{8,2}$, LGBP, and TPCF, but excluding LPQ. Further, it is shown that LCVBP allows for sparse representation, is person-independent, and robust against illumination changes.

Local Gabor Directional Patterns (LGDP) was proposed in 2012 by [21]. LGDP is interesting because it revives the use of Gabor wavelet filters in modern binary pattern features, and attain a high recognition rate in doing so. LGDP combines Gabor wavelet filtering and *Local Directional Patterns* (LDP). This is done by exchanging the eight Kirsch compass kernel filters normally used in LDP by eight Gabor wavelet filters corresponding to eight different angles. The eight filter responses are encoded using the same approach as LDP. This is done at five different scales, creating five descriptors which are concatenated. LGDP is showed to provide higher recognition accuracy than LDP and LBP (among others), but it is not compared to LPQ. It is further showed that it is invariant to illumination changes and random noise.

B. Features from image sequence data

Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) was proposed in 2011 by [15]. LPQ and its derivatives are interesting due to its well established theoretical background, invariance to image blurring and illumination changes, and high level of descriptiveness. It has been showed, that the spatial phase angles of a centrally symmetric *Point Spread Function* (PSF), such as atmosphere and out-of-focus blur, is 0 at frequencies with positive real valued frequency responses. LPQ encodes the phase angles, which makes it blur invariant. The Three Orthogonal Planes framework was developed as an LBP extension by [33]. A given spatial 2D image sequence is considered as a 3D space by letting the movement in time represent the third dimension. From the 3D space, three orthogonal 2D planes are extracted. LPQ descriptors are extracted from each plane and concatenated together to form the sequence descriptor. LPQ-TOP provides higher recognition rates than LBP-TOP and is invariant to image blurring as well as intensity changes. The TOP framework has also been applied to LGBP, forming the LGBP-TOP descriptor by [36]. Unfortunately, LGBP-TOP and LPQ-TOP was not compared.

C. Features from 3D depth data

Curvature Local Binary Patterns (CLBP) was proposed in 2013 by [46]. CLBP is interesting because it is the only binary pattern flavored feature we have encountered which uses 3D depth data. CLBP is invariant to changes in face rotation, face orientation (pose), and changes in illumination. Four types of curvature measures are calculated, namely the two principal curvatures k_1 and k_2, the mean curvature, and the shape index. The four curvature measures are encoded by LBP, and the four resulting feature descriptors are concatenated. It is showed that the recognition accuracy of CLBP is higher than that of two state-of-the-art 3D FER methods, which does not use binary pattern flavored features. It would be interesting to compare this 3D feature to state-of-the-art 2D features with respect to invariance and recognition accuracy. Such a study would provide arguments for choosing one type of input data over another.

VI. RECOMMENDATIONS FOR EVALUATION

To extend the state-of-the-art, one has to benchmark results against those previously reported. Benchmarking can easily be done if all researchers use the same well defined benchmarking datasets along with the same well defined benchmarking procedures. As things are now, quite a few different facial

expression databases and evaluation procedures are used. This includes different databases containing static and sequential images, static and sequential 3D depth data, labeled with basic facial expressions for entity based FER, and/or AUs for AU based FER.

Some 3D depth databases of facial expressions exist, but as was shown in Section V, research on 3D methods is relatively sparse. Therefore, we have not been able to identify a standard benchmarking database and procedure for 3D methods.

For 2D methods in intensity images, the *Cohn-Kanade* (CK) and the *Extended Cohn-Kanade* (CK+) [22] databases are popular. Both databases were primarily created for research in AUs and AU based FER. Therefore, they contain AU labels for all images, but only CK+ contains facial expression labels. As a consequence, researchers who use the original CK database have to manually apply labels to the images, thus potentially creating different labels. To fairly compare results, procedures such as the number of cross-validation groups and the specific identity of samples in each cross-validation groups have to be precisely the same. This process also tends to vary from researcher to researcher.

A good candidate for a unified benchmarking dataset is the well defined dataset and procedure which was used for FERA2011 called GEMEP-FERA2011 [12]. It is composed of two subsets of the original GEMEP dataset, one annotated by FACS for AU based FER and one annotated by emotions for entity based FER. The used emotions are: Anger, Fear, Joy, Relief, and Sadness. Though it has been little used since the competition, it is still maintained and available.

Naturally one usually wants to know how well their FER system recognizes real facial expressions. As mentioned in Section II, true spontaneous expressions are difficult to obtain. Some databases containing spontaneous expressions have been compiled, such as the Rochester/UCSD FacialActionCodingSystem Database 1 (RUFACS1) [23] for 2D intensity data and the BP4D-Spontaneous [24] database for 3D depth data. However, when using only a single database there is the risk of training ones system to recognize some underlying property of the images, other than the actual facial expressions. If so, artificially high recognition rates will generally be obtained when testing on images from the same dataset. To avoid this, [25] showed that cross-database evaluation on multiple facial expression databases will provide a closer-to-real-world recognition accuracy.

VII. AN EXAMPLE STUDY

This section performs an experiment which studies how well LFD recognizes blurred facial expressions. We chose to compare LFD against $\mathrm{LBP}^{u2}_{8,1}$ due to its continued popularity within the field and against LPQ due to its link to LFD as well as its high blur invariance and level of descriptiveness. LFD was developed by [51] as an extension of LPQ which includes magnitude information. They showed that LFD provides higher recognition rates than LPQ for face recognition in blurred images. Thus, our hypothesis was that LFD would also provide higher recognition for FER in blurred images.

We chose to use the CK and the *Karolinska Directed Emotional Faces* (KDEF) [54] databases because of the larger number of images contained in both databases compared to the GEMEP-FERA2011 dataset. Previous research on LPQ have focused on entity based FER. As we wished to benchmark LFD against the state-of-the-art, we also focused on entity based FER. For the same reason we chose still images as our input data.

We chose to base our system on the baseline system defined for FERA2011 because of its simplicity. Our system started by extracting LBP, LPQ, and LFD features using the grid division approach. Then, the dimensionality of each of the feature descriptors was reduced to 140 dimensions by PCA. Lastly, One-Versus-One Support Vector Machines (SVM) [55] were used for classification. Five fold cross-validation was used, and the experiment was executed a total of 10 times with 10 different partitions for the five cross-validation groups. This was done to lower the risk of obtaining bad results due to a bad selection of cross-validation groups.

Every time the experiment was run, the clear images in the four training cross-validation groups were used. The images in the remaining cross-validation group were blurred and used for testing. Gaussian blur was used, and a total of 17 different kernel variances were tried, ranging from 0 to 4 in steps of 0.25. The results of the experiment can be seen in Fig. 3.

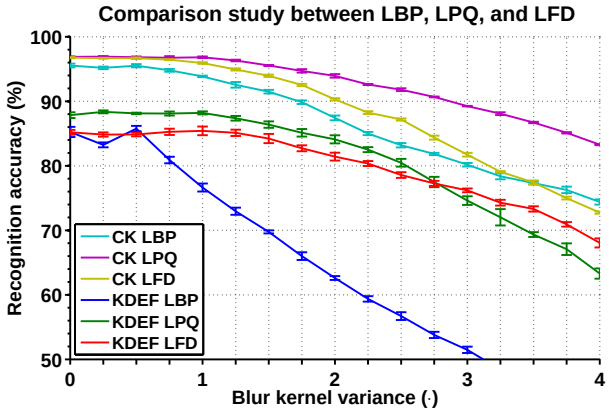


Fig. 3. FER accuracy of LBP, LPQ, and LFD features extracted from blurred images from the CK and KDEF databases. Note that LPQ performs generally better than LFD which again performs generally better than LBP for both databases. Error bars are used to indicate the variance of the recognition accuracy calculated based on 10 runs of five fold cross-validation.

We observe that LPQ generally performs better than LFD, which generally performs better than LBP. A steeper decline in the recognition accuracy of LPQ than LFD is observed for KDEF, which makes LFD better than LPQ at blur-variances of 3 or higher. However, the this effect is not seen for CK. A reason might be, that the CK dataset contains more explicit expressions compared to the KDEF set (which is why the accuracy for CK is generally 10%-points higher than for KDEF). Thus, LFD might have an advantage over LPQ for recognizing subtle expressions in very blurred images. Note that this correlation is highly speculative, and more studies on other datasets containing subtle expressions would have to be done to state anything conclusive.

VIII. CONCLUSION

This paper compiled the most comprehensive overview of the use of binary pattern flavored features for FER to date. We chose to focus on this type of features due to their high popularity and thus large number of recent publications. In addition to the overview, we unified the fundamental FER terminology, elucidated unresolved FER challenges, provided considerations for selecting an FER feature extractor as well as recommendations for scientific evaluation of FER systems, and studied the descriptiveness of the novel LFD feature by comparing it against LBP and LPQ.

Our study of LFD showed that it provides a recognition rate and blur invariance which is higher than LBP, but lower than LPQ.

Despite the large quantity of research in the field, we showed that there are still challenges to overcome. In particular, we showed that multi-robust systems, the use of 3D depth data, the use of image sequences, and FER based on AUs are all more or less open topics. In particular, we see AU based FER as an interesting topic. Recognizing AUs would open up for the possibility of recognizing micro expressions as well as indicating the intensity of recognized expressions. Those properties would be of value to clinical psychology as well as advanced human-machine interfaces. Further, de-identification would be a natural part of an AU system if only the recognized AUs are stored. This could make the system easier applicable to public spaces as no personal identity information would be stored.

IX. ACKNOWLEDGMENTS

We would like to extend our gratitude to ?, ?, ?, and ? for providing travel funds which made our corporation between ? and ?.

REFERENCES

[1] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art", *arXiv.org*, abs/1203.6722, 2012.

[2] N.U. Khan, "A comparative analysis of facial expression recognition techniques", *IEEE Third International Advance Computing Conference (IACC)*, Ghaziabad, 2013, pp. 1262-1268.

[3] C. Caleanu, "Face expression recognition: A brief overview of the last decade", *IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, 2013, pp. 157-161.

[4] C.F. Liew and T. Yairi, "A comparison study of feature spaces and classification methods for facial expression recognition", *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Shenzhen, 2013, pp. 1294-1299.

[5] P. Ekman and E.L. Rosenberg, "What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)", *Oxford University Press*, vol. 1, 1997.

[6] P. Ekman and W.V. Friesen, "Constants across cultures in the face and emotion", *Journal of Personality and Social Psychology*, vol. 17(2), 1971, pp. 124-129.

[7] W.G. Parrott, "Emotions in Social Psychology: Key Readings", *Psychology Press*, vol. 1, 2000.

[8] S. Du, Y. Tao, A.M. Martinez, "Compound facial expressions of emotion", *Proceedings of the National Academy of Sciences*, vol. 111(15), 2014, pp. 1454-1462.

[9] M.F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-Analysis of the First Facial Expression Recognition Challenge", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42(4), 2012, pp. 966-979.

[10] P. Ekman and W.V. Friesen, "Manual for the Facial Action Coding System", *Consulting Psychologists Press*, vol. 3, 1978.

[11] Y. Tian, T. Kanade, and J.F. Cohn, "Recognizing action units for facial expression analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(2), 2001, pp. 97-115.

[12] M.F. Valstar, B. Jiang, M. Mhu, M. Pantic, and K. Scherer, The First Facial Expression Recognition and Analysis Challenge, *IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, 2011, pp. 921-926.

[13] V. Ojansivu and J. Heikkil, "Blur Insensitive Texture Classification Using Local Phase Quantization", *Image and Signal Processing*, vol. 5099, 2008, pp. 236-243.

[14] S. Yang and B. Bhanu, "Facial expression recognition using emotion avatar image", *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, Santa Barbara, 2011, pp. 866-871.

[15] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features", *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, Santa Barbara, 2011, pp. 878-883.

[16] S.M. Lajevardi and H.R. Wu, "Facial Expression Recognition in Perceptual Color Space", *IEEE Transactions on Image Processing*, vol. 21(8), 2012, pp. 3721-3733.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24(7), 2002, pp. 971-987.

[18] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition", *Computer Vision and Image Understanding*, vol. 115(4), 2011, pp. 541-558.

[19] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l1 optimization", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, San Francisco, 2010, pp. 32-39.

[20] K. Yu, Z. Wang, M. Hagenbuchner, and D.D. Feng, "Spectral embedding based facial expression recognition with multiple features", *Neurocomputing*, vol. 129, 2014, pp. 136-145.

[21] S.M.Z. Ishraque, A.K.M.H. Banna, and O. Chae, "Local Gabor directional pattern for facial expression recognition", *15th International Conference on Computer and Information Technology (ICCIT)*, Chittagong, 2012, pp. 164-167.

[22] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, San Francisco, 2010, pp. 94-101.

[23] M. Frank, J. Movellan, M. Bartlett, and G. Littleworth, "Rochester/UCSD FacialActionCodingSystem Database1", Machine Perception Laboratory, U.C. San Diego

[24] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J.M. Girard, "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database", *Image and Vision Computing*, vol. 32(10), 2014, pp. 692-706.

[25] C. Mayer, M. Eggers, and B. Radig, "Cross-database evaluation for facial expression recognition", *Pattern Recognition and Image Analysis*, vol. 24(1), 2014, pp. 124-132.

[26] X. Feng, "Facial expression recognition based on local binary patterns and coarse-to-fine classification", *The Fourth International Conference on Computer and Information Technology (CIT)*, 2004, pp. 178-183.

[27] C. Shan, S. Gong, and P.W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study", *Image and Vision Computing*, vol. 27(6), 2009, pp. 803816.

[28] M.R. Mohammadi and E. Fatemizadeh, "Fuzzy local binary patterns: A comparison between Min-Max and Dot-Sum operators in the application of facial expression recognition", *8th Iranian Conference on Machine Vision and Image Processing (MVIP)*, Zanjan, 2013, pp. 315-319.

[29] R.A. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Framework for reliable, real-time facial expression recognition for low resolution images", *Pattern Recognition Letters*, vol. 34(10), 2013, pp. 1159-1168.

[30] A. Hafiane, G. Seetharaman, and B. Zavidovique, "Median Binary Pattern for Textures Classification", *4th International Conference on Image Analysis and Recognition*, Montreal, 2007, pp. 387-398.

[31] T. Jabid, and O.S. Chae, "Local Transitional Pattern: A Robust Facial Image Descriptor for Automatic Facial Expression Recognition", *Proceedings of the International Conference on Computer Convergence Technology*, Seoul, 2011, pp. 333344.

[32] F. Bashar, A. Khan, F. Ahmed, and M.H. Kabir, "Robust facial expression recognition based on median ternary pattern (MTP)", *International Conference on Electrical Information and Communication Technology (EICT)*, Khulna, 2014, pp. 1-5.

[33] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29(6), 2007, pp. 915-928.

[34] X. Huang, G. Zhao, W. Zheng, and M. Pietikainen, "Spatiotemporal Local Monogenic Binary Patterns for Facial Expression Recognition", *IEEE Signal Processing Letters*, vol. 19(5), 2012, pp. 243-246.

[35] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42(4), 2012, pp. 993-1005.

[36] T.R. Almaev and M.F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition", *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, Geneva, 2013, pp. 356-361.

[37] W. Gu, C. Xiang, Y.V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis", *Pattern Recognition*, vol. 45(1), 2012, pp. 80-91.

[38] T. Ahsan, T. Jabid, U. Chong, "Facial Expression Recognition Using Local Transitional Pattern on Gabor Filtered Facial Images", *IETE Technical Review*, vol. 30(1), 2013, pp. 47-52.

[39] T. Jabid, H. Kabir, and O. Chae, "Robust Facial Expression Recognition Based on Local Directional Pattern", *ETRI Journal*, vol. 32(5), 2010, pp. 784794.

[40] H. Kabir, T. Jabid, and O. Chae, "A Local Directional Pattern Variance (LDPv) Based Face Descriptor for Human Facial Expression Recognition", *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Boston, 2010, pp. 526-532.

[41] F. Ahmed, "Gradient directional pattern: A robust feature descriptor for facial expression recognition", *Electronics Letters*, vol. 48(19), 2012, pp. 1203-1204.

[42] X. Wang, C. Jin, W. Liu, M. Hu, L. Xu, and F. Ren, "Feature fusion of HOG and WLD for facial expression recognition", *IEEE/SICE International Symposium on System Integration (SII)*, Kobe, 2013, pp. 227-232.

[43] J. Li and Z. Ying, "Facial Expression Recognition Based on Rotation Invariant Local Phase Quantization and Sparse Representation", *Second International Conference on Instrumentation, Measurement, Computer, Communication and Control (IMCCC)*, Harbin, 2012, pp. 1313-1317.

[44] J. Pivrinta, E. Rahtu, and J. Heikkil, "Volume Local Phase Quantization for Blur-Insensitive Dynamic Texture Classification", *Proceedings of the 17th Scandinavian Conference on Image Analysis (SCIA)*, Ystad, 2011, pp. 360-369.

[45] S.H. Lee, H. Kim, Y.M Ro, and K.N. Plataniotis, "Using color texture sparsity for facial expression recognition", *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, 2013, pp. 1-6.

[46] Y. Wang and M. Meng, "3D Facial Expression Recognition Based on Curvature Local Binary Patterns", *5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2, Hangzhou, 2013, pp. 123-126.

[47] X. Qi, R. Xiao, C. Li, Y. Qiao, J. Guo, and Xiaoou Tang, "Pairwise Rotation Invariant Co-Occurrence Local Binary Pattern", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36(11), 2014, pp. 2199-2213.

[48] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast Keypoint Recognition Using Random Ferns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(3), 2010, pp. 448-461.

[49] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a Local Binary Descriptor Very Fast", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34(7), 2012, pp. 1281-1298.

[50] J. Zhang, J. Liang, and H. Zhao, "Local Energy Pattern for Texture Classification Using Self-Adaptive Quantization Thresholds", *IEEE Transactions on Image Processing*, vol. 22(1), 2012, pp. 31-42.

[51] Z. Lei, T. Ahonen,M. Pietikainen, and S.Z. Li, "Local frequency descriptor for low-resolution face recognition", *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG2011)*, Santa Barbara, 2011, pp. 161-166.

[52] M. Abdulrahman, T.R. Gwadabe, F.J. Abdu, and A. Eleyan, "Gabor wavelet transform based facial expression recognition using PCA and LBP", *EEE 22nd Signal Processing and Communications Applications Conference*, Trabzon, 2014, pp. 2265-2268.

[53] T. Ahonen, M. Pietikainen, A. Hadid, and T. Maenpaa, "Face recognition based on the appearance of local regions", *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, vol. 3, Cambridge, 2004, pp. 153-156.

[54] D. Lundqvist, A. Flykt, and A. hman, "The Karolinska Directed Emotional Faces - KDEF", *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, Solna, 1998.

[55] C. Chang and C. Lin, "LIBSVM: A library for support vector machines", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2(3), 2011, pp. 27:1-27:27.

[56] Z.-H. Tan, N.B. Thomsen and X. Duan, "Designing and Implementing an Interactive Social Robot from Off-the-shelf Components", *3rd IFToMM Symposium on Mechanism Design for Robotics (MEDER 2015)*, Aalborg, 2015.

Detecting Humans in Videos by Combining Heterogeneous Detectors

Karla Brkić, Tomislav Hrkać and Zoran Kalafatić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10 000 Zagreb, Croatia
{karla.brkic, tomislav.hrkac, zoran.kalafatic}@fer.hr

***Abstract*—Human detection is an actively researched problem in computer vision, and it has driven the development of a number of detectors, both problem-specific and general purpose. We investigate the performance of human detection in video sequences when three commonly used detectors are combined: background subtraction, boosted cascades of Haar-like features and histograms of oriented gradients. We qualitatively analyze the performance of each of the detectors and quantitatively estimate the performance gain obtained by combining all three. We employ sequences from the publicly available TUD dataset. HOG and background subtraction perform reasonably well when used individually, while boosted cascades of Haar-like features performs inadequately in terms of localization accuracy and false positive rate. Our results indicate that overall detection precision and recall can be improved when combining HOG with background subtraction. Adding the boosted cascades of Haar-like features does not contribute to the overall performance. We conclude that combining HOG and background subtraction is an elegant solution for utilizing the advantages of video data to improve detection performance while maintaining a low processing cost.**

I. Introduction

Reliable human detection in image and video data is essential for the advancement of many actively researched problems in artificial intelligence. For example, a good person detector is at the core of pedestrian detection and collision prevention in smart cars, visual modules of robots for help in the home, intelligent surveillance systems that take into account privacy and similar systems. Detecting humans helps in building better, more advanced human-computer interfaces, one recent and well known example being the Kinect sensor.

In this work, we consider building a system for human detection by combining three heterogeneous detectors: (i) boosted cascades of Haar-like features [1], (ii) histograms of oriented gradients (HOG) [2] combined with a support vector machine classifier, and (iii) background subtraction [3]. Our goal is to benchmark how well a combination of these detectors performs on video data, and whether the performance of image-based detectors, namely histograms of oriented gradients and boosted cascades of Haar-like features, can be improved by introducing video-specific cues obtained using background subtraction. Recent research [4], [5] identifies false positive detections as the main problem of state-of-the-art human detectors. Experiments suggest that low miss rates can be obtained at the cost of several false positive detections per image. We show that false positives can be reduced by assuming motion of people in video, and using the motion information obtained by background subtraction as a support for detections reported by the HOG detector and boosted cascades of Haar-like features. We take special care to support detection in cases when there is no motion, due to a human stopping in the scene. To that end we use a history of background subtraction outputs to infer the probability that such a situation had occurred. An outline of our system is shown in Figure 1.

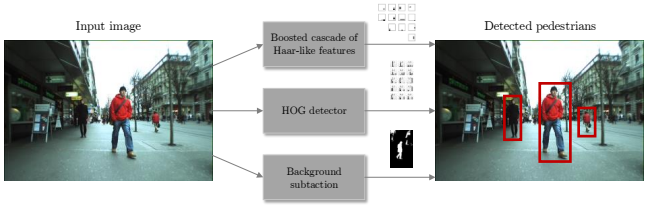


Fig. 1: An outline of the proposed system.

II. Related work

The past decade has been marked by a high interest in pedestrian detection and object detectors in general. In their seminal work [1] Viola and Jones proposed a detector based on a boosted cascade of Haar-like features and applied it to the problem of face detection, obtaining very high performance rates. The work was later extended to the problem of pedestrian detection [6].

Dalal and Triggs [2] proposed a descriptor based on histograms of oriented gradients (the HOG descriptor), that has ever since been used as a standard solution for detecting humans, and has inspired a number of derived and extended approaches.

One of the most promising extensions of HOG is the Deformable Part Model (DPM), proposed by Felzenszwalb et al. [7]. The basic idea of DPM is to model the object of interest with a number of parts. Each part is modeled with histograms of oriented gradients, and can be deformed to a degree. Recent research results [8], [9] suggest that the strength of DPM lies in allowing object deformations rather than in object division into parts, as similar performance can be achieved by using deformations of a single template.

Dollar et al. [10] proposed integral channel features that build upon the ideas of HOG and boosted Haar cascades. The idea is to compute a function of the input image that preserves image layout; the output of such a function is called a channel.

Possible channels include specific gradient orientations, gradient magnitudes, components in RGB, HSV, or LUV color spaces etc. The detector is trained using boosting, similar to the framework of Viola and Jones [1]. The used features are simple sums within rectangles, easily computed using integral images of the considered channels. An extended version of the integral channel features, called the Katamari-v1 [5], has been shown to be the best-performing detector on the Caltech-USA dataset among more than 40 state-of-the-art detectors.

Recently, deep learning has gained considerable popularity in the computer vision community. Several authors [11], [12], [13] propose methods for detecting humans based on deep learning. Although acceptable performance can be obtained by using a convolutional neural network trained on raw pixel values [11], there is a measurable performance improvement when deep architectures are trained using manually crafted features (e.g. edge or color features) [12], [13].

Background subtraction is a method for finding moving objects in videos that can be used as a secondary cue for human detection. For example, Harville [14] proposed combining background subtraction with dedicated higher-level modules for person detection and tracking in order to achieve better overall performance. There are many different background subtraction algorithms that mainly differ in the manner in which the background model is constructed [15], [16], [17]. A comparative overview of several background subtraction algorithms for detecting pedestrians and vehicles in urban scenes can be found in the work of Cheung et al. [15]. Considered algorithms include simple frame differencing, median filtering, linear predictive filtering, non-parametric estimate of the pixel density function, approximated median filter, Kalman filter and mixtures of Gaussians. Experimental evidence suggests that mixtures of Gaussians [18], [19], [20] produce the best results. Similar results can be found in the work of Herrero and Bescós [16], where it is shown that top-performing methods are chi-square modelling, mixtures of Gaussians and simple median filtering.

More details on the current human detection methods can be found in several comprehensive review papers [21], [4], [5]. Dollar et al. [21] introduced the Caltech Pedestrian dataset, benchmarked seven different detectors and proposed to measure false positives per image rather than false positives per detection window in order to obtain a fair estimate of method performance. They followed up this work in [4], by evaluating sixteen state-of-the-art human detectors on six public datasets. Benenson et al. [5] performed an in-depth evaluation of the state of the art in pedestrian detection in the past ten years.

III. The considered detectors

A. Boosted cascade of Haar-like features

The boosted cascade of Haar-like features, proposed by Viola and Jones [1], is a very fast generic object detector. There are several key concepts used in the detector:

1) Haar-like features: A Haar-like feature is a simple rectangular feature commonly drawn as a rectangle consisting of black and white parts, as illustrated in Figure 2. The value of the feature is calculated by summing the pixel intensities under white parts and subtracting them from the sum of pixel intensities under black parts. The computation of the value of a simple 2×2 feature on a 4×4 image is illustrated in Figure 3.

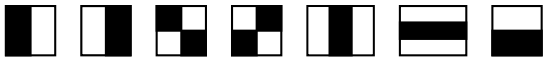


Fig. 2: Examples of Haar-like features.

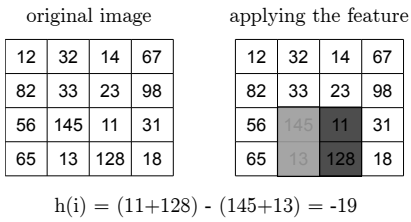


Fig. 3: Computing the value of a Haar-like feature.

Haar-like features are designed to be sensitive to particular image structures, such as edges or angles. In itself, a single Haar-like feature paired with a threshold comprises a very simple classifier. For instance, one might design a classifier by pairing the feature in Figure 3 with a threshold of -5, and designating each image for which the feature takes on a value less than or equal to -5 as belonging to one class, and otherwise belonging to another class.

2) Integral image: An integral image is calculated from the original image in the following way: the value of the pixel (x, y) of the integral image is equal to the sum of the values of all the pixels in the rectangle spanned between points $(0, 0)$ and (x, y) of the original image. Using an integral image, one can compute the value of a Haar-like feature very efficiently, using only four array references.

3) AdaBoost: AdaBoost is a machine learning algorithm that enables combining weak classifiers (i.e. classifiers that are only slightly better than random guessing) into a strong classifier.

In order to build the detector, one needs a training set consisting of object and background images. Based on the training set, the learning process selects a number of very simple classifiers based on Haar-like features that best discriminate object images from background images. The selection is made from a pool of possible Haar-like features, constructed by considering all possible locations and sizes of each feature type. The selected classifiers based on Haar-like features are arranged into a cascade. The cascade is divided into several stages, and each stage comprises of a number of features paired with thresholds. The main idea behind using a cascade structure is the fast rejection of non-objects. Easy non-objects are discarded early on, and harder examples go through to deeper stages in the cascade that are trained on harder examples.

After the training process is complete, the detector is usually applied to a query image in a sliding window fashion, so that sub-windows of the image are scanned for the presence of the object of interest. The detector is very fast and can be used in real time on modern computers.

B. The HOG detector

The HOG detector is a generic object detector comprising of two components: the HOG descriptors and a support vector machine (SVM) classifier. The main idea behind the detector is to represent the image using a grid of histograms of oriented gradients, concatenate the histograms into a feature vector, and then classify the feature vector into object or non-object using a support vector machine classifier.

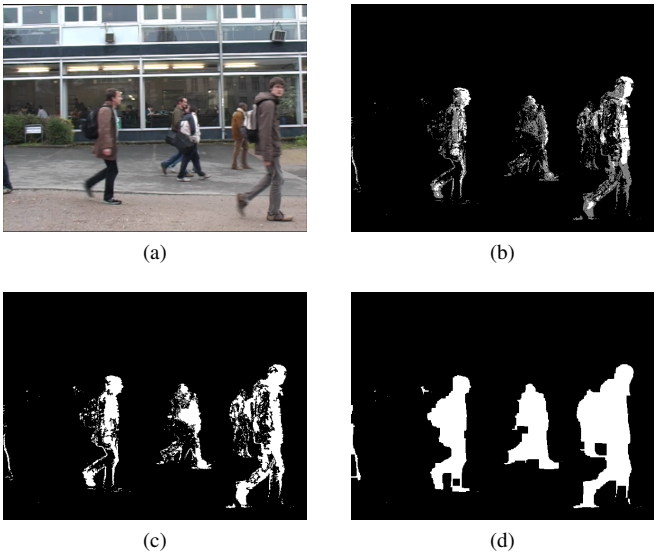
In order to build a feature vector of an image, the image is first divided into a regular grid of cells. The cells are usually rectangular (R-HOG), although Dalal and Triggs [2] also proposed using a circular grid geometry (C-HOG). For each cell, a histogram of gradients is built by calculating the gradient orientation and magnitude in each pixel. The histogram bins represent ranges of gradient orientations, and the weight of the vote of a pixel for a particular orientation is determined by the magnitude of the gradient in that pixel. To reduce the influence of illumination changes, neighboring cells are grouped into blocks, and the contrast in each block is normalized using the total histogram energy computed over the block. The resulting cell histogram probabilities in a block, concatenated into a single vector, are referred to as the HOG descriptor. The final HOG feature vector is obtained by concatenating HOG descriptors of all the blocks.

The detector is trained in a supervised manner, having a dataset of object and background images. Each instance from the training set is represented by its HOG feature vector. Dalal and Triggs [2] suggest using a linear SVM as a classifier. To detect objects in an image, a window is moved across the image and the region of interest spanned by the window is classified as either object or background using the trained SVM classifier.

C. Background subtraction

Background subtraction is a method for finding moving objects in video data. The basic idea of background subtraction is that by observing a sufficient number of frames one can learn a background model. Foreground objects, i.e. objects that are moving in the scene, can then be found by subtracting the learned background model from the query frame.

In this paper we consider the background subtraction algorithm based on mixtures of Gaussians proposed by Zivkovic [20]. The distributions of values of each pixel over time are represented with a mixture of weighted Gaussian distributions. The weights of the Gaussians are updated based on each new frame, and older frames lose influence over time. It is assumed that the distributions of all pixel values will have the highest weights assigned to the Gaussian mixture components corresponding to background, given a sufficiently wide time window. If the time window is wide enough, it can be expected that the number of frames in the window in which the background is being occluded is smaller than the number of frames in which the background is being visible. The Gaussian mixture components of each pixel are ordered according to their weights and thresholded to obtain background and foreground models. The number of Gaussian mixture components is not fixed, but computed dynamically for each pixel based on the available data.



(a) (b) (c) (d)

Fig. 4: Background subtraction in the proposed system: (a) the original frame, (b) the output of the background subtraction algorithm, (c) the extracted and filled contours, (d) the output after morphological closing

Example output of the described background subtraction algorithm is shown in Figure 4 (a) and (b).

IV. Combining the detectors

Our goal is to combine the outputs of the HOG detector, background subtraction and boosted cascades of Haar-like features in order to reduce false positives. We start by post-processing the output of background subtraction. As illustrated in Figure 4(b), the output foreground pixels of the used background subtraction algorithm often do not cover all pedestrian pixels, i.e. some pedestrian pixels are marked as background. As we would like to detect all pedestrian pixels as foreground, we first extract and fill contours of foreground pixels (Figure 4(c)). As some small and unconnected regions still remain after this step, we then apply the morphological operation of closing (dilation followed by erosion), as illustrated in Figure 4(d).

For each HOG detection, we calculate the ratio of foreground and background pixels obtained by background subtraction within the HOG detection rectangle. If the calculated ratio is greater than the predefined threshold called overlap threshold, the detection rectangle is retained, otherwise it is discarded.

The output of boosted cascades of Haar-like features is used as additional evidence for the correctness of detections. We apply several boosted cascades that detect frontal faces, profile faces, lower and upper body and full body. However, in the video sequences that we used the boosted cascades of Haar-like features did not produce satisfactory results and therefore the importance of their output was downweighted. Further details are given in the experimental section.

In combining the output of HOG detector with the output of background subtraction, we need to take into account that some pedestrians in the video can be motionless for some time. Therefore we extend the algorithm logic to take into

Fig. 5: Example frames from the TUD-campus (left) and TUD-crossing (right) sequences

account detections from previous frames. The idea is to retain the detections with weak support from background subtraction if there is evidence of pedestrian presence in recent frames. In our experiments, we retained the HOG detections that do not contain enough foreground pixels if they overlap at least 75% with the detection in the previous frame.

V. Experimental evaluation

We evaluate the performance of the proposed system both qualitatively and quantitatively, in terms of visual inspection of detector output and the achieved precision and recall. For the evaluation, we use two video sequences from the TUD dataset [22]: TUD-Campus and TUD-Crossing. This dataset is chosen over more commonly used ones (e.g. Caltech [21]) due to the fact that it is filmed with a static camera, which is a requirement for applying background subtraction. We use the standard Pascal overlap criterion (intersection over union) [23], counting the detection as correct if its overlap with the annotation is at least 0.5.

The TUD-Campus sequence consists of 71 frames, while the TUD-Crossing sequence consists of 200 frames. Both sequences contain multiple pedestrians. Figure 5 shows example frames from the TUD-Campus and TUD-Crossing sequences.

Qualitative performance of our system is illustrated in Figure 6. It can be seen that the individual outputs of the HOG detector and background subtraction are significantly better than the output of boosted cascades of Haar-like features. We employ the pre-trained HOG detector from the OpenCV library [24], as well as pre-trained boosted cascades. As mentioned previously, we apply several cascades simultaneously, specialized for detecting different pedestrian parts (frontal faces, profile faces, lower and upper body and full body). Although the boosted cascades produce some responses around pedestrians, localization is poor, not all pedestrians are detected and there are a number of false positive detections. The HOG detector detects all pedestrians with some localization errors and two false positives. Background subtraction produces strong responses on the pedestrians and some minor noise, visible at the windows of the building in the scene. Combining the responses of the HOG detector and background subtraction (as explained in the previous section) results in reliable pedestrian detection with no false positives.

Quantitative results of our experiments for combining HOG with background subtraction can be seen in Figures 7 and 8. In each figure, individual curves represent the results of a series of experiments. To obtain each curve we vary the SVM classifier

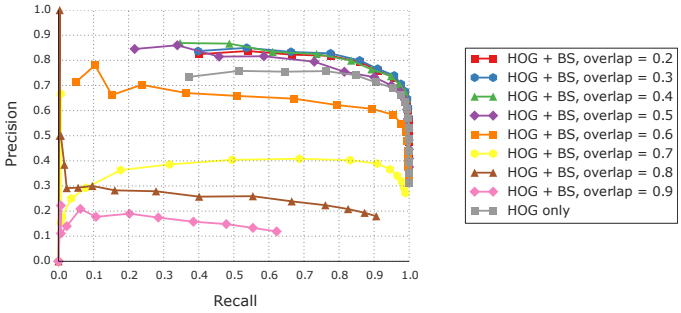


Fig. 7: Precision-recall curves for the TUD-Campus sequence, histograms of oriented gradients (HOG) combined with background subtraction (BS)

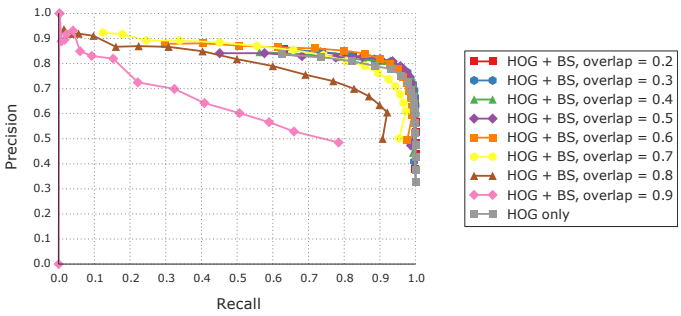


Fig. 8: Precision-recall curves for the TUD-Crossing sequence, histograms of oriented gradients (HOG) combined with background subtraction (BS)

threshold of the HOG detector while fixing the overlap threshold between HOG detections and foreground pixels identified by background subtraction. The SVM classifier threshold takes values from the interval $[-1.6, 1.6]$ with a step of 0.2, while the overlap threshold takes values from the interval $[0.2, 0.9]$ with a step of 0.1.

As can be seen from Figures 7 and 8, applying our method with an adequate overlap threshold and SVM threshold improves performance over HOG only. For a fixed recall we achieve improvements in precision of 5% to 10% in TUD-Campus sequence for overlap thresholds ranging from 0.2 to 0.4, and around 2% to 5% in TUD-Crossing sequence for overlap thresholds ranging from 0.2 to 0.6. Greater values of overlap threshold significantly decrease performance of the detector. This decrease is expected because HOG returns a rectangle around each pedestrian that contains not only foreground but also a considerable amount of background pixels. By increasing the overlap threshold we effectively require more and more foreground pixels in the detection rectangle, which is not a realistic expectation given the average pedestrian shape and motion.

Figures 9 and 10 show detection results obtained by combining all three individual detectors (HOG, background subtraction and boosted cascades of Haar-like features). As already mentioned, this combined detector retains HOG detections that have either enough overlap with the foreground area, or at least some overlap with the detections obtained by the boosted cascades of Haar-like features. Detections that do

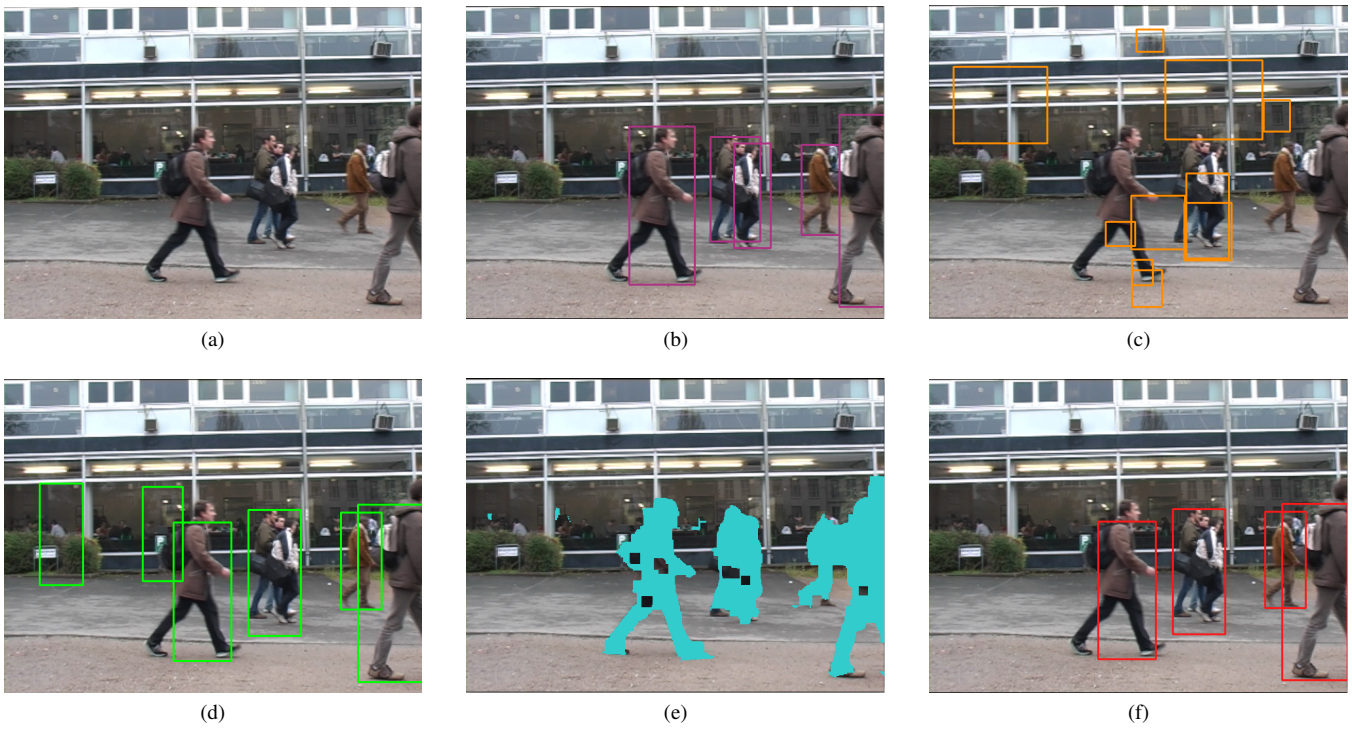


Fig. 6: Performance of our system on an example frame from the TUD-Campus sequence: (a) the original frame, (b) pedestrian annotations, (c) the output of the boosted cascades of Haar-like features, (d) the output of the HOG detector, (e) the output of background subtraction, (f) the output of the combined detector (HOG + background subtraction)

not satisfy neither of the two requirements are rejected as false positives.

The individual precision-recall curves in Figures 9 and 10 were obtained by varying the SVM classifier threshold of the HOG detector while keeping the following two thresholds fixed: (i) the overlap threshold between the HOG detections and foreground area found by background subtraction, and (ii) the overlap threshold between the HOG detections and boosted Haar-like cascade classifier detections. Different curves were obtained by repeating the procedure for different values of the overlap threshold between HOG detection and foreground. However, in this set of experiments we consider only overlap thresholds in the range between 0.2 and 0.6 (with the step of 0.1), since the previous set of experiments showed that larger values of this threshold only degrade the results. The overlap threshold between the HOG detections and boosted Haar-like cascade classifier detections was fixed to a constant small value (0.1) during all of the experiments, as preliminary qualitative evaluation revealed that majority of the detections of boosted cascades of Haar-like features were poorly localized and quite small compared to full HOG detections. The small size of the detections is to be expected, given that we used cascades that detect pedestrian parts rather than full pedestrians. For example, face detections are significantly smaller than full body HOG detections. Hence, we estimated that requiring a greater overlap would effectively trigger no confirmed responses from the boosted cascades.

As can be seen from Figures 9 and 10, adding boosted cascades of Haar-like features to the combined classifier does not result in significant improvement in the classification

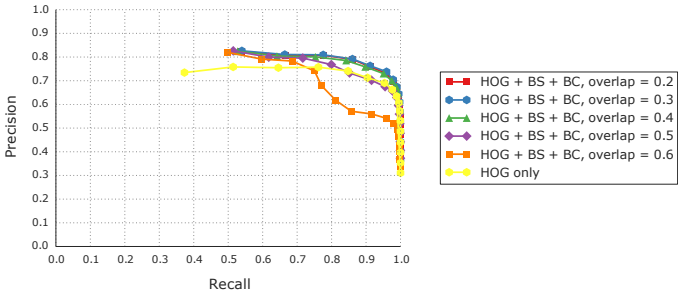


Fig. 9: Precision-recall curves for the TUD-Campus sequence, histograms of oriented gradients (HOG) combined with background subtraction (BS) and boosted cascades (BC)

performance compared to combining HOG and background subtraction only. We attribute this lack of improvement to disappointingly poor detection and localization rate of the used boosted cascades of Haar-like features (see Figure 6(c)).

VI. Conclusion

We proposed a method for detecting humans in videos by combining outputs of three heterogeneous detectors: boosted cascades of Haar-like features, histograms of oriented gradients (HOG) combined with a support vector machine classifier, and background subtraction. The proposed method was succesfully employed to improve performance of the baseline (HOG) detector by reducing false positive rate.

There are several possible improvements of the described

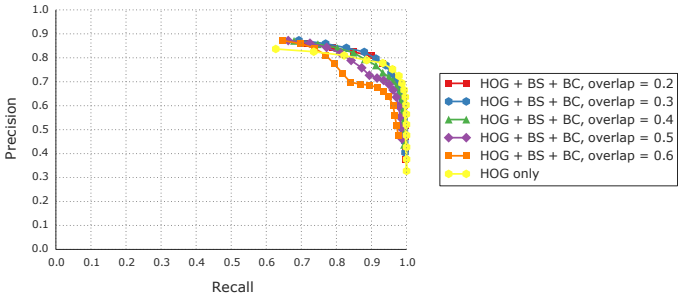


Fig. 10: Precision-recall curves for the TUD-Crossing sequence, histograms of oriented gradients (HOG) combined with background subtraction (BS) and boosted cascades (BC)

idea that will be the basis of our future work. First, better pedestrian detectors can be used instead of HOG + SVM, for example recently proposed pedestrian detectors such as Deformable Part Models combined with latent SVM, or the detector based on Integral Channel Features and boosted ensemble of depth-2 trees. Second, better and more recent methods of background subtraction can be used, less sensitive to stationary foreground objects. Finally, more sophisticated methods of fusion of detector outputs can be considered, such as combining the corresponding feature vectors before classification instead of voting after classification.

ACKNOWLEDGEMENTS

This work has been supported by the Croatian Science Foundation, within the project "De-identification Methods for Soft and Non-Biometric Identifiers" (DeMSI, UIP-11-2013-1544), and by the COST Action IC1206 "De-Identification for Privacy Protection in Multimedia Content". This support is gratefully acknowledged.

The authors would like to thank UNIZG-FER students Martin Blažević and Mateo Srebot who developed parts of the framework used for experiments in this paper.

REFERENCES

[1] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886–893.

[3] M. Piccardi, "Background subtraction techniques: a review," in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 4, 2004, pp. 3099–3104.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2011.155

[5] R. Benenson, M. Omran, J. Hosang, , and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *ECCV, CVRSUAD workshop*, 2014.

[6] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vision*, vol. 63, no. 2, pp. 153–161, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-6644-8

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.167

[8] S. K. Divvala, A. A. Efros, and M. Hebert, "How important are "deformable parts" in the deformable parts model?" in *Proceedings of the 12th International Conference on Computer Vision - Volume Part III*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 31–40.

[9] B. Hariharan, C. L. Zitnick, and P. Dollár, "Detecting objects using deformation dictionaries," in *CVPR*. Computer Vision and Pattern Recognition, June 2014. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=220570

[10] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2009, doi:10.5244/C.23.91.

[11] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning." in *Proc. CVPR*, 2013, pp. 3626–3633.

[12] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 3258–3265.

[13] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 3222–3229.

[14] M. Harville, "A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models," in *ECCV (3)*, 2002, pp. 543–560.

[15] S.-C. S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," *Visual Communications and Image Processing 2004*, vol. 5308, no. 1, pp. 881–892, 2004.

[16] S. Herrero and J. Bescós, "Background subtraction techniques: Systematic evaluation and comparative analysis," in *Proceedings of the 11th International Conference on Advanced Concepts for Intelligent Vision Systems*, ser. ACIVS. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 33–42.

[17] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1937–1944.

[18] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.

[19] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, 1999, pp. 2246–2252.

[20] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *ICPR (2)*, 2004, pp. 28–31.

[21] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 304–311.

[22] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008. [Online]. Available: http://www.mis.tu-darmstadt.de/node/382

[23] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1007/s11263-009-0275-4

[24] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

Identification of authors of documents based on offline signature recognition

Tonćo Marušić*, Željko Marušić, Željko Šeremet**
*Faculty of Science and Education, Mostar, BiH
**Faculty of Mechanical Engineering and Computing, Mostar, BiH
tonco.marusic@gmail.com

Abstract - Handwritten signature is used in various applications on daily basis. Whether one signs a contract, work documents, petition, or wants to approve a check payment, one will use personal signature to do all those things. In this paper we use this daily based biometric characteristic for identification and classification of students' papers and various exam documents used at University of Mostar. In this paper we used OpenCV library as an image processing tool for feature extraction. As regards to classification method, we used Support Vector Machine.

I. INTRODUCTION

In the teaching process, teaching staff (professors, assistants) must organize and manage a lot of documentation. Exam lists, quizzes, written assignments, tests, attendance lists are just some of them. When we bear in mind that these materials must be archived and stored for some time, it is clear that it is necessary to digitize. This heap of documentation can then be used for various analyses, as well as for formation of a knowledge base. In the process of digitizing, we can use different types of scanners or even cameras from smartphones. Once the documents are scanned, they should be grouped, sorted and stored in individual folders and directories. This entire process can be automated.

Generally, signatures and student ID numbers are used for identification of these documents. Handwritten signatures or handwritten ID numbers represent non-invasive and inexpensive identification methods.

Recognition of individual digits of a number is a problem that is relatively well resolved. A much difficult problem is to recognize digits when some of the handwritten digits are connected. A simple solution is keeping digits separated during enrollment, and to write them to bounding boxes. This is illustrated in Figure 1 that represents proposed header of documents used in various student exams.

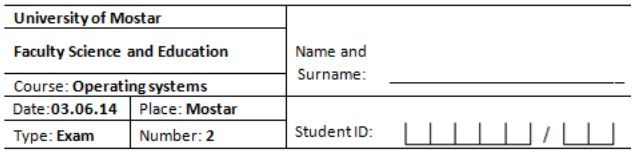


Figure 1. Proposed header of test documents with place for signature and student ID.

During this enrollment, students often type the wrong digit by mistake or they simply cross the bounding box reserved for the individual digits, which will inevitably result in the wrong classification of digits, and thus the misidentification of the author of the document. It would be desirable that the system is resistant to these situations, and potential candidate is to use signatures as a way of verification and/or identification of the author. In this paper, we will check what are the possibilities for identification of various scanned exam documents based on students' signatures. After recognition of the authors, the documents can be automatically classified in the individual directories and linked to document management system (DMS).

As regards to literature, there is much more to deal with the offline verification of signatures in relation to the offline identification of signatures. Terms are similar in that they rely on the database which contains records about user's biometric characteristic features. Then we compare given biometric characteristic features with those stored in the database. The main difference is that verification answers the question: "Is the user really who he or she says to be?'' On other hand, biometric identification answers the question: "Who is the person?"As regards to classification point of view, verification is often referred to as „one-to-one" comparison. Biometric identification is known as „one-to-many" comparison [1]. Since they are similar, papers that deal with verification can be used in our analysis.

Various off-line techniques have been proposed in literature for the signature verification problem. These techniques include: Hidden Markov Models (HMM), Neural classifiers, Wavelets, Dynamic Time Warping [2], Graph matching, Support Vector Machines and so on. Generally, we can say that SVM outperform HMM [5] and Neural networks [3], therefore, we chose SVM as a classifier.

Offline signature verification and identification is a pattern recognition problem and a typical pattern recognition system has the following steps [4]:

- Data Acquisition
- Preprocessing
- Feature Extraction
- Classification
- Performance Evaluation

II. ACQUISITION OF SIGNATURES

There are inevitable variations in the signature patterns produced by the same person (intrapersonal variability) [6]. Some factors that affect our signatures are: a person's physical and psychological state, body position, writing material, purpose of signing (formal vs. informal). In order to build a reliable system, we must obtain a database of signatures extensive enough to capture possible individual variations. We can address this issue by modeling acquisition system that will include different size rectangle areas, pens with different stroke width, different color of imprint, different angle and so on [7].

There are two types of acquisitions in our approach. First, we collected the signatures from 30 students. Each student wrote 22 signatures on a printed sheet A4 paper with the same size rectangles. They used different pens with different color (blue, black, red) and with different stroke width. During the acquisition, we emphasized to students that during the signing the boundary lines should not be touched. With the development of technology for the purpose of acquisition, we can use cameras and mobile phones with additional image processing. After that, the whole A4 document was photographed with a camera from mobile phone Sony Xperia T with a resolution of 4128 * 3096 (13Mpx). For purpose of photography we built improvised tripod with led lamps that illuminate the document. The signatures were manually extracted using an external tool for image processing Gimp. These signatures were then used for our database.

Another type of acquisition included collecting various tests, exam papers of students to whom we performed our classification. In total, we collected 145 documents. We photographed them in an analogous manner with the same resolution as above. However, in this case we did not extract the signatures by manual cropping, but we used computer vision algorithms for doing this automatically. We used mobile phone camera just to capture the documents and all processing was done on PC.

III. PREPROCESSING

Preprocessing is a set of subsequent operations applied for the improvement of quality of signature image and locating ROI (Region of Interest) of signature.

The various sub-processes which can be considered as image preprocessing are: Cropping, Binarization, Noise removal, Morphological Operations, Image normalization, Skeletonization, etc.

A. Binarization

Binarization represents an initial step in most systems for the analysis of document images. It refers to conversion of the gray-scale image into a binary image. It is crucial, because it is the basis for successful segmentation and recognition of characters, words or zones. Generally, we distinguish global and local methods. Global methods use one threshold value to classify the pixels in the image object or to classify the background. On other hand, local schemes can use multiple values based on the information in the local area. Global techniques are not suitable for binarization of documents that are delivered via camera, because such documents suffer from non-unformed lighting, low contrast, shadows [8]. We used local adaptive technique instead.

B. Image denoising

After image acquisition and adaptive threshold, noises like salt and pepper, Gaussian noise were generated during painting. In the documents it is possible to use a series of filters as the median, Gaussian and Wiener filter. We used the median filter which reduces noise with a very small degree of image degradation an also preserves sharp edges unlike the mean filter and considerably less blurs the image in relation to the Gaussian filter.

C. Data Area Cropping and rotation

At the top of each test sheet there is a header containing basic information about the exam. This information consists of the name of the exam, the exam type, number of exam period, the date of the exam, the student's signature, index number, etc. If we are dealing with documents, we must first segment the header, and then in the header we locate the region of interest of signature. For this purpose, the algorithm used is based on the topological structural analysis of the image that has been implemented in OpenCV [9]. This is illustrated in Figure 2 and Figure 3.

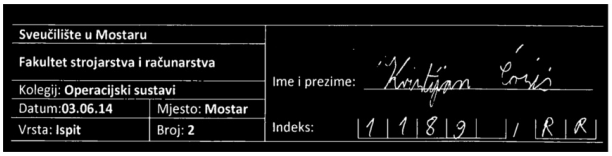


Figure 2.Header document after median filter in negative

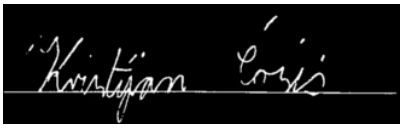


Figure 3. Signature with baseline

In Figure 3 wecan see that base line is extracted with a signature. This line is of great importance to us,we can very easily determine and correct the angle of image acquisition to zero with rotation and using contour segmentation. In this way we improve accuracy of feature extraction.

D. Morphological operations

After binarization, cropping and rotation, we continue processing resulting images in order to improve the quality of text regions and preserve connectivity within certain characters, and meet some possible gaps, holes within the character. For this purpose, we used mathematical morphology operations such as: dilation and image erosion. These operations are usually applied to binary images. Furthermore, in this process we removed the base line with dilation and bitwise operation. After cropping, we found bounding box of signature like in Figure 4

Figure.4. Bounding box of a signature

E. Normalization width

Width of the photographed document is 3096 pxs. Region which the signature occupies is approximately one-third of the header. Once we determine the signature region, each signature will be scaled using bicubic interpolation to constant width of 384 pixels keeping fixed aspect ratio.

F. Skeletonization

When signing the test sheet, a student does not always use the same pen. Thick imprint of pens can greatly affect the value of extracted features. Precise determination of values for certain features in this paper can only be obtained on the skeleton image. Furthermore, this speeds up the execution of algorithms in gathering values for some features. This is illustrated in Figure 5.

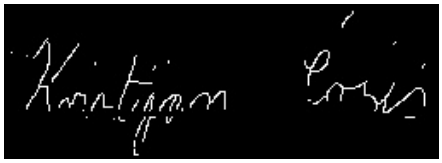


Figure 5. Skeleton signature image

After preprocessing, we can move to determine values of global, grid and SIFT features.

IV. EXTRACTING FEATURES

Features are a key process in achieving high accuracy in identifying signatures. The ideal separation techniques use minimal sets of features that are used to maximize the difference among different people's signatures while minimizing the difference of the same person's signature [10].

Global structural and statistic features of signatures are extracted from the entire image signatures. Grid features are extracted from the grid formed by the parts of the signature [7]. Also we used SIFT feature descriptor in our work.

A. Global features

In literature, there are numerous possible global features [7][10][11][12][13][14][15][16]. After the analysis and testing, we chose 16 global features that proved their efficiency.

1. Aspect ratio - represents the proportionality rate of the skeleton signature image. This is calculated by dividing the height with the width of the signature;

2. Signature occupancy represents the number of pixels that belong to the signature;

3. Baseline shift is a difference between the vertical center of gravity of the left and right side skeleton signature image. At first, counting divided the image signature vertically in two halves and calculated the center of gravity for each half of the CL and CR. Then baseline shift (baseline shift) was defined as BSL = CL – CR;

4. Global Slant Angle - It is the overall direction of line strokes in the skeleton signature. The original signature was rotated from -45degrees to +45 degrees in steps of 5 degrees. For each rotation, the original signature was first pre-processed followed by counting the number of vertical 3-pixels connections from the rotated skeleton image. The global slant angle is the angle which has the maximal number of vertical 3-pixels connections. This is illustrated in Figure 6.

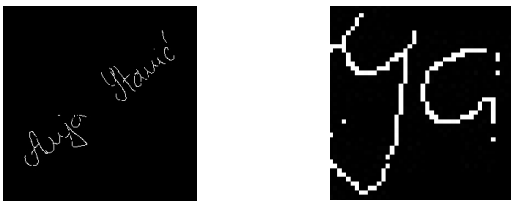


Figure 6. Rotated and skeletonized images for procedure of finding Global slant

5. Number of Edge Points - According to [15] an edge point is a black pixel which has only one 8-neighbour. This is illustrated in Figure 7.

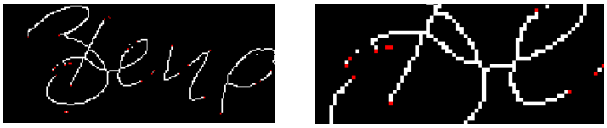


Figure 7. Number of edge points

6. Number of Cross-Points and Spatial Symbols [14]

Every person uses some spatial symbols in his signature, such as: some 'x' marks (cross marks), star

marks or other symbols. The total number of spatial symbols of a person's signature is unique. If we find that one pixel has more than two neighbors, then those pixels will form a spatial symbol. Such types of pixels are shown in Figure 8.

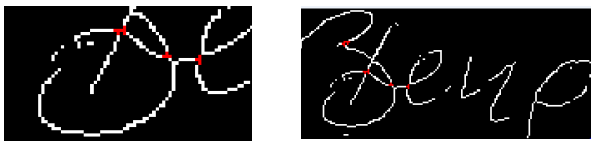


Figure 8. Number of Cross-Points and Spatial Symbols

7. and 8.Horizontal and vertical center of gravity

Center of gravity is pair (X,Y) represented with formula (1)

$$X = \frac{1}{N}(\sum_{i=1}^{n} x_i) \qquad Y = \frac{1}{N}(\sum_{i=1}^{n} y_i) \qquad (1)$$

where x_i and y_i represent numbers of white pixels in columns and rows respectively on binarized image. Part X represents horizontal centre of gravity, and part Y represents vertical centre of gravity.

9. and 10. Length name, Length surname

After analysis of students' papers, we can say that students identify themselves with full first and last name. Generally, there is a considerably larger gap between first and last name in relation to the distance between the individual characters that are often completely connected. This property can be utilized to segment first name and last name, and then retrieve the width for each part. For this purpose, we iterated morphological dilation operation until we got only two segmented parts. During the acquisition we found no instance where all pixels of a signature were completely connected. However in case that there is only one connected contour left during this process, both features would be set to value 0. In case that person has multiple names the largest gap between names is used for division.

11. and 12. Maximal horizontal and vertical histogram

The horizontal histogram is calculated for each row. Row with the highest value is taken as the maximum. Similarly, we calculate the vertical histogram for each column. Column with highest value is taken as maximal vertical histogram.

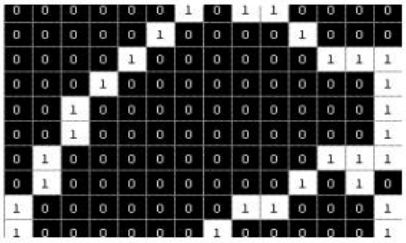
13. The length and ratio of Adjacency Columns [14]

In this stage, we firstly computed the length of the adjacency columns from top and from bottom of the Sample signature image and stored it in a 1D array L_T and L_B. Then we computed the sum of all elements of L_T and all elements of L_B. Overall procedure for this feature extraction is shown in the following Figure6. Upper half of Figure 6 shows a portion of a signature image. Lower half of Figure 6 shows the length of the corresponding adjacency columns from the top and from the bottom.

Now Adjacency Ratio can be calculated (2) as follows:

$$Adjacency\ Ratio = \frac{Sum\ (L_T)}{Sum\ (L_B)} * Si\ gnatureOccup. \qquad (2)$$

Adjacency ratio will be more or less the same for every signature of a person. This is shown in Figure 9.



Length From Top (L_T):													
8	6	4	3	2	1	0	9	0	0	1	2	2	2
Length from Bottom (L_B):													
0	2	4	6	7	8	9	0	1	1	2	3	2	0

Figure 9. Illustration of finding the length and ratio of Adjacency Columns [14]

14. "Heaviness of signature"

Calculation of this feature is based on previous feature. We calculated the absolute difference between the top and the bottom adjacency columns (Figure 10).

Figure10. Illustration of "heaviness" procedure

Then we summed up the difference. After that, we calculated the ratio of difference between sum and background area (black pixels). We used formula (3) based on previous figure.

$$ALR = \frac{sum(height - (L_T + L_U))}{number\ of\ black\ px} \qquad (3)$$

15. Ratio of first vertical pixel to height of image.

In the bounding box image we calculated distance between top left pixel and the y coordinate of the signature first pixel. Then, this distance is divided with height of bounding box. In this way we can find out where the first vertical pixel in our picture is located in relation to height.

16. Number of Closed Loops

It represents the number of closed regions in a signature. We first find the closed region using [9] and then we subtracted with original image and counted contours that left out. In figure we can see 2 closed loops (Figure 11).

Figure 11. On the leftthere is a filled signature, on the right there is a subtracted imagewith two contours

There are a significant number of other global features in literature that we can use, such as: height, pure width, pure height, local slant angle. Since we normalized the width of signature area, some features like width and height were redundant. Instead of a pure width, we used width of two main parts of a signature (name and surname).

B. Grid Features

Each signature image is divided into 18 equal grid regions. From these grid regions, grid characteristics as local features are estimated, such as: width, height, area of signature pixels of each grid region, center of gravity, horizontal and vertical projections, etc. [13]. The global features can also be considered as local features for each grid region. In this implementation we only used number of pixels that belong to a signature for every particular region as a grid feature. We divided signature bounding box to 18 regions so in total we collected 18 features per signature.

C. SIFT Features

SIFT operator is an image local feature descriptor introduced by David G. Lowe in 2004[17]. It is one of the most popular local features. SIFT algorithm detects features in the scale space and confirms the location and scale of key points. Then, it sets the direction of gradient as the direction of the point. Thus the scale and direction invariance of the operator are realized. SIFT is a local feature, which is invariant to rotation, scaling and change of light and stable in a certain extent of changes in visual angle, affine transformation and noise. It ensures specificity and abundance, so it is applicable to fast and accurate matching among mass feature data. Its large quantity ensures that even a few objects can generate a number of SIFT features, high-speed satisfies the requirement of real-time, and extensibility makes it easy to combine with other feature vectors.

We used SIFT features using Bag of Words Model. This model method is extensively used in image classification and retrieval [18].
The main steps are as following:
(1) Detect key points though image division or random sampling etc.
(2) Extract the local features (SIFT) of the image and generate the descriptor.
(3) Cluster these feature related descriptor (usually via K-means) and generate visual vocabulary, in which each clustering center is a visual word.
(4) Summarize the frequency of each visual word in a histogram.
This method was applied to grayscale images. After extracting the features, descriptor was clustered using k - means algorithm where k = 48. In this way we obtained 48 features.

When we obtained a set of global, grid and SIFT features, we combined these feature sets into a feature vector of 82 elements.

V. DATA TRAINING

In this paper, offline identification of signatures was made with the support of vector machine (SVM). The SVM, a learning method introduced by Vapnik et al. [19][20], tries to find an optimal hyper plane for separating two classes. Therefore, the misclassification error of data both in the training set and the test set is minimized. Basically, SVM was used for linear separation of the two classes. When data are non linearly separable, a kernel function is used as polynomial function, radial basis function (RBF) or multi layer perception. The classification based on SVM involves training and testing stages. The training stage was used to find the optimal parameters. Hence, the two parameters should be determined: the kernel parameter and the regularization parameter. These two parameters were found experimentally depending on the dataset. The testing stage allowed evaluating the robustness of the classifier parameters [21]. For finding parameters and for evaluation of robustness of classifier we used signatures from acquisition. That is, we used 15 collected signatures per student for training (15*30=450) and 5 signatures for testing (5*30=150). Various kernels were tried, and the best results were achieved using a RBF kernel. Parameters C and γ for RBF kernel were determined through a grid search in Weka, machine learning software. Overall prediction accuracy for learning model was 98%.

VI. PERFORMANCE EVALUATION

In order to test proposed system with real exam documents, we collected 145 various exam documents of 30 students (from 5 students, we collected 4 documents and for rest it was 5 per student). After feature extraction and learning model on signatures from database, we started identification/ prediction of the authors using SVM as classifier. For measuring performance concerning of the identification of signatures and its affiliation in a particular set of students we used accuracy measurement. The testing was carried out individually for each set of features as well as combination of three selected feature sets.

Results of evaluation are shown in Table 1.

TABLE 1. RESULTS OF IDENTIFICATION OF AUTHORS ON EXAM DOCUMENTS BASED ON SIGNATURE

Method	Correct Prediction	Error Prediction	Average Accuracy (%)
Global Features	104	41	71.72%
Grid Features	90	55	62.07%
SIFT Features	84	61	57.93%
Global + Grid + SIFT features	129	16	88.97%

From the table 1, it can be concluded that global structural and statistic characteristics of signature were the best discriminators with 71.72% accuracy. Set of SIFT and Grid features had a significantly worse accuracy. However, when these individual feature set are joined into a single vector of features we gained significantly in performance and get the best accuracy of the proposed model with 88.97% correctly identified signatures and consequently authors of exam documents.

VII. CONCLUSION AND FUTURE WORK

In this paper we examined the problem of classification of exam documents based on the vector of global, local and SIFT features of students' signatures. The best performance was achieved by using global, local and SIFT features combination, where average accuracy was 88.97 %. Although the learning model showed excellent performance, when we tested the model on real documents it showed significantly lower results. We believe that problem was during acquisition where we emphasized to students that during the signing on rectangular areas, the boundary lines should not be touched. But on real documents they generally crossed the baseline with lower parts of signature.

However the results are promising, considering that our approach is relatively simple. Certainly, the results can be facilitated by using some other feature set (Histogram of oriented gradients, Dynamic Time Warping, Sparse Coding, Features using graph theory) and with some other classifier or with ensemble of classifiers.

In general, we can say that the proposed system will be of great help when used with the system for the recognition digits of the students' ID numbers.

REFERENCES

[1] Fotak T., Bača M., Koruga P., Handwritten signature identification using basic concepts of graph theory , WSEAS TRANSACTIONS on SIGNAL PROCESSING

[2] A. Piyush Shanker, A.N. Rajagopalan, "Off-line signature verification using DTWA. Pattern Recognition Letters 28 (2007) 1407–1414

[3] Enrique Frias-Martinez, Angel Sanchez and Jose Velez, Support Vector Machines versus Multi-Layer Perceptrons for Efficient Off-Line Signature Recognition, Engineering Applications of Artificial Intelligence, Volume 19, Issue 6, September 2006, Pages 693-704

[4] Duda R O and Hart P E. Pattern Classification (2e). Wiley India Private Limited. 2006

[5] E. J. R. Justino, F. Bortolozzi, and R. Sabourin, "A comparison of SVM and HMM classifiers in the off-line signature verification," Pattern Recognition Letters, vol. 26, no. 9, pp. 1377–1385, 2005

[6] E.J. Justino, F. Bortolozzi and R. Sabourin, "The Interpersonal and Intrapersonal Variability Influences of Off-Line Signature Verification using HMM", Proc. XV Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2002), Fortaleza (Brazil), October 2002.

[7] G. Azzopardi, K.P. Camilleri, „Offline Handwritten Signature Verification using Radial Basis Function Neural Network",IEEE International Conference on Electrical and Electronics (INDICON-2008), pp. 17-22, 2008.

[8] Gatos B., Pratikakis I., and Perantoni S., An Adaptive Binarization Technique for Low Quality Historical Documents, In proceeding of: Document Analysis Systems VI, 6th International Workshop, DAS 2004, Florence, Italy, September 8-10, 2004, Proceeding

[9] Suzuki, S. and Abe, K., *Topological Structural Analysis of Digitized Binary Images by Border Following*. CVGIP 30 1, pp 32-46 (1985)

[10] Batista, L., Rivard D., Sabourin R., Granger E., Maupin P. 2007. State of the art in off-line signature verification. In: Verma B., Blumenstein M. (eds.), Pattern Recognition Technologies and Applications: Recent Advances, (1e). IGI Global, Hershey (2007).

[11] McCabe A., Trevathan J., Read W. Neural Network-based Handwritten Signature Verification,Journal of Computers, vol. 3, no. 8, august 2008

[12] Saikia H., Chandra S. K. Approaches and Issues in Offline Signature Verification System, International Journal of Computer Applications (0975 – 8887)Volume 42– No.16, March 2012

[13] D. Kisku, P. Gupta, J. Sing, Offline Signature Identification by Fusion of Multiple Classifiers using Statistical Learning Theory, International Journal of Security and Its Applications Vol. 4, No. 3, July, 2010

[14] Biswas S. T. Kim, D. Bhattacharyya , Features Extraction and Verification of Signature Image using Clustering Technique, International Journal of Smart Home Vol. 4. , No 3. July 2010

[15] H. Baltzakis, N. Papamarkos"A new signature verification technique based on a two-stage neural network classifier", Volume 14, Issue 1,February 2001, Pages 95–103

[16] H. Lv, W. Wang, C. Wang, and Q. Zhuo, "Off-line Chinese signature verification based on support vector machines,"Pattern Recognition Letters, vol. 26, no. 15, pp. 2390–2399,2005.

[17] D. Lowe, Distinctive image features from scale-invariant key points, Proceedings of International journal of computer vision, vol. 60, no. 2, 2004., pp. 91-110

[18] X. Lou, D. Huang, L. Fan, A. Xu, An Image Classification Algorithm Based on Bag of Visual Words and Multi-kernel Learning, Journal of Multimedia, Vol. 9, No. 2, February 2014, pp. 269-277.

[19] T. Mitchell, Machine Learning, McGraw-Hill, 1997.

[20] V. N. Vapnik, The Nature Of Statistical Learning Theory, Springer, 1995.

[21] N Abbas, Y Chibani SVM-DSmT Combination for Off-Line Signature Verification, International Conference on Computer, Information and Telecommunication Systems (CITS) Amman, Jordan, May 13 - 16, 2012

Comparing affective responses to standardized pictures and videos: A study report

Marko Horvat[1], Davor Kukolja[2] and Dragutin Ivanec[3]
[1]Polytechnic of Zagreb, Department of Computer Science and Information Technology
[2]University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electric Machines, Drives and Automation
[3]University of Zagreb, Faculty of Humanities and Social Sciences, Department of Psychology
E-mail: marko.horvat@tvz.hr

Abstract - Multimedia documents such as text, images, sounds or videos elicit emotional responses of different polarity and intensity in exposed human subjects. These stimuli are stored in affective multimedia databases. The problem of emotion processing is an important issue in Human-Computer Interaction and different interdisciplinary studies particularly those related to psychology and neuroscience. Accurate prediction of users' attention and emotion has many practical applications such as the development of affective computer interfaces, multifaceted search engines, video-on-demand, Internet communication and video games. To this regard we present results of a study with *N*=10 participants to investigate the capability of standardized affective multimedia databases in stimulation of emotion. Each participant was exposed to picture and video stimuli with previously determined semantics and emotion. During exposure participants' physiological signals were recorded and estimated for emotion in an off-line analysis. Participants reported their emotion states after each exposure session. The *a posteriori* and *a priori* emotion values were compared. The experiment showed, among other reported results, that carefully designed video sequences induce a stronger and more accurate emotional reaction than pictures. Individual participants' differences greatly influence the intensity and polarity of experienced emotion.

I. INTRODUCTION

Any multimedia file can generate positive, negative or neutral emotions of varying intensity and duration [1]. By observing still images, films, printed text or listening to sounds, music and voices emotional states of affected subjects may be modulated [2] [3]. This spontaneous cognitive process is an important research topic in psychology, neuroscience and cognitive sciences but also in many interdisciplinary domains like Affective Computing and Human-Computer Interaction (HCI).

Multimedia documents with *a priori* annotated semantic and emotion content are stored in affective multimedia databases and are intended for inducing or stimulating emotions in exposed subjects. Because of their purpose such multimedia documents are also referred to as stimuli. Affective multimedia databases are standardized which allows them to be used in a controllable and predictable manner: the emotion elicitation results can be measured, replicated and validated by different research teams [4] [5]. Combined with immersive and unobtrusive visualization hardware in low-interference ambient affective multimedia databases provide a simple, low-cost and efficient means to investigate a wide range of emotional reactions [6] [7].

Compared to static pictures and sound, video is more powerful format for elicitation of emotional states because it can seamlessly and concurrently stimulates visual and auditory senses thereby multiplying their individual impacts through psychophysiological and neurological underlying mechanisms [8] [9]. It has been repetitively experimentally demonstrated that if careful attention is paid to video editing, i.e. the temporal and contextual alignment of multimedia stream relative to personal cognitions of targeted subjects, it is possible to achieve more intense and accurate stimulation of emotional states and related phenomena such as attention, anxiety and stress [7]. In practical terms affective video databases are much more useful tools than picture of sound databases. However, today very few affective video databases exist while the most prevalent are picture databases. Therefore, it is important to explore the possibility and scope of using documents from existing picture and sound databases to construct successful video sequences for fast, accurate and strong stimulation of emotional states. This goal was behind the motivation for the described study.

The remainder of this paper is organized as follows; Section 2 provides background information about the experimental study and its setup. Section 3 brings forward the results of the study which are discussed in Section 4. Finally, Section 5 concludes the paper and outlines future work into this subject.

II. METHOD

The study was performed at University of Zagreb, Faculty of Electrical Engineering and Computing in cooperation with experts from Department of Psychology, Faculty of Humanities and Social Sciences. A homogeneous group of *N*=10 college students (4 males, 6 women) with an average age 23.8 years (*std* = 4.35) participated in the experiment.

Each participant was stimulated with videos and still images taken from the International Affective Picture System (IAPS) [4] and the International Affective Digital Sounds System (IADS) [5]. IAPS and IADS are two of the most cited databases in the area of affective stimulation. These databases were created with three goals in mind: *i*) better experimental control of emotional stimuli, *ii*) increasing the ability of cross-study comparisons of results, and *iii*) facilitating direct replication of undertaken studies [10]. In this experiment a picture from IAPS and a sound from IADS were combined to make one video-clip. Same IAPS pictures were also used as still image stimuli without

†This research has been partially supported by the Croatian Ministry of Science, Education and Sports.

sounds. Some of the pictures used to construct the emotion elicitation sequences are shown in Fig. 1.

Fig. 1. A sample of IAPS pictures used as emotion eliciting video clips and images. Neutral (left column), fear (middle) and happiness dominant emotion stimuli (right).

The dominant emotions purposely elicitated in the experiment were happiness and fear. The stimuli were selected using published research on emotion norms in IAPS and IADS [11] [12] as the most powerful images and sounds for simulation of the two targeted emotions. Firstly, using 200 images and 100 sounds were selected and ranked based on their intensity of happiness and fear emotion norms [11] [12]. Secondly, the sorted list was thoroughly examined by psychology experts and 20 optimal pictures and 20 sounds were manually selected for inclusion in the elicitation sequences. These stimuli were considered the most likely to induce happiness and fear in the participants' population.

Each participant was exposed to four emotion elicitation sequences in two separate sessions or series. Each session consisted of one happiness sequence and one fear inducing sequence, and also of one video sequence and one still image sequence. A single sequence was made up from 10 images or 10 video-clips (Fig. 2). Therefore, in total each participant was exposed to 20 images and 20 video-clips. The length of each stimulus was exactly 15 seconds after which the participant was shown a blank neutral screen and had to write down his affective judgments in a self-assessment questionnaire (SAQ). The expression of subjects' ratings was not time restricted after which the participant could resume the sequence by himself (i.e. with a mouse click). Immediately before the start of the experiment each participant was separately introduced to the stimulation protocol with a neutral sequence. The neutral sequence consisted of one low arousal and valence picture (N Pic) and one video-clip without dominant emotions (N Video). All stimulation sequences are available by contacting the first author.

Half of the participants were first exposed to happiness sequences, and then fear sequences, and also a half of the participants first watched still images and then videos. To prevent the unwanted drift of physiological signals (cardiac and respiratory) before nonneutral sequences participants were exposed to a neutral stimulus until their baseline response was established [13]. The neutral blank screen only showed teal color which − according to [14] − has an optimal ratio of stimulating positive and negative emotions.

The participants' emotional responses were recorded by two methods: 1) self-assessment responses i 2) real-time monitoring of physiological signals. After each exposure to a stimulus participants filled out a self-assessment questionnaire. Each report was unique to a specific stimulus and participants could not see its contents before the stimulus has finished. The test contained following instructions: 1) "Evaluate the type and intensity of emotions" for each of the emotional norms ("happiness", 'sadness", "disgust", "fear" and "some other emotion" if none of previous) on a scale with values 0 − 9, where 0 represented "None", 9 "Extremely" while 5 was a neutral value; and 2) "Evaluate pleasure and arousal" with values -4 – 4 where value -4 was labeled "Extremely unpleasant" and "Extremely calming", and 4 "Extremely pleasant" and "Extremely arousing". Value 0 indicated a neutral sensation of valence or arousal. The report was validated during the preparations for the experiment.

The monitored physiological signals were skin conductance, electrocardiogram (ECG), respiration and skin temperature with a sampling frequency of 1250 Hz. For the acquisition of signals we used BIOPAC MP150 with AcqKnowledge software. The system was synchronized with SuperLab tool for presentation of stimuli to the participants. Emotional states were estimated off-line, with varying levels of certainty, from the recorded signals using previously developed MATLAB software [15] [16]. Before starting the experiment, each participant read the instructions, filled introductory questionnaire and signed informed consent agreement. Additional help, if necessary, was provided by the trained lab assistant who also placed physiological sensors on the participant's body. Participants were seated in a separate part of the lab, 60 – 90 cm before a 19" 4:3 high-definition LCD computer monitor and wore ergonomic headphones. The supervisor station was equipped with a separate computer monitor where the experiment was controlled and participants' psychological signals were monitored in real-time [16].

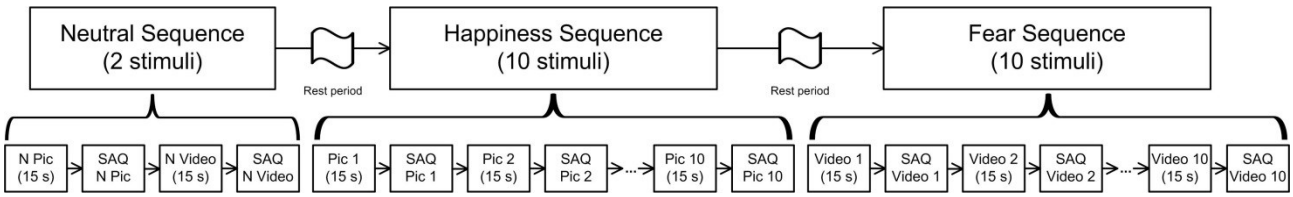


Fig. 2. The timeline of emotion stimulation paradigm. The order of dominant emotion (happiness/fear) and multimedia format (picture/video) were randomized.

†This research has been partially supported by the Croatian Ministry of Science, Education and Sports.

The experimenter and participants were physically separated by a solid artificial wall and had no mutual visual or auditory contact. Additionally, participants did not experience sensory distractions during the experiment. The implemented procedures and the experimental layout were compatible with a setup of a common psychophysiological laboratory [17].

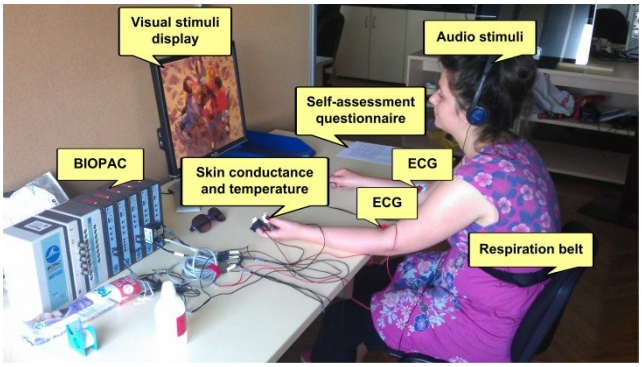


Fig. 3. The layout of participants' station. A person is observing LCD screen with audio headphones and wearing skin conductance, electrocardiogram, respiration and skin temperature contact sensors connected to BIOPAC MP150 system. Self-assessment questionnaire is on the desk. Supervisor is seated on the other side of the wall barrier.

A rest period followed each exposure session during which participants relaxed. This was verified by examining the physiological signal parameters that were visualized in real time. The exposure could resume only after the baseline signal levels were reestablished.

III. RESULTS

The potential of evoking emotional reactions using video clips − constructed from IAPS pictures and IADS sounds − and IAPS pictures was evaluated under equal conditions. Emotional dimensions pleasure and arousal were rated on a scale 1 − 9, and emotion norms (i.e. discrete emotions) on a scale 1 − 10. A lower value (1 − 3) in both scales implies a lesser experience and a higher value (7 − 10) a more intense experience of the particular emotional dimension or norm, respectively. Also, higher frequency of ratings in the upper part of the scale (response > 5) or the highest attainable values (9 and 10) signifies a more intense and powerful stimulation. The frequencies of responses are displayed as frequency distribution diagrams. The aggregated results are shown in Fig. 4.

IV. DISCUSION

Results based on the analysis of participants' self-assessments indicate that the most pronounced reported difference is in arousal emotion dimension. Indeed, videos relative to pictures can more frequently stimulate higher arousal in sequences with dominant happiness and fear. Additionally, video sequences often stimulated higher levels of pleasure in happiness sequences, and lower levels of pleasure in fear sequences, but this distinction is less pronounced than with arousal.

From Fig. 4 it is evident that video sequences were more powerful in stimulation of both emotional dimensions than picture sequences. Happiness-dominant video sequences more often elicited higher levels of happiness basic emotion than happiness-dominant picture sequences. Similarly, fear-dominant video sequences provoked more above average fear ratings than fear-dominant picture sequences. Although frequency distribution differences in basic emotions are present, they are less obvious than the spread in emotional dimensions, especially arousal. In general, the emotion provoking potential of video is more apparent in emotional valence and arousal than in specific emotions happiness, sadness, disgust and fear. Due to reported low stimulation of other emotions except fear and happiness it may be concluded that the sequences were emotionally accurate. This is particularly evident in very low levels of disgust and sadness in happiness sequences and even in fear sequences. In overall video sequences provoked a "cleaner" and more powerful affective response, i.e. with lower reported intensities of emotions different from those targeted, than picture sequences which corresponds well with findings from previous studies [13] [18].

However, because of relatively small number of participants the results analysis is strongly influenced with noticable differences between individual reports. There is a significant variability in the intensity of provoked emotion, polarity (i.e. positive or negative) and discrete category among some participants. Such discrepancies are present in video and picture sequences.

Unfortunately, due to objective reasons it was not possible to include more participants in the study. Subsequently, an independent stimulation protocol could not be implemented and the same visual stimuli had to be used in video and image sessions. If the number of participants was significantly larger different stimuli could be used in videos and images.

Based on the collected results it can be expected that multimedia sequences, carefully prepared for a particular group of participants, will be able to provoke targeted emotional states with the desired intensity. However, construction of optimal sequences proved to be difficult because IAPS, and particularly, IADS databases do not have a wide selection of stimuli with accentuated specific basic emotions. A better choice of provoking visual stimuli is clearly needed which encourages construction of more affective multimedia databases, annotated both with emotional dimensions and discrete emotions, and having a large semantic space. This in turn necessitates development of powerful tools for multimedia stimuli retrieval which can efficiently perform multifaceted search in such databases, along several emotional, semantic and contextual data dimensions, thus assisting researchers in finding optimal stimuli for personalized emotion elicitation sequences [19].

†This research has been partially supported by the Croatian Ministry of Science, Education and Sports.

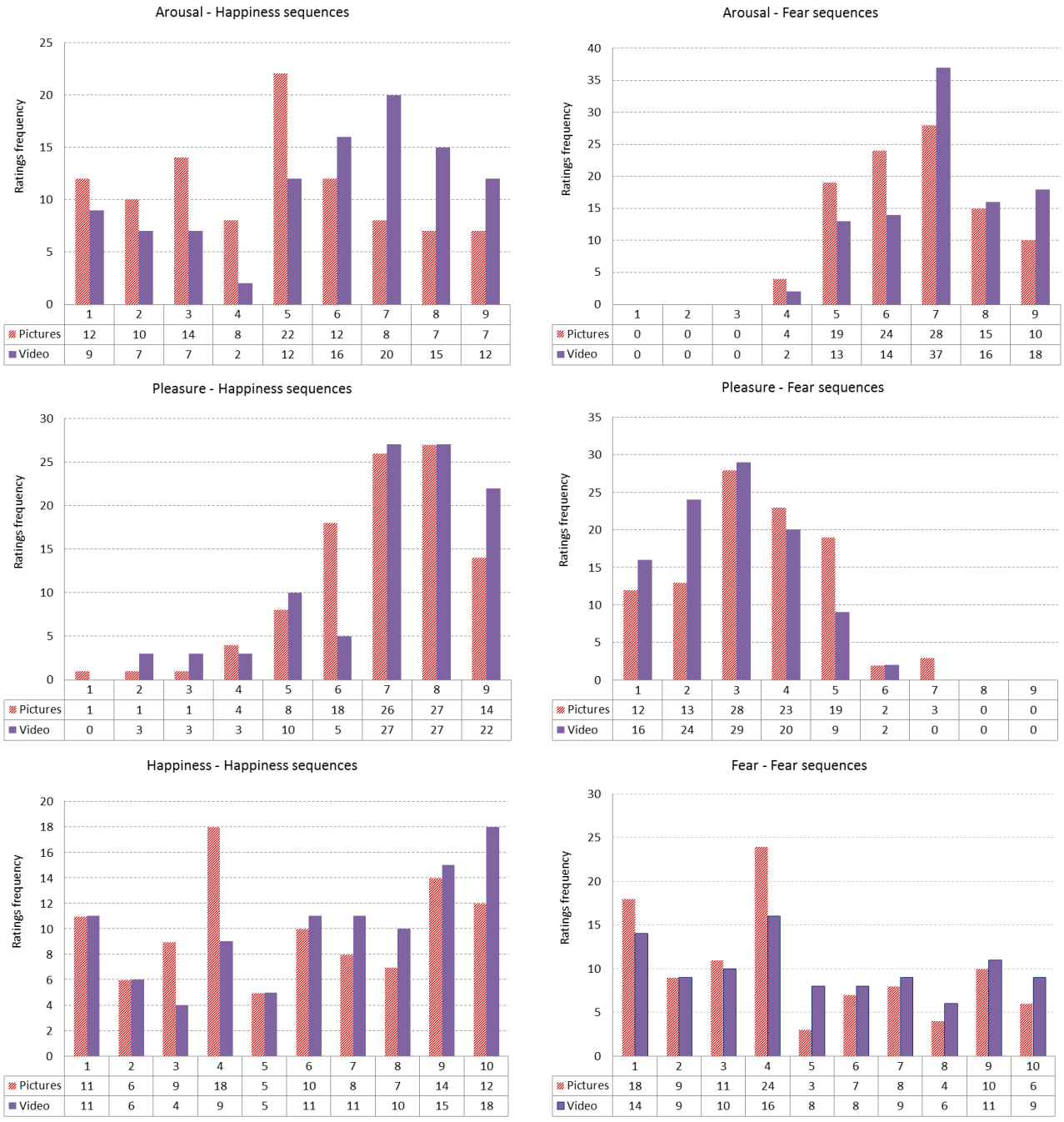


Fig. 4. Frequency distribution of *N*=10 participants' ratings after elicitation with IAPS and IADS video clips ("Video") and IAPS pictures ("Pictures"). Reported emotional dimensions arousal and pleasure (upper and middle rows), discrete emotions happiness and fear (bottom row).

V. CONCLUSION AND FUTURE WORK

Emotional reactions can be induced by virtually any multimedia format: films, pictures, sounds, voice and even text. Participants' responses depend not only on stimuli content, properties and type, but also on a number of intrinsic and extrinsic factors which may be difficult to control in a practical experiment [13] [17]. For example, participants' motivation, attitude, character traits, beliefs, past experiences and the experimental environment (i.e. the setup of the laboratory and the experimental protocol) play an extremely important role in formation of emotion. Therefore, a comparison of emotional reactions induced by pictures and videos may be regarded as a hands-on tutorial for researchers as to which multimedia stimuli properties can lead to more accurate, precise and faster elicitation.

The study showed that standardized picture and sound databases can be joined together and used as videos for elicitation of high-arousal emotional states. Deliberate and practical stimulation of discrete emotions happiness and fear is attainable but it is more difficult and prone to error, especially happiness. The least successful was stimulation of very positive, i.e. high valence, emotional states.

We hope that the presented study could be used in design of emotion elicitation protocols as well as future affective multimedia databases. Additionally, the study's

results may help researchers to find the optimal multimedia format for elicitation of emotion even if appropriate video stimuli are not available.

REFERENCES

[1] S. Brave and C. Nass, "Emotion in human-computer interaction," The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications, CRC Press, Taylor & Francis, Florida, pp. 81−96, 2002.

[2] J. A. Coan and J. J. B. Allen, "The Handbook of Emotion Elicitation and Assessment," Oxford University Press Series in Affective Science, Oxford University Press, USA, 2007.

[3] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," Consciousness and Cognition, vol. 17, pp. 484–95, 2008.

[4] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Technical Report A−8, University of Florida, Gainesville, FL, 2008.

[5] P. J. Lang and M. M. Bradley, "The International Affective Digitized Sounds (2nd Edition; IADS-2): affective ratings of sounds and instruction manual," Technical report B-3, University of Florida, Gainesville, FL, 2007.

[6] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," Cognition & Emotion, vol. 9(1), pp. 87–108, 1995.

[7] D. Villani and G. Riva, "Does interactive media enhance the management of stress? Suggestions from a controlled study," Cyberpsychology, Behavior, and Social Networking, vol. 15(1), pp. 24–30, 2012.

[8] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, and T. A. Ito, "The psychophysiology of emotion," Handbook of emotions, vol. 2, 173–191, 2000.

[9] K. C. Berridge and M. L. Kringelbach, "Neuroscience of Affect: Brain Mechanisms of Pleasure and Displeasure," Current opinion in neurobiology, vol. 23(3), pp. 294–303, 2013.

[10] M. M. Bradley and P. J. Lang, "Measuring emotion: Behavior, feeling and physiology," Cognitive neuroscience of emotion, Oxford University Press, New York, pp. 242–276, 2000.

[11] T. M. Libkuman, H. Otani, R. Kern, S. G. Viger, and N. Novak, "Multidimensional normative ratings for the international affective picture system," Behavior Research Methods, vol. 39(2), pp. 326−334, 2007.

[12] R. A. Stevenson and T. W. James, "Affective auditory stimuli: Characterization of the International Affective Digitized Sounds (IADS) by discrete emotional categories," Behavior Research Methods, vol. 40(1), pp. 315−321, 2008.

[13] J. Rottenberg, R. D. Ray, and J. J. Gross, "Emotion elicitation using films," Handbook of emotion elicitation and assessment, Oxford University Press, New York, pp. 9−28, 2007.

[14] K. A. Y. A. NAz and H. Epps, "Relationship between color and emotion: a study of college students," College Student Journal, vol. 38(3), pp. 396−405, 2004.

[15] D. Kukolja, S. Popović, B. Dropuljić, M. Horvat, and K. Ćosić, "Real-time emotional state estimator for adaptive virtual reality stimulation," Foundations of Augmented Cognition, Neuroergonomics and Operational Neuroscience, Springer Berlin Heidelberg, pp. 175−184, 2009.

[16] D. Kukolja, S. Popović, M. Horvat, B. Kovač, and K. Ćosić, "Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications," International journal of human-computer studies, vol. 72(10), pp. 717−727, 2014.

[17] J. J. Curtin, D. L. Lozano, and J. J. Allen, "The psychophysiological laboratory," Handbook of emotion elicitation and assessment, Oxford University Press, New York, pp. 398−425, 2007.

[18] C. G. Courtney, M. E. Dawson, A. M. Schell, A. Iyer, and T. D. Parsons, "Better than the real thing: Eliciting fear with moving and static computer-generated stimuli," International Journal of Psychophysiology, vol. 78(2), pp. 107–114, 2010.

[19] M. Horvat, N. Bogunović, and K. Ćosić, "STIMONT: a core ontology for multimedia stimuli description," Multimedia Tools and Applications, vol. 73(3), pp. 1103−1127, 2014.

Retrieval of multimedia stimuli with semantic and emotional cues: Suggestions from a controlled study

Marko Horvat[1], Davor Kukolja[2] and Dragutin Ivanec[3]
[1]Polytechnic of Zagreb, Department of Computer Science and Information Technology
[2]University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electric Machines, Drives and Automation
[3]University of Zagreb, Faculty of Humanities and Social Sciences, Department of Psychology
E-mail: marko.horvat@tvz.hr

Abstract - The ability to efficiently search pictures with annotated semantics and emotion is an important problem for Human-Computer Interaction with considerable interdisciplinary significance. Accuracy and speed of the multimedia retrieval process depends on the chosen metadata annotation model. The quality of such multifaceted retrieval is opposed to the potential complexity of data setup procedures and development of multimedia annotations. Additionally, a recent study has shown that databases of emotionally annotated multimedia are still being predominately searched manually which highlights the need to study this retrieval modality. To this regard we present a study with $N = 75$ participants aimed to evaluate the influence of keywords and dimensional emotions in manual retrieval of pictures. The study showed that if the multimedia database is comparatively small emotional annotations are sufficient to achieve a fast retrieval despite comparatively lesser overall accuracy. In a larger dataset semantic annotations became necessary for efficient retrieval although they contributed to a slower beginning of the search process. The experiment was performed in a controlled environment with a team of psychology experts. The results were statistically consistent with validates measures of the participants' perceptual speed.

I. INTRODUCTION

Manual searching of multimedia datasets such as picture repositories and video databases is still predominately based on keyword-based information retrieval. However, range and expressivity of manual data description process are limited. When adding a picture to an Internet repository or social network users are asked to describe the picture, who or what is portrayed, when and where the photo was taken, which emotion does the picture stimulate etc. Frequently users provide contextual information which is insufficient for searching. The retrieved datasets are large and inaccurate [1]. Various semantic categories, available multimedia content and metadata are used to boost retrieval accuracy and improve ranking of results important to the posed query [2] [3] [4] [5]. These procedures are the subject of ongoing research but each of them has restrictions and in practice cannot completely substitute missing expert information.

The missing data problem is especially exacerbated in affective multimedia databases because they implement only sparse annotation with unrestricted keywords [6]. These databases are especially important tool for study of emotion and attention in the areas of psychology, neurology and cognitive sciences [7]. Retrieval of loosely annotated data, with only one expert-defined tag per picture, is even more open to errors than with, for example, social media with community generated tags, hashtags or folksonomies. Although tags are noisy they can still provide useful information if combined with intelligent algorithms for statistical reasoning [8]. As has been shown in a recent survey on usage modalities of emotionally-annotated picture databases [9], the ultimate consequence of such poor semantic annotation scheme is that it is more productive to browse affective multimedia databases visually, although they may contain hundreds or thousands of images, than to search them with keywords. Additionally, this survey showed that 70% of experts need more than 1h to retrieve optimal pictures for a single emotion elicitation sequence and 20% require more than 12h which is extremely impractical. Therefore, it is necessary to establish how much context information is enough to enable efficient keyword and emotion search paradigms in affective multimedia databases. How formal and expressive must the data representation be? Is it necessary to better describe multimedia semantics or emotion, or both? These and other similar questions may be theoretically examined but ultimately only a controlled experiment can provide a definite answer to them. Annotation quality and retrieval efficiency are in a mutual tradeoff. More thorough annotation comes at a cost of better data preparation, but will result in a more accurate search. The experiment described in this paper was motivated to explore optimal methods for annotation of semantics and emotion in affective multimedia databases.

The presented research had two goals; considering the results of the survey it was necessary to determine how fast and accurately experts can find a specific picture stimulus in ideal circumstances. The second goal was to establish minimal emotion and semantic annotation requirements which still enable an efficient searching of affective multimedia databases.

The remainder of this paper is organized as follows; Section 2 provides detailed background information and describes the experimental setup. Section 3 specifies and visualizes all obtained results, and the following Section 4 discusses the results from the perspective of the experiment's goal. Finally, Section 5 concludes the paper and outlines future work into this subject.

II. METHOD

The controlled experiment was performed at the Department of Psychology of University of Zagreb, Faculty of Humanities and Social Sciences. A total of N=75 college students (10 males, 65 women) with an

†This research has been partially supported by the Croatian Ministry of Science, Education and Sports.

average age 19.49 years (*std* = 0.87) participated in the experiment.

All participants performed two tasks; Task 1: find five pictures with particular emotion, and; Task 2: find five pictures with specific emotion and semantics. The search criteria can be considered similar to emotion elicitation sequences typically used in real settings. All participants first performed Task 1 and then Task 2. The results help to estimate the average time to retrieve a stimulus on the basis of the set of semantic and affective criteria, and the number of errors that can occur at the same time.

In both tasks, participants had to, as quick as possible perform visual search tasks in the valence-arousal emotion plane to find the desired stimuli [10] [11]. However, to achieve an optimal search strategy in the second task it was necessary to first input search keywords to reduce the stimuli subset and leave only those pictures that match the Task 2.

A preselected subset of 250 picture stimuli from the International Affective Picture System (IAPS) [12] database was used for both tasks. In total 50 pictures (20% of the subset) matched the search criteria. In Task 1 the participants had to find IAPS pictures with values of emotion dimensions within a given region of valence-arousal plane (*valence* > 5.0 and *arousal* > 5.0), while in Task 2 the pictures had *arousal* > 5.0 and preset semantics (i.e. children, families, specific animals etc.). For Task 2 the participants were given a list of keywords they had to match with the presented stimuli. All tasks were designed with positive, i.e. non-null, results. The tasks' texts and the list of picture stimuli used in the experiment are available by contacting the first author.

The participants used a previously developed software tool "Intelligent stimuli generator" to search for stimuli with emotion and semantic parameters [7]. The tool was specially adapted for the experiment: user interface was greatly simplified for implementation of the experimental protocol. All search paradigms except the manual search and other unnecessary features were removed. However, a comprehensive logging module was added which can enumerate, record and timestamp different user actions, search attempts and categorize them as successful or not. The logs are generated automatically and can be later analyzed with dedicated statistical tools like SPSS or Matlab.

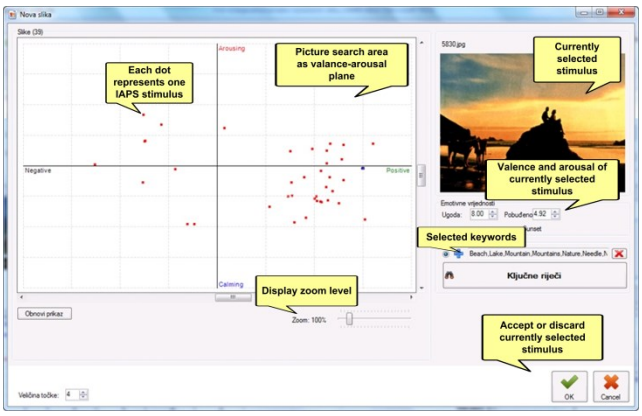


Fig. 1. The main screen of the "Intelligent stimuli generator" tool displaying IAPS stimuli subset in Task 2 defined with selected keywords.

A week before the experiment the college students were given comprehensive written instructions about the "Intelligent stimuli generator" tool so they could familiarize themselves with the user interface and the types of tasks they will have to carry out. The experiment itself took place in the laboratory of the Department of Psychology of University of Zagreb, Faculty of Humanities and Social Sciences. The laboratory contains a number of separate sound-proof cabins with a bench, two seats, PC computer, 17" 4:3 LCD monitor and sound speakers. At a time each cabin seated one lab assistant with psychology training and one experiment participant.

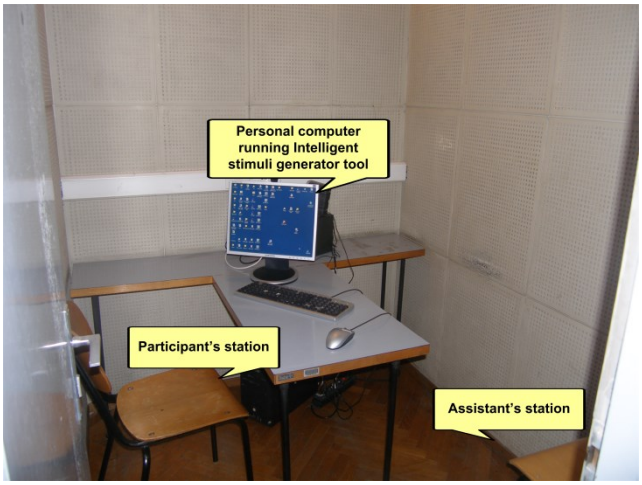


Fig. 2. A soundproof cabin used in the study with two stations for the participant and assistant, and a personal computer for picture retrieval with semantic and emotional cues.

Before the picture search experiment all participants took two standard psychological tests of their perceptual speed [13]. Each test lasted 5 minutes and they were also conducted in controlled conditions with minimum outside sensory interference. The aim of these additional tests was to investigate the relationship of participants' visual search performance and achievement on perceptual speed tests, as a measure of stable individual traits connected to the perceptual ability [14].

III. RESULTS

Complete results of N_{RTOT}=750 pictures retrievals are shown in Fig. 3. Each point in diagrams represents an aggregated mean of N_R=75 records for each of the five searched pictures (numbered 1−5 on x-axis) in Task 1 and Task 2, respectively, with indicated standard deviation. Picture retrieval duration in seconds is provided on y-axis.

The results show that in Task 1 on average 10.74 s (*std* = 8.66) was necessary to find a picture using only valence and arousal emotion constraints. More time was needed to locate files by searching semantics. In Task 2 average 47.48 s (*std* = 15.18) was needed to retrieve a picture with specific keywords and emotion which is more than 4 times longer than the retrieval without semantics.

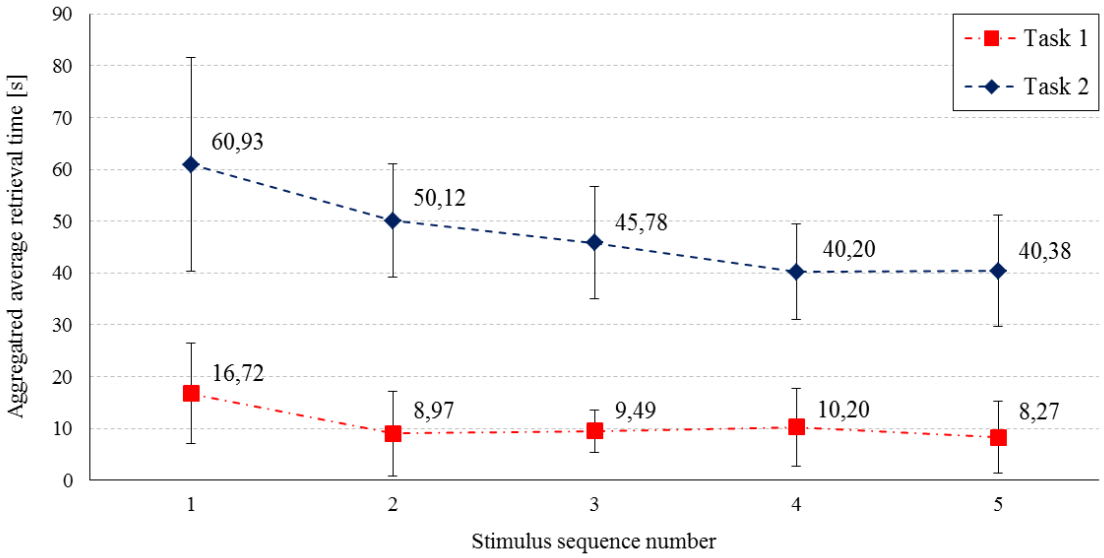


Fig. 3. Aggregated average performance of N=75 participants in Task 1 and Task 2. Individual searched pictures (labeled 1−5) are represented on x-axis, and the recorded search time is y-axis.

In Task 2 the search time monotonically decreases with each found picture and the difference between the first and the last search is over 20 s − from 60.93 s (*std* = 20.67) down to 40.38 s (*std* = 10.73). In Task 1 the average time spent to found one picture was marginally better with 16.72 s (*std* = 9.66) for the first and 8.27 s (*std* = 6.91) for the fifth and last picture.

In Table 1 are the correlations between the two perceptual speed, and Task 1 and Task 2. Both tests are significantly statistically related only with Task 2 but not with Task 1.

TABLE I

Correlation measures (r) among achievements on perceptual speed tests (Test 1, Test 2) and efficiency in picture search tasks (Task 1, Task 2). Each measure was calculated on N_R=75 sample (*p<0.05; **p<0.01).

	Test 1	**Test 2**	**Task 1**	**Task 2**
Test 1	1	0.57**	0.08	-0.27*
Test 2		1	0.11	-0.33**
Task 1			1	-0.07
Task 2				1

The aggregated average number of incorrectly found images (false positives) in Task 1 was 2.75 (*std* = 2.75), while only 1.26 (*std* = 0.53) in Task 2 is an indication of greater search accuracy which was made possible by introduction of semantic constraints. Also, statically very large dispersal of false positives was present in Task 1 suggesting that some participants' search strategy was based on volume, i.e. fast processing of many pictures, rather than on accuracy and precision.

IV. DISCUSSION

The study showed that in ideal conditions, i.e. if the stimuli database is sufficiently small, it is possible to relatively quickly find a specific image just based on affective tagging. But at the same time this approach results in relatively large error rate and participants participant retrieved more false pictures before finding the correct one.

Stimuli retrieval with keywords was initially slower, which is especially apparent by comparing search times for first pictures in both tasks, but once the required keywords were found the displayed set of images contained enough true positives so the required picture could be retrieved relatively quickly and at the same time the participants consistently made less mistakes. In both tasks the average retrieval time asymptotically falls with the increasing sequence number of the image, indicating that respondents needed a period of getting used to work with the "Intelligent stimuli generator" tool, but they could quickly achieve their relative maximum search efficiency.

However, it should be pointed out that in this study the original IAPS database was substantially reduced and the participants searched a subset with only 250 of the original 1182 IAPS stimuli (21.15 %), and in this subset 96 pictures (38.4 %) matched the overall search criteria. It reasonable to assume that the search times and the number of errors being made would be grater if realistic (i.e. larger and more complex) emotionally annotated multimedia databases were used or if the participants could not have the benefit of having a pre-defined set of semantic tags at their disposal.

Perceptual speed is defined as the speed of accurate perception of details in observed structures. It is measured with the speed of accurate recognition of similarities and differences between the observed structures [14].

Correlation among scores on perceptual speed tests and Task 2 (stimuli retrieval with keywords) were obtained (Table 1). Correlations (i.e. Pearson coefficients) have negative values which can indicate that those participants who were better in perceptual tests required less time to find keywords given to them in Task 2. It also demonstrates that stable interindividual traits may, at least to some extent, explain efficiency of keyword-based database retrieval method. Additionally, it could also mean that an intelligence measure could explain part of retrieval efficiency, because the perceptual speed has low to moderate correlation with intelligence which points out to their common probable cause [15]. The obtained correlations are not strong, but stable − both tests have similar correlation to time of retrieval. Indirectly, correlation values indicate that when task was more complex some cognitive individual differences played a role in general retrieval efficiency. This was expected and means that participants were motivated in both tasks.

V. CONCLUSION

Multimedia databases may be searched and retrieved by any combination of available text and visual features. With current knowledge models it is possible to describe emotion and high-level semantics in XML-derived formats thus enabling reasoning about emotional states, anxieties, phobias, stress or mental fatigue that a multimedia file may provoke or be related to [16]. During construction of affective multimedia retrieval information systems the optimal solution is to have the simplest possible requirements on the complexity of data model but at the same time to achieve high search efficiency and retrieval accuracy.

We showed that if emotionally annotated multimedia databases have a few hundred images it is not necessary to annotate their semantics. If databases are relatively small they can be efficiently searched just by using information about multimedia affective content. However, semantic-based search is practical in large multimedia sets. In a two-step retrieval process the search space must be first reduced with semantic constraints to a manageable size which enables multimedia to be searched efficiently with emotion or other metadata tags.

We hope that the presented study could be used in design of future affective multimedia retrieval systems as well as affective multimedia databases. The study's results may help researchers to find the optimal choice between the data model complexity and the required affective multimedia search efficiency.

REFERENCES

[1] W. B. Croft, D. Metzler, D., and T. Strohman, "Search engines: Information retrieval in practice," Reading: Addison-Wesley, pp. 351−358, 2010.

[2] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," Proc. of the international conference on Multimedia, ACM, pp. 251−260, October 2010.

[3] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 34(4), pp. 723−742, 2012.

[4] S. Clinchant, J. Ah-Pine, and G. Csurka, "Semantic combination of textual and visual information in multimedia retrieval," Proceedings of the 1st ACM International Conference on Multimedia Retrieval, pp. 44:1–44:8, April 2011.

[5] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'El, I. Ronen, ... and S. Chernov, "Personalized social search based on the user's social network," Proc. of the 18th ACM conference on Information and knowledge management, ACM, pp. 1227−1236, November 2009.

[6] M. Horvat, S. Popović, N. Bogunović, and K. Ćosić, "Tagging multimedia stimuli with ontologies," Proceedings of the 32nd International Convention MIPRO 2009, Croatia, pp. 203–208, May 2009.

[7] M. Horvat, N. Bogunović, and K. Ćosić, "STIMONT: a core ontology for multimedia stimuli description," Multimedia Tools and Applications, vol. 73(3), pp. 1103−1127, 2014.

[8] F. Abel, Q. Gao, G. J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," The Semanic Web: Research and Applications, Springer Berlin Heidelberg, pp. 375−389, 2011.

[9] M. Horvat, S. Popović, and K. Ćosić, "Multimedia stimuli databases usage patterns: a survey report," Proceedings of the 36nd International Convention MIPRO 2013, Croatia, pp. 993−997, May 2013.

[10] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament," Current Psychology, vol. 14(4), pp. 261−292, 1996.

[11] J. A. Russell, "A circumplex model of affect," Journal of personality and social psychology, vol. 39(6), pp. 1161−1178, 1980.

[12] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Technical Report A−8, University of Florida, Gainesville, FL, 2008.

[13] D. Ivanec, "New Perceptual speed tests. Unpublished tests developed under project: Development, standardization and psychometric validation of cognitive ability tests," Faculty of Humanities and Social Sciences, Zagreb, 2011.

[14] K. Rayner, "Eye movements and attention in reading, scene perception, and visual search," The quarterly journal of experimental psychology, vol. 62(8), pp. 1457−1506, 2009.

[15] L. D. Sheppard and P. A. Vernon, "Intelligence and speed of information-processing: A review of 50 years of research," Personality and Individual Differences, vol. 44(3), pp. 535−555, 2008.

[16] M. Schröder, P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato, "EmotionML–an upcoming standard for representing emotions and related states," Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg, pp. 316−325, 2011.

GWAT: The Geneva Affective Picture Database WordNet Annotation Tool

Marko Horvat[1], Dujo Duvnjak[2] and Davor Jug[1]
[1]Polytechnic of Zagreb, Department of Computer Science and Information Technology
[2]University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Telecommunications
E-mail: marko.horvat@tvz.hr

Abstract - The Geneva Affective Picture Database WordNet Annotation Tool (GWAT) is a user-friendly web application for manual annotation of pictures in Geneva Affective Picture Database (GAPED) with WordNet. The annotation tool has an intuitive interface which can be efficiently used with very little technical training. A single picture may be labeled with many synsets allowing experts to describe semantics with different levels of detail. Noun, verb, adjective and adverb synsets can be keyword-searched and attached to a specific GAPED picture with their unique identification numbers. Changes are saved automatically in the tool's relational database. The attached synsets can be reviewed, changed or deleted later. Additionally, GAPED pictures may be browsed in the tool's user interface using simple commands where previously attached WordNet synsets are displayed alongside the pictures. Stored annotations can be exported from the tool's database to different data formats and used in 3rd party applications if needed. Since GAPED does not define keywords of individual pictures but only a general category of picture groups, GWAT represents a significant improvement towards development of comprehensive picture semantics. The tool was developed with open technologies WordNet API, Apache, PHP5 and MySQL. It is freely available for scientific and non-commercial use.

I. INTRODUCTION

Affective multimedia databases are repositories of multimedia documents with annotated semantics and emotion. They are primarily used as standardized tools for stimulating specific emotional responses in exposed subjects. Affective multimedia databases are frequently employed in psychology and neuroscience for research of subjective, psycho-physiological, behavioral and neurophysiological reactions induced by viewing affective stimuli (per example [1] [2] [3]).

The landscape of affective multimedia databases is very diverse with many different and structurally unrelated databases [4] [5]. Currently, affective multimedia databases are created by manual annotation of semantics and emotion in which volunteers record their personal impressions using standardized questionnaires [6] after being induced by stimuli. Obtaining affective annotations always includes a psychological experiment conducted with a statistically relevant group of participants. Software tools for querying affective multimedia databases and extraction of stimuli are virtually nonexistent [7]. The databases are still being mostly searched manually by experts in a lengthy, repetitive and labor intensive process [7].

Typically, affective multimedia databases have very simple structures constituting of a multimedia repository and a description file [4] [5]. The repository is a container with uniquely indexed multimedia which is referenced in the description file. The repository has a very simple implementation as a file system folder – databases must be acquired from their owners and installed on a user's computer. All multimedia metadata, emotion and semantic annotations are stored in the description file which is also located in the repository as a formatted text or comma separated values (CSV) file. The format of the description file differs between databases; no two databases share exactly the same format.

GAPED is a relatively new and large affective multimedia database with 730 pictures [8]. It has an even simpler architecture than most other databases. Pictures are stored in six separate folders each compromising one semantic category: "snakes", "spiders", "human concerns", "animal mistreatments", "neutral" and "positive". The first four categories are emotionally negative. All pictures have unique names by which they are referenced and emotionally described. However, there are absolutely no semantic annotations and objects or events in pictures are not specified. The semantic content of pictures can only be vaguely implied by their semantic category membership, and a more detailed context can only be obtained by accessing and observing each individual picture.

GWAT was developed with a motivation to improve the GAPED semantically sparse dataset and to provide an opportunity to reuse GAPED content in other domains which require a more expressive multimedia description. The tool is free for academic and nonprofit use and can be obtained by contacting the first author.

The remainder of this paper is organized as follows; Section 2 gives an overview of the tool and its system architecture. Sections 3 and 4 illustrate how the multimedia content is displayed in the user interface and how to browse through pictures, together with already existing GAPED-WordNet annotations. Section 5 provides information on how to retrieve WordNet synsets in GWAT and use them to annotate GAPED pictures. Finally, Section 6 concludes the paper and provides insight into future work.

II. GWAT ARCHITECTURE

The NAPS Search Tool is a thin-client web application written in PHP 5.4 programming language. The application requires a web server (e.g. Apache HTTP Server) configured to run PHP 5.x, or newer versions, and MySQL 5.0 database to be preinstalled on the target computer. The tool is distributed as a single package with PHP pages, presentation files and a SQL script for MySQL database import. The distribution package also contains a help file with instructions on how to install the tool, configure the web server and create a new instance of GWAT relational database.

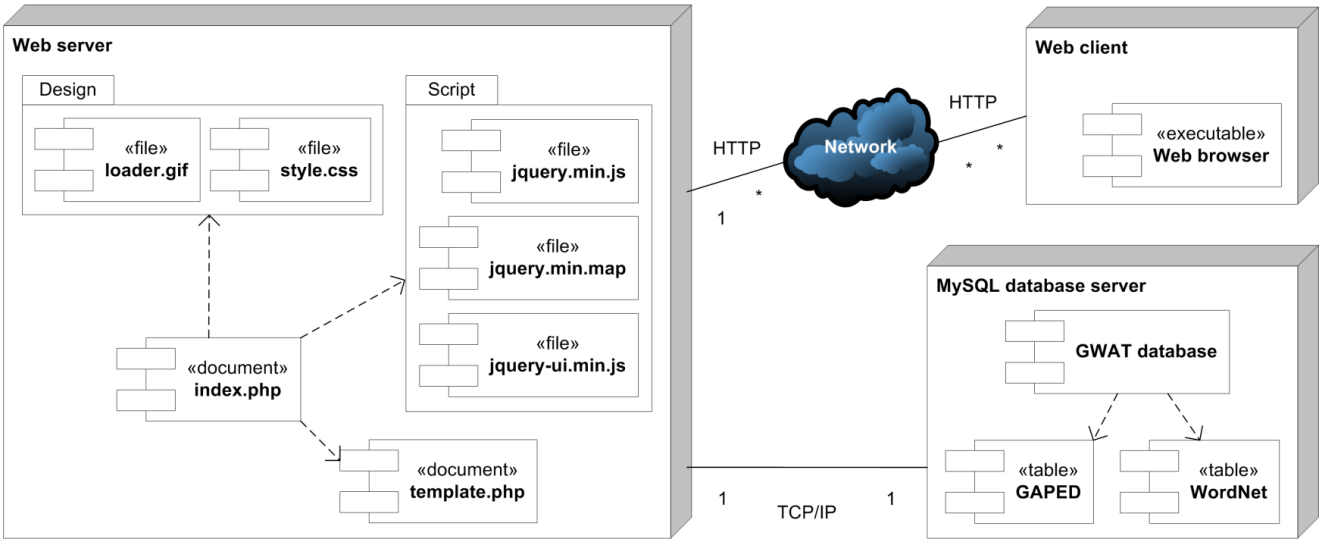


Fig. 1. GWAT's UML component and deployment diagrams.

The architecture is simple and consists of the presentation layer, the logic layer and the data layer. Cascading Style Sheet (CSS), graphics and JQuery scripts form the presentation layer, while PHP5 code files implement application logic and connect the data layer with the presentation. GAPED picture identifiers, WordNet synsets and pictures annotations created with GWAT are stored together in a single MySQL database. Pictures and affective ratings are not provided with the tool and have to be obtained from the Swiss Center for Affective Sciences[1].

WordNet semantic network is employed as a dictionary to describe semantics of GAPED pictures [9]. WordNet represents a very important component of the tool's architecture. WordNet is a large knowledge taxonomy with >100,000 individual concepts or synsets (i.e. synonym sets) connected by different semantic relationships such as IS-A and PART-OF. It is freely available and comes with an extensive software development interface. The glossary, originally in English but subsequently translated to a number of other languages, contains uniquely referenceable nouns, verbs, adjectives and adverbs. WordNet offers many benefits to description of multimedia compared to plain keywords and user defined tags. Firstly, it has very large and sufficiently expressive vocabulary. Since the glossary is controlled and standardized, picture annotations can be transferred to any other application compatible with WordNet. Finally, WordNet glossary enables users to search queries in a linguistically more natural and expressive manner as already demonstrated by existing WordNet stimuli annotating tools (per example [10] [11]).

The GWAT's UML component and deployment diagrams are shown together in Fig. 1. The main PHP page is "index.php". It uses "template.php" as a modular component for presentation of different GAPED pictures, three JQuery files for client-side functionality and CSS definition file with an animated GIF picture for design.

The program user interface has six major components:

1) the selected stimulus component for displaying a single stimulus,
2) GAPED search box,
3) GAPED navigation buttons,
4) WordNet annotations component for displaying the list of synsets describing picture semantics,
5) WordNet component for searching and attaching synsets to pictures,
6) help box with useful information for GWAT users.

The components' functionalities are described in next sections.

III. DISPLAYING GAPED CONTENT

After starting the web tool, i.e. opening the main page "index.php", first GAPED picture A001.bmp is always automatically loaded and displayed in the selected stimulus component for displaying a single stimulus. The component is placed in the central region of the main page. If GAPED database is not installed on the server a busy JavaScript animated icon will be visible instead of the picture. All pictures can be accessed and displayed in the tool's main page together with their existing WordNet annotations. The layout of the main page is explained in the Fig. 2.

The name of the picture file with its extension is shown above and to the left of the displayed picture, and GAPED search box is on the right. User can enter name of a picture, from any of GAPED folders containing different semantic categories ("A", "H", "N", "P" "Sn" and "Sp"), to be loaded and displayed in the main page. It is not necessary to specify a specific folder because the tool will search all folders until the desired picture is found. Picture format extension is obligatory. It is necessary to specify picture names exactly – special search characters such as "*", "?" etc. are not supported in the current version of GWAT.

[1] http://www.affective-sciences.org/

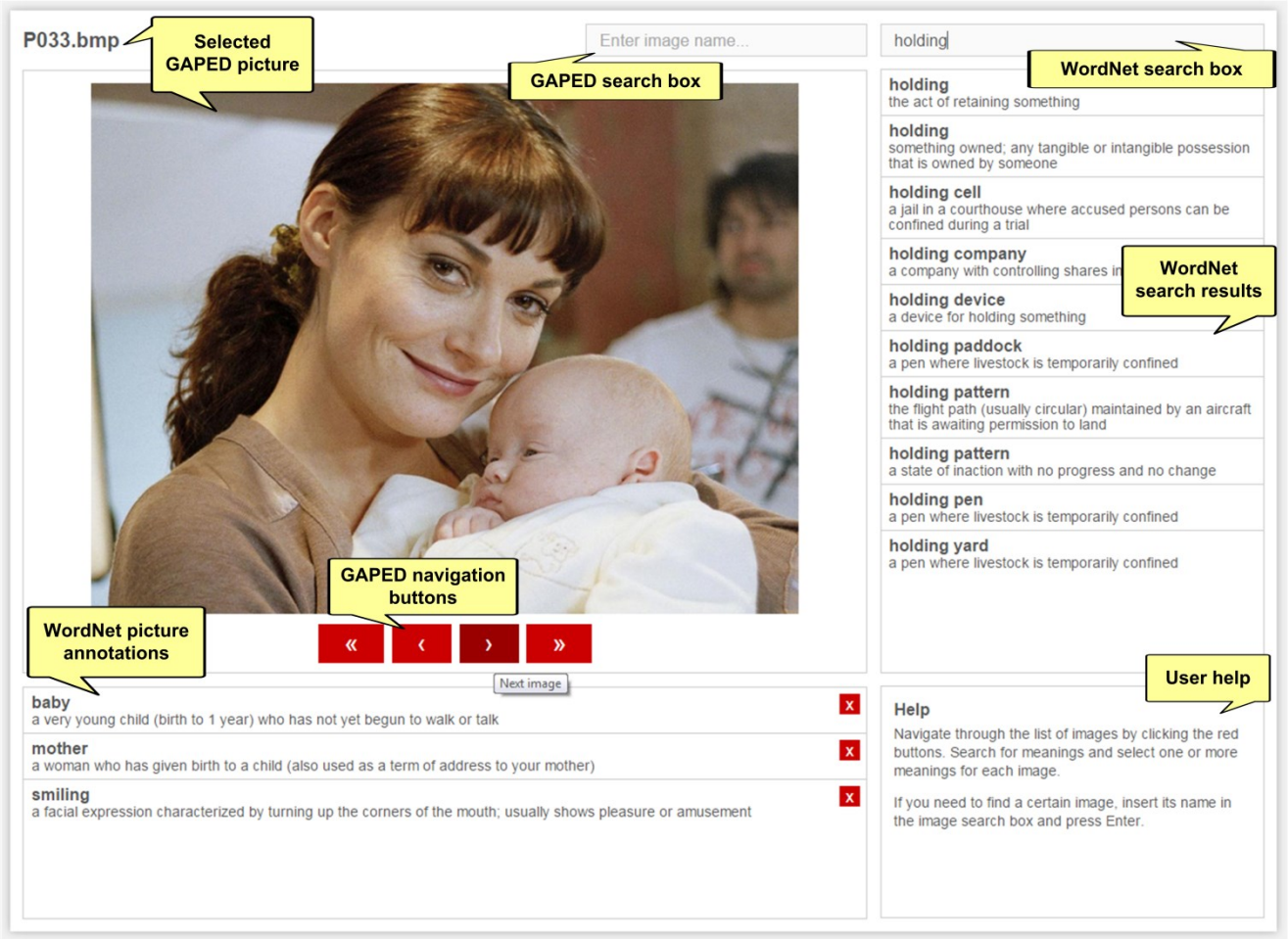


Fig. 2. The GWAT's main page.

The picture search starts only after pressing enter key and not after each letter keystroke in order to minimize resource requirements from quickly repetitive database queries. If a picture cannot be found, i.e. search query returns NULL value, a JavaScript generated (i.e. client-side) error message is shown informing the user that the particular picture cannot be found.

Below the central area are four navigation buttons (Fig. 2). The left most button ("<<") shifts the focus to the first picture in alphabetical order ("A001.bmp"). The right most button (">>") moves the focus to the last picture in the set ("Sp160.bmp"). Buttons ("<") and (">") change current display by one position towards lower position or higher position in the alphabetically sorted sequence, respectively. For example, if picture "P033.bmp" is shown (as in Fig. 2) then by pressing two times button ("<") "P032.bmp" will be displayed first followed by "P031.bmp". After pressing (">") the selection returns to "P032.bmp". The navigation buttons can also be used to circulate the view between both ends of the set: if the last picture "Sp160.bmp" is selected then with (">") the first picture "A001.bmp" will be displayed, and vice versa with ("<"). Tooltips are provided for all buttons so users can get accustomed to the tool's interface more easily.

IV. GAPED-WORDNET ANNOTATIONS

Below the navigation buttons is the WordNet annotations component with a list of WordNet synsets assigned to the currently displayed GAPED picture. The content of the component is refreshed dynamically whenever a new picture is shown in the main page. Synsets are displayed in rows with their name, full description and a large "X" button to the right (Fig. 2). By pressing the red "X" button the particular annotation is immediately removed from semantic description of the currently shown picture. This feature is implemented with JQuery JavaScript framework. There is no undo action so the synset has to be attached again if it was removed in error.

Initially the component is always empty because the GWAT database does not contain any annotations, but as − in the course of working with the tool − new semantic descriptions are generated they will be made visible in the component.

The annotations are displayed together in the component interface, sorted alphabetically and grouped based on their lexical type (noun, verb, adjective and adverb). The database is designed so that, at least in theory, there is no limit to the number of synsets that can be assigned to a single picture.

All attached synsets can be exported from the tool's database as a SQL script containing CREATE TABLE and

ADD ROW statements with identifications (IDs) of GAPED pictures and WordNet synsets associated with SQL foreign key relationship. Using the script it is possible to transfer the generated annotations to other applications or 3rd party systems compatible with SQL standard and WordNet semantic network.

V. USING WORDNET TO ANNOTATE GAPED PICTURES

WordNet search component is located on the right side of the main page. Search box is on the top of the component and a frame with search results below. The component is implemented by almost exclusively relying on JQuery JavaScript framework functionalities. The search starts on each keystroke, i.e. when browser registers keyboard press event, and depending on the user's computer performance several seconds are needed for results to be retrieved from the database. The waiting time is inversely proportional to the length of WordNet query − if only a few letters are used in search the retrieval process takes longer − and becomes progressively shorter as the query gets longer. The busy JavaScript animated icon is visible during retrieval execution indicating that the query has been accepted by the system and is being processed. With regards to the implementation the search is executed by JQuery JavaScript framework dynamically querying MySQL database, i.e. directly from the client to the server without any additional server-side PHP code.

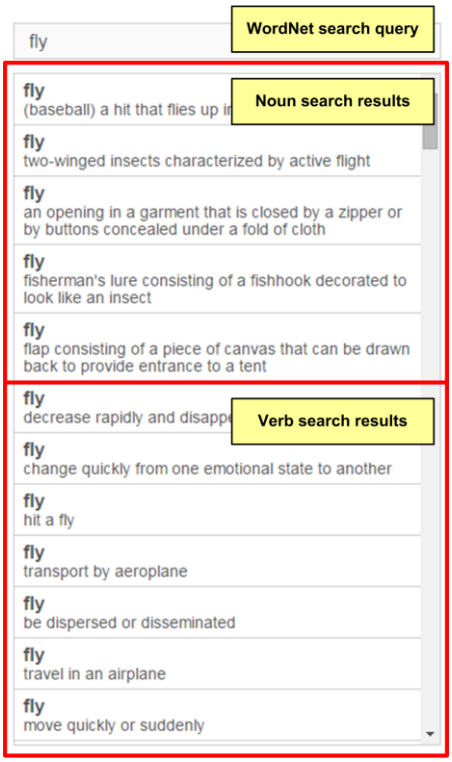


Fig. 3. WordNet search results for query "fly" in GWAT. Lexically identical synsets are grouped together.

The WordNet search results are presented as a vertical list with each row in a separate frame containing the name and description of one found synset. To improve the search component's usability synsets are sorted alphabetically and, additionally, those with the same lexical type are grouped together. If a search returns synsets of different lexical types they will be sorted in the following order: nouns first, adjectives second, verbs third and finally adverbs as the fourth group (Fig. 3). If many synsets are retrieved a vertical scroll bar will appear on the right side of the component allowing users to slide between different parts of the list.

When a synset is found with the desired semantic meaning user must click the row containing the synset with left mouse button to attach it to the currently displayed GAPED picture. This process is almost instantaneous, even on a computer with a modest performance, because the client code is short and on the server side only a single row has to be inserted in the table connecting the synset in WordNet table and the row in the table with GAPED pictures' names.

Once the synset attaching process is completed server-side, and the confirmation is received on the client, a new item is automatically generated in the WordNet annotations component for displaying the list of synsets describing picture semantics. The new item can be removed at any time by clicking on "X" button, if added in error, and a different synset can be attached instead.

VI. CONCLUSION

GWAT is web-based image annotation tool for manual identification of objects and events in GAPED images and annotation with WordNet knowledge taxonomy. Generated semantic annotations are stored in a relational database where they can be modified as needed and later exported to 3rd party tools. The purpose of GWAT is to enable a more detailed semantic description of GAPED pictures with a large and standardized linguistic ontology. There are many potential applications which benefit from rich multimedia semantics. Primarily, these can be used to build multimedia search tools that take advantage of WordNet semantic similarity measures to improve quality of the retrieved dataset. The similarity measures may also be used to rank the results relative to their significance. It has been already demonstrated with other affective multimedia databases that WordNet annotations significantly improve ranking accuracy and precision in picture retrieval [4]. Secondly, GWAT generated data provides an opportunity to study the relationship between semantics and emotion in multimedia. This line of research could establish statistically relevant relationships which can then be used to automatically infer emotion of a picture or video from their semantic description. If successful, such applications could eventually lead to machine-based emotion indexing of multimedia and emotional search engines [5].
However, both of these applications are impossible with the current state of GAPED semantics and can be investigated only with tools such as GWAT.

Web image repositories and tools for semantic annotation of pictures are fairly common, but there are not many tools for construction of affective multimedia databases. Therefore, the development of GWAT is interesting as a relatively new research area in the fields of applied computing and construction of information systems.

The future work will focus on development of new GAPED search tools which use WordNet descriptions generated with GWAT. In this regard we plan to develop a web-based search tool compatible with WordNet described GAPED. Furthermore, we would like to extend GWAT so it is possible to annotate image regions and associate WordNet description to individual blobs, features or objects in pictures. GAPED pictures are high-quality and contain a wide range of semantic content so detailed annotations could be used in computer vision research for automated object recognition. Additionally, it could be expected that such more detailed annotations would also further increase the quality of picture search.

REFERENCES

[1] G. W. Alpers, D. Adolph, and P. Pauli, "Emotional scenes and facial expressions elicit different psychophysiological responses," Int. J. Psychophysiol., vol. 80(3), pp. 173–81, 2011.

[2] M. M. Nielen, D. J. Heslenfeld, K. Heinen, J. W. Van Strien, M. P. Witter, C. Jonker, and D. J. Veltman, "Distinct brain systems underlie the processing of valence and arousal of affective pictures," Brain Cogn., vol. 71(3), pp. 387–96, 2009.

[3] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, "Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli," IEEE Trans. Inf. Technol. Biomed., vol. 14(3), pp. 589-97, 2010.

[4] M. Horvat, N. Bogunović, and K. Ćosić, "STIMONT: a core ontology for multimedia stimuli description," Multimedia Tools and Applications, vol. 73(3), pp. 1103−1127, 2014.

[5] M. Horvat, S. Popović, N. Bogunović, and K. Ćosić, "Tagging multimedia stimuli with ontologies," Proceedings of the 32nd International Convention MIPRO 2009, Croatia, pp. 203–208, May 2009.

[6] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Technical Report A−8, University of Florida, Gainesville, FL, 2008.

[7] M. Horvat, S. Popović, and K. Ćosić, "Multimedia stimuli databases usage patterns: a survey report," Proceedings of the 36nd International Convention MIPRO 2013, Croatia, pp. 993−997, May 2013.

[8] E. S. Dan-Glauser and K. R. Scherer, "The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance," Behavior Research Methods, vol. 43(2), pp. 468−477, 2011.

[9] G. A. Miller, "WordNet: a lexical database for English," Communications of the ACM, vol. 38(11), pp. 39−41, 1995.

[10] A. Torralba, B. C. Russell, and J. Yuen, "Labelme: Online image annotation and applications," Proceedings of the IEEE, vol. 98(8), pp. 1467−1484, 2010.

[11] M. Horvat, A. Grbin, and G. Gledec, "WNtags: A Web-Based Tool For Image Labeling And Retrieval With Lexical Ontologies," Frontiers in artificial intelligence and applications, vol. 243, pp. 585–594, 2012.

Evaluation of the Quality of Voice over IP Telephony Using the Decision Support System "Qualimetric Expertise"

R.S. Dorofeev*, A.S. Dorofeev* and O.A. Kostowskaya*
* Irkutsk National Research Technical University/Department of Computer Engineering, Irkutsk, Russian Federation
rdobermann@list.ru, dorbaik2007@mail.ru, koairk@mail.ru

Abstract – The paper considers the use of qualimetric expertise and ontological approach for evaluation of the quality of voice transmission over IP telephony in order to choose which of operators provides telecommunication services of the best quality. On the base of the suggested approach, an application the of the developed decision support system "Qualimetric Expertise" for automation of the process of voice transmission quality evaluation and identification of the possible causes of the quality loss is shown.

I. INTRODUCTION

Due to telephony, people from different cities and countries can always find out the latest news, the information about health of their relatives and just communicate. Telephony allows connecting people in spite of the distance. IP telephony today is one of the most promising and cheap means of communication in the world.

IP telephony is a telephonic communication over Internet Protocol. It means a number of communication protocols, technologies, and methods that provide the following typical for telephony things: dialing operation, call completion, two-way voice communication, and also video communication over the Internet or over any other IP networks. The signal is transmitted over the link channel in a digital form. As a rule, it is converted (compressed) before the transfer in order to remove the excess of information and to reduce the load on the data transmission network.

IP telephony is an application of a more general VoIP (Voice over IP) technology for two-way communication. The VoIP technology in general means all variants of voice transmission over IP including those not related to telephony and communication of people. For example, the VoIP technology is used for sound transmission in IP video monitoring systems, alarm systems, by broadcasting of webinars, by online film watching and so on [1].

The date of birth of IP telephony (or Internet telephony) is considered to be August 1996. Just then the company Vocal Nec (Israel) has completed a gateway testing. The gateway allows the connecting of the traditional phone network with the data transfer system in the Global Network. In fact, IP telephony is a technology that makes it possible to use the Internet as a means of organization of telephone conversations quite effectively [2].

A large number of operators provide customers with IP telephony services. The main selection criterion is for the end-user a price and quality correlation of provided telecommunication services in a certain direction.

Currently the challenge is how to make a quantitative evaluation of the quality of voice transmission over IP telephony for traffic grooming and selecting of the best operator providing telecommunication services, as well as for finding out of possible causes of quality loss with the aim of early detection and solving of telecommunication problems.

II. TRADITIONAL QUALITY EVALUATION OF IP TELEPHONY

The quality of IP telephony is evaluated today in a number of ways.

- An ACD (Average Call Duration) is a statistical parameter of telephony. It indicates an average call duration in one or other direction. A call detail record (CDR) is usually the basic data for calculation of the ACD value. Operator companies often use the ACD for an evaluation of the demand for directions, and indirectly for a telecommunication quality determination of a certain direction. For example, if in the given direction the average ACD value decreases, this may mean that subscribers do not like the telecommunication quality. The ACD is calculated by the formula: ACD = (the total duration of all calls in the given direction) / the number of completed calls.
- ASR (Answer Seizure ratio) and ABR (Answer Bid Ratio) are statistical parameters. They determine the quality of a telecommunication through a particular communications node (or IP telephony communications node) in a given direction. ASR and ABR are defined by the International Telecommunication Union (ITU) in "SG2 Recommendations: E.411: International Network Management – Operational Guidance". The ASR is calculated as a percentage ratio of the number of answered calls to the total number of

call attempts in the given direction. An engaged condition of the subscriber is considered as a failed call, as well as other situations with a call rejection. The calculated value of the ASR parameter may vary depending on the behavior of the called subscriber [3, 6].

- A MOS (Mean Opinion Score) is evaluation of the service quality by user on the 5-point scale and calculation of the arithmetical average. In practice, MOS – P.800 series can rarely be above 4.5 [4, 5, 6].
- An E-model is a multi-criteria evaluation of the quality (R-factor) in the range from 1 to 100 points [4].
- All these results (except for MOS) are obtained only through CDR (Call Detail Record) analysis [3, 6], after a number of calls in the given direction have been completed to provide objective data analysis and to obtain corresponding results. MOS calculation does not provide information on what criteria the quality does not conform. The authors believe that the mentioned methods of call quality evaluation by IP telephony are not objective as they do not take into account such parameters as disconnection during a call, false billing, use of various electronic secretaries by terminating party, voice disappearance during a call, noises, direction of a call to another subscriber. To eliminate these problems, the authors suggest to apply a quite new science, qualimetry. Qualimetry is applied for the quality evaluation of objects described by number of features.

III. USE OF ONTOLOGICAL APPROACH TO SOLVE THE TASKS OF QUALIMETRIC EXPERTISE

Qualimetry is a science dealing with a quantitative evaluation of an objects quality described by number of features. These features form a hierarchy. It means that one of them in the underlying level includes a number of features of the overlying level. The hierarchy forms a calculation model of an object.

A qualimetric expertise is one of the tools of integrated management system in companies. The difference between qualimetry and qualimetric expertise is that qualimetric expertise needs experts not only in the field of quality evaluation, but in the fields of development and application of the investigated (evaluated) object (variant).

In order to carry out the qualimetric expertise, a team of experts should make a calculation model of an object quality evaluation, that is a description of object properties at different hierarchy levels. The properties of the object have definite priorities for the object quality evaluation. There are morphological, functional, combined, integrated and mixed calculation models of evaluations.

It is usually selected a combined calculation model of an evaluation. It is presented in a form of combination of morphological and functional calculation models. The result of the expertise is a calculated integral performance index (in the range from 0 to 1). The qualimetric expertise is carried out at all stages of the life cycle of any product.

The concept of ontology is widely used in information science and artificial intelligence. The ontology in knowledge engineering is understood to be a precise specification of a certain subject area, which includes a glossary of this area and many logical connections between components of the glossary.

In the centre of ontologies are usually classes that describe concepts of a subject area. Slots (parameters) describe the properties of classes. An ontological analysis is based on a description of a system in terms of essentialities. The ontological analysis begins with a compilation of a glossary used in discussion and study of objects characteristics composing the considered system and also with a creation of a system of terms definitions [7, 8, 9].

As qualimetric expertise is based on the hierarchy of features, and ontology development is a construction of a features system, the authors suggest applying the ontology system as basic data for qualimetric expertise. As evaluated parameters for the qualimetric expertise of IP telephony, the authors take the following most significant criteria, which were found out after conducting a customer survey among customers using IP telephony in everyday life:

- disappearance of words of subscriber B
- sound disappearance during a conversation
- noises
- audibility of subscriber B
- calls control signals
- echo.

In accordance with the notation of IP telephony, the origination subscriber is a subscriber A, and the called subscriber is a subscriber B. As an editor of the ontology was selected the package OntoStudio. Fig. 1 shows the ontology developed for IP telephony quality evaluation. The ontology developed in OntoStudio and presented in markup Object Logic Language (OBL) contains a description of ontology classes, instances, their attributes and relations. The ontology in OBL allows us to make any adjustments, to change names, to add and delete information adhering to the overall structure of the ontology.

Let's explain every feature of the ontology. Call direction means a direction in which a subscriber A calls, for example, a call in the direction of Azerbaijan using the code 994 or in the direction of Kazakhstan using the code 77. Disappearance of words of a subscriber B means a disappearance of speech fragments (words, words ending) during a voice transmission. Sound disappearance during a conversation is an absolute absence of a voice traffic transmission. It means that one or both subscribers cannot hear each other after a time. Noises mean a side tone during a conversation. Audibility of a subscriber B is a level of a transmitting sound during a conversation. Calls control signals mean signals of a telephone exchange,

designed to assure an origination subscriber that a called subscriber's line is ringing, but a subscriber B does not pick up his phone yet. Echo is a repetition of words with a short delay. It makes an echo effect during a conversation. The developed ontology graphically demonstrates list of features and model of the evaluated object.

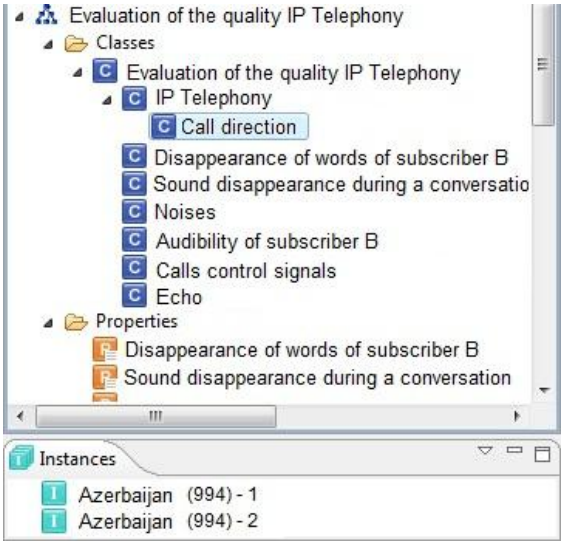


Figure 1. The ontological features system for IP telephony

Subscribers A will score these parameters on the scale from 1 to 5, where 1 means "dissatisfied", 5 – "completely satisfied". The polling will follow the termination of a conversation. It is implemented by means of the IVR menu. IVR is an interactive voice response, a computer system. It gives the possibility to the user, usually subscriber, to select options from a voice menu and interact with the computer system [3]. The authors have developed the decision support system (DSS) "Qualimetric Expertise" for calculation of the total quality evaluation and diagnosis of causes of the quality loss in the automated mode. The ontological system of features is the basic data for the expertise. In order to transform the ontology in OBL, the authors have developed the application ONT-Converter. Further, the result is imported in the DSS "Qualimetric Expertise". Fig. 2 shows the result of the import of characteristics for quality evaluation.

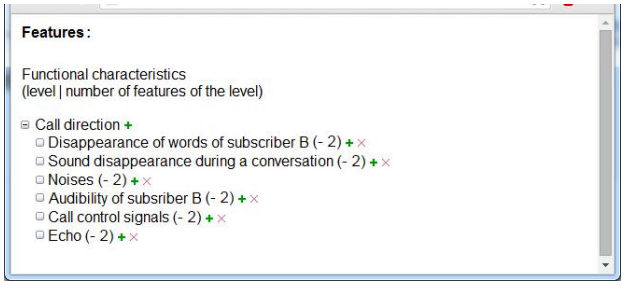


Figure 2. The features ontology in DSS "Qualimetric Expertise".

The results received from subscribers A in every direction are stored as the average values of the features indexes scaled to the qualimetric scale by (1) and (2) and used in the quality evaluation.

$$k_s = (val_s - ng)/vg, \tag{1}$$

$$ks_s = (k_s - ng)/(vg - ng), \tag{2}$$

where *val* = index value entered by the user by means of IVR, *vg* = upper bound (in our case is 5), *ng* = lower bound (in our case is 1).

Subscribers A will make calls to Azerbaijan through two different line operators using the code 994. To provide customers with better telecommunication, it is necessary to identify which line operator provides the best telecommunication quality in this direction.

After calculation of the integral characteristics in hierarchy levels by (3) – (9) and calculation of the resulting integral evaluations by (10) – (13) the following characteristics shown on the Fig. 3 are obtained.

$$znkv_j = \sum_{i=1}^{m} val_i, \tag{3}$$

where *znkv* = sum of index values for all subscribers A, *m* = quantity of subscribers A participated in the survey, *val* = index value on a hierarchy level.

$$sum = \sum_{j=1}^{k} znkv_j, \tag{4}$$

where *sum* = sum over all indexes, *k* = quantity of indexes.

$$znkv'_j = \frac{znkv_j}{sum}, \tag{5}$$

$$QS = \sum_{j=1}^{k} (znkv'_j * ves_j), \tag{6}$$

$$DQS = 1 - QS, \tag{7}$$

where *QS* = quantitative value of the object quality conformance to the requirements calculated as a sum on the one hierarchy level, *DQS* = quantitative value of the object nonconformance to the requirements, *ves* = importance factor of a feature by evaluation; in our case all importance factors have the same value, which is equal to 1/6 = 0.1666.

$$QP = \prod_{j=1}^{k} znkv'^{ves_j}_j, \tag{8}$$

$$DQP = 1 - QP, \tag{9}$$

where *QP* and *DQP* = quantitative value calculated as a product.

$$SQS = \frac{\sum_{g=1}^{f} QS}{f - 1}, \tag{10}$$

$$SDQS = \frac{\sum_{g=1}^{f} DQS}{f-1}, \quad (11)$$

$$SQP = \frac{\sum_{g=1}^{f} QP}{f-1}, \quad (12)$$

$$SDQP = \frac{\sum_{g=1}^{f} DQP}{f-1}, \quad (13)$$

where f = maximal quantity of hierarchy levels $f \neq 1$, *SQS* (*Q*) = total quantitative value of the object quality conformance to the requirements calculated as a sum, *SDQS* (*DeltaQ*) = quantitative value of the object nonconformance to the requirements calculated as a sum, *SQP* (*Q*) = total quantitative value of the object quality conformance to the requirements calculated as a product, *SDQP* (*DeltaQ*) = quantitative value of the object nonconformance to the requirements calculated as a product.

Figure 3. The result of combined characteristics calculation for both operators

Q in this instance is the quantitative value of the object quality conformance to the requirements. *DeltaQ* is the quantitative value of the object nonconformance to the requirements. The calculation of integral characteristics is possible by the sum or by the product.

IV. Analysis of Results

After we have obtained the results, it is possible to carry out an analysis of them to discover deviations in the features values from the features of the object with the best parameters. For this purpose, the function "Analysis of Results" implemented in the application is used. Fig. 4 shows the result of the function.

Analysis of Results

Name of parameters	Level	Azerbaijan (994) -1	Azerbaijan (994) -2	Comment
Disappearance of words of subscriber B	2	0.75	0.84	Azerbaijan (994)-2 provides the better telecommunication quality. Recommended.
Sound disappearance during a conversation	2	0.63	0.8	
Noises	2	0.5	0.91	
Echo	2	0.72	0.88	

Figure 4. The result of analysis of obtained data

As it is seen from the results, the second operator provides the better telecommunication quality in the direction of Azerbaijan. It is reasonable to send a voice traffic using the equipment of this operator. After the comparison of the results, the DSS makes the analogical decision from the knowledgebase by correlation of date from several operators concerning the quality of voice transmission over IP telephony.

In this way, the developed decision support system with use of ontology and qualimetric expertise allows us to make the quantitative evaluation of the voice transmission quality over IP telephony for traffic grooming, as well as to identify the possible causes of the quality loss through the visual analysis of comparative results.

References

[1] Official website of the company «Multimedia Opportunities» [Electronic resource]. Access mode: http://www.prochor.ru/content/uslugi/voip-telefoniya/, free.

[2] Information resource «Computer Technologies Bureau» [Electronic resource]. Access mode: http://www.bcot.ru/prjeimushhjestva_i_njedostatki_ip-tjeljefonii.html, free.

[3] Official website of the company «Setion Ltd» [Electronic resource]. Access mode: http://gsm2.biz/article/definition-voip/, free.

[4] IP telephony organisation scenarios, protocols, voice processing and voice transmission algorithms, QoS [Electronic resource]. Access mode: http://www.seti.sut.ru/admin61/editor_files/file_upload/t15_12.pdf, free.

[5] Information resource «OpenNET» [Electronic resource]. Access mode: http://www.opennet.ru/docs/RUS/voip_asterisk/1.html#_ftn14, free.

[6] ITU-T Recommendation P.564: Conformance testing for voice over IP transmission quality assessment models. - Geneva, 2008.

[7] G.G. Azgaldov, A.V. Kostin, V.V. Sadovov, Qualimetry: basic information. Reference book with an example for ANO «Agency of strategic initiatives for promotion of new projects.», Study Guide, Publishing Company «Vysshaya shkola», Moscow, 143 pgs., 2010.

[8] G.G. Azgaldov, V.M. Marugin, Qualimetric expertise. Handbook of expertise orgnization and carrying out qualimetric calculations. Calculation models of quality. Vol. II, Publishing Company «Russkiy Registr», Saint Petersburg, Moscow, 517 pgs., 2002.

[9] R.S. Dorofeev, «Methodology and software implementation of ontology and qualimetric expertise sharing by machine tools quality evaluation.», Vestnik of Irkutsk State Technical University, No. 3 (74), Irkutsk, pgs. 16 – 23, 2013.

INTELLIGENT SYSTEMS

A New Ensemble of Features for Breast Cancer Diagnosis

İ. Işıklı Esener*, S. Ergin ** and T. Yüksel*
* Bilecik Seyh Edebali University/Electrical-Electronics Engineering, Bilecik, Turkey
** Eskisehir Osmangazi University/ Electrical-Electronics Engineering, Eskisehir, Turkey
idil.isikli@bilecik.edu.tr, sergin@ogu.edu.tr, tolga.yuksel@bilecik.edu.tr

Abstract - In this paper, an automatic Computer Aided Diagnosis (CAD) system is completely designed for breast cancer diagnosis and it is verified on a publicly available mammogram dataset constructed during Image Retrieval in Medical Applications (IRMA) project. This database comprises three different patch types indicating the health status of a person. These types are normal, benign cancer, and malignant cancer and they are labeled by the radiologists for the IRMA project. In the realization of CAD system, all mammogram patches are firstly pre-processed performing a histogram equalization followed by Non-Local Means (NLM) filtering. Then, the Local Configuration Pattern (LCP) algorithm is performed for feature extraction. Besides, some statistical and frequency-domain features are concatenated to LCP-based feature vectors. The obtained new feature ensemble is used with four well-known classifiers which are Fisher's Linear Discriminant Analysis (FLDA), Support Vector Machines (SVM), Decision Tree, and *k*-Nearest Neighbors (k-NN). A maximum of 94.67% recognition accuracy is attained utilizing the new feature ensemble whereas 90.60% was found if only LCP-based feature vectors are used. This consequence obviously reveals that the new feature ensemble is more representative than an LCP-based feature vector by itself.

I. INTRODUCTION

Breast cancer is one of the major causes of female deaths all over the world, but mortality can be reduced with early diagnosis [1]. Although mammography is known as the most effective radiological imaging technique both for breast examination and diagnosis, mammogram images contain little information since X-ray permeabilities of normal and cancerous regions are not very different [2]. Besides, important issues on mammograms may be overlooked during interpretation of mammography depending on breast tissue type or human factors. Hence, usage of a Computer Aided Diagnosis (CAD) system, which detects and diagnoses cancer, gives smoother view to radiologists and allows reconsideration of mammography resulting in increased sensitivity of cancer detection.

A CAD system consists of four stages. The first of these stages is a pre-processing stage where some filtering operations for noise reduction, contrast enhancement operations, and structural improvement operations like label and pectoral muscle removal from mammogram images are realized. Papadopoulos et al. compared five different image enhancement algorithms for microcalcification detection: Contrast Limited Adaptive Histogram Equalization (CLAHE), Local Range Modification (LRM), Redundant Discrete Wavelet (RDW), Wavelet Linear Stretching (WLST), and Wavelet Shrinkage (WSRK), then the best performance is obtained by applying LRM technique with $A_z = 0.932$ where A_z is defined as the area under the Receiver Operating Characteristic (ROC) curve [3]. Wavelet transform [4], polygon modeling [5], Wiener filtering [6], morphological operations [6-8], Hough transform [7], and histogram processing operations [7, 9-10] are widely used image enhancement and noise reduction methods.

In the second stage, suspicious regions for breast cancer, Region of Interest (ROI), are detected from the whole breast mammograms. Wavelet transform [11], segmentation algorithms [12-13], edge operators [14] are the samples of commonly used methods for lesion detection in the past studies. Moayedi et al. detected ROIs by thresholding intensity variances. In this method, regions with lesions have low intensity variance, and the high frequency components of these regions are close to zero beside the energy of lesion boundaries are higher than non-cancerous regions [10].

Efficiency of a CAD system highly depends on feature extraction stage, the third stage, in which mammogram data is represented by the most discriminative features in a low-dimensional feature space. Feature extraction for mammogram images can be performed using statistical methods [4, 15-17], model-based methods [18-19], signal processing methods that compute tissue features due to pixel characteristics [3, 17] or frequency spectrum [9-10, 15, 20] of an image, and mammographic features [5, 7, 21-23] like shape, spicule index, contour, size, density and brightness.

Finally, a classification stage is realized using the features obtained in the third stage. Artificial neural networks [3-4, 18, 20, 22], support vector machines [4, 7-10, 16, 24], subspace learning algorithms [8, 16-17, 24], Bayes, decision tree, and *k*-nearest neighbor classifiers [4, 16] are the most common classifiers used for mammogram classification.

In this paper, it is aimed to prepare a CAD system for breast cancer diagnosis. For this purpose, the mammogram images in the database are firstly pre-processed performing a histogram equalization followed by Non-Local Means (NLM) filtering as described in Section 2. Secondly, feature extraction is realized as explicitly explained in Section 3. The classification methods and the experimental study performed in this paper are given in Sections 4 and 5, respectively. Finally, conclusions are presented in Section 6.

II. PRE-PROCESSING

Radiologists basically face two important problems during analyzing mammogram images. These problems are originated from low-contrast and noise issues. Therefore, contrast enhancement and noise reduction operations are needed in order to investigate mammogram images. In this study, a histogram equalization performed for contrast enhancement is followed by NLM filtering for noise reduction.

A. Histogram Equalization

Histogram equalization maximizes image contrast by applying a gray-level transform that tries to flatten the histogram of an image [25]. Therefore, histogram equalization is an essential step for mammogram classification problem since brightness is one of the most critical concerns for mammograms.

B. Mammogram Image Enhancement

Non-Local Means (NLM) filter is an adaptive smoothing filter that changes the window size according to the similarity between neighborhoods of any two pixels as well as preserves the fine details by computing a weighting function according to the derivatives in the corresponding search window [26]. Given a discrete noisy image $v = \{v(i) \mid i \in I\}$, the filtered value $NL[v(i)]$ of any pixel i is computed as:

$$NL[v(i)] = \sum_{j \in I} w(i,j) \cdot v(j) \tag{1}$$

where $w(i,j)$ refers to the weight coefficient computed utilizing the similarity between the pixels i and j, and satisfies the conditions $0 \le w(i,j) \le 1$ and $\sum_j w(i,j) = 1$.

The above mentioned similarity between pixels i and j is measured as the Gaussian weighted Euclidean distance $\|v(N_i) - v(N_j)\|_{2,\pm}^2$, where $\pm$ ($\pm > 0$) is the standard deviation of the Gaussian kernel; $v(N_i)$ and $v(N_j)$ are the neighborhoods of pixels i and j in the similarity window. The pixels with a similar neighborhood have larger weights computed as in Eq. (2). $Z(i)$ and h in the Eq. (2) refer to the normalizing constant and the degree of filtering, respectively.

$$w(i,j) = \frac{1}{Z_i} \cdot e^{-\frac{\|v(N_i) - v(N_j)\|_{2,\pm}^2}{h^2}} \tag{2}$$

$$Z_i = \sum_j e^{-\frac{\|v(N_i) - v(N_j)\|_{2,\pm}^2}{h^2}} \tag{3}$$

III. FEATURE EXTRACTION

Feature extraction is the most crucial stage in CAD systems because the little differences between any normal and abnormal tissues on mammogram images play very important roles. In this study, Local Configuration Pattern (LCP) algorithm is used for feature extraction with additional statistical and frequency-domain features.

A. Local Configuration Pattern

Local Configuration Pattern (LCP) is a method that describes the local structural information and pattern occurrences together, and it forms rotation-invariant feature vectors [27].

Local structural information is found by Local Binary Pattern (LBP) algorithm that labels every pixels of image I comparing its P neighbors in the circular neighborhood with radius R as:

$$LBP(P,R) = \sum_{i=0}^{P-1} u(g_i - g_c) \cdot 2^i \tag{4}$$

$$u(x) = \begin{cases} 1, x \ge 0 \\ 0, x < 0 \end{cases} \tag{5}$$

where g_i and g_c refer to the intensities of neighboring pixel i and the central pixel (c), respectively. Rotation-invariant LBP vectors are obtained by rotating each bit pattern circularly to a minimum value so that the maximum value will be the last element of the vector. Eq. (6) defines the rotation-invariant LBP vector in where U is the uniformity measure identified by the bit-wise transitions.

$$LBP^{riu2}(P,R) = \begin{cases} \sum_{i=0}^{P-1} u(g_i - g_c) \text{ , } U(LBP(P,R)) \le 2 \\ P+1 \qquad\qquad \text{, otherwise} \end{cases} \tag{6}$$

In some cases, LBP representations may be the same although the neighborhoods are quite different due to the quantization of gray-level differences to binary levels. In order to handle this problem, local variance (VAR) of each pattern is computed, and the joint histogram (O) is obtained.

$$VAR = \frac{1}{P} \cdot \sum_{i=0}^{P-1} (g_i - \mu)^2 \tag{7}$$

$$O = \frac{LBP^{riu2}}{VAR}, \tag{8}$$

where « is the average intensity of the neighboring pixels.

In case of LBP^{riu2} vectors with equal variances, same feature vectors are obtained although patterns have different configurations. In order to resolve this problem, the microscopic configurations of the patterns are modeled. In this model, the optimal weights (A_L) of neighboring pixels are calculated using the least square estimation to form the central pixel. Then, 1D Fourier transform is applied on A_L and H_L values are obtained to make the optimal weight vectors rotation-invariant. The magnitude of H_L is the microscopic feature of the pattern. LCP vector of any pattern is obtained by associating local structures with microscopic configuration as:

$$LCP = \left[\left[H_0\right];O_0\right],\left[\left[H_1\right];O_1\right],\cdots;\left[\left[H_{q-1}\right];O_{q-1}\right], \quad (9)$$

where q is the number of patterns.

B. Statistical Features

In addition to the LCP vectors, some statistical and frequency-domain features are also computed to increase the data representability of the feature vectors. Statistical features utilized in this study are the energy, mean, variance, maximum, minimum, standard deviation, skewness, kurtosis, area-descriptor, mean energy, energy variance and entropy value of any LCP vector. Among all statistical features, area-descriptor is computed as :

$$Area-descriptor = \frac{\pm}{\ll} \quad (10)$$

where ± and « refer to the standard deviation and the mean of an LCP vector, respectively.

C. Frequency – Domain Features

Frequency-domain features are obtained from the 2-level Two-Dimensional Discrete Wavelet Transform (2D-DWT) of the pre-processed mammogram images using Daubechies1 (db1) wavelet and therefore 16 sub-bands are obtained. The energy values of each sub-band are calculated on account of the fact that brightness is one of the most vital issues for breast cancer diagnosis.

IV. CLASSIFICATION METHODS

Four different classifiers which are Fisher's Linear Discriminant Analysis (FLDA), Support Vector Machines (SVM), Decision Tree, and *k*-Nearest Neighbors (k-NN) are performed in this study.

A. Fisher's Linear Discriminant Analysis

FLDA tries to find the projection matrix that projects the training data onto a low-dimensional space which maximizes between-class variance as well as minimizing within-class variance [28]. This is known as the Fisher maximization criterion and defined as:

$$J(\vec{w}) = \frac{\vec{w}^T \cdot S_B \cdot \vec{w}}{\vec{w}^T \cdot S_w \cdot \vec{w}} \quad (11)$$

where $\vec{w}$, S_B and S_w refer to the projection vectors, between-class and within-class scatter matrices, respectively.

On the test stage of FLDA, any test vector is projected via $\vec{w}$ projection vectors and the distances to the training vectors on the low-dimensional space are calculated. Decision criterion for FLDA is given as:

$$K = \arg\min_{1\le c\le S}\left\{\left\|\vec{\Omega}^c - \vec{\Omega}_{test}\right\|\right\} \quad (12)$$

where c is the class index, S is the total number of classes, $\vec{\Omega}^c$ and $\vec{\Omega}_{test}$ are the projected training vector of the c^{th} class and the projected test vector, respectively.

B. Support Vector Machines

SVM, also known as maximum margin classifier, aims to find the optimal hyperplane which maximizes the distance between the hyperplane and the support vectors [29]. Support vectors are the training vectors nearest from each class to the hyperplane [30]. As it can classify linearly separable data, SVM can classify non-linear data by transforming the data to a higher-dimensional space by using an appropriate kernel function [30]. If the training set is $TS = \{(\vec{x}_1, L_1), (\vec{x}_2, L_2), \cdots, (\vec{x}_M, L_M)\}$ for a two-class problem, where $\vec{x}_i$ ($i = 1, 2, \cdots, M$) is the training data and L_i ($L_i \in \{-1, 1\}$) is the class label, test vector is classified according to the sign of the function given as:

$$f(\vec{x}_{test}) = \sum\left\{\sim_i \cdot L_i \cdot \left(\vec{x}_i^T \cdot \vec{x}_{test}\right) + b\right\} \quad (13)$$

where $\sim_i$ ($i = 1, 2, \cdots, M$) are the non-zero quadratic coefficients, and $\left(|b| / \|\vec{w}\|\right)$ is the perpendicular distance between the hyperplane and the origin while $\vec{w}$ is the normal vector of the separating hyperplane.

C. Decision Tree

The principle of the Decision Tree classifier is to cluster any data into sub-groups until all elements of any sub-group have the same class label [31]. The branches and the leaves of a tree in this classifier refer to the classification rules and the class labels, respectively. In the training stage, classification rules are defined by clustering data into the leaves. In the test stage, those rules are applied to any test sample and the leaf which the test sample reaches, gives the class label of test sample.

D. k – Nearest Neighbors

k-NN classifier categorizes any test vector considering the distances between test and training vectors in the feature space [32]. This classifier assigns a test vector to the respective class that its k-nearest neighbors belong at most. Although it is obvious that classification performance is directly related to parameter k, there is no obvious information on the selection of k except that it should be positive and not be a multiple of the total number of classes.

V. EXPERIMENTAL STUDY

In this paper, a CAD system that classifies mammogram images into normal, benign, and malignant classes is proposed. At first, histogram equalization is applied to the mammogram images, then this process is followed by NLM filtering. Secondly, the LCP feature vectors, statistical and frequency-domain features are obtained to construct the feature vectors. Finally, these feature vectors are classified with four different classifiers; SVM (using linear and quadratic kernels), FLDA, Decision Tree, and k-NN.

A. Database

It has great importance to work on the images with their ground truths for medical imaging applications. In this study, a publicly available mammogram dataset constructed during IRMA project is used [33]. This dataset consists of 12 classes defined by Breast Imaging Reporting and Data System (BI-RADS). These classes are four breast tissue classes (fatty, fibro-glandular, heterogeneously dense, and extremely dense) and three health status classes (normal, benign cancer, and malignant cancer) for each breast tissue type. It has 233 mammogram patches at a resolution of 42-50 μm/pixel for each class and therefore totally 2796 patch images are collected in the dataset [33].

In this paper, a database of 600 patches from fatty tissues is used. Each of normal, benign cancer, and malignant cancer cases has 200 patches. Sample patches of size 128x128 of each class on the used part of the IRMA database are shown in Fig.1. The rows of the Fig.1 illustrate five different samples of the normal, benign cancer, and malignant cancer cases, respectively.

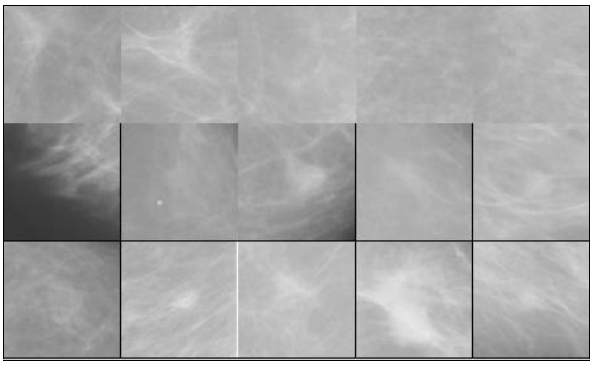


Figure 1. Sample mammographic patches

These patches are classified using 10-fold cross-validation technique. It implicitly means that 180 patches in each class are used for training while the remaining 20 patches are treated as the test patches. This scheme is repeated ten times in the cross-validation technique.

B. Feature Vector Construction

The key idea is to extract rotation-invariant features for obtaining representative feature vectors for breast cancer diagnosis. In this study, the LCP technique is used for rotation-invariant feature extraction. In order to extract these LCP-based feature vectors, firstly, pre-processed mammogram patches are decomposed into four sub-bands that are LL (low-low), LH (low-high), HL (high-low) and HH (high-high) by 1-level 2D-DWT utilizing db1 wavelet. Therefore, 81x1 dimensional LCP vectors of each sub-band are constructed. The endmost values in those LCP vectors are appreciably high, so they are removed to get rid of their domination over other features. The rest 80-dimensional feature vectors of each sub-band are then weighted with the coefficients shown in Table I; and they are summed up to form an 80-dimensional feature vector for each mammogram patch.

TABLE I. COEFFICIENTS OF EACH SUB-BANDS AND CORRESPONDING LCP CODES

LCP Code	LL-LH-HL-HH Coefficients
LCP 1-1-1-0	1-1-1-0
LCP 1,4-1-1-0	1,4-1-1-0
LCP 1,8-1-1-0	1,8-1-1-0

In the weighting process, HH sub-bands are completely ignored by using 0-coefficient since they introduce the noise component of the mammogram patches. Similarly, LL sub-bands, namely approximations in Wavelet theory, are the most meaningful components of the patches. Hence, effect of LL sub-bands is examined by three different coefficients used in [2].

In order to increase the representative power of feature vectors, 12 statistical features computed from the LCP feature vectors, and 16 frequency-domain features are concatenated to the LCP feature vectors. Consequently 108-dimensional feature vectors are obtained for each mammogram patch. Table II summarizes the feature vector construction process.

TABLE II. FEATURE VECTOR CONSTRUCTION PROCESS

108-dimensional Feature Vector Content		
LCP-based Feature Vector	**Statistical Features**	**Frequency Domain Features**
80 x 1	***12 x 1***	***16 x 1***
The 80-dimensional LCP-based Feature Vector	LCP: Energy LCP: Mean LCP: Variance LCP: Maximum LCP: Minimum LCP: Standard Deviation LCP: Skewness LCP: Kurtosis LCP: Area-Descriptor LCP: Mean Energy LCP: Energy Variance LCP: Entropy	LLLL: Energy LLLH: Energy LLHL: Energy LLHH: Energy LHLL: Energy LHLH: Energy LHHL: Energy LHHH: Energy HLLL: Energy HLLH: Energy HLHL: Energy HLHH: Energy HHLL: Energy HHLH: Energy HHHL: Energy HHHH: Energy

In the Table II, the phrase “LCP: Energy” means the energy value of an LCP vector whereas the “LLLL: Energy” is the energy of the LLLL (low-low-low-low) sub-band.

C. Performance Evaluation

Four different classification methods are performed on three different feature vector set for the CAD system in this study. Average classification accuracies of classifiers obtained by 10-fold cross-validation using "LCP 1-1-1-0", "LCP 1,4-1-1-0", and "LCP 1,8-1-1-0" coded feature vectors are shown in Table III.

The phrases "SVM ('p', 1)" and "SVM ('d', 2)" in the Table III are the SVM classifiers using linear and quadratic kernels, respectively.

TABLE III. AVERAGE CLASSIFICATION ACCURACIES OF CLASSIFIERS OBTAINED BY 10-FOLD CROSS-VALIDATION USING "LCP 1-1-1-0", "LCP 1,4-1-1-0", AND "LCP 1,8-1-1-0" CODED FEATURE VECTORS

CLASSIFIERS	FEATURE VECTOR CODES		
	LCP 1-1-1-0	***LCP 1,4-1-1-0***	***LCP 1,8-1-1-0***
SVM ('p', 1)	94.00	94.17	94.67
SVM ('d', 2)	94.33	94.00	84.33
FLDA	94.00	94.17	94.17
Decision Tree	86.33	86.17	84.50
k-NN (k=5)	45.67	49.50	48.83

VI. CONCLUSION

In this paper, a Computer Aided Diagnosis (CAD) system is entirely designed and verified on the mammogram images with their foreknown ground truths for breast cancer diagnosis. The benchmark mammogram database includes three different patch cases indicating the health status of a person. These cases are labeled as normal, benign cancer, and malignant cancer. The LCP algorithm is performed by itself for several image processing applications. Moreover, it is recently used in a biomedical image processing study [2], but LCP is not adequate to more correctly classify mammogram patches since it can be affected from various issues. Therefore the usage of LCP by itself does not result in representative features for a mammogram image. Not only discriminative but also descriptive features are to be computed in the feature extraction stage. Twelve statistical features calculated from the LCP feature vectors, and 16 frequency-domain features are concatenated to the LCP-based feature vectors. A maximum of 94.67% recognition accuracy is attained for LCP 1,8-1-1-0 coded feature vectors whereas 90.60% was obtained [2] if only LCP-based feature vectors are used. This outcome explicitly shows that the new feature ensemble is more representative than an LCP-based feature vector by itself. Furthermore, the linear SVM classifier (94.67%) is more successful than the other ones when the attained recognition accuracies are extensively examined. This consequence is another outcome for this study in order to be taken into consideration. Other feature types and combinations are planned to be exhaustively investigated as future works.

ACKNOWLEDGMENT

The database utilized in this study was used by the courtesy of Thomas M. Deserno, Department of Medical Informatics, Division of Image and Data Management, Aachen, Germany.

REFERENCES

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," CA: A Cancer Journal for Clinicians, vol. 61, no.2, pp. 69-90, 2011.

[2] S. Ergin and O. Kılınç, "A New Feature Extraction Framework Based on Wavelets for Breast Cancer Diagnosis," Computers in Biology and Medicine, vol.51, pp. 171–182, 2014.

[3] A. Papadopoulos, D.I. Fotiadis, and L. Costaridou, "Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques," Computers in Biology and Medicine, vol.38, pp. 1045-1055, 2008.

[4] M. Radovic, M. Djokovic, A.r Peulic, and N. Filipovic, "Application of data mining algorithms for mammogram classification," 13th IEEE International Conference on Bioinformatics and Bioengineering, 2013.

[5] D. Guliato, R. M. Rangayyan, J. D. Carvalho, and S. A. Santigo, "Polygonal modelling of contours of breast tumors with the preservation of spicules," IEEE Transactions on Biomedical Engineering, vol. 55, no. 1, pp. 14-20, 2008.

[6] D.C. Pereira, R.P. Ramos, and M.Z. Nascimento, "Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm," Computer Methods and Programs in Biomedicine, vol. 114, no.1, pp. 88-101, 2014.

[7] W. B. Sampaio, E. M. Diniz, A. C. Silva, A. C. de Paiva, and M. Gattass, "Detection of masses in mammogram images using CNN, geostatistic functions and SVM," Computers in Biology and Medicine, vol.41, pp. 653–664, 2011.

[8] X. Zhang and X. Gao, "Twin support vector machines and subspace learning methods for microcalcification clusters detection," Engineering Applications of Artificial Intelligence, vol.25, pp. 1062–1072, 2012.

[9] P. Agrawal, M. Vatsa, and R. Singh, "Saliency based mass detection from screening mammograms," Signal Processing, vol. 99, pp. 29-47, 2014.

[10] F. Moayedi, Z. Azimifar, R. Boostani, and S. Katebi, "Contourlet-based mammography mass classification using the SVM family," Computers in Biology and Medicine, vol.40, pp. 373–383, 2010.

[11] M. J. Lado, C. Cadarso-Suárez, J. Roca-Pardiñas, and P. G. Tahoces, "Categorical variables, interactions and generalized additive models. Applications in computer-aided diagnosis systems," Computers in Biology and Medicine, vol.38, pp. 475-483, 2008.

[12] Dr. H. B. Kekre, T. Sarode, and S. Gharge, "Tumor detection in mammography images using vector quantization technique," International Journal of Intelligent Information Technology Application, vol. 2,no. 5, pp. 237-242, 2009.

[13] Dr. H. B. Kekre, Dr. T. Sarode, Ms. S. Gharge, and Ms. K. Raut, "Detection of cancer using vector quantization for segmentation," International Journal of Computer Applications, vol. 4, no. 9, pp. 14-19, 2010.

[14] W. Haider, M. Sharif, and M. Raza, "Achieving accuracy in early stage tumor identification systems based on image segmentation and 3d structure analysis," Computer Engineering and Intelligent Systems, vol. 2, no. 6, pp. 96-103, 2011.

[15] R. P. Ramos, M. Z. do Nascimento, and D. C. Pereira, "Texture extraction: an evaluation of ridgelet, wavelet and co-occurrence based methods applied to mammograms," Expert Systems with Applications, vol. 39, pp. 11036-11047, 2012.

[16] N. Vallez, G. Bueno, O. Deniz, J. Dorado, J. A. Seoane, A. Pazos, and C. Pastor, "Breast density classification to reduce

false positives in CADe systems," Computer Methods and Programs in Biomedicine, vol. 113, no. 2, pp. 569-584, 2013.

[17] J.-B. Li, Y.-H. Wang, S.-C. Chu, and J. F. Roddick, "Kernel self-optimization learning for kernel-based feature extraction and recognition," Information Sciences, vol. 257, pp. 70-80, 2014.

[18] N. R. Pal, B. Bhowmick, S. K. Patel, S. Pal, and J. Das, "A multi-stage neural network aided system for detection of microcalcifications in digitized mammograms," Neurocomputing, vol. 71, pp. 2625-2634, 2008.

[19] S. K. Biswas and D. P. Mukherjee, "Recognizing architectural distortion in mammogram: a multiscale texture modeling approach with GMM," IEEE Transactions on Biomedical Engineering, vol. 58, no. 7, pp. 2023-2030, 2011.

[20] A. N. Karahaliou, I. S. Boniatis, S. G. Skiadopoulos, F. N. Sakellaropoulos, N. S. Arikidis, E. A. Likaki, G. S. Panayiotakis, and L. I. Costaridou, "Breast cancer diagnosis: Analyzing texture of tissue surrounding microcalcificatons," IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 6, pp. 731-738, 2008.

[21] A. Keleş, A. Keleş, and U. Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," Expert Systems with Applications, vol.38, pp. 5719–5726, 2011.

[22] B. Verma, P. McLeod, and A. Klevansky, "A novel soft cluster neural network for the classification of suspicious areas in digital mammograms," Pattern Recognition, vol.42, pp. 1845—1852, 2009.

[23] M. M. R. Krishnan, S. Banerjee, C. Chakraborty, C. Chakraborty, and A. K. Ray, "Statistical analysis of mammographic features and its classification using support vector machine," Expert Systems with Applications, vol.37, pp. 470–478, 2010.

[24] K. Ganesan, U. R. Acharya, C. K. Chua, C. M. Lim, and K. T. Abraham, "One-class classification of mammograms using trace transform functions," IEEE Transactions on Instrumentation and Measurement, vol. 63, no. 2, pp. 304-311, 2014.

[25] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 2nd ed., New Jersey: Prentice Hall, 2002, pp. 91-94.

[26] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," Multiscale Modelling and Simulation, vol. 4, no. 2, pp. 490-530, 2005.

[27] Y. Guo, G. Zhao, and M. Pietkäinen, "Texture classification using a linear configuration model based descriptor," British Machine Vision Association, vol. 119, pp. 1-10, 2011.

[28] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol. 7, no. 2, pp. 179-188, 1936.

[29] O. Chapelle, P. Haffner, and V. N. Vapnik, "Support vector machines for histogram-based image classification," IEEE Transaction on Neural Networks, vol. 10, no. 5, pp. 1055-1064, 1999.

[30] K. Özkan, S. Ergin, Ş. Işık, and İ. Işıklı, "A new classification scheme of plastic wastes based upon recycling labels," Waste Management, vol. 35, pp. 29-35, 2015.

[31] S.R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Transactions on Systems Man and Cybernetics, vol. 21, pp. 660-674, 1991.

[32] E. Fix and J. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," Report 4, Project 21-49-004, US Air Force School of Aviation Medicine, Randolph Field, 1952.

[33] T. M. Deserno, J. E. E. de Oliveira, and A. A. Araujo, "Towards computer-aided diagnostics of screening mammography using content-based image retrieval," 24. SIBGRAPI Conference on Graphics, Patterns and Images, pp. 211-219, 2011.

Preliminary study on detection of atrial fibrillation during oscillometric blood pressure measurement

M. Meža*
* Faculty of Electrical Engineering, University of Ljubljana, User-adapted Communications & Ambient intelligence Lab - LUCAMI, Ljubljana, Slovenia
marko.meza@fe.uni-lj.si

Abstract - The paper presents a prototype method for detection of cardiac arrhythmia from pneumatic cuff pressure course, obtained during oscillometric blood pressure measurement. Oscillometric blood pressure measurement is an indirect method for blood pressure measure where pneumatic cuff is placed on patient's extremity, usually upper arm, inflated and then gradually deflated. During deflation phase the pressure oscillations in the cuff caused by squeezed blood vessels are analyzed to determine patient's blood pressure. Since presence of arrhythmia can be detected through observation of heart rhythm, it is possible to detect arrhythmia during blood pressure measurement. We have captured cuff pressure courses – samples from 34 patients of which 6 had arrhythmia present. From captured pressure courses we determined a set of features which we used to train a classifier for discrimination of samples with present arrhythmia from samples where arrhythmia is not present.

I. INTRODUCTION

Atrial fibrillation is one of the most common abnormal heart rhythms. It is relatively common in older patients. It affects about 1 percent of patients younger than 60 years and about 8 percent of patients older than 80 years [1]. It may occur in episodes lasting from minutes to days. It can result in reduced cardiac output or formation of blood clots in the heart [2] [3], which can lead to various serious complications [4]. Patients suffering from atrial fibrillation have symptoms including palpitations, chest pain, dyspnea, and fatigue [3, 5].

However, atrial fibrillation may be asymptomatic, and thus subclinical, causing no alarm to doctor or patient [5].

A standard test used to evaluate atrial fibrillation is usage of 12-lead electrocardiography, which is used to diagnose atrial fibrillation and identify other arrhythmia (e.g., atrial flutter, atrial tachycardia) [1]. Nevertheless, examination using 12-leads electrocardiograph is relatively complicated and therefore not suitable for widespread screening purposes.

It is crucial to have simple to perform and therefore widely available detection method for screening of the patients with atrial fibrillation. Besides 12-lead electrocardiogram several other methods for detection of atrial fibrillation exists. One of simplest and cheapest methods is pulse palpation [6]. Other cost effective and therefore accessible to wide population method for detection is usage of modified blood pressure monitor [6]. Several successful implementations of blood pressure monitors, capable of detecting atrial fibrillation during blood pressure measurement already exist.

A. Motivation

Since we have already developed a blood pressure monitor for monitoring blood pressure using oscillometric method [7], we wanted to adapt the monitor with an algorithm for atrial fibrillation detection. We conducted a pilot study, where we collected pressure course samples from 34 patients and used them to develop a proof of concept version of an algorithm for detection of atrial fibrillation. Since it is possible to detect atrial fibrillation solely by palpation, it is therefore also possible to detect it during using oscillometric blood pressure measurement [6].

II. BACKGROUND

A. Oscillometric blood pressure measurement

The blood-pressure in vessels is not constant. It oscillates from its maximal - *systolic* to the minimal - *diastolic* value. It is caused by the heart muscle beating. To explain the blood-pressure course, the heart-muscle beating cycle is divided into two phases, namely systole and diastole. During the systole phase, the heart muscle contracts and forces the blood into the arteries. During the systole phase, the blood pressure reaches its highest value - the systolic blood pressure. During the diastole, the heart muscle expands and draws blood from the blood-vascular system. During this phase, just before the systolic phase begins again, the blood pressure in vessels is at its lowest value - the diastolic pressure. The blood pressure is reported in unit [*mmHg*].

The arterial blood pressure can be measured in several ways. The most accurate method is a direct measurement, requiring the pressure sensor to be connected over tubing filled with liquid to a cannula needle, placed within the arteries. The method is used in clinical practice. The method is invasive and difficult to perform and therefore unsuitable for our approach. Besides this invasive direct method, there are noninvasive indirect blood-pressure measurement methods [6].

The oscillometric method is an indirect blood-pressure measurement method. It is used in most of commercial and home usage-intended blood-pressure measurement devices [8, 7]. It measures the blood-pressure through air pressure oscillations in the cuff, caused by the compressed arteries pulsations. Using this method, the cuff pressure is gradually released and measured. At a certain air pressure, when the blood pressure in vessels overcomes the air pressure in the cuff and blood starts to flow through the vessels, the air pressure in the cuff starts to oscillate. After the air pressure is further reduced, the oscillation amplitude increases to a certain maximum value and then decreases until oscillations vanish. By measuring the air oscillations amplitude the systolic, diastolic and mean blood pressure can be determined. Analyzing the oscillations, heart pulsation course can also be observed.

III. MATERIALS AND METHODS

A. Data collection

Data for algorithm development was collected in two phases. During the first phase, dataset of three persons was acquired. This step was taken in order to agree on all data items to be captured and to test if selected format of data capture is adequate. In second phase, data from actual patients in hospital was acquired.

Each data sample consisted of several data items for a specific patient. The first data item was raw cuff pressure course from temperature compensated and calibrated pressure sensor SLP33A, sampled with 200Hz, using 16bit A/D converter. Along pressure course, several metadata items were collected and recorded. Metadata consisted of patient's age, height, weight, diagnosed condition and smoking habits.

We acquired data of 34 patients of which 6 had atrial fibrillation present and 28 didn't. 17 patients were males, 17 were females. Patient's average age was 63.03 years with standard deviation 20.2 years. Minimum age was 25 years and maximum age was 89 years. Patient's average weight was 76 kilograms with standard deviation 13.8 kilograms. Minimum weight was 42 kilograms and maximum weight was 112 kilograms. Two patients were smokers.

For each patient we have also recorded 12-lead ECG signal.

All measurements were performed by skilled clinical practitioners who are performing these measurements during their daily clinical practice.

B. Tools used

Data was preprocessed using Matlab version 7.12.0.635 (R2011a). The algorithm required two low-pass filters which were designed using a Matlab filter design tool: *filter design*.

Machine learning was performed using Rapid Miner, version 5.3.015. Using the Rapid Miner Neural Network, SVM, Decision tree, Logistic regression, Random forest, and Naïve Bayes classifiers were tested. Classifier performance was evaluated using leave-one-out cross-validation scheme.

C. Classifier performance measure

For each classifier accuracy and classification error were calculated from classifier's confusion matrix. Actual classifier was selected observing classifier's confusion matrices through googles of domain knowledge. Classifier was selected based on harmonized performance classifying positive and negative samples with regard that available dataset was highly biased.

IV. ALGORITHM DEVELOPMENT

The algorithm consists of several phases. In first phase, the signal is processed to extract features, which are in our opinion relevant to discrimination of pressure course samples where atrial fibrillation is present from samples where atrial fibrillation is not present.

In the second phase, extracted features are used to build a classification model, which classifies samples into two classes: class where atrial fibrillation is present (class 1) and class where atrial fibrillation is not present (class 0). Then classifier results are evaluated and compared and best performing classifier is selected.

A. Data preprocessing

1) Conversion of voltage samples into actual pressure values

In the first stage, raw values of A/D convertor sampling voltage from pressure sensor are converted to pressure values. The relation between the cuff pressure, given in units *[mmHg]*, and the voltage, sampled by the A/D converter from the pressure sensor is given with a set of equations eq. 1, where ADC_{ref}, ADC_0 and ADC_{200} values are calibration values for used pressure sensor and A/D converter [7].

$$\begin{aligned} ADC_{ref} &= 2521 \\ ADC_0 &= 325 \\ n &= \frac{ADC_0 * 65536}{ADC_{ref}} \\ k &= \frac{(ADC_{200} - ADC_0) * 65536}{200 * ADC_{ref}} \\ p &= \frac{adc - n}{k} \end{aligned} \tag{1}$$

2) Elimination of the pressure course caused by the cuff air deflation and noise removal

The raw pressure course contains whole measurement pressure course, sampled by 200Hz. Measurement starts by cuff inflation using a pneumatic pump until the desired air pressure in the cuff is reached. Then the pump is switched off and air is gradually released from the cuff through a pneumatic valve. The deflation part of the pressure course is interesting for further analysis. It is therefore isolated and processed in further process. The goal of processing is extraction of pressure oscillations caused only by pulsating blood vessel.

At first, the course of air pressure in the cuff, caused by air deflation is determined. The course contains cuff

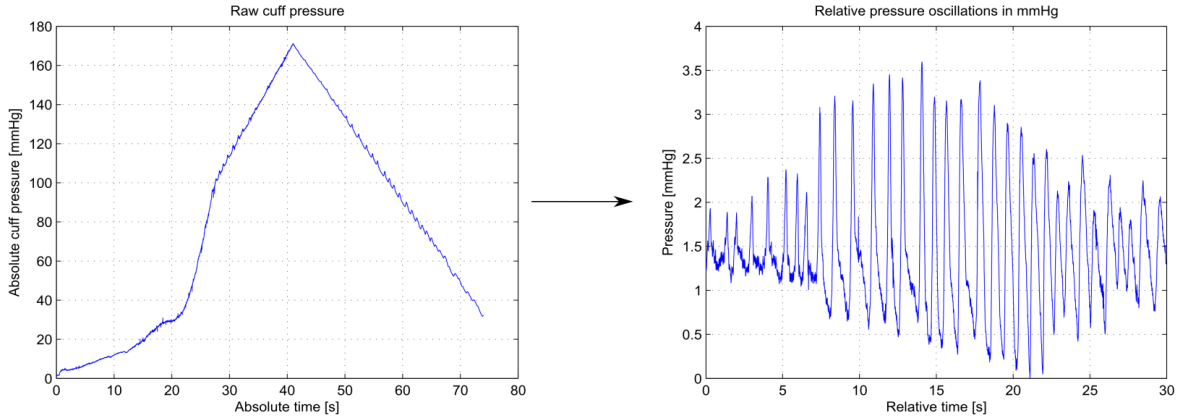


Figure 1. Transformation of raw cuff pressure course signal (left) to filtered signal where only by vessel pulsating caused pressure oscillations in cuff are present (right). Only the deflation part of raw pressure course used in further processing is shown on the left graph.

deflation pressure course without superimposed pressure oscillations caused by pulsating blood-vessels. By subtracting obtained course from the pressure course, only pressure oscillations, caused by the pulsating vessels are obtained. To obtain the deflation pressure course, the pressure data is filtered using a low-pass filter. We used Hamming low-pass FIR filter of order 100 with low-pass frequency F_{cut}=*0.1Hz*. The frequency was chosen empirically and lies well below the expected lowest frequency of the signal, representing pressure pulsations. Obtained cuff deflation pressure course is then subtracted from non-filtered pressure course. The signal is then filtered again to remove high-frequency noise. Used noise removal filter is Hamming low-pass filter of order 100 with low-pass frequency F_{cut}=*20Hz*. Frequency 20 Hz was chosen empirically and is well above the expected frequency of the pressure-oscillation signal [7].

Since filter responses are known, signal obtained after filtering is trimmed in order to remove filter ringing on beginning of the signal. Figure 1 illustrates effect of processing on raw signal, resulting in signal where only by blood-vessel pulsation caused oscillations are present in the signal.

3) Detection of heart beats from pressure course

In order to detect heart beats from pressure course, local pressure oscillations signal extreme search is performed, using the method *peakdet*, proposed by author cited in [9] on in previous step filtered, but still partially noise occluded signal.

To determine maxima and minima of the signal in partially by the noise occluded signal an assumption that local maxima and minima follow each other in an alternating manner is made. The algorithm scans through the data alternating search of local maximums and minimums. To declare the next local maximum, its value must differ from the previous local minimum for at least δ. The value of δ is determined empirically with regard to the analyzed data. When the algorithm finds the next extreme, it records its value and index and moves to the next extreme search. Figure 2 shows local extreme search algorithm result.

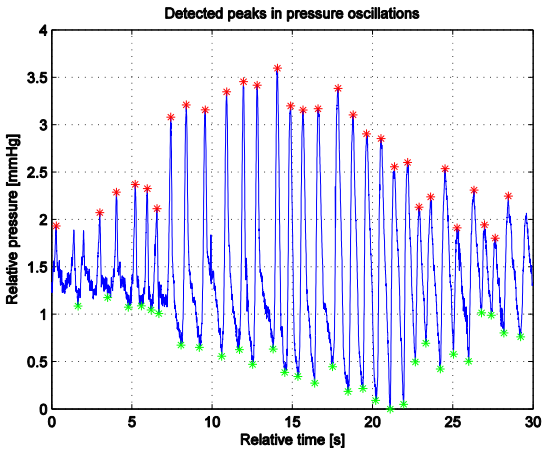


Figure 2. Local extreme search using peakdet [9] algorithm. Detected local maxima and minima of the signal are marked with stars.

B. Feature extraction

1) Detection of irregular heart beats

First step of feature extraction is determination of times of heart beats and beat intervals. In order to determine times of heart beats, previously detected local maxima of pressure course are used. From vector containing maxima, beat intervals are calculated by differentiating the vector containing beat times. In order to determine irregular heartbeats, beats that differ from average beat time by more than through optimization determined threshold are marked. Using our dataset, optimal value of threshold was beat time difference of 45% from average beat time.

To detect irregular heartbeats, first an average beat frequency is determined from beat intervals. In our case, best results for overall heart beat frequency are obtained calculating median of vector, containing beat intervals. Next, beats differing from overall average beat frequency more than threshold are marked. Figure 3 illustrates result of irregular heart beat detection.

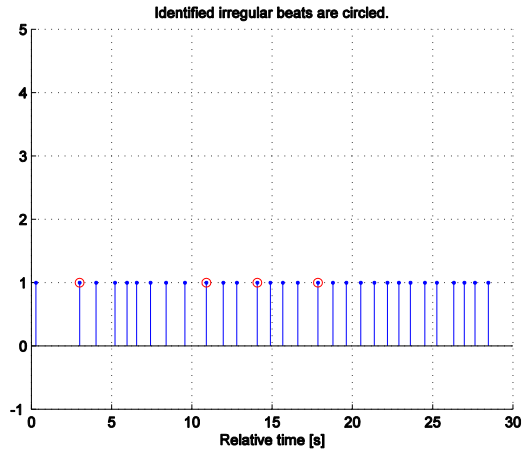


Figure 3. Timeseries of detected heartbeats. Detected irregular heartbeats differing from average frequency are marked by circles.

2) Features determined

From vector containing heartbeat intervals and irregular beats several features relevant for atrial fibrillation determination are computed. Computed features are:

- **Beat frequency (BF)**, which is computed by computing median of beat intervals. Obtained value is then used to calculate beat frequency in beats per minute;
- **Ratio of irregular beats (RIB)**, which is computed by counting irregular beats and dividing the value with number of all detected beats;
- **Average irregular only beat interval (aIRBI)**, which is calculated by computing mean value of all irregular beat intervals;
- **Standard deviation of irregular only beat interval (stdIRBI)**, which is calculated by computing standard deviation of all irregular beat intervals;
- **Number of consecutive irregular beats (NCIB)**, which is calculated by finding maximum value of consecutive irregular beats.

TABLE I. EXAMPLE OF DATASET CONTAINING SET OF FEATURES, USED TO TRAIN CLASSIFIER.

Sample	BF	RIB	aIRBI	stdIRBI	NCIB	AF
1	69.7	0.13	329	143.5	0	1
2	70.6	0.08	316.7	100.8	0	1
...						
7	67.9	0.03	357	0	0	0
8	69.4	0.05	212.5	171.8	0	0
9	63.2	0	0	0	0	0
…						

Determined features for each patient (processed pressure course) are recorded in a single vector, and label about presence of atrial fibrillation (**AF**) is added from metadata. The vector, containing all features, is then exported into a CSV file. Table I illustrates dataset excerpt, containing features of 5 samples.

C. Modelling

Dataset containing features was processed in Rapid Miner. At first, data was imported, using "Read CSV" operator. AF attribute role was set to label.

1) Selection of best classifier

Several classifiers were tested and evaluated. Since available dataset was small, leave one out cross-validation scheme was used for each classifier. For each classifier confusion matrix was computed. Since available dataset was highly biased (6 positive vs 28 negative) overall accuracy of classifier is not representative measure of classifier's performance. Best classifier was selected based on performance of classifying positive samples correctly.

Tested classifiers were: Naïve Bayes, Logistic regression, Decision tree, Random forest, SVM and Neural network. All classifiers were used using their default settings.

V. RESULTS

Several classifiers were trained and evaluated using "leave one out" validation process. Table II illustrates classifier performance. Overall classifiers accuracy ranges from 50% for Naïve Bayes to 85.92% for Neural network. However, observing class recall and class precision, it is evident that classifier accuracy is not effective performance indicator. For example, SVM classifier predicted all sample classifications as negative and therefore scored 82.35% on accuracy regardless it failed to predict any positive class. Observing the confusion matrices, we believe, best classifier for our problem is Neural network classifier, since it was able to find nearly all positive samples (5 of 6) and produced 4 false positives. It also correctly predicted nearly all negative samples (24 of 28) and produced 1 false negative.

TABLE II. EVALUATED CLASSIFIERS PERFORMANCE.

Naïve Bayes (accuracy 50%)			
	true 1	**true 0**	**class precision**
pred. 1	6	17	26.09%
pred. 0	0	11	100.00%
class recall	100.00%	39.29%	
Logistic regression (accuracy: 61.76%)			
	true 1	**true 0**	**class precision**
pred. 1	3	10	23.08%
pred. 0	3	18	85.71%
class recall	50.00%	64.29%	
Decision tree (accuracy: 76.47%)			
	true 1	**true 0**	**class precision**
pred. 1	2	4	33.33%
pred. 0	4	24	85.71%
class recall	33.33%	85.71%	
Random forest (accuracy: 79.41%)			
	true 1	**true 0**	**class precision**
pred. 1	0	1	0.00%
pred. 0	6	27	81.82%
class recall	0.00%	96.43%	
SVM (accuracy: 82.35%)			
	true 1	**true 0**	**class precision**
pred. 1	0	0	0.00%
pred. 0	6	28	82.35%
class recall	0.00%	100.00%	
Neural network (accuracy 85.29%)			
	true 1	**true 0**	**class precision**
pred. 1	5	4	55.56%
pred. 0	1	24	96.00%
class recall	83.33%	85.71%	

VI. CONLUSION

Developed algorithm performed above expectations. This can partially be caused by the fact, that the best performing classifier is able to fit relatively complex functions. Therefore it is possible, that the classifier is over-fitting the data and will not perform so great when confronted with more real data samples. But nevertheless, the algorithm looks promising. We will therefore continue with its development. Our first task will be focused on collection of a larger dataset, on which the algorithm will be furtherly tested and optimized. Main reserves for algorithm improvement lie in more precisely determined feature values, determined from raw data. Since feature extraction process was partially iterative, during which certain feature extraction thresholds were back set based on classifier performance, we believe, that features will be more precise when a larger dataset is available. Further work will be also focused on determination of features, based on which atrial fibrillation can be distinguished from other arrhythmias.

Sometimes atrial fibrillation cannot be distinguished from ectopic beats by only observing heart beats detected from oscillometric method. However the described method is intended as a screening method, based on which, patients with detected rhythm anomaly are referred to further, more invasive examinations.

ACKNOWLEDGMENT

The author would like to thank to Jakob Šušterič, Tomo Krivic and Ana Mlinar of MESI d.o.o.. The author would like to thank to now retired professor dr. Jurij Franc Tasič for his support in this project. This work has in part been financed by Mesi d.o.o. and in part within the research group “Algorithms and optimization methods in telecommunications”.

BIBLIOGRAPHY

[1] C. Gutierrez and D. G. Blanchard, "Atrial Fibrillation: Diagnosis and Treatment," Am Fam Physician., vol. 83, no. 1, pp. 61-68, 2011. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] E. Antman and e. al., "Atrial fibrillation: Current understandings and research imperatives," J Am Coll Cardiol., vol. 22, no. 7, pp. 1830-1834, 1993.

[3] Wikipedia, "Atrial fibrillation," 19 January 2015. [Online]. Available: en.wikipedia.org/wiki/Atrial_fibrillation. [Accessed 9 January 2015].

[4] W. Verberk and P. de Leeuw, "Accuracy of oscillometric blood pressure monitors for the detection of atrial fibrillation: a systematic review.," Expert Rev Med Devices, vol. 9, no. 6, pp. 635-40, 2012

[5] S. Sinha, "Asymptomatic Atrial Fibrillation: Should You Be Concerned?," Medpage Today, 13 07 2012. [Online]. Available: http://www.medpagetoday.com/resource-center/atrial-fibrillation/Asymptomatic-Atrial-Fibrillation-Should-You-Be-Concerned/a/33607. [Accessed 19 01 2015].

[6] K. Harris, D. Edwards and M. J, "How can we best detect atrial fibrillation?," The Journal of the Royal College of Physicians of Edinburgh, vol. 42, no. Suppl, pp. 5-22, 2012.

[7] M. Meža, J. Šušterič, T. Krivic and J. F. Tasič, "An algorithm to determine the ankle-brachial pressure index using the oscillometric blood pressure measurement method," Elektrotehniški vestnik, vol. 78, no. 3, pp. 159-164, 2011.

[8] G. Geršak, A. Žemva and J. Drnovšek, "A procedure for evaluation of non-invasive blood pressure simulators.," Med Biol Eng Comput, vol. 47, no. 12, pp. 1221-8, 2009.

[9] E. Billauer, "peakdet: Peak detection using MATLAB," [Online]. Available: http://www.billauer.co.il/peakdet.html. [Accessed 2 February 2015].

[10] E. e. a. Antman, "Atrial fibrillation: Current understandings and research imperatives," J Am Coll Cardiol., vol. 22, no. 7, pp. 1830-1834, 1993.

[11] D. G. B. Cecilia Gutierrez, "Atrial Fibrillation: Diagnosis and Treatment," Am Fam Physician., vol. 83, no. 1, pp. 61-68, 2011.

[12] d. L. P. Verberk WJ, "Accuracy of oscillometric blood pressure monitors for the detection of atrial fibrillation: a systematic review.," Expert Rev Med Devices, vol. 9, no. 6, pp. 635-40, 2012.

[13] E. D. M. J. Harris K, "How can we best detect atrial fibrillation?," The Journal of the Royal College of Physicians of Edinburgh, vol. 42, no. Suppl, pp. 5-22, 2012.

Hybride neuro-fuzzy expert system for assessing diabetes risk

Zeljko Knok*, Zikrija Avdagic** and Samir Omanovic**
* Polytechnic of Medimurje / Department for Computing, Cakovec, Croatia
** Faculty of Electrical Engineering / Department for Computing and Informatics, Sarajevo, Bosnia and Herzegovina
zknok@mev.hr, {zikrija.avdagic, samir.omanovic}@etf.unsa.ba

Abstract – There is an increasing number of diabetic diseases in the population. Early diagnosis increases possibility of healing and decreases healing expenses. This paper is focused on modeling of an expert system for assessing diabetes risk using artificial intelligence methods. The model is hierarchical with neuro-fuzzy blocks and voting on the output. Model optimization (learning) is done using the data of real patients acquired in a one public health institution. The expert system is implemented in the Matlab/Simulink environment. Validation of the system shows high sensitivity of 100% which is important for early diagnoses. Specificity is lower – only 90% which means that some patients are sent to further diagnoses although they are healthy. Results indicate that this modeling approach is applicable on assessing diabetes risk. The medical domain knowledge and experience contained in the real data is successfully transferred in the solution model – the proposed expert system. Validation of the proposed expert system indicates that such a system can be used as an auxiliary expert for early diagnoses of diabetes and improve the quality of health systems with lower diagnostic expenses.

I. INTRODUCTION

Human life quality is related to health. Improving health is related to reducing risks of diseases. Best way to reduce risks of diseases is to increase the quality of diagnoses and provide early diagnosis which gives better healing possibilities. Disease early diagnostic and risk estimation reduces treatment costs which is very important for healthcare institutions.

Nowadays, public each healthcare institution has information system which enables monitoring of the health status of patients and keeping records of their treatments. However, these data are not sufficiently used to support and help doctors in primary health care decision making. On the other side, possibilities to exploit these data are increasing, especially in the area of artificial intelligence [1].

There are a number of papers related to the diagnostics of diabetes mellitus and its association with risk factors, through the use of artificial intelligence and data mining. T. Jayalskshmi and A. Santhakumaran in their work [2] discuss various data processing techniques. Artificial neural networks (ANN) [1] represent a significant area of interest for solving health problems because of their capability to learn from data with the high accuracy (near 99% accuracy in many applications). Marcos. A. González-Olvera et al in their work [3] presented data-driven modeling of the insulin dynamics in different patients using a recurrent ANN with the output feedback. GG Gardner et al in their work [4] investigated usage of neural networks in detection of diabetes through images of the retina of the eye and showed the comparison in the pictures of the patients (the sample of 147 diabetics and 32 normal). By comparing the results of ophthalmologists and the network, the network has achieved 88.4% sensitivity and specificity of 83.5% for the detection of diabetic retina. Compared to the ophthalmologist, the network has reached good accuracy, and the system can be used to support the screening of diabetic patients.

II. METHODS AND DATA SOURCES

One of the most frequently used combinations of soft computing methods is the one that combines fuzzy logic and neural computing, usually in the form named neuro-fuzzy system. The most popular neuro-fuzzy system is the adaptive-network-based fuzzy inference system (ANFIS) [5] that is capable to learn from data, i.e. to optimize fuzzy rules using a neural network as an optimization mechanism.

The overall model (see Fig.1) of the expert system for assessing diabetes risk is divided in two parts. The first part is based on ANFIS and the second part uses simple neural network. Results from these two parts are combined (ensemble) in voting block to produce final diabetes risk assessment. Using two different models and combining results is for safety reasons, i.e. decreasing possibility of the wrong assessment. In medical terminology this is like second opinion of another doctor.

In the first part, input data are divided into three groups and fed into three ANFIS. This way ANFIS execution is simplified because of the lower dimensionality. It would be better to use all nine input variables in one ANFIS but that makes the problem much complex for execution and output fuzzy model is much harder to understand. In this way fuzzy models of all three blocks are simple, but price is necessity to ensemble results by using one voting block.

In the second part all input data are fed into one neural network. Training, testing and validation are performed for complete model, including both parts, simultaneously. That means that training data are fed in the first part (three ANFIS) and in parallel to the second part (neural

network). Output from the system represents a percentage risk for diabetes disease development.

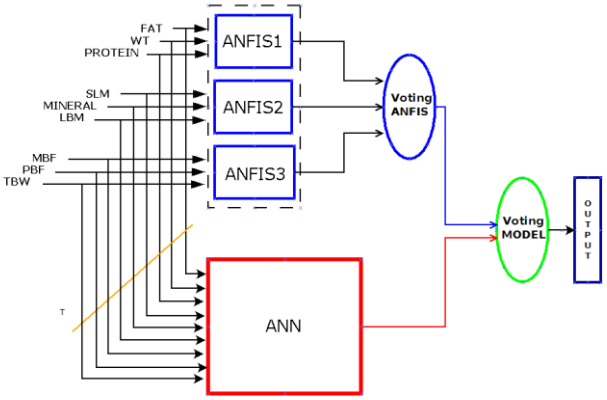


Figure 1. Hierarchical structure of the neuro-fuzzy system for assessing the risk of diabetes

TABLE II. DESCRIPTION AND ABBREVIATIONS OF THE MEASURED VALUES

Abbreviation	Description
FAT Fatness (%)	Fatness measure in percentages compared to the standard body weight FATNESS = ((*real weight – standard weight*) / *standard weight*) * 100
WT Weight	Measured body weight
PROTEIN Protein (kg)	Weight that remains when water is eliminated from muscle mass
SLM S.L.M.	Mass of lean tissue (*weight – fat mass – mass of bones*)
MINERAL Mineral (kg)	Weight of bones and electrolites
LBM L.B.M. (kg)	Weight body weight less fat. Shows the mass: water, muscle, minerals and other lean tissue
MBF M.B.F. (kg)	Weight of body fat in kilograms
PBF P.B.F. (%)	Percent of Body Fat - weight ratio of body fat compared to body weight
TBW T.B.W. (kg)	Weight of water in the body (intracellular and extracellular)

Data used in this research are real measurements of parameters collected by Croatian Institute of Public Health of the Medimurje County, using body composition analyzer Avis 333. The dataset contains samples for 165 adults of different ages and sexes.

Nine collected measurements are described in table I. Classification of samples, those at risk for diabetes (class 1) and those, in which there is no risk (class 0), was conducted by the domain expert. This way is prepared the dataset for supervised [1] training of the model.

A. *ANIFS modeling*

After the dataset preparation, it is split into subsets for training, testing and validation. Structures of Sugeno fuzzy systems are created from data using ANFIS tool, which also enables training and testing of the created fuzzy system.

In order to break the complex multidimensional data structure into a smaller dimensional structure, data segmentation was performed in the way that nine input attributes – variables are split in three sets, so that three FIS structures are created.

B. *Neural network*

In this research is experimented with 12 different architectures of the neural network. The 12 architectures differ in the number of hidden neurons, transfer functions in the hidden and output layer, and error function. It was also experimented with different dataset divisions on subsets for training, testing and validation to obtain valid statistics of model creation.

III. RESULTS

Obtained results were assessed, and finally one neural network, that gives the best result, was selected (see table II). Final model is presented on Fig.2. It is a Simulink block model. Voting bloc VOTER_ANFIS is created as perceptron type network, while voting bloc VOTER_MODEL is created as single-layer network with a linear transfer function (see Fig.3). Decision about network type used in voting blocs is made experimentally.

TABLE I. RESULTS OF TRAINING, VALIDATION AND TESTING. THE MOST EFFECTIVE BACKPROPAGATION NEURAL NETWORK IS THE ONE WITH 5 NEURONS IN THE HIDDEN LAYER (NET10)

Neural Network	Net10
Sample division Training% Validation% Testing%	70% 15% 15%
Number of neurons in the hidden layer	5
Transmission function in the hidden layer	tansig
Transmission function in the output layer	purelin
Number of epochs	18
Lowest error (MSE)	$1{,}8405*10^{-11}$

IV. DISCUSSION, CONCLUSION AND FUTURE WORK

Validation of the system is done by using 20 samples not used in system creation and learning (optimization). Half of them are of class 1 (positive) and half of them are of class 0 (negative). System showed high sensitivity (100%) which means that all patients with the risk (class 1) are recognized by the system properly, but specificity of the system is lower (90%) which means that some of patients that have not diabetes risk are recognized as patients with the diabetes risk. In reality that means that they would be send to further medical examinations. Additional check for each patient does not present a burden, and it can eliminate the doubts. That produces less damage than in the opposite case (when the specificity is higher than the sensitivity).

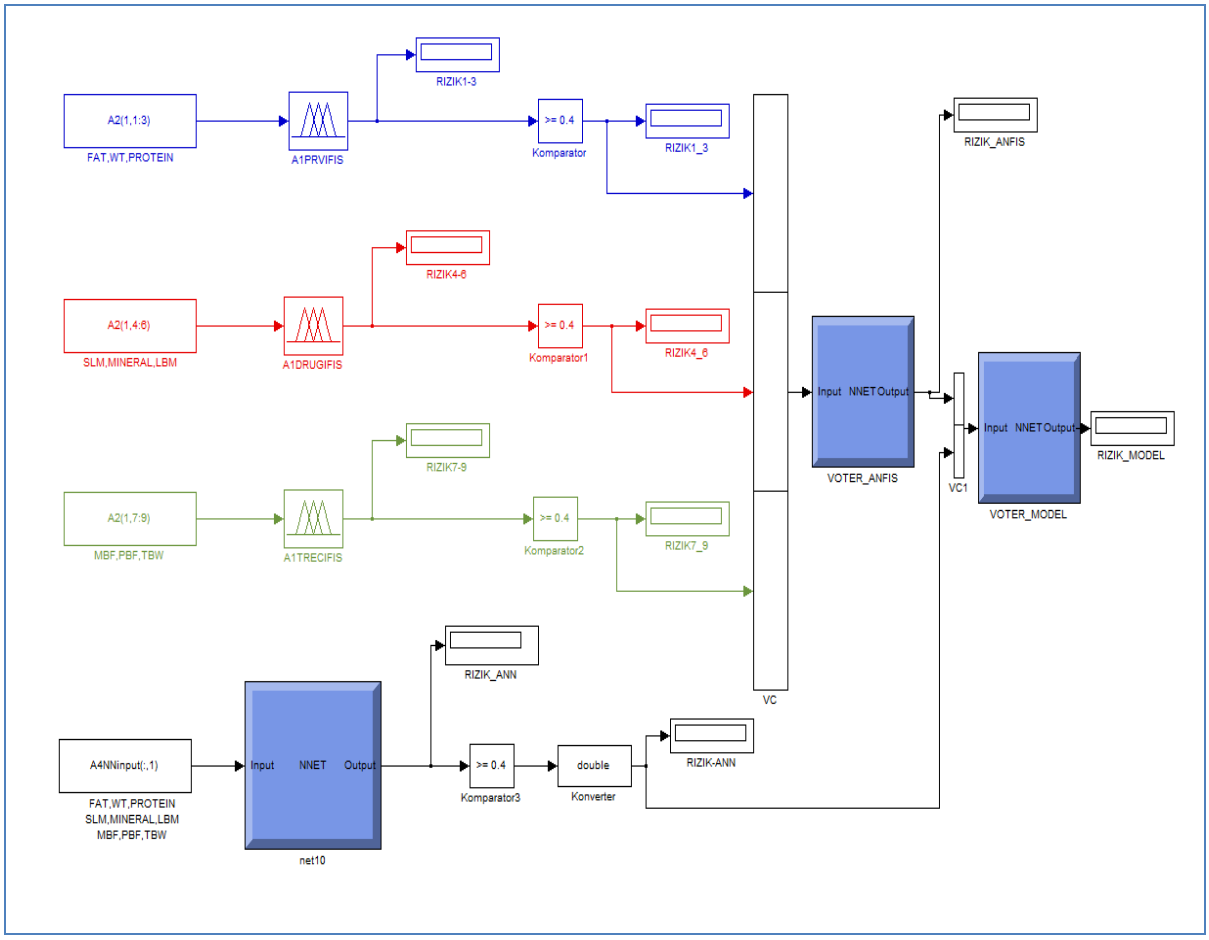


Figure 2. Simulink application

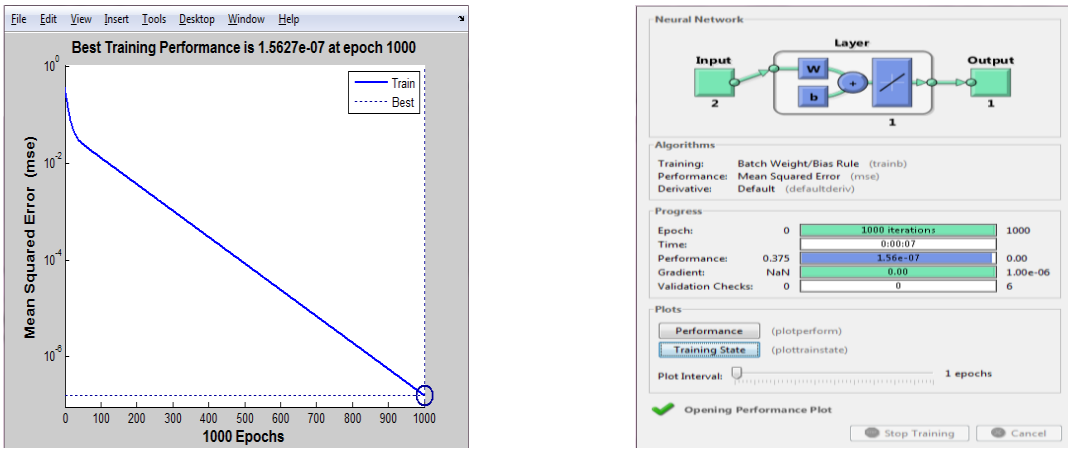


Figure 3. VOTER_MODEL performances and neural network

The aim of this study was to develop a model of intelligent system which takes a number of attributes, input parameters, and based on them provide risk assessment of a disease. It is implemented as two models that complement each other. Both models use same input data. The first model (FIS structures) is developed and implemented using ANFIS tool, and for easier construction it is divided into three parts with three input parameter attributes within each FIS structure. Neural network, as the second model, showed slightly lower results than the first model, considering the fact that it is a relatively small data set for training, testing and validation.

After system creation, validation of the system showed good results However, for a more complete and better assessment of the risk of disease, it is necessary to have a larger number of samples available, and the greater number of attributes for the model to be more representative for decisions making.

It should be taken into consideration that only nine attributes where used, and that for complete and comprehensive diagnostics the glucose levels for all cases

should be included. This can be one of directions for a future work.

REFERENCES

[1] Z. Avdagic, Vestacka inteligencija i fuzzy-neuro-genetika, GrafoArt, Sarajevo, 2003.

[2] T. Jayalskshmi, A. Santhakumaran, "Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks", 2010 Second International Conference on Machine Learning and Computing (ICMLC), pp.109,112, 9-11 Feb. 2010, doi: 10.1109/ICMLC.2010.65.

[3] M.A. González-Olvera et all, "A discrete-time recurrent neurofuzzy network for black-box modeling of insulin dynamics in diabetic type-1 patients", Int J Neural Syst. 2010 Apr; vol.20 no.02, pp. 149-58, doi: 10.1142/S0129065710002322.

[4] G.G. Gardner et all, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool", Br J Ophthalmol 1996; vol. 80 pp.940-944, doi:10.1136/bjo.80.11.940.

[5] J.-S.R. Jang, "ANFIS: adaptive-network-based fuzzy inference system". IEEE Transactions on Systems, Man and Cybernetics, vol.23, no.3, pp.665,685, May/Jun 1993, doi:10.1109/21.256541.

Solving Strip-Packing Problem Using Sequence Pair

D. Domović*, T. Rolich*
* University of Zagreb, Faculty of Textile Technology, Zagreb, Croatia
{daniel.domovic, tomislav.rolich}@ttf.hr

Abstract - Strip-packing problem is a combinatorial optimization problem with a goal of finding optimal positions of cutting patterns that meet the following conditions: each pair of cutting patterns must not overlap and they must be placed within the boundary of rectangular material area. An optimal placement that utilizes material usage needs to be created. Strip-packing problem is a common textile industry problem where it is also known as automated marker making problem. In this paper a sequence pair abstract representation is used to obtain a non-overlapping placement. For that reason cutting patterns are approximated with a rectangular bounding box and placement of rectangles is obtained first. Non-approximated cutting patterns' positions and material dimensions are then decoded using longest common subsequence method. These methods are combined with simulated annealing algorithm to find an optimal placement of cutting patterns. Experiments were conducted on 4 benchmark datasets: MARQUES and ALBANO (24 cutting parts), DAGLI (30 cutting parts), and MAO (20 cutting parts) with material utilization of 83.54%, 82.02%, 69.86%, and 71.26% respectively.

Keywords - sequence pair, longest common subsequence, simulated annealing, irregular items, strip packing problem.

I. INTRODUCTION

In order to remain competitive on modern market a part of the budget needs to be invested in the improvement of production. Referring to textile and garment industry specifically it has been recognized that the industrial advance lies in the investment in new technologies, developing new products, specialization and export [1].

Considering the fact a large part of production process in garment industry is still done manually, by applying new technological achievements into the production process it is possible to improve productivity and production speed.

Collaboration between computer science and textile industry can contribute to this process. In that context a strip-packing problem is introduced, also known as the automated marker making problem in garment industry.

This research is conducted to create an algorithm that will automatically create markers consisted of cutting patterns optimally allocated on a material. Since the material greatly influences final garment price, it is important to rationally nest cutting patterns, and to speed up the production process to reduce its cost.

In computer science a problem of automatic marker making is referred to as 2D-irregular strip-packing problem and is described as follows. In a container i.e. rectangular shaped material of fixed width and infinite length, cutting patterns i.e. polygons of convex and non-convex shape need to be placed on material to create a placement that utilizes the material length. Two conditions have to be met: any of the two items should not overlap and polygons must not exceed the boundary of a container. When items are irregularly shaped the strip-packing problem is commonly referred to as nesting problem [2].

Formally, a set of items $I = \{i_1, i_2, \ldots, i_n\}$, where n is the number of items in I ($|I| = n$) needs to be placed in the container C so the following conditions are satisfied: (1) $i_i \cap i_j = 0$, and (2) $i_i \cap C = i_i$ [3].

Strip-packing problem belongs to a larger field of packing problems. An attempt to categorise the subproblems has been made by Dyckfoff [4] and more recently Wäscher [5]. Wäscher differs packing problems by the criteria of dimensionality, problem goal, large objects, small objects and shape of small objects.

Packing problems can be examined from three perspectives: container count minimization, optimal subset selection, and container size minimization (area minimization). In this research the width of a container as well as the length have arbitrary size. Therefore we deal with finding a minimal circumscribed rectangle in which all items can be placed.

The paper is organized as follows: in Section II a literature overview is given. In Section III methods used in this paper are described. In Section 0 the results are presented alongside a short discussion. The paper is concluded in Section V.

II. LITERATURE OVERVIEW

Strip-packing problem is known to be NP-complete which means that globally optimal solution is unlikely to be found using polynomial algorithms [6]. As a result solution methodologies predominantly utilize heuristics [2]. Since obtaining a feasible rectangle packing is a core of our approach we present some rectangle packing methods used in the literature.

The authors in [7] cope with a layout in physical design of VLSI which is in its core a rectangle packing problem. All circuit elements in a chip have to be packed

without violating design rules. A packing representation called sequence-pair has been proposed where each sequence pair corresponds to a packing. A method to transform a sequence pair into a horizontal and vertical constraint graph is also described. A solution space is searched using simulated annealing.

Since block placement algorithms usually use simulated annealing, a fast algorithm is needed to transform a sequence pair into a corresponding placement. In [8] an approach that uses longest common subsequence (LCS) to evaluate a sequence pair in $O(n^2)$ and an improved algorithm in $O(n\log n)$ running time is presented. LCS algorithm is able to record and reconstruct the position of each block and dimensions of a placement.

In [9] the author interprets the sequence pair as a packing order. Generated placements are semi-normalized, i.e. each item is moved as far left and down as possible according to the current packing contour. A transformation of sequence pair to placement in $O(n\log\log n)$ time has been presented. For large packing problems it is able to find solutions quickly that waste no more than 3%–5% of the space.

As seen from the literature overview, to our best knowledge, sequence pair has only been applied to obtain rectangle packing. For that reason and because of the complexity irregular items introduce to the implementation authors have tried to approximate items using convex hull (as in [10] and [11]) or circumscribed rectangle (often referred to as bounding box). The rectangular approximation is used in this paper so the sequence pair method can be applied.

III. METHODS USED

In this section we describe our proposed methodology and concepts related to its behavior, involving polygon approximation, sequence pair, longest common subsequence and simulated annealing.

A. Polygon approximation

Since sequence pair methodology applies only on rectangle packing problem, irregular polygons used in this research are approximated using bounding box representation. A bounding box of a polygon is a rectangle circumscribed around a polygon, i.e. the smallest rectangle that contains all points of a polygon as seen in Fig. 1. Each bounding box rectangle is represented with its reference point (coordinate of a lower left corner is used), width and length.

B. Sequence pair

A placement can be presented directly as a list of polygons. Each polygon is thereby presented with coordinates of its vertices which can potentially result with an overlapping placement. Therefore, many heuristics use abstract representation techniques whose goal is to present a placement by describing the relationship between elements rather than dealing with elements' coordinate positions. The abstract representation is later subjected to a decoding algorithm to reconstruct a placement [3]. In this research a sequence

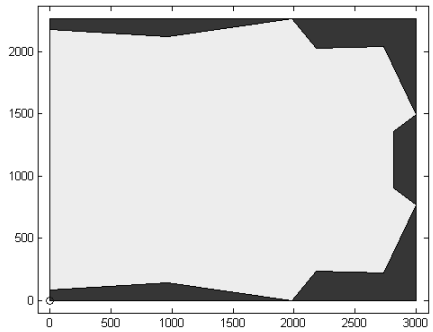


Figure 1. Bounding box representation

pair abstract representation is used and applied on rectangle packing.

Sequence pair is a pair of permutation lists consisted of m indexes where $m = \{1, 2, ..., n\}$ and n is the number of rectangles. Each index m represents one rectangle.

The relationship between polygons is decoded using the following rules [8]:

$$(< ..m_i..m_j..>, < ..m_i..m_j.. >) \Rightarrow m_i \text{ is left of } m_j \quad (1)$$

$$(< ..m_j..m_i..>, < ..m_i..m_j.. >) \Rightarrow m_i \text{ is below of } m_j \quad (2)$$

Example of a packing for a sequence pair (X, Y) is shown in Fig. 2. In this paper a packing is reconstructed with longest common subsequence method.

C. Longest common subsequence

Longest common subsequence (LCS) method is used to reconstruct a rectangle packing from a sequence pair. By applying it, we are able to find the position of a reference point of a rectangle and to find its position in a placement [8]. Using LCS method the dimensions of an overall placement can also be calculated [3].

For two weighted sequences X and Y, a sequence Z is a common subsequence of X and Y if Z is a subsequence of both X and Y. In example in Fig. 2 a subsequence (5 1 4) would result as a LCS of (X, Y).

To obtain the length of a packing a LCS(X, Y) needs to be performed. Height of a packing is calculated as LCS(X^R, Y), where X^R represents inverted X permutation. Based on this relation coordinates of a rectangle can be calculated.

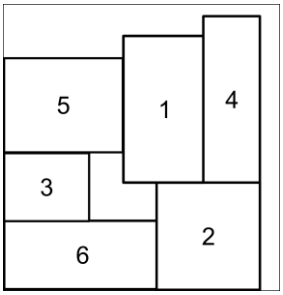


Figure 2. Packing represented by sequence X = (5 3 1 4 6 2) and sequence Y = (6 3 2 5 1 4)

```
Initialize P and L to zeros
FOR each element i
  b ← X(i)
  p ← index of element b in sequence Y
  P(b) ← L(p)
  t ← P(b) + width of element b
  FOR j = p to n
    IF t>L(j)
      L(j) = t
    ELSE
      Break
    END IF
  END FOR
END FOR
LCS width = L(n)
```

Figure 3. LCS algorithm for obtaining a width of a packing based on sequence pair (X, Y)

If sequence pair (X, Y) is represented as (X_1bX_2, Y_1bY_2), b is the rectangle whose coordinates are calculated, X_1 and Y_1 are the predecessor of rectangle b in a sequence pair and X_2, Y_2 are successors of rectangle b in a sequence pair. The x-coordinate of polygon b can be reconstructed by calculating LCS(X_1, Y_1).

If sequence pair (X, Y) = (X_1bX_2, Y_1bY_2), then (X^R, Y) = ($X_2^RbX_1^R$, Y_1bY_2), so y-coordinate of a rectangle b in a packing is reconstructed by calculating LCS(X_2^R, Y_1).

The pseudo code of a LCS algorithm implemented in this research is shown in Fig. 3.

D. Simulated annealing

Simulated annealing is a metaheuristic local search technique inspired by a process of cooling metal which is known as annealing. Simulated annealing searches a set of neighborhood solutions to a certain problem in a randomized manner. Each solution is said to have a neighborhood, i.e. a set of solutions in the close area of a current solution. Neighborhood of a solution is expressed by a criteria depending of a problem. For e.g. if a solution is presented by a permutation, a neighborhood of a solution is a set of permutations where distance between permutations is two (permutations differ in positions of two elements).

In simulated annealing the neighborhood of a solution is searched in a randomized manner. If a better solution is found in the neighborhood it is always accepted and the search process continues from that solution. This method is also used in hill climbing algorithm. One of the main drawbacks of this heuristic is that it can get easily stuck in a local optimum. To escape local optimum some worse solutions from the neighborhood are also accepted, but with a certain probability.

The probability of a downward move is based on the temperature factor T. Factor T is decreasing over the course of the algorithm according to a cooling schedule. If a new solution X_{i+1} has worse fitness value than the current solution X_i, X_i is replaced with X_{i+1} if a random number r $\in$ [0, 1] is lower than probability $e^{-D/T}$. Otherwise, it is always accepted. Temperature T is decreasing in each iteration according to the problem specified formula.

Downward move serves as a mechanism to better explore the search space because sometimes the best solution can be found in the neighborhood of the worst solution. Initially, the probability of allowing a downward move will be relatively large, but later in the algorithm, closer it gets to the optimum, downward moves are permitted with smaller probability [12].

E. Proposed methodology

The proposed methodology is described by pseudo code in Fig. 4. After initializing SA parameters, bounding boxes are created for a set of irregular polygons. The rectangle packing is obtained using sequence pair and SA.

After randomly choosing the initial sequence pair, a neighborhood solution needs to be found. A neighborhood of a sequence pair is a set of permutation pairs in which at least one sequence within a pair differs in positions of two elements from its neighbor. For e.g. if a sequence pair is X = (1 2 3 4 5), Y = (5 4 3 2 1) then a possible neighbor can be X_1 = (5 2 3 4 1), Y_1 = (5 4 3 2 1). Function that calculates (X_1, Y_1) permutes only two elements in a permutation in sequence X, or in sequence Y, or in both sequences X and Y. The option is randomly selected by the algorithm. After selection solutions are evaluated by calculating the fitness of each solution according to (1). A polygon packing is obtained by removing bounding boxes from a layout.

```
Initialise: SA parameters t0, ts, number of iterations
Create bounding boxes from a set of polygons
polyArea ← area of polygons
Randomly choose initial sequence pair (X, Y)
Calculate the fitness of (X, Y)
FOR each iteration
  Find the solution (X1, Y1) in the neighbourhood of (X, Y)
  Calculate fitness f(X1,Y1) to evaluate the (X1, Y1)
  IF f(X1,Y1) is better than f(X,Y)
    Accept it
  Else
    Choose random number r
    Calculate temperature factor T = 1 / (t0 + ts * a)
    Calculate fitness ratio factor D = (f(X1, Y1) − f(X, Y)) / f(X, Y)
    IF r < e^(−D/T)
      Accept solution
    END IF
  END IF
  IF solution is accepted
    X ← X1, Y ← Y1
    a = a + 1
    bbArea ← area occupied by rectangles using LCS algorithm
    IF bbArea is smaller than minimum area so far
      Minimum area ← bbArea
    END IF
  END IF
END FOR
```

Figure 4. Proposed algorithm

TABLE I. EXPERIMENTAL RESULTS

		RECT				*POLY*				*DIFF*
DATASET	***ITER***	***MIN***	***AVG***	***MAX***	***STDEV***	***MIN***	***AVG***	***MAX***	***STDEV***	***AVG***
Albano	10k	82.78	87.73	92.69	2.79	70.44	74.66	78.88	2.38	13.08
	50k	89.60	92.13	95.99	2.16	76.24	78.40	81.68	1.83	13.73
	100k	91.48	93.23	95.43	1.30	77.85	79.34	81.20	1.10	13.89
	500k	91.69	93.81	96.38	1.37	78.02	79.83	82.02	1.16	13.98
Dagli	10k	80.13	83.92	87.91	2.37	58.74	61.52	64.45	1.74	22.40
	50k	87.95	90.99	94.12	2.19	64.47	66.70	69.00	1.60	24.29
	100k	90.08	92.72	94.19	1.32	66.04	67.97	69.05	0.97	24.75
	500k	91.15	94.27	95.30	1.16	66.82	69.11	69.86	0.85	25.16
Mao	10k	86.83	89.98	93.43	2.41	63.92	66.24	68.78	1.77	23.74
	50k	89.88	93.14	95.45	1.74	66.17	68.57	70.27	1.28	24.57
	100k	91.77	93.89	96.18	1.43	67.56	69.12	70.81	1.06	24.77
	500k	92.11	95.19	96.79	1.68	67.81	70.08	71.26	1.24	25.11
Marques	10k	84.09	87.21	89.43	1.94	72.84	75.54	77.45	1.68	11.68
	50k	89.43	90.92	94.10	1.36	77.45	78.75	81.50	1.18	12.18
	100k	88.40	90.90	93.14	1.46	76.56	78.73	80.67	1.27	12.17
	500k	91.76	93.75	96.46	1.54	79.47	81.20	83.54	1.33	12.55

IV. RESULTS

In this section experimental results and discussion on obtained results are presented.

A. Environment

The experiments have been carried out on a personal computer with the following configuration: Intel Core i5-2400@3.1 GHz, 10 GB DDR3, 500 GB HDD. An application has been implemented in MATLAB environment.

B. Parameters

Simulated annealing parameters initial temperature t_0 and temperature step t_s are initialized to empirically chosen values 0.1 and 0.01 respectively.

A series of experiments have been conducted on 4 benchmark datasets available on EURO Special Interest Group on Cutting and Packing (ESICUP) website [13]: ALBANO, MARQUES, MAO and DAGLI all of which are real instances from textile industry. The data sets are described in Table II. Although rotations of polygons for these data sets can be allowed, we have not included experiments with rotation in our research.

TABLE II. DATA SET CHARACTERISTICS

Data set	**Number of different pieces**	**Total number of pieces**	**Orientation (degrees)**
Albano	8	24	0, 180
Dagli	10	30	0, 180
Mao	9	20	0, 90, 180, 270
Marques	8	24	0, 90, 180, 270

For each data set a series of ten experiments with 10000 (10k), 50000 (50k), 100000 (100k) and 500000 (500k) iterations of SA have been conducted.

C. Discussion

Experimental results for packing density of a rectangular packing (column RECT) obtained by sequence pair, and its corresponding polygon packing (column POLY) are presented in Table I. For each data set (column *DATASET*) and for each iteration count (column *ITER*) a minimum and maximum value of ten experiments is recorded in column MIN and MAX respectively while an average value of ten experiments is recorded in column AVG. The values are calculated in percentage.

From Fig. 5 a visualization of a fitness function can be interpreted. The blank space between items and a rectangular bounding box that represent a container is considered as waste area. The goal of a fitness function is to decrease the amount of waste area.

Since the waste between items wants to be minimized, fitness function is defined as:

$$f(X,Y) = bbArea - polyArea \quad (1)$$

$$polyArea = \sum_{i=1}^{n} area(p_i) \quad (2)$$

$$bbArea = \mathrm{LCS}(X,Y)*\mathrm{LCS}(X^R, Y) \quad (3)$$

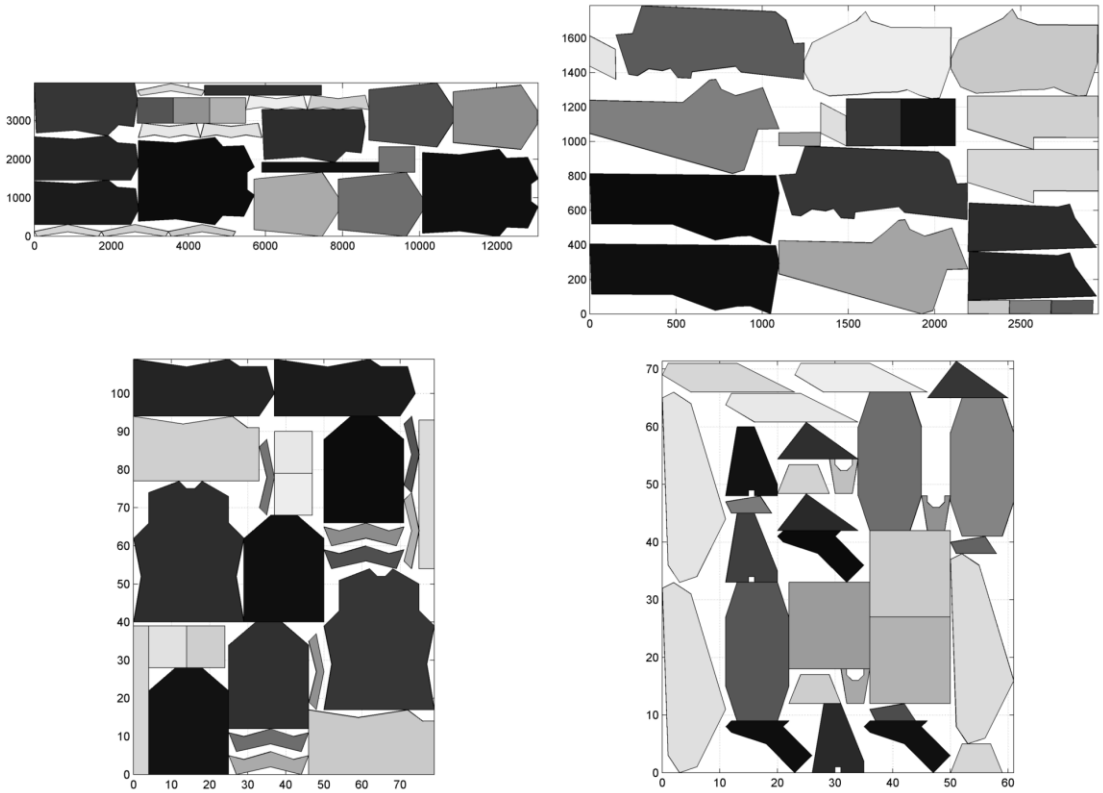


Figure 5. Best packing obtained for ALBANO (upper-left), DAGLI (upper-right), MARQUES (bottom-left) and MAO (bottom- right) dataset

The (*X*, *Y*) is a sequence pair, *polyArea* is the sum of area of all items in a data set, and *bbArea* is an area of a bounding box that is surrounding current packing. Value of *bbArea* is calculated using LCS algorithm.

In this research we tried to examine the efficiency of a rectangular polygon approximation and the ability of applying a rectangular packing method on 2D irregular polygon packing problem. We performed experiments with various iterations count in order to examine the rate in which the number of iterations influences packing density.

The best obtained packing density for the dataset ALBANO is 82.02%, for DAGLI 69.86%, for MAO 71.26%, and for MARQUES 83.54%.

Some of the best results for current algorithms are presented in [6] and [14]. For ALBANO dataset, packing density of 89.96% has been obtained by SAHA algorithm, 84.60% by BLF, 86.96% for 2DNest, 87.88% for BS, and 88.40% for AM.

For DAGLI dataset 87.15% has been obtained by SAHA, 83.70% by BLF, 87.05% by 2DNest, 87.99% by BS.

For MAO dataset 82.54% has been obtained by SAHA, 79.50% by BLF, 85.15% by 2DNest, 84.07% by BS, and 82.69% by AM.

For MARQUES dataset 88.14% has been obtained by SAHA, 86.50% by BLF, 89.17% by 2DNest, 88.92% by BS, and 87.80% by AM.

Some of our results are not comparable with the ones in literature since they allow rotation of polygons and we do not. We would also like to note that any compaction method has not been performed on our final packing so the results are expectedly worse than the ones in literature.

When analyzing the quality of a bounding box approximation we conclude that the effectiveness of an approximation depends on the shape of items in a dataset. The results for data sets ALBANO and MARQUES versus data sets MAO and DAGLI indicate this hypothesis. As it can be seen in Fig. 4, in data sets MAO and DAGLI the degree of irregularity of elements is higher than in MARQUES and ALBANO in terms that the elements are more similar to rectangle in MARQUES and ALBANO than in MAO and DAGLI.

The results in column DIFF record the average decrease in packing density of 10 experiments between rectangular packing and its corresponding polygon packing. These values confirm the above mentioned hypothesis that the waste contribution is greater the more the items are not similar to a rectangle. The DIFF value for MAO and DAGLI is roughly 24% while for MARQUES and ALBANO roughly 13%.

Overall we can conclude that a bounding box representation produces a lot of waste area between rectangular and polygonal packing. To make improvements some other sort of approximation might be considered such as convex hull, but the sequence pair method can only be applied on rectangles so we would not be able to implement this method.

Regarding the influence of iteration count on packing density, from Table I it can be seen that the most significant leap in packing density occurs between 10k and 50k iterations. Experiments with 50k and 100k

obtained similar results. Significant improvement in results between 100k and 500k iterations is not seen for any data set but there is significant time savings in time since the algorithm after with 100k iterations finishes 5 times faster.

Execution time for our experiments with 10k iterations varies from 20 to 30 seconds. With 50k iterations the execution time varies from 100 to 150 seconds, for 100k iterations from 200 to 300 seconds and for 500k iterations from 1000 to 1500 seconds. All results are dependent of the number of polygons in a set. In [15] and [16] the execution time of MATLAB code is compared with other programming languages. The execution of MATLAB code is from 10 (in [16]) to 500 times slower (in [15]) than the execution time of the same code in C++. This is seen as an opportunity for future performance optimization.

Overall it can be concluded that iterations count influences packing density. With the increase of iterations count the denser the packing is obtained. The best packings were obtained for 500k iterations for all data sets. These packings are shown in Fig. 4.

D. Future work

As it can be seen in Fig. 4, if items are irregular, bounding box approximation produces more waste, therefore leaving a lot of gaps between items. Applying some compaction method might be able to solve the problem. A compaction method based on non-linear programming (ILSQN) as described in [17] that solves the overlap optimization problem and controls the container size based on the success in each iteration might be used.

In [18] a group of sliding algorithms (*BLi*, *BLPi*, *BLDPi*, *BLFi*) commonly known as bottom-left which packs polygons by a series of downwards and leftwards movements is described.

In our experiments rotation of items has not been taken into account. We believe that allowing items to rotate would obtain better results even though the search space will be significantly increased.

A further adjustment of our proposed model might be implemented in terms of making the width of a container a fixed value as defined by the strip-packing problem.

V. CONCLUSION

In this paper an approach that uses sequence pair abstract representation is used to solve strip packing problem. Irregular polygons are approximated using rectangular bounding box and simulated annealing is then used to obtain a rectangle packing from a sequence pair. A polygon packing reconstruction is obtained by removing bounding box approximations.

Proposed methodology is test on 4 benchmark datasets from textile industry and the following best packing densities (in brackets) have been obtained: ALBANO (82.02%), DAGLI (69.86%), MAO (71.26%) and MARQUES (83.54%). The improvements to obtain a tighter packing can be implemented such as bottom-left heuristic and compaction strategies.

ACKNOWLEDGMENT

This work has been fully supported by Croatian Science Foundation under the project number 3011 Application of mathematical modeling and intelligent algorithms in clothing construction.

REFERENCES

[1] R. Zelenika and A. Grilec Kaurić, "Ocjena ekonomskog položaja tekstilne i odjevne industrije u Republici Hrvatskoj," *Ekon. Misao Praksa*, no. 2, pp. 543–566, 2011.

[2] J. A. Bennell and J. F. Oliveira, "The geometry of nesting problems: A tutorial," *Eur. J. Oper. Res.*, vol. 184, no. 2, pp. 397–415, Jan. 2008.

[3] J. Egeblad, "Heuristics for Multidimensional Packing Problems," Københavns Universitet, Faculty of Science, Datalogisk Institut, Department of Computer Science, 2008.

[4] H. Dyckhoff, "A typology of cutting and packing problems," *Eur. J. Oper. Res.*, vol. 44, no. 2, pp. 145–159, Jan. 1990.

[5] G. Wäscher, H. Haußner, and H. Schumann, "An improved typology of cutting and packing problems," *Eur. J. Oper. Res.*, vol. 183, no. 3, pp. 1109–1130, Dec. 2007.

[6] B. A. Júnior, P. R. Pinheiro, R. D. Saraiva, and P. G. C. D. Pinheiro, "Dealing with Nonregular Shapes Packing," *Math. Probl. Eng.*, vol. 2014, pp. 1–10, 2014.

[7] H. Murata, K. Fujiyoshi, S. Nakatake, and Y. Kajitani, "VLSI module placement based on rectangle-packing by the sequence-pair," *Comput.-Aided Des. Integr. Circuits Syst. IEEE Trans. On*, vol. 15, no. 12, pp. 1518–1524, 1996.

[8] X. Tang, R. Tian, and D. F. Wong, "Fast Evaluation of Sequence Pair in Block Placement by Longest Common Subsequence Computation," in *Design, Automation and Test in Europe Conference and Exhibition 2000. Proceedings*, Paris, 2000, pp. 106–111.

[9] D. Pisinger, "Denser Packings Obtained in $O(n \log \log n)$ Time," *Inf. J. Comput.*, vol. 19, no. 3, pp. 395–405, Aug. 2007.

[10] D. Domović, T. Rolich, D. Grundler, and S. Bogović, "Algorithms for 2D Nesting Problem Based on the No-Fit Polygon," in *Proceedings of the 37th International Convention MIPRO 2014: CIS - Intelligent Systems*, Opatija, Croatia, pp. 1344–1349.

[11] W. Han, J. A. Bennell, X. Zhao, and X. Song, "Construction heuristics for two-dimensional irregular shape bin packing with guillotine constraints," *Eur. J. Oper. Res.*, vol. 230, no. 3, pp. 495–504, Nov. 2013.

[12] D. L. Kreher and D. R. Stinson, "Combinatorial Algorithms: Generation, Enumeration, and Search," *SIGACT News*, vol. 30, no. 1, pp. 33–35, Mar. 1999.

[13] "ESICUP : ESICUP." [Online]. Available: http://paginas.fe.up.pt/~esicup/tiki-index.php. [Accessed: 29-Jan-2015].

[14] M. A. Shalaby, "A Particle Swarm Optimization Algorithm for a 2-D Irregular Strip Packing Problem," *Am. J. Oper. Res.*, vol. 03, no. 02, pp. 268–278, 2013.

[15] T. Andrews, "Computation time comparison between matlab and C++ using launch windows," 2012.

[16] S. B. Aruoba and J. Fernández-Villaverde, "A Comparison of Programming Languages in Economics," National Bureau of Economic Research, Working Paper 20263, Jun. 2014.

[17] T. Imamichi, M. Yagiura, and H. Nagamochi, "An iterated local search algorithm based on nonlinear programming for the irregular strip packing problem," *DISCRETE Optim.*, vol. 6, no. 4, pp. 345–361, Nov. 2009.

[18] E. Hopper, "Two-dimensional packing utilising evolutionary algorithms and other meta-heuristic methods," Doctoral Thesis, University of Wales. Cardiff, 2000.

Predicting Stock Market Trends Using Random Forests: A Sample of the Zagreb Stock Exchange

T. Manojlović* and I. Štajduhar*
* University of Rijeka, Faculty of Engineering, Department of Computer Engineering, Rijeka, Croatia
teo.manojlovic@gmail.com

Abstract - Stock market prediction is considered to be a challenging task for both investors and researchers, due to its profitability and intricate complexity. Highly accurate stock market predictive models are very often the basis for the construction of algorithms used in automated trading. In this paper, 5-days-ahead and 10-days-ahead predictive models are built using the random forests algorithm. The models are built on the historical data of the CROBEX index and on a few companies listed at the Zagreb Stock Exchange from various sectors. Several technical indicators, popular in quantitative analysis of stock markets, are selected as model inputs. The proposed method is empirically evaluated using stratified 10-fold cross-validation, achieving an average classification accuracy of 76.5% for 5-days-ahead models and 80.8% for 10-days-ahead models.

I. INTRODUCTION

Consider the following scenario. You are currently in possession of a substantial amount of money. Additionally, you do not completely approve the behavior of the late W.C. Fields, actor and writer, who has claimed to have spent half of his money on gambling, alcohol and wild women, and the other half he has wasted. Quite the opposite - you would like to save some of that money (the wasted bit, at least). What would be the best choice of investment for the future? You cannot just keep it in a drawer, because the inflation will reduce its actual value. One of the investment possibilities is buying specific stocks, bonds, or mutual funds from certain markets. But, how to be (almost) certain, that this investment will pay off? How to choose markets and exact paper assets, and put up an exact portfolio investment plan for the near(-er) future?

These questions are the reason why the analysis of stock market price movement has been in the focus of both investors and researchers for quite some time. Predicting stock price movement is considered to be a rather challenging task, according to the efficient-market hypothesis [1], which states that one cannot consistently achieve returns in excess of average market returns, given the information at the time the investment is made. To beat the average, in the absence of expert knowledge, we resort to using machine learning for predicting stock price movements. We are, of course, interested in the technical analysis of the markets, i.e. using historical market data to create actionable intelligence, which could be used to make a portfolio investment plan.

There are two common ways of predicting stock market behavior. The first one is based on the prediction of future *price values* of a stock. This approach usually requires treating the historical data as time series data, feeding the distinct time frame signals to an algorithm and trying to model the future time points in the signal, e.g. using ARIMA models. The second is based on predicting the future *price direction* of a stock, i.e. guessing whether the price will rise or fall the next day, or in a couple of days (trend forecasting). There are considerably less studies dealing with the prediction of future price direction of a stock [2].

Previous research showed promising results on the possibility of correctly predicting the price direction of a stock or market index. We mention some of that work here. Kim [3] was the first to apply support-vector machine (SVMs) to predict the next day stock price direction of Korea composite stock price index (KOSPI), using twelve technical indicators. Kara et al. [4] applied SVMs and artificial neural networks (ANNs) for building the model used to predict future movement on the Istanbul Stock Exchange (ISE) National 100 index. Hajek [2] proposed a 5-day-ahead prediction model of NASDAQ Composite Index using prototype generation classifiers. Masoud [5] applied artificial neural networks (ANNs) for building the model used to predict future direction of the Libyan Stock Exchange.

Further improvements in stock direction prediction research were made more recently, using hybrid machine learning models, selecting fundamental indicators [6] (price/earnings ratio, return on equity, etc.) and by discretising technical indicators. Choudhry and Garg [7] proposed a hybrid genetic-algorithm SVM-model for predicting the direction of stock price movement in the Indian stock market. Chandwani and Singh Saluja [8] used a combination of both fundamental and technical indicators and applied genetic-algorithm SVM- and ANN-hybrid models to predict the direction of 25 companies listed on the Bombay Stock Exchange. Patel et al. [9] discretised technical indicators by representing them as trend deterministic data and then trained predictive models using random forests, ANNs, naïve Bayes and SVM models on CNX Nifty, S&P BSE Sense indices and

The work presented in this paper is partially supported by the University of Rijeka research grant 13.09.2.2.16.

Reliance Industries and Infosys Ltd companies on the National Stock Exchange of India.

Concerning the prediction of future behavior of the Croatian stock market, most of the reported work involved is based on forecasting the future value of the stock prices and indices [10,11]. We are unaware of any studies dealing with the prediction of their future direction. In this paper we propose using a random forests model for predicting the trend direction of stocks and index values. We demonstrate its efficiency on the CROBEX index and on several companies listed at the Zagreb Stock Exchange (ZSE) from various sectors, namely HT-R-A, KRAS-R-A and PBZ-R-A. Predictive models were built and evaluated using stratified 10-fold cross-validation.

Structure of this paper is as follows. In the following section we describe the historical data and input parameters used for building the proposed models. In section 3 we give details on the machine learning model used. In section 4 we present the results of empirical evaluation. Finally, in section 5 we give a conclusion and present several ideas for future work.

II. RESEARCH DATA

Data used in this paper was gathered following the CROBEX index and a few companies listed at the Zagreb Stock Exchange from different sectors, namely HT-R-A (telecommunications), KRAS-R-A (confectioner) and PBZ-R-A (finance). They consist of several calculated technical indicators and 5-days-ahead / 10-days-ahead direction for each respective company or the index. Selected datasets cover the period from January 2008 to December 2013. Each time point in the selected datasets consists of the following features, used to calculate the technical indicators: open, close, high and low price. Their movement is shown in Figure 1.

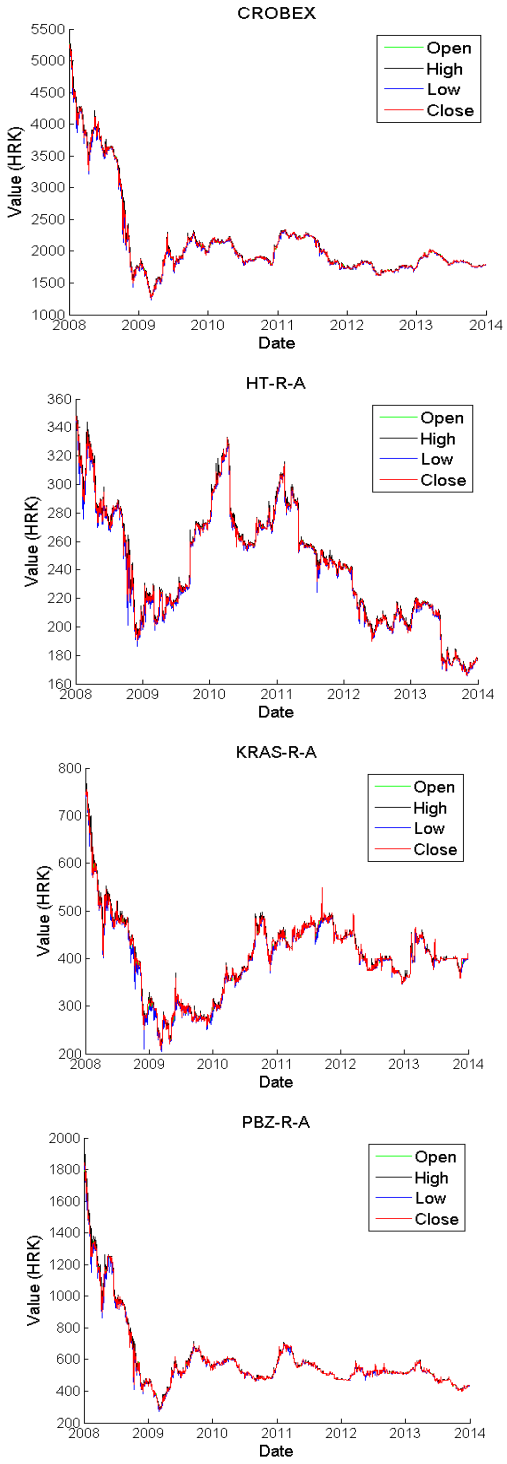


Figure 1. Price movement (open, high, low and close) of the CROBEX index and of the selected stocks: HT-R-A, KRAS-R-A and PBZ-R-A.
Can you spell "recession"?

Twelve different technical indicators were selected for model input, based on prior research [2,3,4,5]. Formulas for calculating the selected indicators are described in Table I. Some technical indicators used in [2,3,4,5] (such as Williams R%) are not used in our model because they show similar properties with respect to class as already used indicators. First two indicators are moving averages. They are used to smooth the stock price signal, making it easier for an investor to identify trends. Next two indicators are stochastic oscillators (*%K* and *%D*). Stochastic *%K* is used to compare the closing price at time *t*, relative to high-low range, over a period of *n* days, and stochastic *%D* is a simple moving average of *%K*, over a period of *n* days. Value of stochastic oscillators is varying from 0 to 100. If their value is higher than 80, it is generally believed that this means that the stock price is overbought. On the other hand, if their value is lower than 20, this means that the stock is oversold. Another indicator, *MACD*, is calculated by subtracting the slow exponential moving average from the fast exponential moving average, and it is commonly used in combination with the technical indicators mentioned above to suggest buying or selling of specific stocks. *CCI* is another indicator, used to detect whether the selected stock is overbought or oversold. It is calculated by subtracting the *n*-days simple moving average from the typical price, and then dividing it by the mean absolute deviation of the typical price. If the *CCI* value is greater than 100, then the stock is considered to be overbought. Otherwise, if the *CCI* value is lower than -100, selected stock is considered to be oversold. *RSI* is yet another indicator, used to detect whether a stock is overbought or oversold.

TABLE I. TECHNICAL INDICATORS USED AS MODEL INPUTS

Names	***Formulas***	***Parameters***
5-days and 10-days moving average ($SMA(5)_t$, $SMA(10)_t$)	$\frac{C_t + C_{t-1} + \cdots + C_{t-n+1}}{n}$	$n = 5$, $n = 10$
5-days and 10-days weighted moving average ($WMA(5)_t$, $WMA(10)_t$)	$\frac{(n)C_t + (n-1)C_{t-1} + \cdots + C_{t-n+1}}{(n + (n-1) + \cdots 1)}$	$n = 5$, $n = 10$
Stochastic %K ($\%K$)	$\frac{C_t - LL_n}{HH_n - LL_n}100$	$n = 14$
Stochastic %D (%D)	$\frac{\sum_{i=0}^{n-1}(\%K_{t-i})}{n}$	$n = 3$
Moving Average Convergence Divergence ($MACD$)	$EMA(fast)_t - EMA(slow)_t$	$fast = 12$, $slow = 26$
Commodity channel index (CCI)	$\frac{M_t - SM_t}{(0.015D_t)}$	$n = 20$ (see description below)
5-days and 10-days disparity	$\frac{C_t}{SMA(n)_t}100$	n=5, n=10
Price oscillator ($OSCP$)	$\frac{SMA(fast)_t - SMA(slow)_t}{SMA(fast)_t}100$	$fast = 5$, $slow = 10$
10-days Rate of change (ROC)	$\frac{C_t}{C_{t-n}}100$	$n = 10$
10-days momentum	$C_t - C_n$	$n = 10$
Relative strength index (RSI)	$100 - \frac{100}{1 + (\frac{\sum_{i=0}^{n-1} Up_{t-i}}{n})/(\frac{\sum_{i=0}^{n-1} Dw_{t-i}}{n})}$	$n = 14$
5-days standard deviation	$\sqrt{\frac{\sum_{i=0}^{n-1}(C_{t-i} - SMA(n)_t)^2}{n}}$	$n = 5$

The original time series inputs description: C_t denotes the closing price at day t, H_t denotes the high price at day t, L_t denotes the low price at day t, HH_n denotes the highest high price in n days, and LL_n denotes the lowest low price in n days. Calculated terms: $EMA(n)$ is the exponential moving average, $EMA(n)_t = EMA(n)_{t-1} + \alpha(C_t - EMA(n)_{t-1})$, α is a smoothing factor $\frac{2}{n+1}$, n is the number of data points used to calculate the exponential moving average, M_t is a typical price, $M_t = \frac{C_t + H_t + L_t}{3}$, SM_t is a simple moving average of typical prices for n days, D_t is a mean absolute deviation of typical prices for n days, and Up_t is the upward and Dw_t the downward price change at day t.

Its value ranges from 0 to 100. If the *RSI* value is above 70, this indicates that the stock is overbought, and if the *RSI* value is below 30, this indicates that the stock is oversold. *Disparity* and *ROC* are similar indicators. *Disparity* is the ratio of closing price at time *t* and the *n*-days simple moving average, and *ROC* is the ratio of the closing price at time *t* and the closing price at time *t-n*. If the values of these two indicators are both above 0, this is a signal for an investor to buy the selected stock. On the other hand, if the values of these two indicators are below 0, this is a signal for an investor to sell the selected stock. *Price oscillator* is a technical indicator that displays the difference between the fast and the slow moving average, divided by the fast moving average. It generates the buy signal if the value of the fast moving average rises above the slow moving average, and generates the sell signal otherwise. *Momentum* is calculated as a difference between the stock price at time *t* and the stock price at time *t-n*. Negative momentum is often used by investors to generate the buy signal, and positive momentum to generate the sell signal. Last indicator used is *standard deviation*. It is used to measure volatility in a period of *n* days. Technical indicators, used in this paper, were calculated using MATLAB and R.

5-days-ahead direction of selected stocks and the index used is labeled "1", if the closing price at time *t*+5 is greater than the closing price at time *t*. 10-days-ahead direction of selected stocks is labeled "1", if the closing price at time *t*+10 is greater than the closing price at time *t*. Otherwise, the directions are labeled "0". The ratio of 5-days-ahead directions is shown in Table II and the ratio of 10-days-ahead directions is shown in Table III.

III. MODELS AND METHODS

Predictive model used in this paper is based on the random forests algorithm, firstly introduced by Breiman [12]. Random forests algorithm belongs to ensemble learning methods. The main idea behind ensemble learning methods is that a single classifier is sometimes not sufficient for correct classification of test data and that multiple classifiers can increase model correctness.

The principle of random forests is simple. First, a set of *k* bootstrap samples *D* is created. Bootstrap samples are samples of same size that are created using sampling with replacement. This means that every sample can be used again, even if it was already used. Then for every bootstrap sample, a decision tree is built using *n* randomly chosen features. The most common algorithms for building decision trees are ID3 [13], C4.5 [14] and CART [15]. When inferring its class, every new instance is classified by *k* trees and the prediction that gains the

TABLE II. PERCENTAGE OF 5-DAYS-AHEAD INCREASE/DECREASE

Symbol	***Increase***	***Decrease***
CROBEX	45.5%	54.5%
HT-R-A	45.6%	54.4%
PBZ-R-A	43.1%	56.9%
KRAS-R-A	44.8%	55.2%

TABLE III. PERCENTAGE OF 10-DAYS-AHEAD INCREASE/DECREASE

Symbol	***Increase***	***Decrease***
CROBEX	45.6%	54.4%
HT-R-A	45.8%	54.2%
PBZ-R-A	43.2%	56.8%
KRAS-R-A	45.7%	54.3%

majority of votes is selected as the class label.

There are several advantages of random forests. For one, they are easily parallelised, which means they can be used for large scale machine learning. Their performance is comparable to other ensemble learning algorithms, such as Adaboost and often it performs even better. Random forests are also faster than boosting and bagging. Finally, they perform well on noisy data containing outliers.

In this paper, we used the random forests algorithm, implemented in the WEKA (Waikato Environment for Knowledge Analysis) toolkit [16], for performing the classification task. WEKA uses a combination of simple random trees as simple classifiers, which are then used for voting to obtain class label. WEKA uses a combination of a bagging algorithm and random trees for the random forests algorithm. Bagging algorithm is used to perform bootstrapping and random tree algorithm is used to train decision trees with randomly chosen features.

For building the predictive models used in this study, we tested multiple parameters for WEKA random forests. Best results were obtained by setting the maximum depth of trees to unlimited, setting the number of features used in random selection to 5 and setting the number of trees to 100. Model used in this paper is illustrated in Figure 2.

IV. EMPIRICAL EVALUATION

For evaluation of the used predictive models, classification accuracy and F-measure were used. Classification accuracy is calculated by dividing the number of correctly classified instances with the overall number of instances. In other terms, formula for calculating the classification accuracy is given by

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (1)$$

where TP stands for the number of true positives, TN stands for the number of true negatives. True positives are positive instances classified as positive and true negatives are negative instances classified as negative. FP stands for the number of false positives and FN stands for the number of false negatives. False positives are negative instances classified as positive and false negatives are positive instances classified as negative.

For calculating the F-measure, it is necessary to calculate precision and recall:

$$Precision_{positive} = \frac{TP}{TP+FP}, \quad (2)$$

$$Recall_{positive} = \frac{TP}{TP+FN}, \quad (3)$$

$$Precision_{negative} = \frac{TN}{TN+FN}, \quad (4)$$

$$Recall_{negative} = \frac{TN}{TN+FP}. \quad (5)$$

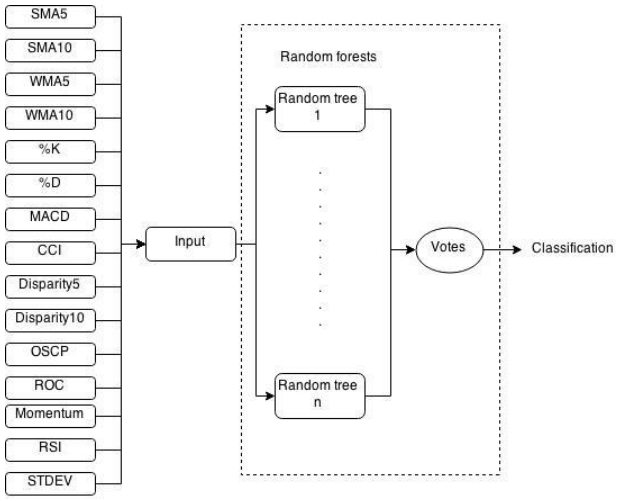


Figure 2. Visual representation of the model used in this paper

F-measure, also named the F-score, is defined as the harmonic mean of precision and recall:

$$F_{positive} = 2\frac{precision_{positive} * recall_{positive}}{precision_{positive}+recall_{positive}}, \quad (6)$$

$$F_{negative} = 2\frac{precision_{negative} * recall_{negative}}{precision_{negative}+recall_{negative}}. \quad (7)$$

Given the negative and positive F-measure, we can now calculate the weighted average F-measure.

Proposed models were evaluated using stratified 10-fold cross-validation. In stratified 10-fold cross-validation, data is divided into 10 randomly partitioned subsamples of approximately equal class distributions. Nine out of ten subsamples are then used to train the model and the remaining subsample is used for validation. This procedure is repeated 10 times, every time using a different partition for validation. The result is then calculated as the average of results obtained for every model. Results of 5-days-ahead direction predictions are shown in Table IV and the results of 10-

TABLE IV. PERFORMANCE OF 5-DAYS-AHEAD MODELS

Symbol	*Accuracy*	$F_{positive}$	$F_{negative}$	*Weighted average F-score*
CROBEX	0.767	0.726	0.797	0.765
HT-R-A	0.772	0.739	0.798	0.771
PBZ-R-A	0.761	0.711	0.797	0.76
KRAS-R-A	0.76	0.721	0.789	0.759

TABLE V. PERFORMANCE OF 10-DAYS-AHEAD MODELS

Symbol	*Accuracy*	$F_{positive}$	$F_{negative}$	*Weighted average F-score*
CROBEX	0.816	0.79	0.836	0.815
HT-R-A	0.825	0.804	0.842	0.825
PBZ-R-A	0.81	0.767	0.84	0.808
KRAS-R-A	0.787	0.758	0.81	0.787

days-ahead predictions are shown in Table V.

Experimental results suggest that random forests achieve significant performance accuracy in stock market trend prediction. Average accuracy of 5-days-ahead models is 0.765 (76.5%) and weighted average F-measure is 0.763. Average accuracy of 10-days-ahead models is 0.808 (80.8%) and weighted average F-measure is 0.808. If we compare our models with those of Hajek [2], who has trained models for predicting 5-days-ahead direction of NASDAQ Composite Index, we can see that our models perform somewhat better than most of the alternative models. Inspecting the calculated results, we can safely assume that the proposed predictive models can be successfully applied for portfolio management on the ZSE market.

V. CONCLUSION AND FUTURE WORK

In this paper, we used the random forests algorithm to build a model used to predict 5-days-ahead and 10-days-ahead directions of the CROBEX index and selected stocks. Our results show that random forests can be successfully utilised for building predictive models for predicting the direction of stock market trends.

From the cited papers, mentioned in introduction, it can be seen that empirical prediction results vary greatly. For example, if we compare different attempts of building 1-day-ahead models using SVMs, on one side, we have Kim [3] with very bad average accuracy (less than 60%) and, on the other side, we have Kara et al. [4], with much higher average accuracy (71.52%). We too have tried to build 1-day-ahead models on CROBEX and selected stocks using SVMs, ANNs and random forests, but only managed to achieve accuracy similar to that reported by Kim [3]. This could be explained with the inherent market volatility and with the fact that the influence of some technical indicators can be seen only a couple of days in the future.

There are several research directions that should be pursued in the future. One of them involves additionally improving the predictive accuracy of the model. We believe that this can be done in two ways. First, other technical indicators, not mentioned in [2,3,4,5], should be explored, especially the indicators that include volume. Next, hybrid models could be used to further improve the model accuracy. Regardless, we plan to continue our work on constructing an algorithm for automated trading, based on our predictive model, and use it for portfolio management.

REFERENCES

[1] E. F. Fama, Random walks in stock market prices. Financial Analysts Journal, 21, 55–59, 1965.

[2] P. Hajek, Forecasting Stock Market Trend using Prototype Generation Classifiers, WSEAS Transactions on Systems, Vol.11, No. 12, pp. 671-80, 2012.

[3] K. Kim, Financial time series forecasting using support vector machines. Neurocomputing, 55, pp. 307–319, 2003.

[4] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan, Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange, Expert systems with Applications, 38, pp. 5311–5319, 2011 .

[5] N. Masoud, Predicting direction of stock prices index movement using artificial neural networks: the case of Libyan financial market, British Journal of Economics, Management & Trade, 4(4): 597-619, 2014.

[6] S. Bareša, S. Bogdan, and Z. Ivanović, Strategy of stock valuation by fundamental analysis, Special issue, UTMS Journal of Economics 4 (1): 45–51, 2013.

[7] R. Choudhry and K. Garg, A hybrid machine learning system for stock market forecasting, World Academy of Science, Engineering and Technology, 39, 315−318, 2008.

[8] D. Chandwani and M. Singh Saluja, Stock direction forecasting techniques: an empirical study combining machine learning system with market indicators in the Indian context, International Journal of Computer Applications (0975–8887) Volume 92, No.11, 2014.

[9] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques, Expert Systems with Applications, Volume 42, Issue 1, Pages 259-268, ISSN 0957-4174, 2015.

[10] Z. Ivanović, S. Bogdan, and S. Bareša, Forecasting Croatian stock market index: Crobex, UTMS Journal of Economics 4 (2): 79–91, 79, 2013.

[11] I. Svalina, V. Galzina, R. Lujić, and G. Šimunović, "An adaptive network-based fuzzy inference system (ANFIS) for the forecasting: The case of close price indices", Expert Systems with Applications, 40, pp. 6055-6063, 2013.

[12] L. Breiman, Random forests, Machine Learning, 45(1):5–32, 2001

[13] J. R. Quinlan, Induction of decision trees, Machine Learning 1, 1 (Mar. 1986), 81-106, 1986.

[14] J. R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann, 1993.

[15] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and regression trees, Wadsworth and Brooks, Monterey, CA, 1984.

[16] M. Hall et al., The WEKA data mining software: an update, SIGKDD Explorations, Volume 11, Issue 1, 2009.

GA-ANFIS Expert System Prototype for Detection of Tar Content in the Manufacturing Process

Lejla Begic Fazlic*, Zikrija Avdagic** and Ingmar Besic**
* *Tobacco Factory Sarajevo,* Sarajevo, BiH
**University of Sarajevo – Faculty of Electrical Engineering, Sarajevo, BiH
lejla.begic@fds.ba, zikrija.avdagic@etf.unsa.ba, ingmar.besic@etf.unsa.ba

Abstract - The purpose of this study is to present novel GA-ANFIS expert system prototype for tar detection in cigarettes during manufacturing process. The proposed system combines capabilities of Adaptive Neuro-Fuzzy Inference System (ANFIS) and Genetic Algorithm (GA).The data recorded for different type of cigarettes are collected by special control quality equipment in real conditions inside cigarette factory. GA-ANFIS system performs optimization in two steps. In the first step it generates six different ANFIS structures, and after that, we have second level of GA optimization using given ANFIS structures resulting in optimal fuzzy model structure. Modeling and validation of the GA-ANFIS system approach is performed in MATLAB environment using validation data set that were not used in the process of training. Our earlier research results based on two different approaches (ANFIS and High-performance liquid chromatography (HPLC)) are also introduced. Performances of these three approaches are compared and novel expert system prototype shows better result related to training, testing and validation errors. It also more precisely shows that low yield tar cigarettes contain similar levels of nicotine opposite to high yield tar cigarettes while benzene, toluene, and xylene (BTX) levels rise along with increasing tar yields.

I. INTRODUCTION

Three main tobacco smoke components harmful to health are nicotine, carbon monoxide and tar. Tar is the term used to describe the toxic chemical residue left behind from burning cigarettes. In 2010, a law that bans tobacco manufacturers from using the terms “light,” “low,” or “mild” in tobacco product labeling, went into effect. Tar is present in all cigarettes and tends to increase as the cigarette is burnt down, which can mean that the last puffs on a cigarette may contain as much as twice the amount of tar as the first puffs. The exact composition of the tar that smokers inhale is uncertain. Produced cigarette is not consumed in the state it is presented, but in the form of smoke which is the basic product being formed in a complex combusting process. When lighting a cigarette, at the end of smoke pillar is formed combusting zone. The material within cigarette is exposed to high temperatures, with a presence of certain amount of oxigen, so it does not burn completely. Two burning processes take place while the cigarette is "burning" [11]. One process takes place when inhaling smoke (Main Stream), while the second one appears in the intervals between two smoke inhaling - so called smolder (Side Stream). Combusting starts with the presence of oxigene, while reactions are reflected in creating certain gasses. Chemical characteristics are now measured on "smoker's machine" that operates under standardize conditions. When the smoke is "cought' according to ISO method, components being fully deposeted on the filter are marked as a total smoke condensate. By removing water and nicotine from the total smoke condensate we get a substance that is marked as TAR. When analyzed by special High Performance Liquid Chromatography (HPLC), the smoke from a so-called light cigarette has a lower yield of tar than the smoke from a regular cigarette. However, a machine cannot predict how much tar a smoker inhales and need 12 hours to determine tar content in any cigarette. Cigarette quality process has a lot controllable variables and we use in our study four: porous cigarette paper diameter, ventilating filters, nicotine and CO.

There have been suggested many approaches to nonlinear system control. However, the problem gets more complex when uncertainties are considered.

In research paper [1], authors used ANFIS in their recent works, to train the mathematical models and provide short-term load forecast. At the same time, GA was responsible to analyze database and possible correlations with consumption. Some authors suggested combination of Adaptive Neuro Fuzzy System, Genetic Algorithm and Singular Value decomposition [9]. The incorporation of genetic algorithm into ANFIS model design starts by presenting the real value parameters as antecedent part of fuzzy system. Authors in research study [12] suggested hybrid GA-ANFIS model in which GA optimizes structure and the number of fuzzy IF-THEN rules by finding best value of sub-clustering method. On the other hand, ANFIS is used to optimize initial FIS based on the clustering parameter generated by GA.

A novel GA-ANFIS expert system prototype for tar prediction in cigarettes is presented in our study. The purpose of this study is to reach smaller prediction errors (in regard to ANFIS model) by developing new GA-ANFIS expert system. Our novel GA-ANFIS system approach is trained with a subset of different parameters of cigarette data. The effectiveness of the GA-ANFIS

system approach is tested and verified by the use of validation data. Modeling and validation of the novel GA-ANFIS system approach is performed in MATLAB environment. The results are compared with ANFIS and HPLC [7], and they show the effectiveness of the proposed approach.

GA-ANFIS system performs optimization in two steps. In the first step it generates six different ANFIS structures, and after that, we have second level of GA optimization using given ANFIS structures resulting in optimal fuzzy model structure. The time spent in searching for the correct configuration by GA-ANFIS and also the amount of training samples utilized to train the ANFIS are optimal and less than time spent in ANFIS..

The paper is set up as follows; in section one, we have described the process of different cigarette components and described process of tar detection. In section two, we have explained the methods used in our research along with the description of Adaptive Neuro Fuzzy Inference architecture (structure of each layer) and Genetic algorithm covered in this work. Section three is about novel GA-ANFIS system approach where we have described GA-ANFIS system design architecture. We have explained the steps taken in using GA-ANFIS Matlab user interface that contains several subsystems. Section five gives the numerical results about training, testing and validation error compared with ANFIS and HPLC results. Discussion is summarized in section five. Finally, the conclusion is summarized in section six.

II. MATERIALS AND METHODS

A. Data preparation

A different kind of laboratory test stations for cigarette parameters measuring are used in different phases of cigarette production. Chemical characteristics are measured on "smoker's machine" that operates under standardize conditions. When the smoke is "cought' according to ISO method, components being fully deposeted on the filter are marked as a total smoke condensate. By removing water and nicotine from the total smoke condensate we get a substance that is marked as TAR. Each data sample consists of five values: the diameter of a cigarette, filter ventilation, nicotine and CO.

The total number in the base was 1020 data samples. The group is divided in two parts. In the developing phase, we used 950 data that were used for training, testing and checking. 70 separated data did not participate in the training and testing, but were applied on the finished model in the application phase. We used as an output parameter tar results collected from about 1020 cigarette samples that Quality control measured during the last two years by HPLC.

B. Anfis

Adaptive Neural-Fuzzy Inference System (ANFIS) was proposed by Roger Jang in 1992 [4][5]. We used ANFIS architecture with four - input - one output and sixteen rules (Fig.1).

ANFIS has five layers within which node functions of the same layer have the same function type as described below.

Layer 1: Every node i in this layer is an adaptive node with node function refer to equation in the text as (1) - (4):

$$O_{1,i}=\mu_{Ai}(D),\ i=1,2 \quad (1)$$

$$O_{1,i}=\mu_{Bi-2}(FV),\ i=3,4 \quad (2)$$

$$O_{1,i}=\mu_{Ci-4}(N),\ i=5,6 \quad (3)$$

$$O_{1,i}=\mu_{Ci-6}(C),\ i=7,8 \quad (4)$$

where D (diameter), FV (filter ventilation), N (nicotine), C (carbon monoxide) are premise parameters and A_i, B_i, C_i, D_i, $i=1,2$ are the linguistic label associated with node function.

Layer 2: Every node i in this layer is a fixed node labeled Π, whose output is the product of all the incoming signals refer to equation in the text as (5).

$$O_{2,i}=\omega_i=\mu_{Ai}(D)\times\mu_{Bi}(FV)\times\mu_{Ci}(N)\times\mu_{Di}(C),\ i=1,..16 \quad (5)$$

Layer 3: Every node i in this layer is a fixed node labeled N. The i-th node calculates the ratio of the i-th rule's firing strength to the sum of all rules' strengths refer to equation in the text as (6).

$$O_{3,i}=\omega_i=\omega_i/(\omega_1+\omega_2+...+\omega_{15}+\omega_{16}),\ i=1,..16 \quad (6)$$

Layer 4: Every node i in this layer is adaptive node with node function refer to equation in the text as (7),

$$O_{4,i}=\omega_i(p_i(D)\times q_i(FV)\times r_i(N)\times s_i(C)+z_1,\ i=1,..16 \quad (7)$$

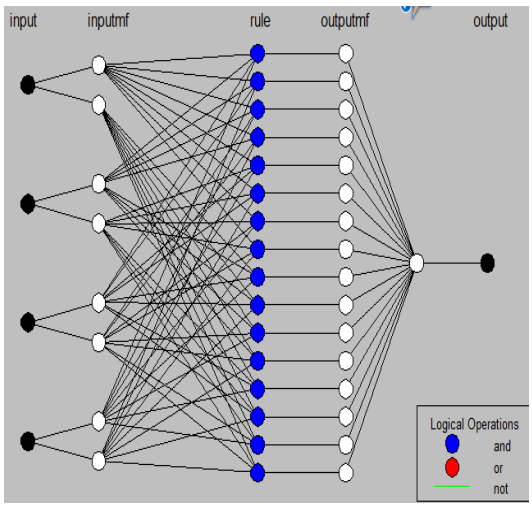


Figure 1. ANFIS architecture utilized in this work.

where { *pi qi ri si zi*}are consequent parameters [3,4], updated through Recursive Least-Squares Estimation.

Layer 5: The single node in this layer is a fixed node labeled Σ, which computes the overall output as the summation of all the incoming signals refer to equation in the text as (8) (*i=1,..16*).

$$O_{5,i}=\sum_i \omega_i f_i=(\omega_1 f_1+\ldots+\omega_{16} f_{16})/(\omega_1+\ldots+\omega_{16}), \quad (8)$$

C. Genetic Algorithm

Genetic algorithm (GA) is an optimization algorithm that incorporates the process of evolution. GA uses a population of chromosomes, each representing a solution to the problem that has been solved. It is assumed that a potential solution to a problem may be represented by set of parameters [10].

After initialization, the population is evaluated and stopping criteria is checked. If none of the stopping criteria is met, new population is generated again and the process is repeated until one or more of the stopping criteria are met. Crossover is a genetic operator that combines two chromosomes to produce a new chromosome.

The idea behind crossover is that the new chromosome may be better than the chromosomes of both parents, if it takes the best characteristics from each of the parents. Crossover occurs during evolution, according to a user-definable crossover probability.

We used different types of crossover: one point, two point, heuristic and arithmetic. We use Gaussian mutation to increase the performance of genetic algorithm.

Fig. 2 shows the genetic algorithm implemented in the experiment.

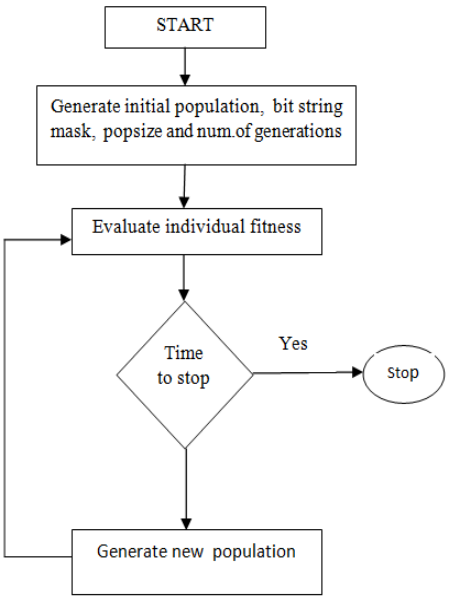


Figure 2. GA flowchart

III. GA -ANFIS SYSTEM DESIGN AND MATLAB IMPLEMENTATION

The GA-ANFIS system application was developed in MATLAB. The size of chromosome structure- fuzzy structure was generated by ANFIS computational complexity (Table I)[8].

For NumIn=4 (input variables) and NumMf=2 (number of membership functions) we have concrete chromosome structure (Fig. 3). The process flow chart is shown in Fig. 4.

TABLE I. ANFIS COMPUTATIONAL COMPLEXITY

Layer type	Nodes	Parameters
$O_{0,i}$ Inputs	NumIn	0
$O_{1,i}$ Values	(NumMf*NumIn)	3*(NumMf*NumIn)
$O_{2,i}$ Rules	$NumMf^{\wedge NumIn}$	0
$O_{3,i}$ Normaliz.	$NumMf^{\wedge NumIn}$	0
$O_{4,i}$ Lin.func..	$NumMf^{\wedge NumIn}$	$(NumIn+1)*NumMf^{\wedge NumIn}$
$O_{5,i}$ Sum	0	1

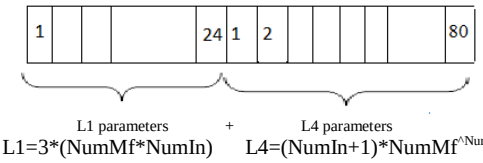


Figure 3. Ga - Anfis Chromosome structure

First part of the system is data base initialization where four different cigarette parameters data sets are created (no.1 in Fig.5). We used one set data for training and another for testing. Half of testing data are used for checking. Validation set of data is used for prediction.

Second step (no.2 in Fig. 5) is ANFIS modeling where we made initialization parameters for ANFIS network. Different fuzzy structures are generated by choosing one of the following partitioning techniques: Sub-Clustering and Grid Partitioning, in combination with two optimization methods (hybrid and back propagation).

Third part (no.3 in Fig. 5) of the system is GA-ANFIS modeling where we defined GA parameterization: defined population size, number of generation, fitness limit and type of crossover operation.

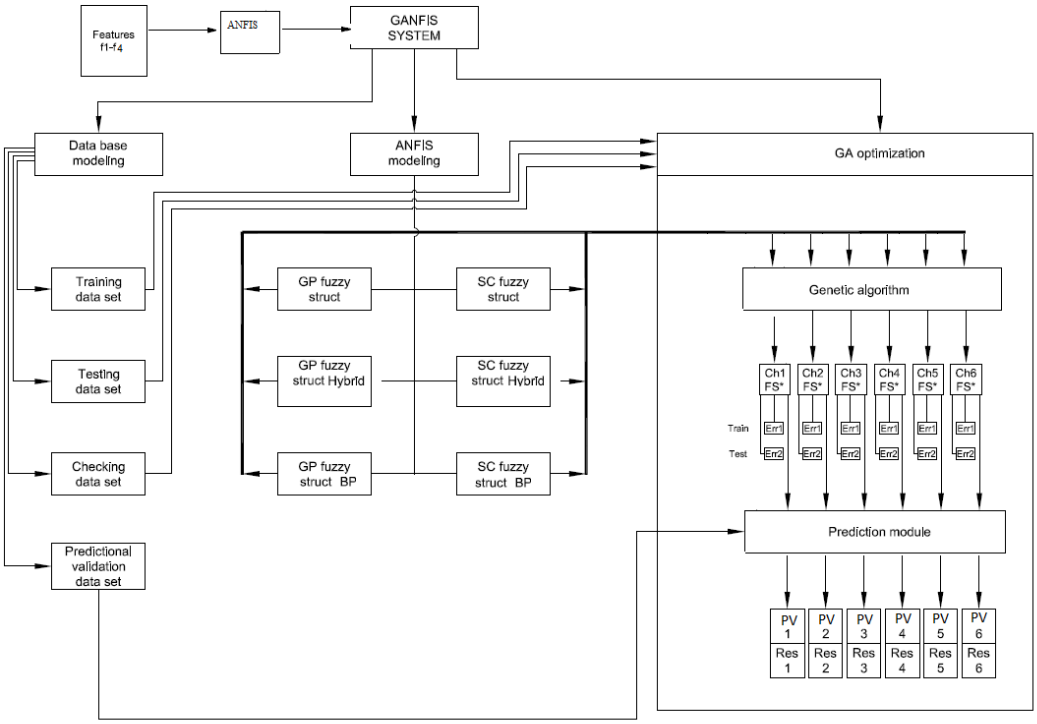


Figure 4. Process flow chart GA-ANFIS system algorithm; GP-Grid Partitioning, SC-Subtr.Clust CH-Chromosome, FS- Fuzzy structure, PV-Predicted

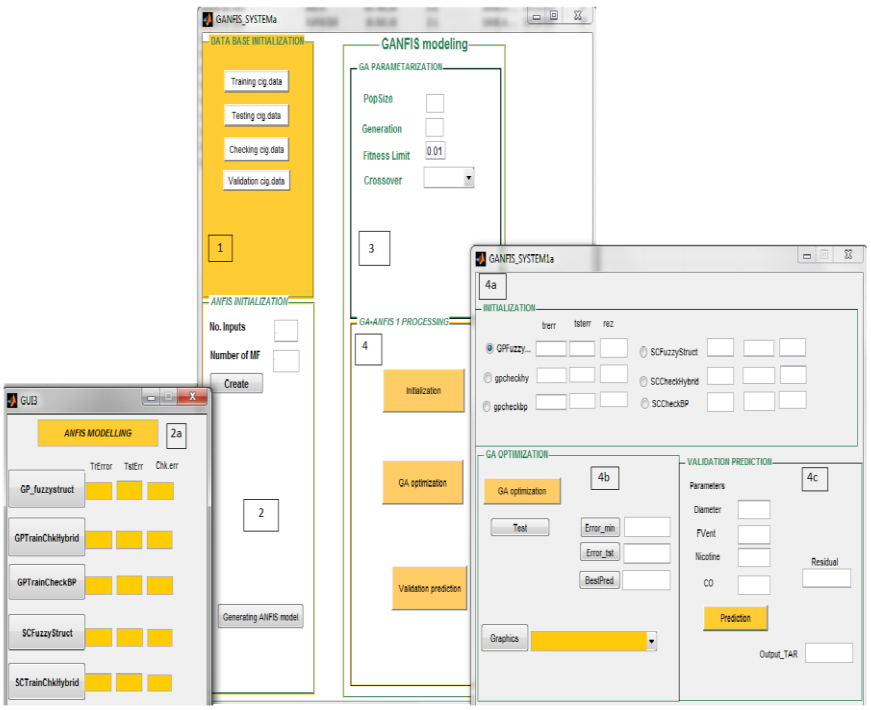


Figure 5. GA- ANFIS System GUI

Fourth part of the system (no.4 in Fig. 5) is GA-ANFIS processing where we initialized fuzzy structures (4a) and optimized them by genetic algorithm (4b). Different chromosome-fuzzy structures are created on GA-ANFIS system output, and used for testing and obtaining training and testing errors.

Validation of the novel GA-ANFIS system approach is performed in MATLAB environment by using validation data set (4c).

IV. TESTING AND RESULTS

In data base processing, we have divided sample data in four groups: 1020 different kind of cigarette samples is used for training, testing and checking (from 2013 till now) and one independent set of data (70 individuals) is used for model validation. In each experiment were 20 individuals for the GA initial random populations, 200 generation for the GA and fitness limit 0.01. When process training is completed, real prediction of the system takes place. As you can see the prediction is very accurate in most of the cases (Fig.6). RMSE error is shown in Fig. 7.

The results of the collected data give the best residual of prediction 0.02. Table II-IV shows the results of the test success of GA-ANFIS versus ANFIS and High Performance Liquid Chromatography (HPLC) for different optimization method used in ANFIS.

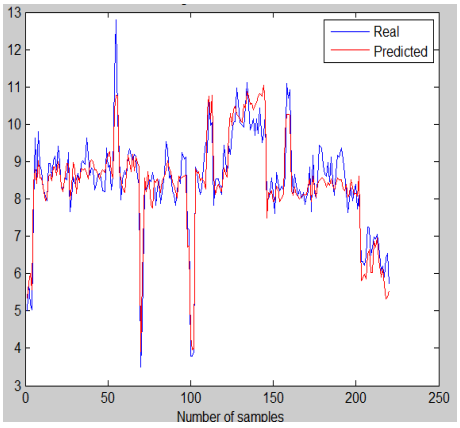


Figure 6. Real vs Predicted values

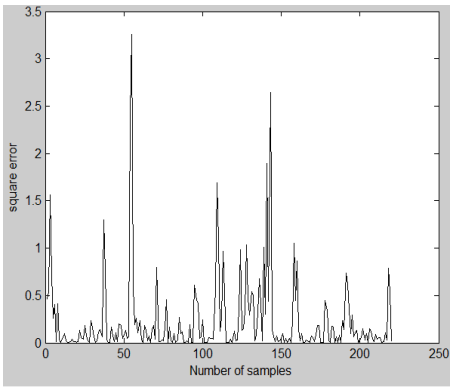


Figure 7. Square error

TABLE II. GA -ANFIS TRAINING ERROR VS TRAINING ERROR IN ANFIS AND HPLC [7]

Method	GP backProp	GP Hybrid	SC BackProp	SC Hybrid
GA ANFIS	1,23	1.23	0.54	0.57
ANFIS	4.53	0.32	0.62	0.63
HPLC	3,56	3,56	3,56	3,56

TABLE III. GA -ANFIS TESTING ERROR VS TESTING ERROR IN ANFIS AND HPLC [7]

Method	GP backProp	GP Hybrid	SC BackProp	SC Hybrid
GA ANFIS	0.27	0.69	0.62	0.57
ANFIS	0.36	3.84	0.38	0.68
HPLC	1,07	1,07	1,07	1,07

TABLE IV. GA-ANFIS PREDICTION ERROR VS PREDICTION ERROR IN ANFIS [7]

Method	GP backProp	GP Hybrid	SC BackProp	SC Hybrid
GA - ANFIS	0.16	0.34	0.07	0.02
ANFIS	0.36	1.50	0.61	0.38

V. DISCUSSION

The comparison results show that our novel approach has better performances than approach based on single ANFIS model in our previous study [1]. It is an approach based on cascade optimization. Physical cigarette parameters have great impact on the TAR content in the cigarette smoke. As the amount of tobacco in cigarette increases, so the TAR production slightly rises. By decreasing TAR amount, concentration of harmful substances is also decreased, because of which is tar subject of legal restrictions and standards. The amount of nicotine and tar is linked with ventilation of filter and cigarette paper, so it could be even two or three times smaller in ventilated cigarettes. Analysing various cigarrete samples, we determined that nicotine in the main smoke stream (inhalation) did not much differ from the nicotine in the mixture, which confirms the fact that the tar depends more on technological parameters and cigarette construction than on the amount of nicotine.

On the other hand, the results we have achieved by this novel approach are more useful for the Quality control during the manufacturing process, because they can get information about tar during processing.

With these results, one can make some changes on physical parameters being used as input variables for ventilation control inside range, in order to increase the reduction of smoking components level, primarily tar.

The training of the Root Mean Squared Error via GA-ANFIS is 0.57; the checking of RMSE is 0.02. Prediction results of the amount of tar in manufacturing process require a

number of additional chemical analyses. The results of nicotine and tar measured via HPLC are detectable with 1,07 % error. GA-ANFIS access enables faster prediction of tar amount with a permitted deviation which is less than 0.02%.

VI. CONCLUSION AND FUTURE WORK

In this paper we built a novel GA-ANFIS system approach using ANFIS modeling and Genetic optimization technique for tar detection. The experiments have demonstrated that this novel algorithm is suitable for the training, testing and prediction for tar detection in different types of cigarette data acquired from different types of machines.

The study shows that tar reduction in light and non-light cigarettes remains identical by increasing values of one of the inputs (cigarette filter ventilation). We also concluded, by analysing impacts and other input variables being trained and tested through our model, that low yield cigarettes contain similar levels of nicotine as compared to high yield cigarettes while BTX levels increase with increasing tar yields.

The following can be inferred from the experiments:

- The basic cigarettes features are used as inputs to classifiers based on Adaptive Neuro - Fuzzy Inference Systems (ANFIS) for first level of fuzzy model optimization.
- GA performs optimization in two steps: in the first step it generates six different ANFIS structures. Each of these ANFIS structures combines one of partitioning techniques: Sub-Clustering and Grid Partition, one of two optimization methods: hybrid and back propagation, and training or checking data set, which is six combinations in total. The results from ANFIS models are inputs to the second level of the optimization that outputs and optimized fuzzy model structure.
- Optimized fuzzy structure, which is a result of the novel GA-ANFIS system approach, is validated in MATLAB environment by using validation data set. Training, testing and model validation errors are collected during fuzzy model structure creation.
- Three functions are created to give answer for best result - least training, testing and validation error.

In this study, we tried to predict tar concentration in different type of cigarettes. We did not have any information about tobacco or other additive components in cigarette rod. We compared our GA-ANFIS system with ANFIS and HPLC.

In our approach we use two-stage cascade optimization, where six types of ANFIS were parallel processed. Originality is in the approach that uses global optimization of the genetic algorithm, so to find ANFIS structure whose parameters give the best preciseness, from the aspect of model accuracy.

If it had worked without GA, finding of ANFIS optimal structure would have taken very long time, and we would not have objective results (because we did not do through research, as GA did).

The next phase of our research will be focused on fusion GA-ANFIS model with chromosome configuration composed of three parts: six ANFIS parameters, L1 parameters as in presented approach, and L4 parameters as in presented approach.

ACNOWLEDGMENTS

The support of "Tobacco Factory Sarajevo" company in data collection is acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the company mentioned above.

REFERENCES

[1] K.Kampouropoulus, F.Andrade, A. Garcia, L. Romeral, "A Combined Methodology of Adaptive Neuro-Fuzzy Inference System and Genetic Algorithm for Short-term Energy Forecasting", Advances in Electrical and Computer Engineering, Volume 14, Issue 1, 2014, pp. 9 – 14, ISSN: 1582-7445, e-ISSN: 1844-7600, DOI: 10.4316/AECE.2014.01002.

[2] Z. Avdagic, L.B. Fazlic, "Detection of Nicotine Content Impact in Tobacco Manufacturing using Computational Intelligence", EFMI special topic conference, Stud Health Technol Inform.165, 2011, pp. 37-42.

[3] Z. Avdagic, L. B. Fazlic, H. Fatkic, "Automatic detection of CO Content in Smoke Condensate using Adaptive Neuro Fuzzy Inference Systems", Inista 2007, June 20-23, 2007.

[4] J.-S.R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system", IEEE Trans. Syst. Man Cybern. 23 (3), 1993, pp. 665–685.

[5] J.-S.R. Jang, "Self-learning fuzzy controllers based on temporal backpropagation", IEEE Trans. Neural Networks 3 (5), 1992, pp. 714–723

[6] Muhamed Jalali Vernamkhasti, "ANFISGA- Adaptive Neuro-Fuzzy Inference System Genetic Algorithm", Global Journal of Computer Science and Technology, 2011, Volume 11 Issue 1 Version 1.0.

[7] Z. Avdagic, L. B. Fazlic, S. Konjicija, "Optimized Detection of Tar Content in the Manufacturing Process Using Adaptive Neuro -Fuzzy Inference Systems", Proceedings of MIE 2009: Medical Informatics in a United and Healthy Europe, 2009, ISBN 978-1-60750-044-5.

[8] Pierro Bonissone, "Adaptive Neural Fuzzy Inference System", Analysis & Application, http://homepages.rpi.edu/~bonisp/fuzzy-course/Papers-pdf/anfis.rpi04.pdf .

[9] M. Soleimani, K. Salmalian, "Genetic Algorithm Optimized ANFIS Networks for Modeling and prediction of Energy Absorption rate of Brass Energy Absorbers", Journal of Applied Mathematics Oct. (2008), Islamic Azad University of Lahijan, Vol.8, No.4(31), pp. 29-45, ISSN 2008-6083.

[10] J.H. Holland, "Adaptation in Natural and Artificial Systems, University of Michigan Press", Ann Arbor, Michigan, 1992, re-issued by MIT Press.

[11] T. Adam, J. McAughey, C. Mocker, "Influence of filter ventilation on the chemical composition of cigarette mainstream smoke", Analytica chimica acta, 2010, pp. 36-44.

[12] M.Zanganeh, S.J. Mousavi, A.E.-Shahidi, "Genetic Algorithm based Fuzzy inference system based in prediction of wave parameter", Computational Intelligence, Theory and Application, 2006,Vol. 38, pp. 741-750.

A review of feature selection methods with applications

A. Jović*, K. Brkić* and N. Bogunović*
* Faculty of Electrical Engineering and Computing, University of Zagreb / Department of Electronics, Microelectronics, Computer and Intelligent Systems, Unska 3, 10 000 Zagreb, Croatia
{alan.jovic, karla.brkic, nikola.bogunovic}@fer.hr

Abstract - Feature selection (FS) methods can be used in data pre-processing to achieve efficient data reduction. This is useful for finding accurate data models. Since exhaustive search for optimal feature subset is infeasible in most cases, many search strategies have been proposed in literature. The usual applications of FS are in classification, clustering, and regression tasks. This review considers most of the commonly used FS techniques. Particular emphasis is on the application aspects. In addition to standard filter, wrapper, and embedded methods, we also provide insight into FS for recent hybrid approaches and other advanced topics.

I. INTRODUCTION

The abundance of data in contemporary datasets demands development of clever algorithms for discovering important information. Data models are constructed depending on the data mining tasks, but usually in the areas of classification, regression and clustering. Often, pre-processing of the datasets takes place for two main reasons: 1) reduction of the size of the dataset in order to achieve more efficient analysis, and 2) adaptation of the dataset to best suit the selected analysis method. The former reason is more important nowadays because of the plethora of developed analysis methods that are at the researcher's disposal, while the size of an average dataset keeps growing both in respect to the number of features and samples.

Dataset size reduction can be performed in one of the two ways: feature set reduction or sample set reduction. In this paper, the focus is on feature set reduction. The problem is important, because a high number of features in a dataset, comparable to or higher than the number of samples, leads to model overfitting, which in turn leads to poor results on the validation datasets. Additionally, constructing models from datasets with many features is more computationally demanding [1]. All of this leads researchers to propose many methods for feature set reduction. The reduction is performed through the processes of feature extraction (transformation) and feature selection. Feature extraction methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Multidimensional Scaling work by transforming the original features into a new feature set constructed from the original one based on their combinations, with the aim of discovering more meaningful information in the new set [2]. The new feature set can then be easily reduced by taking into consideration characteristics such as dataset variance coverage. Feature selection, on the other hand, is a process of taking a small subset of features from the original feature set without transformation (thus preserving the interpretation) and validating it with respect to the analysis goal. The selection process can be achieved in a number of ways depending on the goal, the resources at hand, and the desired level of optimization.

In this paper, we focus on feature selection and provide an overview of the existing methods that are available for handling several different classes of problems. Additionally, we consider the most important application domains and review comparative studies on feature selection therein, in order to investigate which methods perform best for specific tasks. This research is motivated by the fact that there is an abundance of work in this field and insufficient systematization, particularly with respect to various application domains and novel research topics.

Feature set reduction is based on the terms of feature relevance and redundancy with respect to goal. More specifically, a feature is usually categorized as: 1) strongly relevant, 2) weakly relevant, but not redundant, 3) irrelevant, and 4) redundant [3,4]. A strongly relevant feature is always necessary for an optimal feature subset; it cannot be removed without affecting the original conditional target distribution [3]. Weakly relevant feature may not always be necessary for an optimal subset, this may depend on certain conditions. Irrelevant features are not necessary to include at all. Redundant features are those that are weakly relevant but can be completely replaced with a set of other features such that the target distribution is not disturbed (the set of other features is called Markov blanket of a feature). Redundancy is thus always inspected in multivariate case (when examining feature subset), whereas relevance is established for individual features. The aim of feature selection is to maximize relevance and minimize redundancy. It usually includes finding a feature subset consisting of only relevant features.

In order to ensure that the optimal feature subset with respect to goal concept has been found, feature selection method has to evaluate a total of $2^m - 1$ subsets, where m is the total number of features in the dataset (an empty

feature subset is excluded). This is computationally infeasible even for a moderately large *m*. Therefore, putting completeness of the search aside, many heuristic methods have been proposed to find a sufficiently good (but not necessarily optimal) subset. The whole process of finding the feature subset typically consists of four basic steps: 1) subset generation, 2) subset evaluation, 3) a stopping criterion, and 4) validation of the results [5]. Feature subset generation is dependent on the state space search strategy. After a strategy selects a candidate subset, it will be evaluated using an evaluation criterion in step 2. After repeating steps 1 and 2 for a number of times depending on the process stopping criterion, the best candidate feature subset is selected. This subset is then validated on an independent dataset or using domain knowledge, while considering the type of task at hand.

II. CLASSIFICATION OF FEATURE SELECTION METHODS

Feature selection methods can be classified in a number of ways. The most common one is the classification into filters, wrappers, embedded, and hybrid methods [6]. The abovementioned classification assumes feature independency or near-independency. Additional methods have been devised for datasets with structured features where dependencies exist and for streaming features [2].

A. *Filter methods*

Filter methods select features based on a performance measure regardless of the employed data modeling algorithm. Only after the best features are found, the modeling algorithms can use them. Filter methods can rank individual features or evaluate entire feature subsets. We can roughly classify the developed measures for feature filtering into: information, distance, consistency, similarity, and statistical measures. While there are many filter methods described in literature, a list of common methods is given in Table I, along with the appropriate references that provide details. Not all the filter features can be used for all classes of data mining tasks. Therefore, the filters are also classified depending on the task: classification, regression or clustering. Due to lack of space, we do not consider semi-supervised learning feature selection methods in this work. An interested reader is referred to [16] for more information.

Univariate feature filters evaluate (and usually rank) a single feature, while multivariate filters evaluate an entire feature subset. Feature subset generation for multivariate filters depends on the search strategy. While there are many search strategies, there are four usual starting points for feature subset generation: 1) forward selection, 2) backward elimination, 3) bidirectional selection, and 4) heuristic feature subset selection. Forward selection typically starts with an empty feature set and then considers adding one or more features to the set. Backward elimination typically starts with the whole feature set and considers removing one or more features from the set. Bidirectional search starts from both sides - from an empty set and from the whole set, simultaneously considering larger and smaller feature subsets. Heuristic selection generates a starting subset based on a heuristic (e.g. a genetic algorithm), and then explores it further.

The most common search strategies that can be used with multivariate filters can be categorized into exponential algorithms, sequential algorithms and randomized algorithms. Exponential algorithms evaluate a number of subsets that grows exponentially with the feature space size. Sequential algorithms add or remove features sequentially (one or few), which may lead to local minima. Random algorithms incorporate randomness into their search procedure, which avoids local minima [17]. Common search strategies are shown in Table II.

TABLE I. COMMON FILTER METHODS FOR FEATURE SELECTION

Name	Filter class	Applicable to task	Study
Information gain	univariate, information	classification	[6]
Gain ratio	univariate, information	classification	[7]
Symmetrical uncertainty	univariate, information	classification	[8]
Correlation	univariate, statistical	regression	[8]
Chi-square	univariate, statistical	classification	[7]
Inconsistency criterion	multivariate, consistency	classification	[9]
Minimum redundancy, maximum relevance (mRmR)	multivariate, information	classification, regression	[2]
Correlation-based feature selection (CFS)	multivariate, statistical	classification, regression	[7]
Fast correlation-based filter (FCBF)	multivariate, information	classification	[8]
Fisher score	univariate, statistical	classification	[10]
Relief and ReliefF	univariate, distance	classification, regression	[11]
Spectral feature selection (SPEC) and Laplacian Score (LS)	univariate, similarity	classification, clustering	[4]
Feature selection for sparse clustering	multivariate, similarity	clustering	[12]
Localized Feature Selection Based on Scatter Separability (LFSBSS)	multivariate, statistical	clustering	[13]
Multi-Cluster Feature Selection (MCFS)	multivariate, similarity	clustering	[4]
Feature weighting K-means	multivariate, statistical	clustering	[14]
ReliefC	univariate, distance	clustering	[15]

TABLE II. SEARCH STRATEGIES FOR FEATURE SELECTION

Algorithm group	Algorithm name
Exponential	Exhaustive search Branch-and-bound
Sequential	Greedy forward selection or backward elimination Best-first Linear forward selection Floating forward or backward selection Beam search (and beam stack search) Race search
Randomized	Random generation Simulated annealing Evolutionary computation algorithms (e.g. genetic, ant colony optimization) Scatter search

B. Wrapper methods

Wrappers consider feature subsets by the quality of the performance on a modelling algorithm, which is taken as a black box evaluator. Thus, for classification tasks, a wrapper will evaluate subsets based on the classifier performance (e.g. Naïve Bayes or SVM) [18,19], while for clustering, a wrapper will evaluate subsets based on the performance of a clustering algorithm (e.g. *K*-means) [20]. The evaluation is repeated for each subset, and the subset generation is dependent on the search strategy, in the same way as with filters. Wrappers are much slower than filters in finding sufficiently good subsets because they depend on the resource demands of the modelling algorithm. The feature subsets are also biased towards the modelling algorithm on which they were evaluated (even when using cross-validation). Therefore, for a reliable generalization error estimate, it is necessary that both an independent validation sample and another modelling algorithm are used after the final subset is found. On the other hand, it has been empirically proven that wrappers obtain subsets with better perfomance than filters because the subsets are evaluated using a real modelling algorithm. Practically any combination of search strategy and modelling algorithm can be used as a wrapper, but wrappers are only feasible for greedy search strategies and fast modelling algorithms such as Naïve Bayes [21], linear SVM [22], and Extreme Learning Machines [23].

C. Embedded and hybrid methods

Embedded methods perform feature selection during the modelling algorithm's execution. These methods are thus embedded in the algorithm either as its normal or extended functionality. Common embedded methods include various types of decision tree algorithms: CART, C4.5, random forest [24], but also other algorithms (e.g. multinomial logistic regression and its variants [25]). Some embedded methods perform feature weighting based on regularization models with objective functions that minimize fitting errors and in the mean time force the feature coefficients to be small or to be exact zero. These methods based on Lasso [26] or Elastic Net [27] usually work with linear classifiers (SVM or others) and induce penalties to features that do not contribute to the model.

Hybrid methods were proposed to combine the best properties of filters and wrappers. First, a filter method is used in order to reduce the feature space dimension space, possibly obtaining several candidate subsets [28]. Then, a wrapper is employed to find the best candidate subset. Hybrid methods usually achieve high accuracy that is characteristic to wrappers and high efficiency characteristic to filters. While practically any combination of filter and wrapper can be used for constructing the hybrid methodology, several interesting methodologies were recently proposed, such as: fuzzy random forest based feature selection [29], hybrid genetic algorithms [30], hybrid ant colony optimization [31], or mixed gravitational search algorithm [32].

D. Structured and streaming features

In some datasets, features may exhibit certain internal structures such as spatial or temporal smoothness, disjoint/overlapping groups, tree- or graph-like structures. In these datasets, features are not independent. Therefore, it is a good idea to employ specific algorithms to deal with the dependencies in order to increase performance of the selected feature subsets. Most of the algorithms dealing with feature structures are recent, and are based on some adaptation of the Lasso regularization method to accomodate different structures. Good overviews of these methods can be found in [2,33].

Streaming (or dynamic) features are features which size is unknown in advance; they are rather dynamically generated, they arrive as streamed data into the dataset and the modelling algorithms has to reach a decision whether to keep them as useful for model construction or not. Also, some features may become irrelevant over time and should be discarded. This scenario is common in social networks such as Twitter, where new words are generated that are not all relevant for a given subject [2]. The most important feature selection methods in this category are: the Grafting algorithm [34], the Alpha-Investing algorithm [35] the OSFS algorithm [36], and dynamic feature selection fuzzy-rough set approach [37].

III. Feature Selection Application Domains

The choice of feature selection methods differs among various application areas. In the following subsections, we review comparative studies on feature selection pertaining to several well known application domains. Table III summarizes the findings from the reviewed studies.

A. Text mining

In text mining, the standard way of representing a document is by using the bag-of-words model. The idea is to model each document with the counts of words occurring in that document. Feature vectors are typically formed so that each feature (i.e. each element of the feature vector) represents the count of a specific word, an alternative being to just indicate the presence/absence of a word without specifying the count. The set of words whose occurrences are counted is called a vocabulary. Given a dataset that needs to be represented, one can use all the words from all the documents in the dataset to build the vocabulary and then prune the vocabulary using feature selection.

It is common to apply a degree of preprocessing prior to feature selection, typically including the removal of rare words with only a few occurrences, the removal of overly common words (e.g. "a", "the", "and" and similar) and grouping the differently inflected forms of a word together (lemmatization, stemming) [38].

Forman [38] performed a detailed experimental study of filter feature selection methods for text classification. Twelve feature selection metrics were evaluated on 229 text classification problem instances. Feature vectors were formed not as word counts, but as Boolean representations of whether a certain word occurred or not. A linear SVM classifier with untuned parameters was used to evaluate performance. The results were analyzed with respect to precision, recall, F-measure and accuracy. Information gain was shown to perform best

TABLE III. SUMMARIZED FINDINGS OF RELEVANT FEATURE SELECTION METHODS IN VARIOUS APPLICATION AREAS

Application area	Subfield	Datasets	Feature selection methods	Evaluation metrics	Best performing	Study
Text mining	Text classification	229 text classification problem instances gathered from Reuters, TREC, OHSUMED, etc.	Accuracy, accuracy balanced, bi-normal separation, chi-square, document frequency, F1-measure, information gain, odds ratio, odds ration numerator, power, probability ratio, random	Accuracy, F-measure, precision, and recall	Information gain (precision), bi-normal separation (accuracy, F-measure, recall)	[38]
	Text clustering	Reuters-21578, 20 Newsgroups, Web Directory	Information gain, chi-square, document frequency, term strength, entropy-based ranking, term contribution, iterative feature selection	Entropy, precision	Iterative feature selection	[39]
Image processing / computer vision	Image classification	Aerial Images, The Digits Data, Cats and Dogs	Relief (R), K-means (K), sequential floating forward selection (F), sequential floating backward selection (B), various combinations R + K + F/B	Average MSE of 100 neural networks	R+K+B / R+K+F / R+K, depending on the size of feature subset	[40]
	Breast density classification from mammographic images	Mini-MIAS, KBD-FER	Best-first with forward, backward and bi-directional search, genetic search and random search (k-NN and Naïve Bayesian classifiers)	Accuracy	Best first forward, best first backward	[41]
Bioinformatics	Biomarker discovery	Three benchmark datasets deriving from DNA microarray experiments	Chi-square, information gain, symmetrical uncertainty, gain ratio, OneR, ReliefF, SVM-embedded	Stability, AUC	Chi-square, symmetrical uncertainty, information gain, ReliefF	[42]
	Microarray gene expression data classification	Two gene expression datasets (Freije, Phillips)	Information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, t-statistics, one-dimensional SVM	Accuracy	Consensus of all methods	[43]
Industrial applications	Fault diagnosis	Wind turbine test rig dataset	Distance, entropy, SVM wrapper, neural network wrapper, global geometric similarity scheme	Accuracy	Global geometric similarity scheme with wrapper	[22]

with respect to precision, while the author-introduced method bi-normal separation performed best for recall, F-measure and accuracy.

Liu et al. [39] investigated the use of feature selection in the problem of text clustering, showing that feature selection can improve its performance and efficiency. Five filter feature selection methods were tested on three document datasets. Unsupervised feature selection methods were shown to improve clustering performance, achieving about 2% entropy reduction and 1% precision improvement on average, while removing 90% of the features. The authors also proposed an iterative feature selection method inspired by expectation maximization that combines supervised feature selection methods with clustering in a bootstrap setting. The proposed method reduces the entropy by 13.5% and increases precision by 14.6%, hence coming closest to the established baseline, obtained by using a supervised approach.

B. Image processing and computer vision

Representing images is not a straightforward task, as the number of possible image features is practically unlimited [40]. The choice of features typically depends on the target application. Examples of features include histograms of oriented gradients, edge orientation histograms, Haar wavelets, raw pixels, gradient values, edges, color channels, etc. [44].

Bins and Draper [40] studied the use of filter feature selection methods in the general problem of image classification. Three different image datasets were used. They proposed a three-step method for feature selection that combines Relief, K-means clustering and sequential floating forward/backward feature selection (SFFS/SFBS). The idea is to: 1) use the Relief algorithm to remove irrelevant features, 2) use K-means clustering to cluster similar features and remove redundancy, and 3) run SFFS or SFBS to obtain the final set of features. The authors found that using the proposed hybrid combination of algorithms yields better performance than when using Relief or SFFS/SFBS alone. In cases when there are no irrelevant or redundant features in the dataset, the proposed algorithm does not degrade performance. When the goal is to select a specific number of features, it is suggested to use the R+K+B variant of the algorithm if the number of relevant and non-redundant features is less than 110, and otherwise R+K+F. If the number of selected features is allowed to vary, authors suggest using

R+K. The authors also note that Relief is good at removing irrelevant features, but not adequate for selecting the best among relevant features.

Muštra et al. [41] investigated the use of wrapper feature selection methods for breast density classification in mammographic images. Five wrapper feature selection methods were evaluated in conjunction with three different classifiers on two datasets of mammographic images. The best-performing methods were best-first search with forward selection and best-first search with backward selection. Overall, the results over different classifiers and datasets were improved between 3% and 12% when using feature selection.

C. Industrial applications

Feature selection is important in fault diagnosis in industrial applications, where numerous redundant sensors monitor the performance of a machine. Liu et al. [22] have shown that the accuracy of detecting a fault (i.e. solving a binary classification problem of machine state as faulty vs. normal) can be improved by using feature selection. They proposed to use a global geometric model and a similarity metric for feature selection in fault diagnostics. The idea is to find feature subsets that are geometrically similar to the original feature set. The authors experimented with three different similarity measures: angular similarity, mutual information and structure similarity index. The proposed approach was compared with distance-based and entropy-based feature selection, and with SVM and neural network wrappers. The best performance was obtained by combining the proposed geometric similarity approach with a wrapper, so that top 10% of feature subsets were preselected by geometric similarity, following by an exhaustive search-based wrapper approach to find the best subset.

D. Bioinformatics

An interesting application of feature selection is in biomarker discovery from genomics data. In genomics data, individual features correspond to genes, so by selecting the most relevant features, one gains important knowledge about the genes that are the most discriminative for a particular problem. Dessì et al. [42] proposed a framework for comparing different biomarker selection methods, taking into account predictive performance and stability of the selected gene sets. They compared eight selection methods on three benchmark datasets derived from DNA microarray experiments. Additionally, they analyzed how similar the outputs of different selection methods are, and found that the outputs of univariate methods seem to be more similar to each other than to the multivariate methods. In particular, the SVM-embedded selection seems to select features quite distinct from the ones selected by other methods. When jointly optimizing stability and predictive performance, best results were obtained using chi- square, systematic uncertainty, information gain and ReliefF.

Abusamra [43] analyzed the performance of eight different filter-based feature selection methods and three classification methods on two datasets of microarray gene expression data. The best individually performing feature selection methods varied depending on the dataset and the classifier used. Notably, using Gini index for feature selection improved performance of an SVM classifier on both datasets. Some feature selection methods were shown to degrade classification performance. However, Abusamra demonstrated that classification accuracy can be consistently improved on both datasets using a *consensus* of all feature selection methods to find top 20 features, by counting the number of feature selection methods that selected each feature. Seven features were selected by all the methods, and additional 13 features were randomly selected from a pool of features selected by seven out of eight methods.

IV. Conclusion

The current research advancement in this field is identified in the area of hybrid feature selection methods, particularly concerning the methodologies based on evolutionary computation heuristic algorithms such as swarm intelligence based and various genetic algorithms. Additionally, application areas such as bioinformatics, image processing, industrial applications and text mining deal with high-dimensional feature spaces where a clever hybrid methodology design is of utmost importance if any success is to be obtained. Therein, features may exhibit complex internal structures or may even be unknown in advance.

While there is no silver bullet method, filters based on information theory and wrappers based on greedy stepwise approaches seem to offer best results. Future research should focus on optimizing the efficiency and accuracy of feature subset search strategy by combining earlier best filter and wrapper approaches. Most research tends to focus on small number of datasets on which their methodology works. Larger comparative studies should be pursued in order to have more reliable results.

Acknowledgements

This work has been supported in part by the Croatian Science Foundation, within the project “De-identification Methods for Soft and Non-Biometric Identifiers” (DeMSI, UIP-11-2013-1544). This support is gratefully acknowledged.

References

[1] F. Korn, B. Pagel, and C. Faloutsos, "On the „dimensionality curse‟ and the „self-similarity blessing‟," *IEEE Trans. Knowl. Data Eng.* vol. 13, no. 1, pp. 96–111, 2001.

[2] J. Tang, S. Alelyani, and H. Liu, "Feature Selection for Classification: A Review," in: C. Aggarwal (ed.), Data Classification: Algorithms and Applications. CRC Press, 2014.

[3] L Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," J. Mach. Learn. Res., vol. 5, pp. 1205–1224, 2004.

[4] S. Alelyani, J. Tang, and H. Liu, "Feature Selection for Clustering: A Review," in: C. Aggarwal and C. Reddy (eds.), Data Clustering: Algorithms and Applications, CRC Press, 2013.

[5] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, 2005.

[6] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method", *Expert Systems with Applications*, vol. 41, issue 14, pp. 6371–6385, 2014.

[7] I. H. Witten and E. Frank, Data mining: Practical machine learning tools and techniques, San Francisco CA, USA: Morgan Kaufmann, 2011.

[8] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in: *Proc. 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, AAAI Press, pp. 856–863, 2003.

[9] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection-A Filter Solution," in: *Proc. 13th International Conference on Machine Learning (ICML-1996)*, Bary, Italy, Morgan Kaufmann, pp. 319–327, 1996.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification, Wiley-interscience, 2012.

[11] M. R. Sikonja and I. Kononenko, "Theoretical and empirical analysis of Relief and ReliefF," *Mach. Learn.*, vol. 53, pp. 23–69, 2003.

[12] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 713–726, 2010.

[13] Y. Li, M. Dong, and J. Hua. "Localized feature selection for clustering," *Pattern Recognition Letters*, vol. 29, no. 1, pp. 10–18, 2008.

[14] D. S. Modha and W.S. Spangler, "Feature weighting in k-means clustering," *Mach. Learn.*, vol. 52, no. 3, pp. 217–237, 2003.

[15] M. Dash and Y.-S. Ong, "RELIEF-C: Efficient Feature Selection for Clustering over Noisy Data," in: *Proc. 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, Roca Raton, Florida, USA, pp. 869–872, 2011.

[16] Z. Xu, I. King, and M. R.-T. Lyu, "Discriminative Semi-Supervised Feature Selection Via Manifold Regularization," *IEEE Trans. Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[17] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, London: Kluwer Academic Publishers, 1998.

[18] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in: *Proc. 15th International Conference on Machine Learning (ICML-1998)*, Madison, Wisconsin, USA, Morgan Kaufmann, pp. 82–90, 1998.

[19] S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Information Sciences*, vol. 286, pp. 228–246, 2014.

[20] Y. S. Kim, W. N. Street, and F. Menczer, "Evolutionary model selection in unsupervised learning," *Intelligent Data Analysis*, vol. 6, no. 6, pp. 531–556, 2002.

[21] J. C. Cortizo and I. Giraldez, "Multi Criteria Wrapper Improvements to Naive Bayes Learning," LNCS, vol. 4224, pp. 419–427, 2006.

[22] C. Liu, D. Jiang, and W. Yang, "Global geometric similarity scheme for feature selection in fault diagnosis," *Expert Systems with Applications*, vol. 41, issue 8, pp. 3585–3595, 2014.

[23] F. Benoît, M. van Heeswijk, Y. Miche, M. Verleysen, and A. Lendasse, "Feature selection for nonlinear models with extreme learning machines," *Neurocomputing*, vol. 102, pp. 111–124, 2013.

[24] M. Sandri and P. Zuccolotto, "Variable Selection Using Random Forests," in: S. Zani, A. Cerioli, M. Riani, and M. Vichi (eds.), *Data Analysis, Classification and the Forward Search*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, pp. 263–270, 2006.

[25] G. C. Cawley, N. L. C. Talbot, and M. Girolami, "Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation," in: B. Schölkopf, J. C. Platt, and T. Hoffmann (eds.), Advances in Neural Information Processing Systems, MIT Press, pp. 209–216, 2007.

[26] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Briefings in Bioinformatics*, vol. 9, no. 5, pp. 392–403, 2008.

[27] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[28] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in: *Proc. 18th International Conference on Machine Learning (ICML-2001)*, San Francisco, CA, USA, Morgan Kaufmann, pp. 74–81, 2001.

[29] J. M. Cadenas, M. C. Garrido, and R. Martínez, "Feature subset selection Filter–Wrapper based on low quality data," *Expert Systems with Applications*, vol. 40, pp. 6241–6252, 2013.

[30] I. S. Oh, J. S. Lee, and B. R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1424–1437, 2004.

[31] S. I. Ali and W. Shahzad, "A Feature Subset Selection Method based on Conditional Mutual Information and Ant Colony Optimization," *International Journal of Computer Applications*, vol. 60, no. 11, pp. 5–10, 2012.

[32] S. Sarafrazi and H. Nezamabadi-pour, "Facing the classification of binary problems with a GSA-SVM hybrid system," *Mathematical and Computer Modelling*, vol. 57, issues 1-2, pp. 270–278, 2013.

[33] J. Zhou, J. Liu, V. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in: *Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Beijing, China, ACM, pp. 1095–1103, 2012.

[34] S. Perkins and J. Theiler, "Online feature selection using grafting," In: *Proc. 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, AAAI Press, pp. 592–599, 2003.

[35] D. Zhou, J. Huang, and B. Schölkopf, "Learning from labeled and unlabeled data on a directed graph." In: *Proc. 22nd International Conference on Machine Learning (ICML-2005)*, Bonn, Germany, ACM, pp. 1041–1048, 2005.

[36] X. Wu, K. Yu, H. Wang, and W. Ding, "Online streaming feature selection," In: *Proceedings of the 27th international conference on machine learning (ICML-2010)*, Haifa, Israel, Omnipress, pp. 1159–1166, 2010.

[37] R. Diao, M. N. Parthalain, and Q. Shen, "Dynamic feature selection with fuzzy-rough sets," in: *Proc. IEEE International Conference on Fuzzy Systems (FUZZ IEEE 2013)*, Hyderabad, India, IEEE Press, pp. 1–7, 2013.

[38] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.

[39] T. Liu, S. Liu, and Z. Chen, "An evaluation on feature selection for text clustering," in: *Proc. 20th International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, AAAI Press, pp. 488–495, 2003.

[40] J. Bins and B. A. Draper, "Feature selection from huge feature sets," in: *Proc. 8th International Conference on Computer Vision (ICCV-01)*, Vancouver, British Columbia, Canada, IEEE Computer Society, pp. 159–165, 2001.

[41] M. Muštra, M. Grgić, and K. Delač, "Breast density classification using multiple feature selection," *Automatika*, vol. 53, pp. 1289–1305, 2012.

[42] N. Dessì, E. Pascariello, and B. Pes, "A Comparative Analysis of Biomarker Selection Techniques," *BioMed Research International*, vol. 2013, article ID: 387673, DOI: http://dx.doi.org/10.1155/2013/387673

[43] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma," *Procedia Computer Science*, vol. 23, pp. 5–14, 2013.

[44] K. Brkić, "Structural analysis of video by histogram-based description of local space-time appearance," Ph.D. dissertation, University of Zagreb, Faculty of Electrical Engineering and Computing, 2013.

Using similarity-based selection in evolutionary design of decision trees

L. Bošnjak*, S. Karakatič* and V. Podgorelec*
* University of Maribor, Faculty of electrical engineering and computer science/Institute of Informatics, Maribor, Slovenia
leon.bosnjak@um.si, saso.karakatic@um.si, vili.podgorelec@um.si

Abstract - When evaluating the process of building classification decision trees, it is necessary to assess the performance of constructed trees, as well as the speed and efficiency of the algorithm. Top-down induction algorithms are relatively simple and can quickly generate good solutions, however their deterministic nature often prevents them from finding globally optimal solutions. On the other hand, the evolutionary approach to decision tree building has yielded promising results by exploring and exploiting the entire search space. However, the standard evolutionary method of building decision trees uses the fitness-based selection of two trees for crossover, which can lead to premature convergence to a local, often sub-optimal solution. In order to maintain the diversity of the population over the course of evolution, we propose a novel method of selection that takes into consideration the similarity of trees in the crossover process, to prevent inbreeding. Several different approaches to evaluate the similarity between trees were designed and implemented. The approaches of both similar and diverse tree crossover were compared to the standard induction algorithm on twenty different data sets to determine the impact of similarity on the effectiveness and efficiency of the genetic algorithm.

I. INTRODUCTION

As one of the most important machine learning tasks, classification is widely used in the context of many real-world problems, both in the fields of social and natural sciences. In literature, many classification algorithms have been presented, having developed from simple extensions of statistical models to complex algorithms.

Decision trees, being one of the first and even to this day most commonly used classification techniques, have become popular due to several advantages. Among them, the concept of their building can be exposed as a relatively simple process, where understanding how the trees are built is trivial matter. The transparency of the constructed trees is another prominent advantage responsible for their popularity, especially in fields like medical diagnosis, as black-box solutions like neural networks and support vector machines cannot yield as interpretable and easy-to-understand results [1]. Furthermore, their accuracy is comparable to other classifiers', though it has been shown to be affected by the method of tree construction. Whereas the traditional, top-down approach excels at the speed of tree construction and is insensitive to missing data, its deterministic nature often leads to non-optimal solutions.

Alternative approaches of tree building attempted to overcome the aforementioned disadvantages of induction algorithms. Evolutionary decision tree design has been shown to yield promising results, with the trees gradually and incrementally approaching the optimal classification accuracy, using genetic mechanisms such as selection, crossover and mutation.

An important prerequisite for successful evolutionary design is the constant development of trees. Often, fitness-based functions are used to guide the evolution by giving more favorable trees a higher chance of being selected in the next generation. Despite such an approach being relatively successful at a continuous progression towards the global optimum, it can be quite resource- and time-consuming. In addition to that, diversity between the trees begins to deplete with the number of generations, effectively slowing down and eventually stopping further development, as breeding very similar or even identical trees cannot produce significantly better offspring.

To prevent only the best (e.g. most accurate), though often very similar, decision trees from being selected for crossover, the proposed solution introduces an alteration to the selection process: while the first tree remains to be selected by the fitness function, the second tree is chosen depending on its level of similarity with the first tree. By breeding mutually different trees, we aimed to preserve the population diversity by allowing unique and promising genes to advance to the next generation and consequently prevent the evolution from getting trapped in local optima.

II. EVOLUTIONARY DECISION TREES AND THE ROLE OF SIMILARITY

Inspired by the process of natural evolution, following the key principles of natural selection and genetics, the evolutionary algorithms provide a promising framework for the non-trivial task of optimal decision tree creation. As opposed to the traditional induction algorithms, which build trees using complex statistical models or information theory, the evolution-based building approach allows for the development of a population of individuals, depicting final solutions, and then gradually adapting them to the environment [2].

Fig. 1 presents the basic idea of evolutionary decision tree building, using a genetic algorithm. After the initial generation of the population that the given algorithm will operate on, each individual tree is evaluated by a fitness

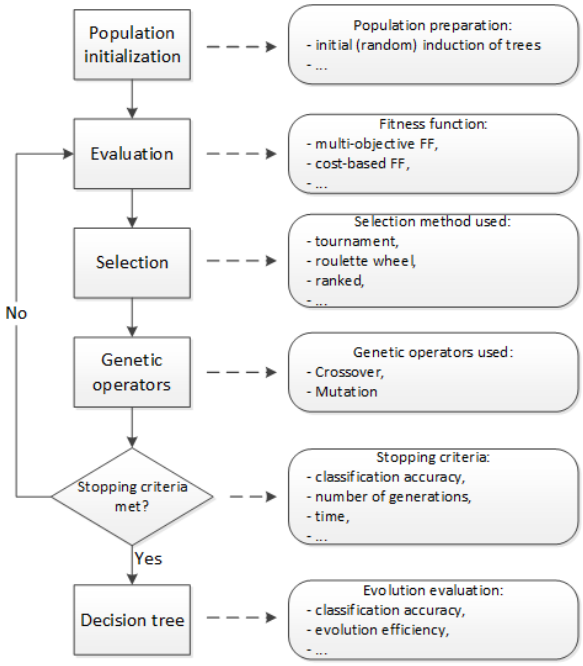


Figure 1. General structure of a genetic algorithm

function – a measurement taking into consideration one or several criteria, trying to assess the general quality of the said individual (i.e. chromosome). Most often, the fitness function considers the tree's classification accuracy and complexity, though it has to be emphasized that this is by no means the best evaluator; with quality being an elusive concept, fitness functions should attempt to measure it by capturing all desired characteristics of a decision tree that are included in the globally optimal solution as well [3].

In every consecutive generation, the individuals that have been evaluated as the best by the fitness function, have a higher chance of being selected for the production of new offspring. That indicates how crucial the selection process is in determining the further course of evolution. Without a good mechanism for selecting the parents, the generated offspring will likely not be as successful as their predecessors, effectively stagnating the evolution through the generations.

The selected individuals are then subjected to various genetic operations: **crossover**, which is responsible for the exchange of genetic material between two parents, and the optional **mutation**, which results in the substitution of a part of the child's genetic material with a new, randomly generated genetic material. The execution of mentioned operations produces new individuals, which replace their parents and consequently create a new generation of individuals. The entire process is then iteratively repeated until the stopping criteria is met [4].

The general idea of evolutionary algorithms is as such very simple, and can be drawn on a parallel with the well-known concept of natural evolution. In nature, species are constantly evolving, which allows them to progressively adapt to the environment. In the context of evolutionary algorithms, the individuals can be described as decision trees, whereas the environment denotes the classification problem they are solving [4]. That reasoning exposes the key difference between the induction- and evolution-based decision tree designs: while the former follows a specialized deterministic algorithm which focuses on building a single solution based on pre-determined rules, the latter has the potential of exploring the entire available search space of solutions.

Unfortunately, two aspects of evolutionary algorithms – selection pressure and gene flow – can quickly cause the nature of the evolution to lean more towards exploitation rather than exploration of search space. In turn, the focus of the algorithm is shifted to a particular subset of the entire search space, effectively ignoring the remainder of the possible solutions. While ideally, we indeed want the algorithm to be able to determine the best solutions in a given set, doing so too early in the course of the evolution can prevent it from considering potentially promising solutions that could represent the global optimum.

Because of that, it is essential to maintain a healthy balance between exploration and exploitation, allowing for the algorithm to *explore* and *identify* promising subsets of the entire search space, and then *exploit* and *select* the optimal solutions. Doing so requires an appropriate level of selection pressure and controlled gene flow. Too strong selection pressure greatly increases the probability of the best solutions being selected for advancement to the next generation. Consequently, after several generations of the same or very similar trees being selected for breeding, the population is quickly filled with similar variations of the successful individuals, preventing further improvement. Likewise, high gene flow also leads to eventual stagnation of the evolution. Algorithms that do not implement a more refined type of selection but instead allow each individual to breed with any other individual in the population, most often cause the same genes to quickly spread throughout the population. As a result, the previously diverse genetic material begins to dilute and as such decreases the range of possible solutions the parents could produce [5].

Both unsuitable selection pressure and high gene flow, resulting in the problem of imbalance between exploration and exploitation, cause progressive loss of gene diversity, hinting at the importance of similarity in the process of evolution. When the individuals of the population become too similar to each other, the population begins to quickly converge towards one of the possible solutions. This phenomenon, known as premature convergence, usually brings the search for the best decision tree to a merely locally optimal solution [6]. Any further exploration of the available search space is from thereon unnecessary – mutually similar parents will continue to produce similar offspring, whose traits will not significantly differ from those of their predecessors [7].

III. SIMILARITY-BASED SELECTION

Typical crossover methods used in genetic algorithms combine parents that had been selected based on their fitness. By considering only the measured "quality" of solutions, such traditional genetic algorithms simulate a process that is in biology more commonly known as random mating: a type of selection, ignoring differences in genotype and phenotype between the individuals [8].

As random mating is not the only existent mechanism of sexual reproduction in nature, the idea of non-random mating has also been considered in the context of genetic

algorithms. It introduces a certain level of intelligence into the selection process by basing it on certain characteristics of the parents. Such non-random and preference-oriented mating schemes are capable of introducing new as well as preserving existing genetic material in the population to maintain or increase diversity between the individuals. By doing so, they lengthen the exploration phase in order to consider all possible neighborhoods of good solutions. A thorough initial exploration of the search space allows the algorithm to avoid premature convergence towards one of local optima traps due to ignoring the subset of promising solutions in which the globally-optimal tree is included. Such approach, focusing on the sustainment of genetic diversity within the population, is particularly relevant for dynamic environments, in which adaptation to change is often more important than convergence towards the currently optimal solution [8].

To preserve genetic diversity, various restricted mating techniques have been introduced, including: assortative and dissortative mating, seduction, evolving agents, tabu genetic algorithm and correlative tournament selection. They incorporate the ideas of *fitness sharing* and *incest prevention*. The former assumes the maintenance of several niches within the search space, each including high-fitness solutions clustering around local optima. By promoting inbreeding between the individual members of each niche, these solutions simultaneously evolve until the globally optimal solution is selected from all sustained niche local optima. On the other hand, incest prevention encourages the recombination of mutually different individuals, as breeding two very similar instances cannot produce solutions containing new genes. In order to select the parents depending on their genetic dissimilarity, this method quantifies the genotypic or phenotypic difference and select the most diverse solutions for crossover [9].

Several studies examining the impact of similarity in the process of selection have been conducted. De, Pal and Ghosh have compared the selection of parents, based on their genotypic or phenotypic similarity, to the traditional, fitness-based selection method, showing the superiority of the novel approach on the problem of complex function optimization [10]. Another study supports the notion of a genetic algorithm, using adaptive dissortative mating scheme, outperforming several other genetic algorithms on a variety of dynamic optimization problems [11]. To an extent, these and other dissortative mating techniques have been inspired by the CHC algorithm, which restricts the crossover between similar solutions, whose Hamming distance is not above a certain threshold [12]. The honey-bee mating optimization algorithm and its enhancements, presented in [13], have shown the algorithm's capability to explore and exploit the search space simultaneously while maintaining population diversity. A comprehensive study describing the most prominent non-random mating techniques, focusing on dissortative mating, was presented by Fernandes and Rosa, emphasizing their effectiveness in maintaining population diversity in comparison to random mating schemes [8]. The results of the study complement Goldberg's, as well as Eshelman and Schaffer's findings showing a more effective exploration of the entire search space by similarity-based non-random mating approaches, both in fitness sharing and incest prevention [9].

A. *Similarity-based mating strategy*

Numerous other similarity-based restricted mating approaches have also been presented in literature [14-16], however none have been observed in the context of classification with decision trees. The purpose of this paper is as such to apply a particular non-random mating scheme based on similarity and diversity between decision trees to an evolutionary algorithm, and compare its performance to a traditional evolutionary algorithm on a given set of classification problems.

A similar mating strategy that was implemented in the context of classification with decision trees and is shown in fig. 2, has already been presented on the problem of vector quantization, indicating the significance of this research [14]. It differs from the conventional genetic algorithm in the way parents are selected for crossover. The first parent tree remains to be selected depending on its fitness, in order to promote the transition of successful genes into the next generation. Any traditional selection method can be used, including but not limited to: ranked, roulette wheel and tournament-based selection. The latter is one of the most commonly used selection methods, due to its efficiency, as well as simple implementation [17], and was therefore used in our case as well. From the entire population, a set of α decision trees was then randomly selected. Each of the candidate trees was evaluated by the fitness function, which quantified the fitness of the trees based on their classification accuracy and size. Of the subset of trees, the one with the highest fitness value was selected as the first mating parent.

Usually, the second parent would have been selected in a similar manner, either as the second-best tree from the already executed tournament, or as the best tree from a new tournament of β randomly selected trees (where, in most cases, but not necessarily, $\alpha = \beta$). Whereas a β-sized tournament could still be held to randomly select a subset of candidate trees, the evaluation and subsequent selection of the second parent was based on the similarity between the first tree and the candidate solutions, rather than the fitness, however. Firstly, similarity indices measuring the difference and/or similarity between two given trees, were calculated for all β pairs. Then, depending on whether we wanted to use assortative or dissortative mating scheme, the most similar or different decision tree from the first one was chosen, respectively.

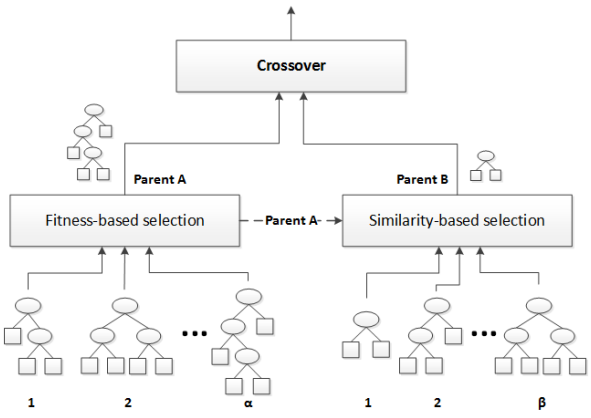


Figure 2. Similarity-based selection of decision trees

Measuring similarity between two trees is not a trivial task. There are several possible characteristics of trees that can be considered, and depending on these contradictory perspectives, different estimates of similarity between decision trees can be derived. Furthermore, any observed differences have to be appropriately quantified, producing a single numeric value as an output – that is necessary so that all similarity scores can be compared to each other, making it easy to determine the most similar or diverse pairs. While fitness score can be computed with relatively small effort, as it is simple to identify the most accurate or the smallest trees as the best (even in the case of several criteria being considered simultaneously, simple weighted formulas can provide a rather straightforward approach to multi-objective optimization) [18], it is very difficult to arbitrate how different two trees are from each other.

Identical trees can be unambiguously labelled with a score of 0, whereas the measure of difference between two trees can theoretically rise to infinity, since we can always find a pair of trees that is more different from each other than the previous one. For computational purposes and ease of comparison and understanding, most studies have placed the limit of difference at 1, with the more different trees gradually approaching the limit without ever actually reaching it, as two trees can never be the *most* different [19, 20]. Even though normalizing the differences in order to limit the score in between the designated boundaries can distort the severity of differences between two given trees, we have decided to follow the above mentioned approach because we are ultimately interested only in the highest and lowest-rated pairs.

B. Implementation of similarity-based evaluation

There is a number of possible ways similarity between trees can be measured. For example, decision trees can be transformed into rule sets, following a particular pattern: "*IF* $attr_A > x$ *AND* $attr_B <= y$ *AND … THEN* $Class_1$". The similarity measure can then be obtained by comparing substructures of particular rules composing two observed decision trees, and finding identical patterns in them [19]. Another study investigates the issue of splitting attribute selection, proposing a technique that takes into account the similarity between any two instances in a given subset, choosing the attribute with the greatest similarity score as the splitting attribute [20]. Finally, with the intention of observing semantical rather than structural differences between the trees, allowing for possible meta-analysis and meta-mining, a semantically-driven similarity measure has been developed [21].

In our work, we present five independent measures of similarity that were designed for the purpose of capturing several different, possibly contradictory features that trees could differ in:

Tree complexity. Representing one of the aspects that affect the tree's performance, complexity should generally follow the Ockham's razor principle, stating phenomena should be explained with as few variables as possible. Smaller and simpler trees are not just easier to interpret, but achieve a high level of generalization and are prone to overfitting [22]. As such, we aimed to breed similarly-sized trees in order to maintain a low level of complexity. We observed the complexity of decision trees on two levels: horizontal and vertical. The former constitutes the tree's comprehensibility and interpretability, and can be easily measured as the number of all nodes. The latter, on the other hand, denotes the maximum number of decisions a tree has to make in order to classify an arbitrary instance into one of the classes, and is represented by the tree's depth. The implemented similarity measure considers both as a balanced weighted difference between two trees.

Number of used attributes. Selection of numeric and nominal attributes, based on which decision trees classify instances, is already performed by the induction algorithm building said trees. Nonetheless, the complexity of trees is, to some degree, dependent on the number of attributes used: more complex trees often use most, if not all of the available attributes describing instances, whereas simple trees classify instances, considering only a few attributes. The metric evaluates similarity by calculating the absolute difference between the numbers of unique attributes used in both decision trees. Despite that being another measure of complexity, it was not included in the previous metric because it is not directly related to tree size. While we are aware the two metrics could have been combined, our aim was to consider several simple measures of similarity, although multi-criteria similarity-based selection methods remain as a possible research direction for future work.

Attribute values. As pointed out in ref. [20], splitting criteria affects the tree's classification accuracy, as well – despite having two equal trees in terms of size and even used attributes, the instances can be classified in different classes depending on the attribute values. The idea behind the implemented metric is therefore the crossover of trees with dissimilar attribute values, to create offspring that would, despite similar tree structure, classify instances in a different (and hopefully better) way. Naturally, it cannot be expected of the descendant decision trees to outperform their parents in every generation. However, since the principles of natural evolution speak in favor of the more successful trees advancing into the subsequent generation, they will only be replaced by their children when the latter surpass them, while the diversity within the population can be maintained by the similarity-based crossover that considers differences in attribute values.

Attributes can be either numeric or nominal, meaning that the differences in both types of attributes had to be considered. Numeric attributes were evaluated simply by calculating the absolute difference between the two tree's attribute values. In case a tree splits the data set using the same attribute multiple times, the sum of that attribute's values is taken instead. To determine the differences in nominal attributes, the different values for that specific attribute were counted. Since every attribute was assessed separately, partial similarity scores were produced, finally being added up to the final score.

The most prominent flaw of this method can be seen on data sets containing attributes, whose values have an extensive range. In such cases, the final scores are heavily influenced by just a single attribute. In order for every attribute to have the same amount of impact on the final score, feature scaling method was used to normalize every sub-score into the closed interval between 0 and 1 [23].

Confusion matrix. The implemented similarity metric focuses on the matrix's diagonal. For every class, the difference in the number of correctly classified instances was calculated, with the final similarity index represented by the sum of all partial scores.

Instance classification. A rather promising approach considers the differences in the way two trees classify the instances. For every instance in the data set, classified into two different classes by the compared trees, the similarity score is incremented by 1. The final score is represented by the number of differently classified instances. With the differences in instance classification, it can be expected that the two parents are structurally different as well. By combining two decision trees that classify the instances in a different manner, the created offspring could inherit the parts of the parent trees that are responsible for correctly classified instances, making the newly created tree even more accurate in turn. It should be noted, however, that such a scenario is greatly dependent on coincidence, due to the crossover points being selected randomly. Since the evolution runs through many generations, the probability of favorable children being constructed in such a manner, is still relatively high, however.

IV. EXPERIMENT AND RESULTS

The similarity-based mating strategies described in the previous section were implemented within a Java program "GaTree", which served as a framework for the execution of our experiment. The program was previously developed with the intention of conducting studies in the field of evolutionary-based decision tree construction. In addition to the standard evolutionary building approach, it offers a wide variety of functionalities, including several selection, crossover and mutation methods, as well as novel fitness functions. An extensive range of the available control parameters was especially useful for proper configuration of the environment, in which the experiment was run. The used parameters, crucial for the experiment, are presented in table 1.

Since we aimed to ensure the validity of the executed experiment, we compared the conventional, fitness-based selection method to the five implemented similarity-based selection methods on twenty benchmark datasets, obtained from the UC Irvine Machine Learning Repository. The selected data sets differed in the number of attributes and instances, type of attributes, existence of missing data and the field of study. Additionally, due to the initial decision tree induction, as well as several genetic operators being based on randomness, the experiment was repeated ten times on each of the twenty data sets. Thus, the probability of the results being as such due to random chance, was accordingly mitigated.

The experiment was run on the 64-bit server operating system Windows Server 2008 R2 Standard, with a 2.4 GHz Intel Xeon E5645 processor and 25 GB of RAM, taking 439,644 seconds (or 5 days, 2 hours, 7 minutes and 24 seconds) to complete. The resulting data consisted of simple metrics, measuring the effectiveness and efficiency of the standard evolutionary algorithm and five similarity-based approaches for all ten repetitions on all twenty data sets. Efficiency was thus measured by the number of generations necessary to obtain the best decision tree in the evolution, whereas effectiveness was determined by the best tree's accuracy, size and F-score.

TABLE I. USED CONTROL PARAMETERS AND THEIR VALUES

Control parameter	*Value*
Number of repetitions on a single data set	10
Size of the population	200
Maximum number of generations	2000
Maximum tree depth in the initial population	10
Maximum number of possible non-final nodes in the initial population	20
Mutation probability for a particular node	0.1
Parent tree selection method	Tournament
Number of randomly selected trees for the execution of the tournament	5
Evaluation of selected parent trees	1st tree: fitness 2nd tree: similarity
Crossover method	Random

In the analysis phase, the obtained data associated with the standard approach was compared to the data from each of the other five, similarity-based approaches. Due to the data not being normally distributed ($D(200) \approx 0.152$, $p < 0.001$, for all six approaches), the Kruskal-Wallis non-parametric test was performed, although no significant difference between the standard and any one of the five implemented approaches has been observed. Since these tests were run on all data sets at once, we repeated them on each data set and noted significant differences on 6 out of 20 individual data sets. On four of these data sets, the standard approach was significantly better than the others in terms of classification accuracy and F-score, while the similarity-based approaches considering tree complexity and the number of used attributes were significantly better than it on another four of the data sets in terms of tree size.

Statistical significance greatly depends on the sample size (which was, in our case, only 10 repetitions per data set, due to computational complexity of the process), so we calculated the effect size that represents a standardized measure of the observed difference's magnitude [24]. On fig. 3, the graph shows the number of data sets, on which at least a weak effect size has been observed. Individual columns represent the implemented approaches that were more successful than the standard approach, with respect to the observed metrics. Yellow color denotes weak, green

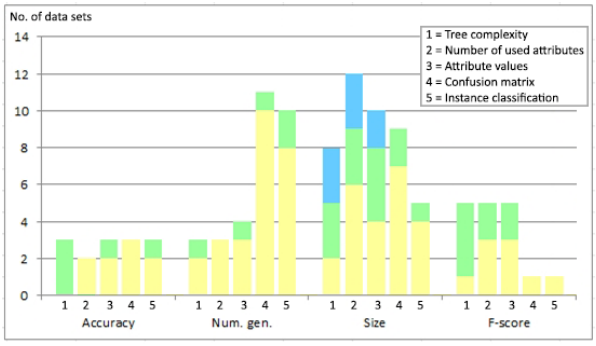


Figure 3. No. of data sets on which one of the similarity-based approaches was better than standard with at least weak effect size

denotes medium and blue denotes strong effect size. As it can be seen, the implemented approaches seemed to be better than the standard approach mostly in respect to the size of the tree, which was consistent with our previous findings that pointed at the most significant improvements being in relation to the size. On ten and eleven data sets, the similarity-based approaches considering differences in the instance classification and confusion matrix appeared to find the most successful tree faster than the traditional approach. Improvements of classification accuracy and F-score were less common, with weak and medium effect size of the implemented approaches' improvements being observed on only one to five data sets.

V. CONLUSION

Evolutionary algorithms, with the idea of mimicking natural evolution and simulating the development and reproduction of organisms in nature, present a promising framework for construction of successful decision trees. Despite that, their features have been observed to lead to premature convergence. With the implemented similarity-based crossover approaches, we aimed to address the problem of diversity loss, one of the main reasons for the evolution getting trapped in one of the local optima.

Although the five selection approaches that consider similarity between trees do not significantly outperform the standard evolutionary approach, at least weak effect sizes, hinting at possible improvements of the similarity-based methods in comparison to the standard one in regards to the tree size, have been noted on up to twelve data sets. Furthermore, the mating strategy considering differences in the confusion matrix and the classification of instances, have been shown to improve the number of generations to the final solution with weak to medium effect size on over half of the data sets. We believe that the significance of these findings would have been shown, had the sample size been bigger. One of the possibilities for future work is thus to observe one of the implemented approaches (or a combination of them) on fewer data sets, by executing a greater number of repetitions. Likewise, new approaches evaluating similarity between trees could also be developed.

It has been established that, while the similarity-based selection does increase diversity within the population, the final result still depends on chance, due to the crossover points being selected randomly. As such, it would have also been plausible to consider adding an additional level of intelligence into the process by determining the most optimal crossover points in both trees to breed even more successful offspring.

REFERENCES

[1] A. A. Freitas, "Comprehensible classification models – a position paper," in ACM SIGKDD Explorations Newsletter, vol. 15, no. 1, pp. 1-10, 2013

[2] P. Kokol, M. Šprogar, M. Zorman, V. Podgorelec, and T. Kokol, "Computerised consulting with vector decision trees in textile engineering," [Knowledge-based Intelligent Information Engineering and Allied Technologies, pp. 1077-1082, 2001]

[3] C. C. Palmer and A. Kershenbaum, "An approach to a problem in network design using genetic algorithms," in Networks, vol. 26, no. 3, pp. 151-163, 1995

[4] A. Freitas, "A review of evolutionary algorithms for data mining," Soft Computing for Knowledge Discovery and Data Mining, pp. 79-111, 2008

[5] R. Ursem, "Diversity-guided evolutionary algorithms," London: Springer, 2002 [Proceedings of the 7th International Conference on Parallel Problem Solving from Nature, pp. 462-474]

[6] K. Zhu and Z. Liu, "Population diversity in permutation-based genetic algorithm," Pisa: Springer, 2004 [European Conference on Machine Learning, pp. 537-547]

[7] A. Nsakanda, W. Price, M. Diaby, and M. Gravel, "Ensuring population diversity in genetic algorithms: A technical note with application to the cell formation problem" in European Journal of Operational Research, vol. 178, no. 2, pp. 634–638, 2007

[8] C. M. Fernandes and A. C. Rosa, "Evolutionary algorithms with dissortative mating on static and dynamic environments," Advances in Evolutionary Algorithms, InTech, pp. 181-206, 2008

[9] S. F. Galán, O. J. Mengshoel and R. Pinter, "A novel mating approach for genetic algorithms," in Evolutionary Computation, vol. 21, no. 2, pp. 197-229, 2013

[10] S. De, S. K. Pal and A. Ghosh, "Genotypic and phenotypic assortative mating in genetic algorithm," in Information Sciences, vol. 105, no. 1-4, pp. 209-226, 1998

[11] C. M. Fernandes, J. J. Merelo and A. C. Rosa, "A comparative study on the performance of dissortative mating and immigrants-based strategies for evolutionary dynamic optimization," in Information Sciences, vol. 181, no. 20, pp. 4428-4459, 2011

[12] L. Eshelman and J. Schaffer, "Preventing premature convergence in genetic algorithms by preventing incest," San Mateo, CA: Morgan Kauffman, 1991, [Proceedings of the fourth International Conference on Genetic Algorithms, pp. 115-122]

[13] N. R. Sabar, M. Ayob, G. Kendall and R. Qu, "A honey-bee mating optimization algorithm for educational timetabling problems," in European Journal of Operational Research, vol. 216, no. 3, pp. 533-543, 2012

[14] C. Fernandes, T. Tavares, C. Munteanu, and A. Rosa, "Using assortative mating in genetic algorithms for vector quantization problems," ACM, 2001 [Proceedings of 2001 Symposium on Applied Computing, pp. 361-365]

[15] J. Gao and L. Fang, "An experimental study of similarity-based mating on a weight-based multiobjective immune genetic algorithm," DC, USA: IEEE, 2009 [Intelligent Computation Technology and Automation, pp. 153-156]

[16] H. Ishibuchi, K. Narukawa, N. Tsukamoto, and Y. Nojima, "An empirical study on similarity-based mating for evolutionary multiobjective combinatorial optimization" in European Journal of Operational Research, vol. 188, no. 1, pp. 57-75, 2008

[17] N. M. Razali and J. Geraghty, "Genetic algorithm performance with different selection strategies in solving TSP," [Proceedings of the World Congress on Engineering, vol. 2, pp. 156-161, 2011]

[18] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho and A. A. Freitas, "A survey of evolutionary algorithms for decision tree induction," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 42, no. 3, pp. 291-312, 2012

[19] P. Perner, "Decision tree induction methods and their application to big data," Modeling and Processing for Next-Generation Big-Data Technologies, vol. 4, pp. 57-88, 2015

[20] X. Zhang and S. Jiang, "A splitting criteria based on similarity in decision tree learning" in Journal of Software, vol. 7, no. 8, pp. 1775-1782, 2012

[21] I. Ntoutsi, A. Kalousis and Y. Theodoridis, "A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees," [Proceedings of the SIAM International Conference on Data Mining, pp. 810-821, 2008]

[22] V. Podgorelec, M. Šprogar, and S. Pohorec, "Evolutionary Design of Decision Trees", in Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, no. 2, pp. 63-82, 2013

[23] H. Bozdogan, Statistical Data Mining and Knowledge Discovery, CRC Press, 2003

[24] A. Field, Discovering Statistics Using IBM SPSS Statistics, 4th ed., 2012 London: Sage, 2012

A Semantic Markup Technique Based on Ontology Polysystem

Evgeny Cherkashin*†‡, Kristina Paskal†‡, Igor Bychkov*
*Institute of System Dynamics and Control Theory at SB RAS, Irkutsk, Lermontov str., 134, 664033, Russia
†National Research Irkutsk State Technical University, Irkutsk, Lermontov str., 83, 664074, Russia
‡Irkutsk State University, Irkutsk, Gagarina blvd., 664003, Russia
eugeneai@icc.ru, bychkov@icc.ru

***Abstract*—The described investigation proposes an approach to overcome known problem of site developers' participation in the real development of Semantic Web content. The site integration in a way of the Semantic Web is too hard to be used in a regular site maintenance. This results in the necessity of development of site content management software integrated by means of semantic web as knowledge acquisition systems, where developers and users have to play a role of an information source for a system-integrated decision-making engine inducing formal knowledge on the base of data analysis.**

The logical layer is generated on the base of analysis of changes introduced by user, which are analyzed and interpreted. The variant of the interpretation (e.g., error correction of a value or a new statement definition) is determined by means of user interview and other factors. The theoretical basis of the technique is based on a polysystem representation of ontologies describing the domain. The presentation is a hierarchically fibered structure of concepts and relations, which are mapped between fibers by means of interpretation relations.

I. Introduction

In 2001 Tim Berners-Lee proposed a blueprint [1] of web development that is aimed at the implementation of network services with reasonable integration of logical layer of the information presented. The information is to be marked up semantically and the program agents are to take advantage of the markup as the logical layer for the information consumption and processing. The blueprint is called Semantic Web (SW).

One of the main problems of SW is the fact, that the regular users of web resources are not fond of the complex technological aspects of SW. They are interested in their practical problems solutions. In order to involve users in SW content development these aspects must be completely hidden. This results in the necessity of development of document and site content management software exploited within SW as knowledge acquisition systems, where user is to play a role of an information source for a system-integrated decision-making engine inducing formal knowledge on the base of data analysis.

To be more comprehensible consider a content of a legal document, which in most cases contains meaningful information on legal entities relationships that is usually passed to other documents in a derivative forms. This suggests an idea to develop a form of the logical layer representation in a reasonably detailed form, which can be rendered as a regular text pages by means of context-depended (in sense of a legal document) text templates. The SW approach could supplement the idea with data formats and technologies of their storage and processing. SW represents logical layer as a network graph of notions and relations between the notions. For example, individuals mentioned in a legal document relate to the document as "part-whole", unless to speak even more explicitly in a meaningful context.

At present most of the use cases of domain ontology models are refining search results. Automatic procedures of ontology extraction from the documents are based on crawling documents in a warehouse, running data mining procedures on text content and metadata attributes of the documents. Simple observation of human behavior in the process of a document preparation will result in the confidence that meaningful parts of the document, which are usually expressed with the logical layer, are located at the points of document changes. Hence, a programming system automating the document preparation should track the changes and analyze them in time to extract ontology structure, i.e., objects, concepts and relation between them.

The logical layer is induced during content edition by means of data mining and user interview. The context of the knowledge acquisition consists of a polysystem [2] of ontologies describing the domain, the source document (its text representation and current logical structure), its list of modifications in transaction, user's action history, and answers to the interview questions clarifying meaning of text content changes and properties of the relations constructed. As a result, new or modified triples `<subject, relation, object>` are constructed. Collected triple data, metadata and knowledge of the logical layer can also be crawled on a regular basis to grasp patterns and functional relations between attributes (triple's objects). In the last case a relational table could be induced to rise the efficiency of data storage and processing.

The present trends of internet information system development show that the systems become a web-services and are oriented to support social networks [3]. The data flow processing there in most cases is input, storage, filtering and transmitting data, i.e., the social networks integrate rather than aggregate data, e.g., to produce reports. Hence, the usage of special design techniques for the middle-ware layer of the software, such as object-oriented design, is not a significant technological advantage. Social network (SN) agent software are interconnected with standardized protocols and data structures, and have the same way of integration with other social networks.

Let us emphasize the properties of social network software:

- there are no predominant common global task to be solved by agents (computer and human) of the SN;
- each agent solves its special task acquiring data from other data sources and agents, hence, the agent's APIs and supported data format must be standardized;
- human users of the social network do most of the aggregation tasks personally including subconscious joint processing of unstructured and semantic information;
- a good practice is to use cheap virtual hosting to run the a SN node, therefore, the information system itself must be written according to the restrictions of a server-side programming language, like PHP, and utilize the computational and data storage resources of client workstation and its browser environment.

In this paper, we continue the development of the approach [4] of document and site content management and integration within a general study [5], and briefly consider the procedure of polysystem ontology usage in the process of user dialog maintenance.

II. Document Markup

Resource description framework (RDF) standard describes informational resources with triples `<subject, relation, object>` in a context. A set of triples forms a graph (network) of data and relations reflecting knowledge. It is convenient to divide graphs to subgraphs and construct hierarchic complexes [5], resulting a hierarchy of contexts. In a general case, a context affects the interpretation of the set of triples coupled with it. For example, in a legal notary document family name and passport data are presented as texts in different parts of the document, but they related to the single person in the context defined by the document. So the parts of the document form the contexts of text rendering for the passport data.

The rendering engine we considered in [4]. According to this approach, all the template data for document rendering is also stored in the ontology graph. We also represent the views and algorithms implementing controllers in the sense of MVC (Model View Controller) [6] pattern with triples. The content of the document from a general point of view is based on a tree graph, where almost each node is both subject and object of their corresponding relations. The root node in a document can be only a subject, and the leaf nodes are objects only. The document root node becomes object node as being referred from document sets and hierarchies organized in a warehouse.

III. The Outline of Markup Scheme

The proposed way of semantic document markup is to track and interpret the results of document editing. Let's suppose that a document body change affects the meaning of the document, hence, alters its logical layer. The following trivial modifications that are the elementary error correction are not considered to be a valuable information for the markup: a field value change (object of a triple), paragraph text editing that might imply its original template correction. The changes of our interest result in a new triple relation construction between context subjects and old/new object value. The content management system and a built-in editor play an active role in the markup process, and user is a source of supplementary information.

The first question that user have to be asked is to determine weather a change is an above mentioned trivial one. If it is not, then the semantic layer enrichment procedure is carried on. The source of data to decision-making procedure of a triple formation/modification is an environment containing a) logic structure of the previous version of the document; b) the source version of the document content as text, c) a text modification expressed in diff format [7]; d) user answers to the question stated by the system, refining the semantics and structure of the modifications; e) user actions prior the modification done. The list of the previous user's actions can reveal a general intention, e.g., if user made a copy of document and changed an individual name, this would imply the necessity to fill in the passport data, which, in turn, implies filling in a number of triples with new values representing the new individual (subject) of the document.

At the next step of markup the extent of the modified value is determined. It could be a word, a phrase, a sentence, the whole object value, or an extent of a text markup (e.g. HTML tag value). A resulting variant is specified by user.

The simplest way of the triple formation procedure is to allow the user to choose subject and relation from list of all relevant subjects mentioned in the document and their relations and construct a new triple `<subject,relation,old object>`. The list of subjects is constructed from a tree of all the subjects of the edited document and the edited object itself. After the first choice is made, a list of all available relation is constructed from all known relation of the chosen subject, its class and parent classes. User must choose one relation to form the triple. In the case, when the list contains no adequate relation, a new one could be defined. New relation is always a subclass of an already existing on, which is also chosen from the list. New relations defined by inexperienced users must be periodically analyzed by knowledge engineers to get rid of semantic inaccuracy, contradictions, redundancy to the equivalents, and, hence, be refactored.

If a value of a triple is removed, then the triple must be removed too. The situation is acceptable if the minimal structural and semantic completeness of all the subjects of the documents holds, otherwise either the user delete action is prohibited or the chain recursive deletion of the subjects is initiated. To control this behavior each subject class have to be accompanied with a list of minimal valuable triples that define the basis of the subject meaning. Partial text removal, if it is not an error correction, is processed analogously to modification, with removed part being the object value. Addition of characters to text is an action similar to modification.

Addition of a triple might result in extending the document with new subjects and relations. For example, let one construct a new tripartite agreement from an existing bilateral one. In the new agreement the third individual appears, so the addition of a new family name of the individual results in construction of a) subject for the individual, b) filling in the necessary triples, as well as c) definition of a new relation between document and

the third individual as a subclass of `structuralElement`: relation. This stage is necessary as all new subjects must be in a relation.

IV. Theoretical Generalization of the Acquisition Process

The above described approach is a simplified step-wise implementation of the procedure of polysystem analysis and synthesis [2] (PPAS). This is a general procedure for system analysis of domains. The essence of the approach is that domain and any its object or process can be fibrated and represented in multidimensional space of fiber-coordinates. The fiber consists of concepts and morphisms between them. This approach is used in decomposition, analysis and synthesis of complex systems. For each concepts of a fiber there exists another concept in all the fibers, which is identical to the former concept via an interpretation (e.g., substitution mapping). All the fibers are element-wise mapped to the abstract system fibers. The same is true for the morphisms. In comparison to the widely known system approach, where the principle of interconnection of the elements is fundamental, polysystem analysis assumes the hypothesis of fibration, i.e., it supposes a possibility of representation of the object under investigation as disjoint fibers, thus, mutually unbound fibers (subdomains). So, in a generalized insight a polysystem is a system itself, and polysystem analysis is a new form of system analysis.

Now then, according to PPAS [2] each concept of an aspect should be presented via interpretation morphisms in all other fibers as well as its relationships with adjacent concepts. For each of the fibers, a complete theory of the domain can be constructed. Each theory differ by their fundamental concept, but they are similar to all other via concept substitution via interpretation. This allows us to induce new axiomatic theories of domains in the image and likeness of the known ones. Each system domain (fiber) of knowledge fibrates multiply and sequentially, giving raise of the polysystem of representation of an object under investigation. All the system theories are combined in the unified model describing the domain and the object under investigation.

The application of the PPAS to the process of data and knowledge acquisition in the documents expressed as simultaneus polysystem fibering the domain in various coordinate systems (aspects) such as follows: structural organization of a document (mereological aspect), expressing domain of activity area, hierarchical inheritance of various kind of documents (domain system concept inheritance), people and organization relationship, deontic logic representation of the document meaning, etc. Having established new class or new relation between two classes, the concept or relation should be reflected as a substituted (interpreted with) one in other ontologies (e.g., mereological). To make the polysystem to be easily understandable to users, definitions for the concepts and relations in a natural language must be supplemented in terms of those ontologies. The traditional requirement of affiliation of an object to a class, and inheritance between classes and relations reflects the idea of hierarchical polysystem fibering. The relations of inheritance and affiliation are the examples of concept interpretation between the fibers.

The conceptual methodological basis of PPAS allows us to a) control the completeness of the system of objects in a document with respect to their interpretation to a predefined abstract system fiber, which we roughly call “template fiber”; b) verify the soundness with respect to the semantics of the relation defined again via interpretations; and in a distant future c) automatically produce program implementing the fiber theories via interpretation and combination of the existing algorithms of other fibers.

V. Maintaining a Dialog With User

The through interpretation feature of the polysystem layers could be used to maintain user interview dialog on the base of an inference by analogy represented as movement along relations, morphisms and interpretation in a polysystem of ontologies.

In order to realize the dialog procedure, there should already exist system fibers representing “part-whole” relations, class hierarchies of the abstract concepts, correspondence of objects to classes. The aspects of the domain is represented by set theories and necessary domain layers, e.g., the fiber defining that “man” and “woman” are disjoint concepts. The user dialog is used to determine that the object belongs to a context (fiber), recognize its interpretations into concepts of other fibers, and to provide all the context to be complete and sound with respect to their template fibers. The dialog is maintained on the base of analysis of known properties and restrictions (relations) of the subject: type of data representing the subject, its location in the context hierarchy, existing relations with subjects and objects, etc.

Let’s consider simple family relationships in fig. 1, where most of the obvious relations are hidden to make the figure more readable. There showed a part of the domain polysystem consisting of four main layers, which model various aspects of Bob’s family. Layer 1 represents the family itself, it is the only layer having objects as concepts; layer 2 represent main roles of individuals in a family; layer 3 shows genders as two opposites; and later 4 corresponds to a rough mereological model “part-whole”.

Let’s construct a sequence of questions to determine a role of Jul in Bob’s family. It is supposed that we know only that Jul is a woman, and she has been mentioned in a document related to Bob’s family. To put Jul in layer 1, in a edited context of a document, we must ask a general question: “Is Jul a part of Bob’s family?” The question is constructed according to a following intuitive inference. All members of a family are “partOf:” the family. This fact is shown as an arrow from relation “hasFather:” of Layer 2 to “partOf:” of Layer 4, other four arrows are hidden. So, in order to determine the fact that Jul is a member of Bob’s family, it is sufficient to ask “Is Jul partOf: Bob’s-family?”. The identifier of the example family is devised from the tradition to name families after their master of housekeeping (MOH) person, that is denoted by arrow from individual Bob (layer 1) to concept MOH (layer 2). Note, that layer 1 interpreted by layer 2 by means of a functor of theory of categories.

If user answered positively, i.e., Jul is a part of the family, then we must refine her role further, as role “partOf:” is not allowed in layer 1, and, hence, layer 1 is not complete yet with respect to template layer 2. Jul is a female by definition, so she cannot be father and son. If she is, she must be male,

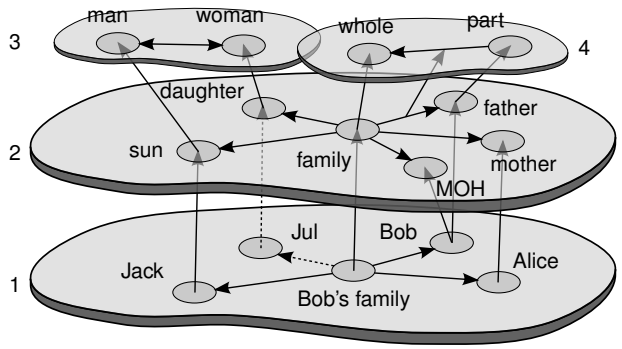


Figure 1. A Usage Example of an Ontology Polysystem

but it is impossible as male and female are formally opposite concepts. If multiplicity of relation "hasMother:" in the layer are defined as one, then we will also know, that the family already has mother. The only possible choice is Jul is daughter, resulting in another general question:"Is Jul a daughter?" If the answer to the question is also positive, finally a new triple constructed <Bob's-family, hasDaughter:, Jul>. Now the layer of Bob's family is complete with respect to the template layer 2.

If, e.g., Jul's gender is unknown, then the second question is constructed as a special question "What role does Jul have in Bob's family?" and a list of two possible answers is shown: "hasDaughter:" and "hasSon:". Another variant of the question can be synthesized if only two alternatives exist: "Which of the following two roles does Jul play in Bob's-family: hasDaughter: or hasSon: ?".

In fact in the original example the question about a role of Jul is redundant as there exist only one role for a woman being a part of Bob's family. If we remove mother Alice from the family then the question have to be asked, even for incomplete families. The general questions are asked when we could narrow a list of alternatives in further questions. The queue of the questions being asked could be defined at runtime by analogy to decision tree construction procedure, e.g., by analyzing entropy gain [8].

VI. A Notary Office Usage of the Technique

One of the applications of the technology under development is document preparation automation of a notary office. Notary office in Russia generate vast amount of printed documents. The documents contain both formalized and unformalized data in a balance reasonable for our study. For example, formalized data are the passport data of individuals, registration data of vehicles. This kind of data are normally stored in rational databases, passed from one document to other documents mostly structured and unchanged, raising a bundles of related documents and patterns of document flows. The final document content structure significantly depends on entry fields data filled in, e.g., form of notary signature field depends on legal capacities of the individuals mentioned in the main document body. Unformalized data are various enumerations of legal empowerments of the individuals, article codes, and explanatory text.

At present most of notary legal documents are generated

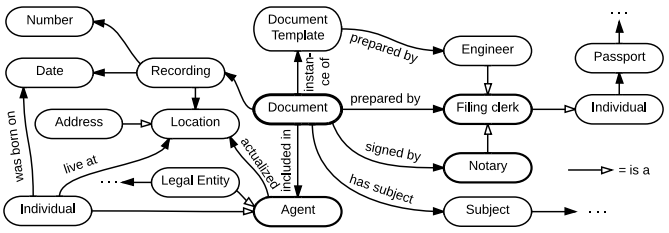


Figure 2. Upper Layer of Notary Office Ontology

from templates. In experimental office, the document templates are prepared by engineer in collaboration with a secretary having necessary experience of use of the existing software. Templates are HTML forms dispersed in the text of document. Logical structure of a document is defined by field names, and the procedures of information processing are implemented as simple form processing routines. The output document is rendered by means of value substitution instead of fields. The field names are defined by identifiers according to a special rules so that fields implicitly combine into a subject. This simple formalization allows one to carry on basic transformation operations on roles of the individuals by means of renaming or exchanging values of the field sets. Application of semantic approach to the logical layer representation of the notary document content is aimed at involvement of the secretaries (office stuff with low engineering skills) into development of the templates and constructing hierarchical arrangements of documents.

In new approach, logical layer of a notary document consists of hierarchy of subjects. The subjects relate to each other on various abstraction levels, e.g., in a letter of attorney there are at least two individuals, the first one is principal, and the second one is trusted (proxy). For each individual passport, data and place of residence are defined. The letter can contain other person data, e.g., for partially capable children. As we just noted, all the individuals are in explicit relation to the document. The triples <letatt998 containsPrincipal: indiv_78> (commas are omitted) and <letatt998 containsProxy: indiv_79> could define a principal and proxy of the document. The mentioned relations are specifications of abstract hasIndividual: and structuralElement: relations. This polysystem approach allows us visually represent hierarchical structures of the document, having interpreted the relations as structural elements, and in the same time to associate individuals with other roles in new documents, swap roles during editing. The role swap function is implemented as user interface widget generated as a result of recognition of certain subject-to-subject relations indistinguishable in a fiber, or specific patterns of relations in the document body. A general graph structure of the abstract level of a notary office automation ontology is drawn in fig. 2.

VII. The Software

The software developed within the project will include two major versions. The first one will be a client and server combination, where client will drive the data mining and visualization procedures, and server will control storage and access rights to a graph of the domain. The second version of

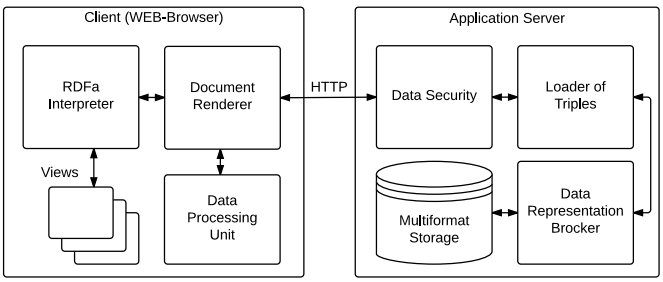


Figure 3. Program System Architecture

the software is a desktop application that is a fusion of the client and server functionality. The application is intended for corporate and personal off-line usage. The following are the requirements of the software to be fulfilled in our investigation:

1) the software should be developed as an open-source project, and we hope to involve in our project a free community of engineers and users;
2) server software must be not too consumptive of computational and memory resources of a shared hosting service;
3) in accordance to 2, a support of various database back-ends are to be implemented to store triple data.

We orient our efforts of implementation to a popular hosting environments that include PHP as programming language and operational environment, MySQL- and PostgreSQL-servers as a storage, functioning under Linux operation system. As a primary client, Firefox web browser has been chosen, and its JavaScript implementation is being considered as a client programming language.

Architecture of the first variant of the software is presented in fig. 3. Server side deals with data storage and security. “Multiformat Storage” module supports various techniques of data representation, it could be a SQL-server. The purpose of “Data Representation Broker” module is to convert triples in a format that is better support special tasks, such as aggregations, and pass the conversion result to the storage. In the backward direction, the broker restores triples from the storage as RDF entities. In order to render a template the data model instance should be loaded through module “Loader of Triples”. The instance is represented as a set of triples. The security is supplied by Data Security module.

At client side the triples are rendered as HTML texts and combined into a document by “Document Renderer” module. The result is either interpreted further by “RDFa interpreter” or processed in “Data Processing Unit”. New concepts and relations are sent back to server as RDF- and RDFa-content.

A. Implementation details

Representation of domain graph and PHP’s access to graph components is implemented on the base of EasyRdf library. The library supports loading RDF-files and represent them as PHP’s objects. The objects can be accessed in a usual object-oriented way and by means of the special standard query language SPARQL.

Storing RDF-triples in MySQL is to be implemented with ARC2 library, which also supports SPARQL-queries. ARC2 is a development platform with SW functionality, allowing one to extend the standard set of routines by means of plug-in modules as well as to create trigger functions fired as soon as SAPRQL query finished.

Text composition library Zope Page Template (ZPT) [9] is used as a generic text rendering engine out of the logical layer. ZPT is cross-platform and allows representing templates and rendering the triples inside. The templates also stored in RDF. ZPT technology and its descendants, like Chameleon [10], allow programmers and designers to share their common work on a dynamic HTML page because of ZPT is based on a declarative approach to template definition.

ZPT engine includes interpreters of several query expression languages to a logical structures of model and template: TAL (Template Attribute Language) and METAL (Meta TAL). There are also tools for internationalization (I18N) support. Developers of ZPT successfully solved the following standard web problems:

- separation application logic and user interface;
- utilization of existing HTML-page design software and editors (e.g., Dreanweaer) for manipulation of template content with conserving TAL and METAL expression;
- trivial cycle programming structures are replaced with a declarative constructions;
- thanks to METAL templates are manipulated in a powerful manner — cutting and pasting subtrees of a HTML page — using model data.

To date there are many implementations of TAL and METAL for various programming systems, including PHP (PHPTAL). The same templates can be used in all of them if programmer uses only the basis syntax and semantics of TAL. For client-side language JavaScript, ZPT has been implemented by Distal library, which can also dynamically load model data into rendered template, resulting consumption more CPU resources of client workstation and less of server.

The storage module of desktop application should not be too resources consuming, that’s why we propose usage of in-process database libraries which are also do not require special configuration and engineer’s periodical maintenance. One of the outstanding in-process database engine is Kyoto Cabinet library [11], which is high-performance implementation of BerkeleyDB standard. The library maps blocks of bytes, i.e., a key to a value, and the mapping is realized by means of B+-trees and hash tables. Kyoto Cabinet supports protection against data loss even due to partial physical damage of hard disks. The protection is achieved by special file format and data replication among a cluster of workstations.

Thus, the nowadays open-source software of internet application development and relatively cheap server computational resources completely cover all the requirements of the software under development.

VIII. Similar Projects

Among existing projects of ontology based document management the project Semantic MediaWiki can be emphasized. A basic Wiki text editing is extended with semantic annotations represented in a special markup format. The annotations are used by search engine of the Wiki pages [12].

In comparison to Semantic MediaWiki the project OntoWiki [13] obtained the similar results from an opposite starting point. OntoWiki is based on the predominantly usage of logical representation of the information as semantic network. The logical structure is edited with entry forms generated on the base of predefined vocabularies (term sets). User is allowed to change just one text property lod:content, which can contain HTML markup. The HTML markup does not relate to the logical structure of the subject to be visualized as web-document. The text is modified with WYSIWYG editor integrated into OntoWiki. The project is aimed at social networks developed under control of Linked Data technology.

Our project can be positioned as a further development if the OntoWiki engine to support a natural representation of the document content, visual editing of the content, conserving the logical layer; implementation of the data and knowledge acquisition on the base of modification analysis. The templates for OntoWiki subjects rendering are stored beyond the ontology and separately from the main content of the document. In our case the templates are auxiliary elements of ontology, it can be logically inferred from inheritance. Most of the interrelations between text content and its logical structure are expressed in RDFa.

IX. Conclusion

The research is aimed at support of Semantic Web with algorithms and software automating text document markup and its logical model induction. The approach is based on polysystem representation of the domain and document content (e.g., web site) as an evolving ontology. The usage of polysystem of ontologies and the procedure of polysystem decomposition [2] allows us to solve a number problems in the activity related to knowledge acquisition in the domain. New knowledge are represented as fibers of categories with property of through cross-fiber interpretation. In the future, the possibility will arise for automatic construction of algorithms and programs implementing a fiber theory in the image and likeness of an already constructed ones. At present the polysystem of ontologies allows us to rate and categorize facts on the domain by analogy with the already known structures, as well as obtain new interpretations of new recognized concepts.

The main part of the paper is devoted to a general description of a technique of maintenance of an interactive process of logic model induction of document content on the base of data mining of changes of various versions of the document. A problem solution outline for user dialog conduction aimed at modification and extension of the logical model with new objects and relations has been considered. The dialog is maintained by analyzing the completeness of the constructed model and logical inference in the polysystem representation of the domain ontology. A pilot software architecture has been presented as well as brief overview of used modules and libraries, their properties and features have been assessed.

As a testing ground for the technologies under development we decided to automate a notarial office, where the legal document representations has both well formalized and unformalized text fragments. On the base of the technology a social network of document data exchange can be devised. The security of the document transfer can be provided as off-line data streams: each physical document is accompanied with its bar or QR-code encoding the corresponding RDF-data of the document.

Acknowledgment

The research is carried on under support of Integration multidisciplinary project of Siberian Branch of Russian Academy of Sciences N 17 Development of services and infrastructure of scientific spatial data for supporting complex multidisciplinary scientific research of Baikal nature territory.

References

[1] T. Berners-Lee, J. Hendler, O. Lissila. "The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," Scientific American, May 17, 2001, pp. 1-18. URL: http://sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[2] A. K. Cherkashin. "Polysystem Analysis and Synthesis. Application in Geography," Novosibirsk, Russia "Nauka" Publ. co. 1997. – 502 p. (In Russian)

[3] Social network – Wikipedia, the free encyclopedia. URL: http://en.wikipedia.org/wiki/Social_network (access date: 09.01.2015).

[4] E. Cherkashin, P. Belykh, D. Annenkov, K. Paskal, "A Document Content Logical Layer Induction on the Base of Ontologies and Processing Changes," Procs. of International Conference on Applied Internet and Information Technologies, October 25, 2013, University of Novi Sad, Technical Faculty "Mihajlo Pupin", Zrenjanin, Republic of Serbia, pp. 252–257. URL: http://www.tfzr.uns.ac.rs/aiit/archives/AIIT2013/Proceedings.pdf.

[5] E. Cherkashin, V. Paramonov, et al, "Model Driven Architecture is a Complex System," E-Society Journal Research and Applications. Volume 2, Number 2, 2011, pp. 15–23. URL: http://www.tfzr.uns.ac.rs/esociety/issues/eSocietyVol2No2.pdf (access date: 09.01.2015).

[6] Model–view–controller — Wikipedia, the free encyclopedia. URL: http://en.wikipedia.org/wiki/Model-view-controller (access date: 09.01.2015).

[7] D. MacKenzie, P. Eggert, R. Stallman. "Comparing and Merging Files," URL: http://www.gnu.org/software/diffutils/manual/diffutils.pdf (access date: 09.01.2015).

[8] Information gain in decision trees — Wikipedia, the free encyclopedia. URL: http://en.wikipedia.org/wiki/Information_gain_in_decision_trees

[9] "Zope Page Templates Reference," The Zope2 Book. URL: http://docs.zope.org/zope2/zope2book/index.html (access date: 09.01.2015).

[10] Chameleon – Chameleon 2.10 documentation. URL: http://chameleon.readthedocs.org/en/latest/ (access date: 09.01.2015).

[11] Kyoto Cabinet. URL: http://fallabs.com/kyotocabinet/ (access date: 09.01.2015).

[12] Semantic MediaWiki. URL: http://semantic-mediawiki.org/ (access date: 20.08.2013).

[13] N.Heino, S.Tramp, N.Heino, S.Auer. Managing Web Content using Linked Data Principles Combining semantic structure with dynamic content syndication. Computer Software and Applications Conference (COMPSAC), 2011 IEEE 35th Annual. pp. 245 - 250. URL:http://svn.aksw.org/papers/2011/COMPSAC_lod2.eu/public.pdf (access date: 30.05.2013).

A Descriptive Specification Tools for Information System Design and Configuration

Alexey Hmelnov*†, Evgeny Fereferov*, Roman Fedorov*†, Evgeny Cherkashin*†‡, Igor Bychkov*,
*Institute of System Dynamics and Control Theory at SB RAS, Irkutsk, Lermontov str., 134, 664033, Russia
†Irkutsk State University, Irkutsk, Gagarina blvd., 20, 664003, Russia
‡National Research Irkutsk State Technical University, Irkutsk, Lermontov str., 83, 664074, Russia
{alex,fereferov,fedorov,eugeneai,bychkov}@icc.ru

***Abstract*—Software development technologies evolved from concrete algorithm design and their programmatic implementation to modeling the whole information systems (IS). The modeling approaches allows developers to describe the system in terms of a domain, software design, properties of implementation platforms, and standard ways of the IS development such as programming traditions in a concrete firm. In the case if a new IS is being constructed, having an already implemented database, Data Driven Engineering (DDE) is used. The model for IS is constructed by importing database structure and adding a semantic layer description reflecting the database structure to the domain terms.**

We describe a DDE-approach based on declarative specifications of existing database schema and properties of the target application, as well as interpretation of the description into IS components. The interpretation in various aspects allows developer to construct user interface for database modification, subsystem of SQL query construction in terms of domain, etc. An example of the technique application is presented.

I. Introduction

Software development technologies evolved from concrete algorithm design and their programmatic implementation to modeling the whole information systems (IS). The modeling approaches allows developers to describe the system in terms of domain, engineering and software design, features of implementation platforms and tools, and standards of the IS development of a software firm such as programming traditions, used modeling and instrumental tools, etc. Very simplified techniques of the modeling usages are infiltrated into nowadays production developer "philosophies", e.g., referred to as DRY (Do not Repeat Yourself), where each concept in a system must be defined only once. DRY is a basis of many open-source web development frameworks such as Django [1].

The theoretical basis of the modeling approach is model transformation theories and their software implementations, which are developed in the field of Model Driven Engineering (MDE) [2]. The transformations construct ultimately either source code modules realizing modeled structures or an operator interpretation of the models of an intermediate level like virtual machine byte code, SQL queries execution plan synthesis, and executable UML. Model transformation is based on recognition of a pattern structures in the source model, relating them to the model of development platform rules of transformation. The results of transformation are structures of a model of more specific (concrete) level of abstraction. At the lowest level the code generation routines are executed, which use source code templates. Thus, from the point of view of generative programming, one of main function of the transformation process is an intelligent configuration generation for the source code generators.

In the case if new IS is being constructed, having already designed and deployed database, Data Driven Engineering (DDE, a kind of MDE) is used. The model for IS is constructed by importing database structure, e.g. DDL-queries, analyzing it, and adding a semantic layer of the database structure description in IS's domain concepts and relations. In Laboratory of Complex Information Systems of Institute of System Dynamics and Control Theory at Siberian Branch of Russian Academy of Sciences, a number of DDE approaches has been developed based on declarative specifications and their run-time interpretations, as well as techniques automating the description process. The interpretation in various aspects allows developer to construct user interfaces for database modification, subsystem of SQL-queries construction in terms of domain, plugin integration, etc.

The report describes an DDE-approach of IS development based on a special specification language describing database properties, interpretation of the description in various IS components, resulting in automatic user interface construction and layouting; a configuration generation for constraint-oriented database editor, SQL-query builder, OLAP subsystem, etc.

II. The Description Language

The DDE-modeling of an IS starts either from description of a general view of data structures used by business-logic or from loading schema of already existing database. The second case has important value in practice, when customer requires a new implementation of an exiting IS with new features and with conserving exiting data in its warehouse.

The structural description allows developers implementing routine functionality via transformation and generation, e.g., data input forms; user interface and business-logic interaction; import, export and integration operations. Adding a domain description allows one to refine user interface with notions familiar to user, as well as construct various query-answering subsystems like SQL-query builders and OLAP operating in terms of the domain.

In Institute of System Dynamics and Control Theory at SB RAS a number of DDE technologies has been developed within economic contracts activities supported by necessary theoretical investigations. One of the recent description format

is based on mix of the `.ini`-file and SQL-query languages. The first format is used to describe database structure and semantic of attributes and their relations, the second part is used to describe the rules of user interface generation, and interaction with geographical information systems and other applications.

A. Development Stages

In the proposed approach, in a general case on the first stage of IS development, system engineer carries on structural analysts of the domain, resulting in description of concepts and their basic relations, business-processes and software requirements. After that, at the second stage, the database design follows, which produces database schema ad its implementation as database server. Common design tools (such as Sybase Power Designer, IBM Rational Rose) are used on this stage. The second stage is carried on if IS is designed from scratch, and there are no database already implemented.

On the third stage, the model of IS database is constructed by loading the implemented database schema and extending it with the description of the following special database object classes: `TABLE`, `VIEW`, `RULE`, and so called *superstructure*. The `TABLE` objects describe IS database structure and relations, attribute names and data types. These object are automatically constructed by loading and analyzing IS database schema, if it is literally devised (it must be at least in third normal form). The lack of schemata data is compensated manually with the description language and the following interpretation.

The `VIEW` objects, which are more complicated than `TABLE` ones, used to generate user interface. Here we define properties of the attribute and table relations, for example, for each foreign key a look-up widget can be used to select a record from a foreign table (a reference book). For a complex reference book and transitive references, the key attribute sets must be specified to increase generated user interface usability. There are other special cases of foreign key interpretations, like "master-details", and all of them can be described with the specification language.

The superstructure objects support interaction of developed IS with foreign modules and libraries, processing information with methods beyond the standard of rational databases and SQL, e.g., computer simulation and complex data visualization.

The `RULE` objects are responsible for expression of functions implementation in generated IS specific to its business-processes. By means of the rules, one can describe input field logic (enabling and disabling them), implement uniqueness check routines and plug-in modules interaction, i.e., rules provide declarative means of describing and implementing relationships between objects and components in business-logic layer of IS. The rules also affect the generated queries to IS database during editing, as well as creating special widgets in user interfaces, which connect input data with spatial tables and render cartographic. For all the object classes a formal set model is devised and investigated, resulting in a theoretical basis of the technology under development.

On the fourth stage, the IS functionality is generated at run time as a result of the specification interpretations. It integrates foreign software subsystems according to the rules.

B. The Description Language

Our specification language is reasonably simple whereas complete with respect to expressive abilities. The language has LL(1) grammar class. Its syntactic structures describe elements of IS and a general options like database connection properties. The specifications written in the language are *precise*, *understandable* and *complete* according to [3], [4]. The language clauses have the following general structure:

```
<Start keyword> <List of mandatory statements>
[ list of optional statements ]
```

The `Start keyword` can be one of the following: `ADO`, `BDE`, `CFG`, `TABLE`, `VIEW`, `PLUGINS`, `USES`, `INCLUDE`, `MENU`. First two are used to describe properties of database connection. Keyword `CFG` used to describe general application settings of IS. Keywords `TABLE`, `VIEW` describe the database schema and view widgets. Keyword `PLUGINS` describes interaction with outer systems; `USES` and `INCLUDE` allow developer to break an IS description into source modules. Keyword `MENU` controls the main menu of IS application.

Let's consider a part of an example description.

```
TABLE ADR FOR SP_ADDR AS 'Address' READONLY
```

This statement describes table `SP_ADDR` of IS, the table will have an alias `ADR` for usage in the rest of the specification. The table will have also a user-friendly alias `'Address'` in user interface forms, table editor, etc. The following language construction describes table fields.

```
FIELDS (ID_BUIL_G P AS 'Building Code',
        ID_NREG ^SP_NREG AS 'Region Code',
 ID_MK_GM ^SP_MK_GM AS 'Microdistrict Code',
        ID_STREET ^vSP_STR AS 'Street Code',
        BUIL=S AS 'Building',
        ANN=S AS 'Annotation');
```

Identifiers `ID_*`, `BUIL` and `ANN` are table fields. Sign `^` creates a reference to a table or a view, in particular, `^SP_NREG` is a table referenced by field `ID_NREG`, and `^vSP_STR` is a reference to a view. Expression `ANN=S` denotes that `ANN` is a string field. At the first line of the example, `ID_BUIL P` denotes that the field is a key field.

There are also more sophisticated ways of various reference denotation, expressed with a general `REFS(...)` construction. Its parameters allows one to define references of various IS business-logic interpretations including "master-details", special cases of "reference book" and so on.

III. Interpretation of Database Description

All current implementations of the interpretation of the specification language are run-time. After preparing the specification of the IS database, the IS developer must run the IS environment with the specification an the database as its input. The specification is translated, and the translator creates two data structures. One is stored in process memory as complex data structures and used by form and query builders. Another one is a technical database of tables, describing IS database

structures. The technical database is used by database editor and OLAP modules. The translator also checks whether the specification is changed and corrects the database accordingly.

A. User Interface Generation

One of the important IS quality characteristics is its input forms, and layouting the widgets in the forms is the question of usability. From one hand side the widgets must efficiently consume the area of a form, and on another hand side, the user interface must be convenient and fit to application functionality requirements.

We considered a number of approaches of user input form constructions, such as [5] and [6], and come to the conclusion: in order to improve usability of user input forms, we must account the domain properties of attributes. For the purpose, a rule-based engine has been developed, which uses the database specification to access the attribute properties. The engine interprets the database specification, adjusts Database Editor widget, and generates form for database records modification.

a) Database Editor: Database Editor controls a dynamic user interface operating over IS database records. Its layout includes three areas (panels): menu of entities, operations, and data table itself. Menu of entities (fig. 1, left panel) consist of database and view table names described in the specification. The entities are joined in semantic groups according to their roles and functions.

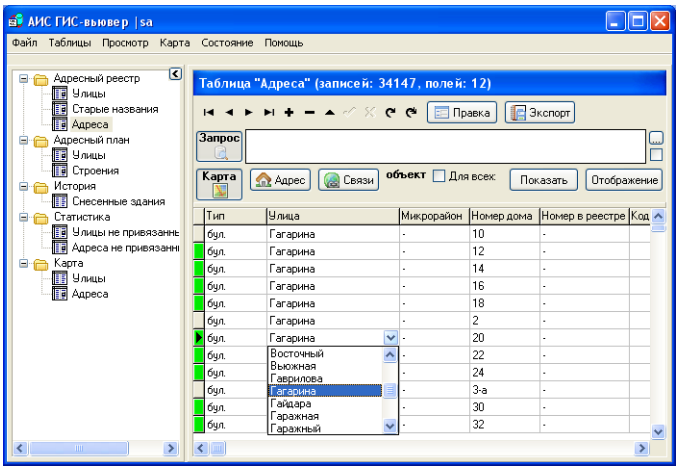


Figure 1. Database editor

The operation area (fig. 1, top panel) displays widgets for construction and execution of data processing operations. The set of widgets for each table (view) is controlled by the specification. The minimal set contains table navigation buttons, “Edit”, “Export” and “Query” buttons. Optional buttons are “Options” and “Cartographic”. The layout of the widgets is under control of layout manager.

The table area (fig. 1, bottom-right panel) displays the chosen entity. The data is represented in two views: tabular and as a form. Tabular form supports standard data editing operations, but they adjusted according to specification: reference fields are choice boxes and look-up widgets. Dependable (in sense of the specification) fields are changed as soon as their ascendant become edited. For example, in fig. 1 street name in an address record is chosen from a look-up list of streets. The street `ID` is stored in an address reference field, this results in changing address `TYPE` (boulevard or street) field according to the type of chosen street.

The order and widths of the table columns can be adjusted according to user’s preferences. The customization is stored in user profile `.wsf`–file.

b) Form Mode: In form mode, user input form is generated. The form widget set is created in accounting table’s references to other tables, views and visual components. The form is activated from Database Editor with a mouse click on “Edit” button, or when a creation of new record is requested. The layout of the widgets of the form is generated according to the database specification and run time context: whether this record belongs to a main table or a reference one. Each attribute editing widget is accompanied with a label showing user-friendly attribute alias from the specification. The label can signify attributes uniqueness, non-null condition and a membership in the primary key.

For each widget, placement, order, widths and width adjustment are controlled by layout manager component. The order and layout of the form must conform to criteria of usability. Upon form size change or screen resolution the widgets must be reordered, resized, and so on, conversing the usability criteria. The layout manager extends ideas of FlowLayout [7] (Java): the widgets a placed in rows from left to right.

FlowLayout manager implies that each widget has fixed size. We took over the restriction, and our layout manager takes advantage of some heuristics based on minimal (`szMin`), best (`szBes`), and maximal (`szMax`) sizes of a widget. Maximal size is equal to field data size, e.g., for `VARCHAR(150)` field, the `szMax` is equal to 150. Experience shows that it is more ergonomic to display about 70 % of `szMax`, and it is `szBest`. Minimal size for strings is taken about 10 characters.

The layout manager arranges the widgets according to the following strategy. At the first stage all three sizes are queried from the widgets, and the manager tries to figure out a layout,

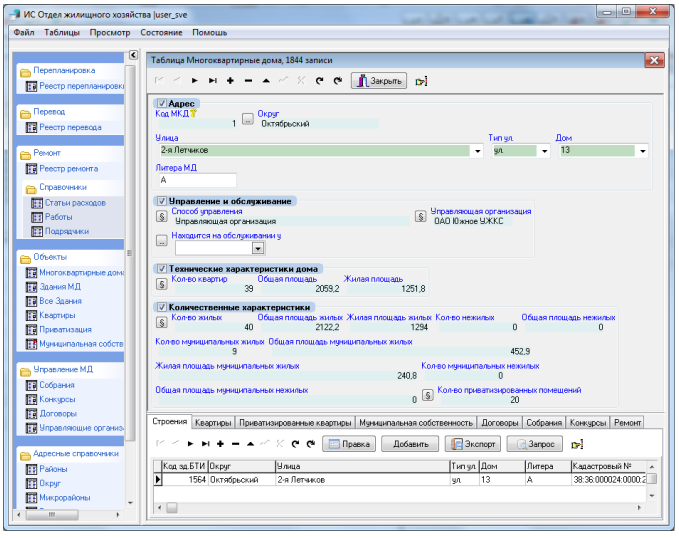


Figure 2. An example of autolayout form

where all the widgets have maximal sizes and placed in rows from left to right. If a widget is not shown in the form, the layouting immediately restarts the process, with `szBest` used as widgets sizes. If the layout process finished successfully and there are some free space, then the layout manager tries to extend significant widgets to their maximal sizes. In the wort case, when the second try has failed, the manager uses `szMin`.

Attributes can also be grouped according to a logical record structure, which a also defined in the specification. The groups are placed in the form as tab containers or group boxes, arrangement of the fields inside tabs is made analogously. An form generation example from "Address Plan" IS instance (described further) is shown in fig. 2.

B. Query Builder

Most of IS applications require components for data search and filtering. We developed an expressive user-friendly search engine module that allows one to quickly jump to a specific record and also allow a subject specialist to query record data from database without programmers assistance.

Query constructor of Microsoft Access allows user to design a complex SQL-queries for aggregating and modifying data. The query builder supports definition of complex conditions for the `WHERE` and `HAVING` sections of the query. The query builder also does not account the domain, and, thus, it is intended for programmers and experienced users accounted with database technologies. Another problem is the availability of SQL-query designer utilities for web projects.

In our work we proposed an engine that supports interactive specification of field value criteria and construction of combinations of logical and filter operations sufficient to describe common user queries. Query builder is a form that is configured in accordance to the specification and the table especially. The logical condition is a tree structure, which is drawn horizontally to allow long field names being displayed correctly without overlapping.

To control the tree editing, a special widget has been developed. The widget placed inside a dialog window is shown in fig. 3. The widget allows to construct tree from elementary operations, move the operations between various levels of nesting, operand switching, editing criteria values.

Extended mode of Query Builder widget supports referencing subordinal tables. Processing such references requires recursive call of the builder. For example, the query "Which streets of Leninsky and Sverdlovsky city districts [of Irkutsk] has more than two gas stations?" can be expressed by means of Query Builder as the following sentence: "City district is in 'Leninsky', 'Sverdlovsky' and Table 'Address' row count [is taken] from (Table 'Address building' has records from (Building type is in 'GS')) >=2." In this example two level of referencing is used: streets include addresses, at the same address can be situated several buildings. This results in generation of correlated subqueries interpreted as "master-detail" relations. The query builder generates SQL-query as follows:

```
SELECT T.id Id, T.reg_id reg_id, T0.type ST,
  T.name name, T1.name Distr, T2.name mk,
  T.BTI_id BTI_id, T.IsNew IsNew,
  T.IsChanged IsChanged from dbo.street T
LEFT OUTER JOIN dbo.street_type T0 ON
  (T0.Id=T.street_type_id)
LEFT OUTER JOIN dbo.distr T1 ON
  (T1.id=T.distr_id)
LEFT OUTER JOIN dbo.mkr T2 ON
  (T2.id=T.mkr_id)
WHERE (T1.id IN (8,10) AND (select count(*)
  from dbo.buildings S
WHERE (T.id=S.street_id) AND (exists
  (select * from dbo.bld R
WHERE (S.Id=R.ADDRESS_ID) AND (UPPER
  (R.HOUSE_TYPE) IN ('GS'))))))>=2)
ORDER BY T.id
```

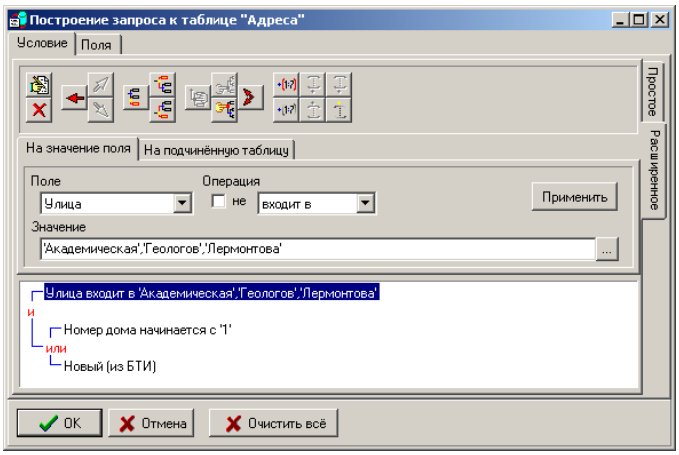


Figure 3. Subject specialist controlled query builder

C. On-Line Analitical Processing

Another valuable usage of specifications is to implement import routines of IS database data for On-Line Analitical Processing unit (OLAPu). OLAPu allows user to store, select and analyze numeric indicators. The indicators are considered as funtions of a set of attributes. OLAPu user interface aids selection of indicators in various attribute projection combinations.

OLAPu's design is a classic client-server architecture, where server is a rational data storage, and client provides a flexible access to the indicator data by means of intuitive user interface. The interface is configured specially to support the multidimensional OLAP. User selects and adjusts projection and indicator aggregation configuration by means of interface widgets, the results of the selection are presented as interactive tables and can be stored in Microsoft Excel spreadsheet files.

To convert IS data into a database of indicators suitable to be processed by OLAPu, `TABLE`, `VIEW` and superstructure objects are used. The data is converted into a tree of indicators. Each node of the tree corresponds to data selected in a projection. Each projection is characterized by a set of attributes. The sets are referred to as dimensions. They could describe regions, industries, types of ownership, etc. A special case of the dimensions is time intervals, such as year, half a year, quartal, month. Most of the indicators are tied to time intervals.

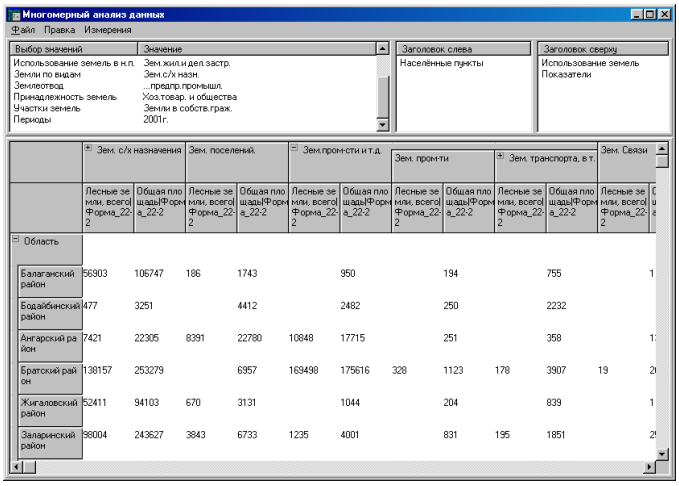


Figure 4. Main window of OLAP unit

The indicator data usually have more complex structure than just a linearly ordered set. In a general case, the set is a tree structure, where one indicator aggregates data of a set of subordinal indicators, e.g., state indicator aggregates state regional data. The same is true for industries. The indicator database includes all the index dimension attribute references to the original tables of the IS database and the index value itself.

The main windows of OLAP unit is shown in fig. 4. The window has been divided on two parts. The upper part consist of three panels with dimension lists, and the lower part is the resulting table. User must select an indicator from the left list of dimensions as well as the dimensions associated with columns and rows of the resulting table. The headers of rows and columns may contain combinations of dimensions. In this case for all the combination a row (column) is inserted into the resulting table widget. The combinations of dimensions and attribute value filters are adjusted with mouse clicks.

Time intervals are selected with a special pop up dialog widget on mouse click. User can drag dimensions between lists and reorder them, this results in regeneration of the resulting table. When editing is allowed, user can change values of indicators directly in the resulting table cells. The changes are immediately stored into the intermediate database of IS and written into a journal for eventual import into original IS database. The table widget allows exporting data and table structure into system clipboard in a table format recognizable by Microsoft Excel, the same is implemented in the opposite direction.

D. Report Builder

Report Builder unit is a plug-in module, that has access to internal representation of the specification via API. Report Builder functionality is based on Microsoft Word documents. The report templates are constructed as texts with special placeholder identifiers (mark points), which are replaced with data. Report Builder templates are listed in the specification as well as the associated `TABLE` and `VIEW` objects. The mark point language is as follows:

```
<Mark Point> ::= <# <Data Domain>
                 <Place Holder> [ <Format> ] >
<Data Domain> ::= TF | PN | IMG | DETAIL
<Place Holder> ::= <Field Name> | <PN> |
                 <Detail Reference>
<FN> ::= SN=<Field Name> FN=<Field Name>
         MN=<Field Name>
<Field Name> ::= N=<Char>+ | <Digit>*
<Detail Reference> ::= <Field Name>
<Format> ::= <Case> | SHOW
<Case> ::= CASE= ( N | G | D | A | I | P )
```

In the report engine, a conditional processing has been implemented. The following special case of mark point construction is used:

```
"<#" IF <Field Name> <Value> ">"
     <Output> "<#ELSE>" <Output> "<#END>"
<Value> ::= V=<Char>* | <Digit>*
<Output> ::= <Mark Point> | (<Char>*|<Digit>*)
```

IV. GIS Integration: a Usage Example

The developed technologies are applied in GeoAW (Geological Automatized Workplace) environment, which is aimed at construction of GIS for municipal authorities. Its derivative GIS systems allow one to represent data accumulated in departments as GIS cartographic. In order to construct such system developers must go through the steps described early to describe data sources and filter data to be represented in new GIS applications. The main attention should be paid to spatial related attributes, as the attributes are used in connecting the table rows and graphics shapes of topographic base.

GeoAW environment is an instrumental tool to create stand-alone Windows GIS applications. Its architecture is shown in fig. 5. GeoAW consists of the following main software components: Specification Management, System Kernel, Database Editor, user Query Builder, Plugin Interface, and GIS as external application.

The GeoAW GIS module is implemented on the base of authors' extension of GIS Toolkit, which is popular in Russian government "Panorama" [8]. Panorama is a COM-server incorporated in applications. The COM-server object can be configured so that the user of IS could reach only allowed GIS data on the cartographic base, namely by running under a special Windows user account. The IS data are transferred to the COM-server process and shown as cartographic (fig. 6).

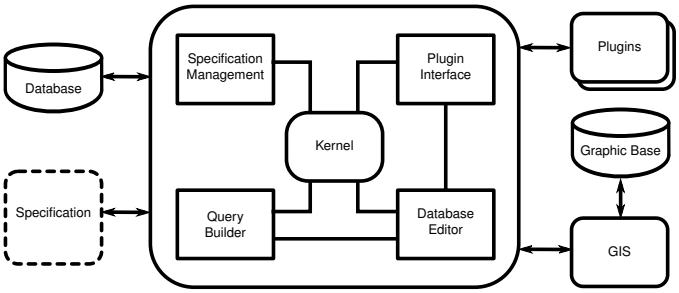


Figure 5. GeoAW instrumental environment architecture

The resulting integration of Address Plan IS and GIS Toolkit has functions relating to selection of objects on the map and showing tabular data with main IS module. At the same time, filter queries over tables having spatial attributes can affect the semantic layer presentation on cartographic.

Therefore, the example showed that the instrumental GeoAW environment fully supports all the functions for specification development, their interpretation as configurations of all modules related to IS data aggregation, storage and processing. The developed technique and software technologies allows one to organize robust and productive IS development. The resulting application is generated dynamically without compilation of any source codes.

V. Further Development

In further work the description should be extended with an additional superstructure, which could be used to extend the capabilities of SQL-query constructor to allow user formulating the query in a natural language in terms of modeled domain. The superstructure expands the description of the database entities and their relationships with abstract notions of domain and relations, resulting in fibered representation of the domain of two fibers: abstract domain and database structure [9], [10].

Another direction of development is further extension of the interpreter and creating a code-generating version of the description language on the base of nowadays techniques of language-oriented programming [11].

VI. Conclusion

The nowadays techniques of application software development are aimed at close cooperation of developer and customer. To shift some responsibility to the customer's users and to extend the maintainability and usability of software, it is necessary to develop techniques, which produce some processing according to a domain description. This description is to be connected to the models of various aspects of the software, in particular, database schema.

In this report we considered such description language, which allows programmer to express database schema in terms of domain via human-readable aliases for the database structure elements, references (foreign keys) can be classified as "master-detail", "reference book" and other relationships. The language is interpreted into configurations of modules for database editing, record adding and modification, query construction, report generation, OLAP and so on. The language expressiveness is sufficient to support development of municipal GIS software. As a part of programmer efforts within the proposed approach are aimed at extending language interpreter, user interfaces generated out of the description have good usability characteristics. Very frequently it is much better than the original software.

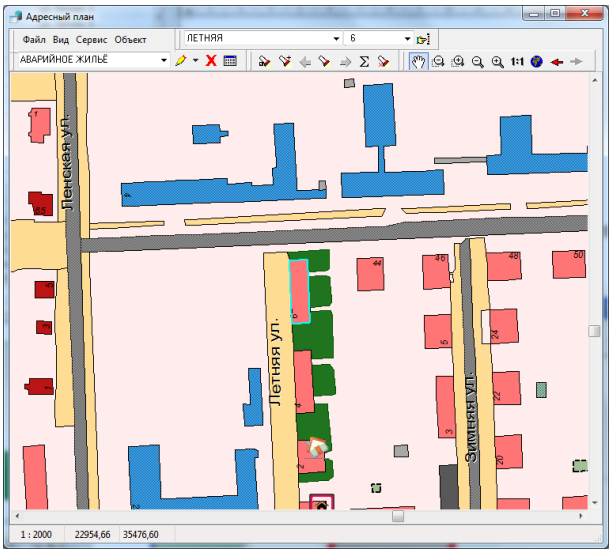


Figure 6. Address plan GIS window

Acknowledgment

The research is carried on under support of Integration multidisciplinary project of Siberian Branch of Russian Academy of Sciences N 17 "Development of services and infrastructure of scientific spatial data for supporting complex multidisciplinary scientific research of Baikal nature territory".

References

[1] The Web framework for perfectionists with deadlines — Django. URL: https://www.djangoproject.com/ (access date: 10.02.2015)

[2] M.Brambilla, J.Cabot, M.Wimmer, "Model-Driven Software Engineering in Practice." Morgan & Claypool Publ. Co., 2012, 182 p.

[3] V.N.Agafonov. "Program specification: concepts and their organization." Novosibirsk "Nauka" publ. co. 1987. 240 p. (in Russian)

[4] V.N.Agafonov. "Requirements and specifications in program development." Moscow "Mir" publ. co. 1984. (in Russian)

[5] Qt Project. URL: http://qt-project.org/. (access date: 02.02.2015)

[6] The GTK+ Project. URL: http://www.gtk.org/. (access date: 02.02.2015)

[7] Class FlowLayout. Java Platform Standart Ed.7. URL: http://docs.oracle.com/javase/7/docs/api/java/awt/FlowLayout.html (access date: 05.07.2013).

[8] GIS Panorama. URL: http://www.gisinfo.net/. (access date: 21.02.2015)

[9] A. K. Cherkashin. "Polysystem Analysis and Synthesis. Application in Geography," Novosibirsk "Nauka" publ. co. 1997. – 502 p. (in Russian)

[10] E. Cherkashin, V. Paramonov, et al, "Model Driven Architecture is a Complex System," E-Society Journal Research and Applications. Volume 2, Number 2, 2011, pp. 15–23. URL: http://www.tfzr.uns.ac.rs/esociety/issues/eSocietyVol2No2.pdf (access date: 09.01.2015).

[11] D.V.Annenkov, E.A.Cherkashin. "Generation technique for Django MVC web framework using the stratego transformation language." Procs of 36th International Convention MIPRO, 2013, 20-24 May 2013, Chroatia, Opatia, 2013, pp. 1084–1087.

Importance of Network Nodes for Navigation with Fractional Knowledge

Denis Helic, Florian Geigl
Knowledge Technologies Institute
Graz University of Technology
Graz, Austria
Email: {dhelic,florian.geigl}@tugraz.at

***Abstract*—Due to sheer size of today's information and social networks people and algorithms are limited in their knowledge about those networks. Since this knowledge is important in many processes taking place on such network, e.g. search and navigation, an interesting question is how this fractional knowledge should be constituted and structured. For example, when agents navigate an information network the question is what are the most important nodes for an efficient navigation. In this paper we apply machine learning techniques to learn an optimal division of network nodes into two classes: nodes that are important for navigation and nodes that are not important for navigation. We learn this optimal classification by maximizing the correlation between global and fractional knowledge about a network. For learning we resort to MCMC methods such as Simulated Annealing and a simple approach for random assignment of nodes to classes. We perform a series of experiments by applying our optimization algorithm to synthetic networks with different community structures to shed more light into the relation between network structure, the node structural properties such as various centrality measures, and node classes produced by the algorithm.**

I. Introduction

Search in a network is a dynamical process in which an agent (a human or an algorithm) moves along the links between the nodes in search for a given target node or a given set of target nodes. In an informed search the agent possesses global or fractional background knowledge of the network, which guides it towards selecting specific links. In both cases, background knowledge is a certain numerical representation of the quality of a link (or the destination node incident to that link) with respect to reaching given target node(s). The distinction between global and fractional knowledge can be summarized as follows. In case of the global knowledge the numerical representations are calculated from the complete knowledge of the network, i.e. the agent knows about every node and every link in the network, whereas in the case of fractional knowledge the agent knows only about some selected nodes and links.

Due to sheer size of today's information and social networks people as well as algorithms are typically limited in their knowledge about those networks. For a typical network it is only possible to have knowledge about certain nodes and links, e.g. in a social network people only know their local friend circle, and additionally a couple of very famous and popular people. In a typical case they will lack knowledge about other people's friends with whom they do not share friends. Similarly, in a information network such as Wikipedia users have knowledge about their topic of expertise or interest, and about some general and well-known concepts, however they will typically lack knowledge in the areas that they are unfamiliar with.

This situation raises an interesting research question. In a situation with limited time and resources to learn about the networks, their nodes, links and structure how can we optimally invest those resources. In other words, what nodes and/or links are the most important ones to know about to efficiently navigate the networks even with a fractional knowledge about those networks. For example, if an agent has resources to learn about a few of currently unknown network nodes, which nodes should she learn about? Which strategies should we apply when learning about a network? Should we first learn about e.g. high-degree nodes, or about nodes with a high local clustering?

In this paper we set out to answer this question by:

(i) Defining a global criteria for assessing the quality of fractional knowledge,
(ii) Applying machine learning techniques to learn an optimal division of network nodes into two classes: nodes that are important for navigation and nodes that are not important for navigation, and
(iii) Analyzing the properties of important nodes and generalizing our findings to a heuristic for learning about new network nodes.

The paper structure follows approximately the presented questions. In the next Section we give a short overview of the related work in this field. In Section III we describe the objective function for the quality of fractional knowledge. Section IV gives the details about our optimization experiments with fractional knowledge of synthetic networks. In addition, we present the experimental results and shortly discuss them in that Section. Finally, we conclude the paper and provide directions for further research.

II. Related Work

Search has been an important concept in theoretical and empirical studies of networks. For example, in Web search the famous PageRank calculation [1], which determines the importance of Web pages is based on a random walk on

the Web network. In a so-called Random Surfer model a hypothetical user navigates the Web by clicking randomly on the Web links. In the limit of large number of clicks the fraction of visits to each Web page gives its PageRank.

Research on decentralized search in social networks started with Milgram's seminal small world experiment [2] in which he aimed to study the connectedness of the US society. Milgram found that people in such a large social network are connected by short chains of acquaintances. Moreover, he found that people are able to navigate large social networks efficiently, i.e. that they are able to find those short chains even if they only possess the local knowledge of the network. In the subsequent research, Kleinberg analyzed the second result of the Milgram's experiment – the ability of humans to find a short path when there is such a path between two nodes [3]–[5]. Kleinberg concluded that social networks possess certain latent properties that humans are aware of. This background knowledge of network structure allows humans to find a short path between two arbitrary network nodes efficiently. Independently, Watts [6] introduced the notion of social identity as a membership in a number of social groups organized in hierarchies and showed the existence of efficient decentralized search algorithms by simulation. Finally, Adamic [7] and others applied such a model to explain search in e.g. professional social networks.

All of these models have been based on the assumption that an agent navigating the network possesses a complete background knowledge on distances or similarities between pairs of network nodes. Contrary to that, in [8] the authors argued that distances between nodes are typically available only for some of the node pairs and that the assumption that an agent possesses a complete background knowledge is unrealistic. Thus, the authors analyzed an informed search algorithm that bases its decisions on the node degree (the number of nodes adjacent to a given node), a quantity that is always available locally. The algorithm performed better in comparison with a random walk but could not reach the performance of an agent having a complete background knowledge.

Additionally, decentralized search can be used to model user navigation in an information network, e.g. the Web or Wikipedia [9]. In that model the user navigates from a Web page to a Web page by following her own intuition and her own knowledge about the topics of various Web pages.

A crucial element of an informed search is the notion of distance or similarity between the pairs of network nodes. Mathematically, we may represent the background knowledge with a distance or a similarity matrix. The rows and columns of the similarity matrix represent the network nodes and the elements of the matrix give the similarity score between corresponding network nodes. If the similarity function is symmetric we obtain a symmetric similarity matrix. Partial background knowledge may be easily represented by blank elements in the matrix for those node pairs where the agent's knowledge is missing.

Different similarity measures have been previously discussed in the literature for various tasks. For example, [10] gives an overview and evaluation of various similarity measures for the link prediction task. In general, similarities in networks can be calculated from external or exogenous information about the networks (e.g. textual content in an information networks), or from internal or endogenous information (internal linking patterns). There are two possible similarity measures based on the network linking patterns: structural equivalence measures and regular equivalence measures. Structural equivalence is based on the simple idea that two nodes are similar if they share many of the same network neighbors, whereas regular equivalence between two nodes is given if they have neighbors who themselves are similar. Since absolute numbers of common neighbors are hard to interpret, this number is typically normalized. Different normalizing constants give rise to different similarity measures such as cosine similarity, Pearson correlation coefficients, or Euclidean distance ([11] gives a nice overview and comparison of various normalization methods). In contrast, regular equivalence between nodes is established when two nodes do not necessarily share same neighbors but rather they have neighbors who themselves are similar, or in a more simple case two network nodes i and j are similar if j is similar to other neighbors of i. In past, several simple algebraic measures have been developed to measure regular equivalence [12], [13].

III. Methodology

A. Katz Similarity

In this paper we represent the knowledge about network as a similarity matrix. We denote the similarity matrix with $\boldsymbol{\sigma}$ – the element σ_{ij} of the matrix gives the similarity between nodes i and j from the network. Further, we concentrate on a particular measure of similarity called Katz similarity [10].

Let us now first introduce the basic quantities relevant for measuring this similarity. Let us denote the adjacency matrix of a graph $G(V, E)$ with $\boldsymbol{A}$:

$$A_{ij} = \begin{cases} 1 \text{ if nodes } i \text{ and } j \text{ are connected by a link} \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

We define Katz similarity recursively as [14]:

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}, \tag{2}$$

where δ_{ij} is Kronecker delta ($\delta_{ij} = 1$ if and only if $i = j$, otherwise $\delta_{ij} = 0$).

The idea is that the term σ_{kj} is large if the neighbor k of i is similar to j. We then sum over all neighbors of i, which is ensured by the product $A_{ik}\sigma_{kj}$ since A_{ik} equals 1 only if i and k are neighbors. Finally, the term δ_{ij} assigns a high value of "self-similarity" of a node to itself. The constant α weights the influence of neighbors as compared to the influence of the self-similarity to the nodes similarity.

We can write the expression for the Katz similarity in the matrix form as:

$$\boldsymbol{\sigma} = \alpha \boldsymbol{A}\boldsymbol{\sigma} + \boldsymbol{I}, \tag{3}$$

where $\boldsymbol{I}$ is the identity matrix.

We can now solve this equation for $\boldsymbol{\sigma}$, which gives us the following:

$$\boldsymbol{\sigma} = (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1}. \quad (4)$$

In case of large networks with e.g. millions of nodes, inverting a matrix is computationally expensive. Therefore, we may want to evaluate the Equation 3 recursively. We would typically start by assigning initial values for similarity to zero, e.g. $\boldsymbol{\sigma}^0 = 0$. We then obtain (for the first three iterations):

$$\boldsymbol{\sigma}^0 = 0, \boldsymbol{\sigma}^1 = \boldsymbol{I}, \boldsymbol{\sigma}^2 = \alpha \boldsymbol{A} + \boldsymbol{I}, \boldsymbol{\sigma}^3 = \alpha^2 \boldsymbol{A}^2 + \alpha \boldsymbol{A} + \boldsymbol{I}. \quad (5)$$

Thus, in the limit of large number of iterations m we have:

$$\boldsymbol{\sigma} = \sum_{m=0}^{\infty} \alpha^m \boldsymbol{A}^m. \quad (6)$$

The elements of matrix $\boldsymbol{A}^m$ at index i and j, i.e. $[\boldsymbol{A}^m]_{ij}$ give the total number of paths of length m between nodes i and j. The Katz similarity can be seen as a weighted count of all paths between the nodes i and j with paths of length r getting weight α^r. As long as $\alpha < 1$ longer paths obtain less weight than shorter ones.Thus, two nodes will be similar if they are connected either by few short paths or if they are connected by many longer paths.

The inverse of $(\boldsymbol{I} - \alpha \boldsymbol{A})$ does not exist for $det(\boldsymbol{I} - \alpha \boldsymbol{A}) = 0$, which is exactly the characteristic equation for eigenvalues of $\boldsymbol{A}$, i.e. in the matrix form $\boldsymbol{\lambda} = \mathbf{1}/\boldsymbol{\alpha}$. Thus, the inverse does not exist whenever $1/\alpha$ equals λ_i. If we start with $\alpha = 0$ and gradually increase it, we first hit the equality $1/\alpha = \lambda_i$ for the largest eigenvalue λ_1, and then again every time when we approach the next eigenvalue. Thus, if we pick an $\alpha < 1/\lambda_1$ we ensure that the inverse exists (or that $\sum_{m=0}^{\infty} \alpha^m \boldsymbol{A}^m$ converges).

B. Correlation between global and fractional knowledge

A nice property of the Katz similarity is that with the parameter α we can control the amount of knowledge that we have about a particular network. For example, setting α closely to the upper bound $1/\lambda_1$ weights also even very long paths significantly. For example by setting e.g. $\alpha = 0.5$ paths of length 4 are weighted with 0.0625 and even paths with with length 8 are weighted with 0.00390625. Thus, we may argue that the similarity is calculated with a complete global information about the network. On the other hand by setting a small value for α, e.g. $\alpha = 10^{-8}$ even path of length 2 are negligible since their weights are 10^{-16}. In other words for smaller values of α the information kept in the Katz similarity matrix is only local. In an extreme case where only paths of length 1 obtain a significant weight the Katz similarity matrix is equal to the adjacency matrix up to a constant and the diagonal entries.

Thus, a possible global criteria for comparing the global knowledge (Katz similarity matrix with a large α, e.g. $\alpha = 0.99/\lambda_1$) and a local knowledge (Katz similarity matrix with a small α, e.g. $\alpha = 10^{-12}/\lambda_1$, or even the adjacency matrix) can be a correlation (Pearson correlation coefficient ρ) of those two matrices. The correlation would be high if we observe links between two very similar nodes according to the Katz similarity and it would be low if that is not the case.

For example, suppose that we have two nodes in the network that are not directly connected to each other but are connected by many paths of length 2 and more. Those two nodes will have a high Katz similarity with a strongly weighted long paths. Because they are very similar we would expect that they are connected to each other where in fact they are not. If we have many such pairs in a network than the network would be non-navigable with the local information only – we would need the global information to navigate this network efficiently. For such network the correlation between the Katz similarity matrix and the adjacency matrix would be low.

Let us now formalize the notion of the correlation between the global and the fractional knowledge. Let $\alpha_{max} = 0.99/\lambda_1$ be the α parameter for calculating the Katz similarity matrix with the global knowledge. Let us call this global knowledge matrix $\boldsymbol{\sigma}^g$. With $\boldsymbol{\sigma}$ being a similarity matrix for an arbitrary α from the permited interval we define the Pearson correlation coefficient ρ as:

$$\rho_{\boldsymbol{\sigma}^g, \boldsymbol{\sigma}} = \frac{cov(\boldsymbol{\sigma}^g, \boldsymbol{\sigma})}{\sqrt{var(\boldsymbol{\sigma}^g) var(\boldsymbol{\sigma})}}, \quad (7)$$

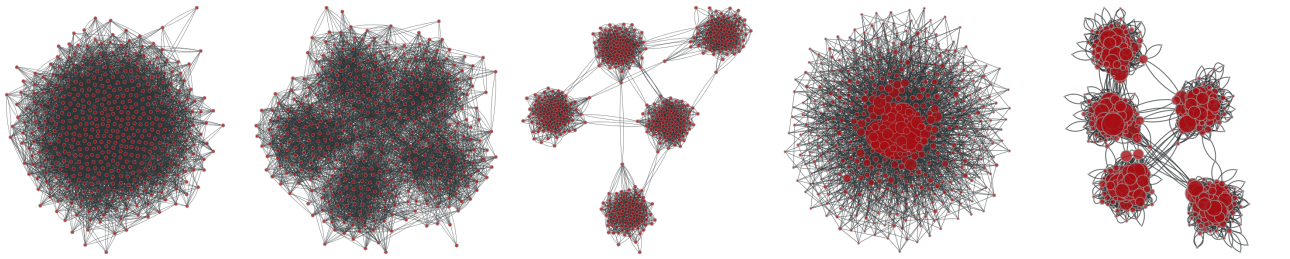
where we first vectorize the similarity matrices to calculate covariance and variances.

After we calculate the correlation between the global knowledge and adjacency matrix we may calculate further correlations by adjusting the parameter α. Thus, we start with a small α and slowly increase it until we reach the α for the global knowledge. The correlation will slowly start to increase until we reach 1. By comparing the current value of the parameter α with the maximal one we can deduce the fraction of knowledge that we need to have to reach a certain level of correlation with the global knowledge.

C. Illustration on synthetic networks

We now illustrate this correlation coefficient on synthetic networks. The goal is to generate networks that structurally closely reflect a typical social or an information network. Although the real networks exhibit a wide range of interesting structural properties we concentrate on two which have been observed in the majority of networks from different domains: (i) degree heterogeneity (where we typically have only a few of high degree and mid degree nodes that keep the network connected), (ii) and modular or community structure (with various groups of nodes that are tightly connected with each other and only loosely connected to other communities – typically via high and mid degree nodes). Such networks reflect also typical search scenarios such as search for a similar node within the same community and search for a distant target node in another community. We generate the following networks:

(i) random networks using the standard Erdős–Rényi [15] or Bernoulli graph model as a random baseline,



(a) Random Network (b) Weak Communities (c) Strong Communities (d) Power-Law Network (e) Power-Law Net. with Comm.

Fig. 1. **Synthetic networks.** The networks have 500 nodes and 2500 links. The structural differences between networks are clearly visible. In the visualizations of the power-law network as well as the power-law network with the communities the node size corresponds to its degree clearly depicting the heterogeneity of the node degrees in these networks.

(ii) networks with weak communities using Stochastic Block Models [16],
(iii) networks with strong communities again using Stochastic Block Models,
(iv) power-law networks using Barabási–Albert model [17],
(v) networks with strong communities with degree-corrected Stochastic Block Models [18].

We generate 6 groups of networks: the first three groups with 500 nodes and 2500, 5000, 10000 links respectively and the second three groups with 1000 nodes and 5000, 10000, and 20000 links. Networks with communities have always 5 communities of equal sizes. Additionally, we repeat generation 10 times and report always average values with the standard error of the mean in our diagrams. Figure 1 visualizes examples of the generated networks.

In Figure 2 we report the results for the networks with 1000 nodes and 10000 links. We observe similar results in all other networks. The results may be summarized as follows. The random networks exhibit consistently the lowest correlation between global and local or between global and fractional knowledge. This corresponds to our intuition about the random networks. Since the links are created at random there are no structural clues in these links that could steer our local decisions during the navigation. The correlation between global and local as well as fractional knowledge in networks with weak communities is slightly higher than for the random networks. The weak communities represent the structural clues that we can utilize when navigating locally, e.g. within a community we will have more links to other similar nodes from the community. The situation improves even more if the communities become stronger – the correlation is higher in networks with strong communities. A similar phenomenon can be observed in power-law networks. The existence of highly connected hubs provides structural clues that correspond to the links in the network, i.e. a hub is similar to many other nodes in the network and a hub has indeed many links to other nodes. Finally, the highest correlation can be observed in networks with degree-corrected communities. Here we have structural clues related to the node degrees as well as to the similarity of nodes within a community. For example, the correlation between the adjacency matrix and the global knowledge is already around 0.6 and improves steadily to e.g. correlation of almost 0.8 with only 50% of the global knowledge.

IV. Experiments and Results

In our experiments we are now interested in the following question. Suppose that we have available resources to learn global information about a given fraction of nodes in a network, e.g. 1%, 5%, or even 10% of nodes. The question is then which nodes are "good" or "optimal" nodes to know about. Having defined the correlation between the global and fractional knowledge ρ in the previous section we may use this quantity as a reasonable global criteria for selecting informative nodes to learn about. We can now think about this node selection procedure as an optimization problem. Thus, we want to know about global information for a given number or fraction of nodes such that the correlation between the global and the fractional knowledge constituted with the selected nodes is maximized.

Let us now formalize the optimization problem in the following way. We first introduce a division vector $\boldsymbol{s}$ as:

$$s_i = \begin{cases} 1 \text{ if we have global information about node } i \\ 0 \text{ otherwise.} \end{cases} \tag{8}$$

With k we denote the number of nodes about which we have global information, i.e. it is the number of 1s in $\boldsymbol{s}$.

Further, we define $\boldsymbol{\sigma^s}$ as the fractional knowledge matrix. This matrix is constructed in the following way. We start with the adjacency matrix $\boldsymbol{A}$ and replace those rows and columns in the adjacency matrix with the corresponding rows and columns from the global Katz similarity matrix $\boldsymbol{\sigma}^g$ for those nodes where we have global information. Formally, we have:

$$\boldsymbol{\sigma}^{\boldsymbol{s}}_{ij} = \begin{cases} A_{ij} \text{ if } s_i, s_j = 0 \\ \boldsymbol{\sigma}^g_{ij} \text{ otherwise.} \end{cases} \tag{9}$$

Finally, we may formalize the optimization problem as follows:

$$\boldsymbol{s}_{max} = \underset{\boldsymbol{s} \in D(k)}{\operatorname{argmax}} \rho_{\boldsymbol{\sigma}^g, \boldsymbol{\sigma}^{\boldsymbol{s}}}, \tag{10}$$

where $D(k)$ is the set of all divisions with k 1s in $\boldsymbol{s}$. Now, D has the cardinality of $\binom{n}{k}$, where n is the total number of nodes in a network. With increasing k the number of possible divisions grows rapidly and reaches the exponential limit with

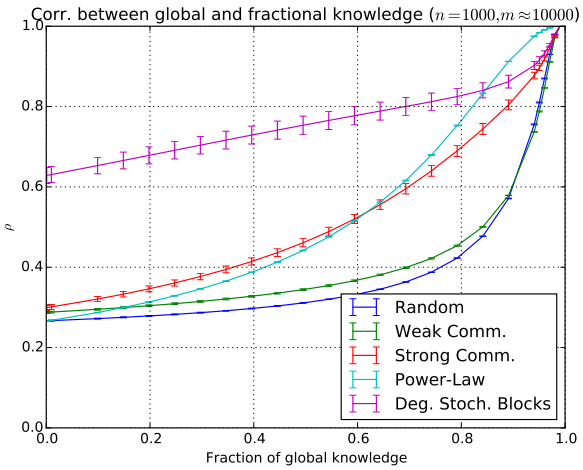


Fig. 2. **Correlation between global and fractional knowledge.** The diagram depicts the correlation between global knowledge and increasing fractional knowledge. The correlation is low in networks where the structural clues in links are weak such as random networks or networks with weak and blurred community structure. In networks with stronger clues such as power-law networks or strong community structures the correlation increases quickly with increased amount of knowledge. The highest correlation can be observed in networks which are rich in structure in that they exhibit both: heterogeneity in node degrees as well as a rich and strong community structure.

e.g. $k \approx n/2$. Thus, to solve the optimization problem we will need to resort to a heuristic optimization strategy.

For our experiments we adopt a stochastic approach, in particular a well-known Markov Chain Monte Carlo optimization algorithm called Simulated Annealing [19]. In this algorithm we start be exploring the configuration space by randomly moving around that space. In our case a random move corresponds to randomly selecting a given number of nodes (e.g. 1 to k) and setting their corresponding s_i to 1. To satisfy the condition that the total number of nodes that we know about is always k we then randomly select the exact same number of known nodes and unlearn them by setting their s_i to 0. We then evaluate the Pearson correlation coefficient (ρ_{new}) with the new matrix $\boldsymbol{\sigma}^{\boldsymbol{s}}$ and accept the move if the correlation (ρ) has increased from the previous iteration step. Otherwise if the correlation is smaller than previously we can still accept it with some small probability. In our experiments we use the standard Metropolis-Hastings acceptance scheme and accept with the probability:

$$p_{accept} = min(1, e^{\frac{-\Delta E}{T}}), \tag{11}$$

where:

$$\Delta E = \rho - \rho_{new}, \tag{12}$$

and T is the control parameter or temperature. By slowly adapting the control parameter, i.e. by cooling the system we become more restrictive and accept less and less random moves concentrating in this way around an optimal solution. In general, this optimum is only a local maximum, however by a proper configuration and multiple starts we almost always find the global maximum in practice. The only drawback of the method is that it is typically very slow as it requires a large number of iteration steps and a large number of different temperature levels.

We generate 10 networks with strong communities with degree-corrected Stochastic Block Models with 500 nodes and approximately 2500 links as described previously. By Simulated Annealing we then select optimal $\boldsymbol{s}$ with 5 nodes for which we have the global knowledge. The progress and the results of the optimization algorithm, as well as the comparison with the $\boldsymbol{s}$ where 5 nodes are selected according to a specific centrality measure are shown in the Figure 3. We show examples of two networks with all other networks having similar results.

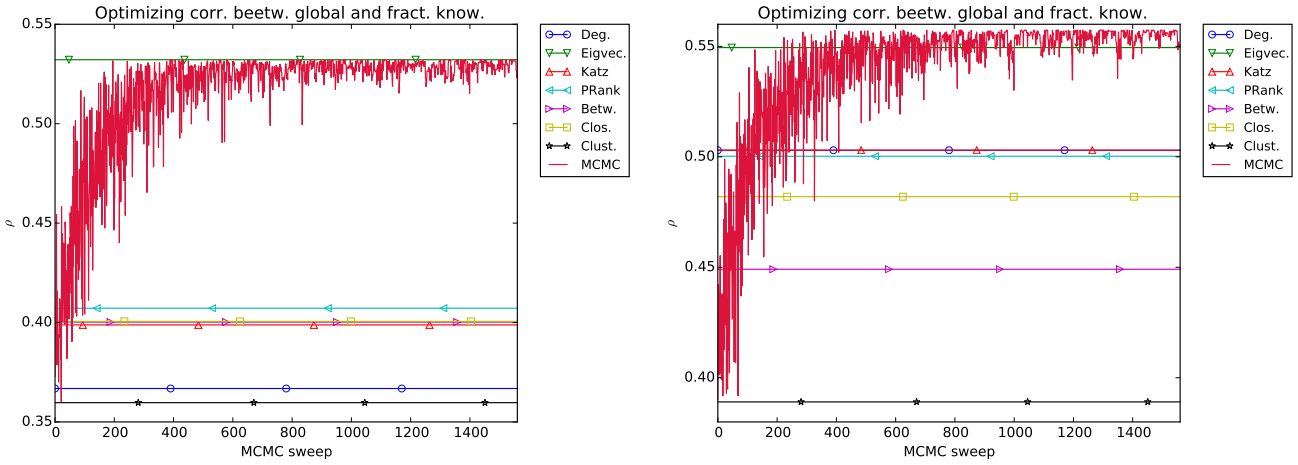
The results may be summarized as follows. With Simulated Annealing we are able to find an optimal solution that increases the Pearson correlation coefficient significantly. For example, the correlation between the global and local knowledge is typically in the interval [0.3, 0.35]. By optimizing for only 5 globally known nodes we are able to raise the correlation coefficient to range [0.55, 0.65], which seems as a great improvement over a local knowledge. Thus, we may assume that by knowing only 1% of "important" nodes we may already have a 60% fraction of the global knowledge. This assumption would need to be confirmed experimentally, e.g. by simulating navigation with the optimal fractional knowledge.

Another interesting experimental result is the comparison of the optimal results with 1% of nodes selected by a given selection criteria. For selection criteria we resort to a wide range of well-known and established centrality measures including degree, eigenvector, Katz, PageRank, betweenness, and closeness centralities, as well as clustering coefficient. Our results suggest that only the eigenvector centrality used as a criteria for node selection can match the optimal solutions obtained by our MCMC method. Thus, selecting 5 top nodes according to their eigenvector centrality will either provide the optimal solution or come very close to the optimal solution (cf. Figure 3). None of the other centrality measures can provide consistently good correlations with the global knowledge. On the other hand, clustering coefficient consistently produces the worse results and should be avoided as a selection criteria.

These results can be explained in the following way. Eigenvector centrality of a given node is high either if this node has a lot of neighbors or if that node is connected to other nodes with high eigenvector centrality. In other words, a node has many links that can steer navigation in various directions and therefore this node is important for navigation, or a node has a few links but all of them to other "important" nodes that may have many links themselves. Thus, either a node can lead a search agent into a proper direction or it can lead an agent to a neighbor that herself has an appropriate link.

V. Conclusions

In this paper we: (i) defined a global criteria (Pearson correlation coefficient between the Katz similarity matrix with the global knowledge and a Katz similarity matrix with a



(a) Random Network

(b) Weak Communities

Fig. 3. **Optimizing fractional knowledge.** A typical correlations obtained by Simulated Annealing. With only 5 optimal nodes globally known (that is with 1% od nodes globally known) we are able to reach correlation values above 0.55 in those two networks. We observe similar or even better results in other simulated networks. We compare the correlation obtained by Simulated Annealing with correlations with 5 globally known nodes which are 5 top nodes according to a given centrality measure. Thereby, the eigenvector centrality perform much better than other centrality measures, i.e. by selecting 5 top nodes corresponding to their eigenvector centrality we are in almost all cases at the optimal (left) or nearly optimal (right) solution. Other centrality measures can not be categorized consistently with the clustering coefficient as being always the worse criteria for selecting globally known nodes.

fractional knowledge) for assessing the quality of fractional knowledge for navigation, (ii) applied MCMC techniques (Simulated Annealing) to select nodes that are important for navigation, and (iii) compared the important nodes according to various centrality measures with the nodes selected by the optimization algorithm. The eigenvector centrality is consistently the best indicator for the quality of nodes for the navigation task.

The main limitation of our work is the theoretical nature of the analysis. Specifically, we performed all of our experiments on synthetic networks. The next step would be to conduct experiments with empirical networks and compare the results with the synthetic networks. However, we leave these experiments for the future work.

ACKNOWLEDGMENT

This research was in part funded by the FWF Austrian Science Fund research project "Navigability of Decentralized Information Networks" (P 24866-N15).

REFERENCES

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the seventh international conference on World Wide Web 7*, ser. WWW7. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998, pp. 107–117.

[2] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, pp. 60–67, 1967.

[3] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, August 2000.

[4] J. Kleinberg, "The small-world phenomenon: an algorithm perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, ser. STOC '00. New York, NY, USA: ACM, 2000, pp. 163–170.

[5] ——, "Small-world phenomena and the dynamics of information," in *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA, USA: MIT Press, 2001, pp. 431–438.

[6] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," *Science*, vol. 296, pp. 1302–1305, 2002.

[7] L. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, no. 3, pp. 187 – 203, 2005.

[8] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Physical Review E*, vol. 64, no. 4, pp. 046 135 1–8, Sep 2001.

[9] R. West and J. Leskovec, "Human Wayfinding in Information Networks," in *Proceedings of the 21th international conference on World Wide Web - WWW '12*. New York, New York, USA: ACM Press, 2012.

[10] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the twelfth international conference on Information and knowledge management*, ser. CIKM '03. New York, NY, USA: ACM, 2003, pp. 556–559.

[11] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Phys. Rev. E*, vol. 73, p. 026120, Feb 2006.

[12] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. V. Dooren, "A measure of similarity between graph vertices: Applications to synonym extraction and web searching," *SIAM Rev.*, vol. 46, no. 4, pp. 647–666, Apr. 2004.

[13] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 538–543.

[14] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[15] P. Erdos and A. Renyi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, pp. 17–61, 1960.

[16] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: first steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.

[17] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. [Online]. Available: http://www.sciencemag.org/content/286/5439/509.abstract

[18] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Phys. Rev. E*, vol. 83, p. 016107, Jan 2011.

[19] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

Using latent features to measure the diversity of recommendation lists

M. Kunaver*, Š. Dobravec* and A. Košir*
* University of Ljubljana/ Faculty of Electrical Engineering, Ljubljana, Slovenia
matevz.kunaver@fe.uni-lj.si

Abstract - One of the well-known issues with content recommender system is that they tend to become over-specialized, which often has a negative influence on user experience. This can be solved by diversification of the recommendation list, a process that implements a tradeoff between accuracy and diversity of recommended items. Normally, item metadata is used in the diversity measure. In certain cases however, the item metadata may not be available thus a different approach to measure diversity is required. The aim of this preliminary study is to determine whether latent features can be used to measure the diversity of recommended items. In order to resolve this we generated recommendation lists for 43 different users using the LDOS-CoMoDa dataset. We then evaluated the diversity of these lists using the standard intra-list diversity measure. In addition we calculated the diversity of each list by comparing the latent features (calculated using the matrix factorization approach) of each item on the list. The comparison of both value sets showed that they show similar characteristics which implies that latent feature space offers an alternative method of evaluating item diversity when no metadata is present.

I. INTRODUCTION

The rapid development of recommender systems (RSs) has led to a change of focus from simply generating recommendations for the user to generating recommendations that consider the user's context [1][2] and personality in order to improve the recommendation results[3]. In addition most of the existing recommender systems are encountering the problem of overspecialization which can have a negative impact on the user experience.

This led to the search for improvements that would help RSs present the user with a selection (a subset) of results that would be the most appropriate for the situation in which the user desires to review the selected results.

One of such improvements is the diversification [4][5][6][7] of recommendation lists that specifically aims to correct the overspecialization problem.

A. Motivation and goal

One of the problems with the diversification of recommendations lists is the measure used to determine the diversity of these lists. One of the more frequently encountered diversity measures is the intra-list diversity (ILD) [7] that was for instance used as one of the comparison parameters in the ESWC diversity challenge [8].

After participating in the challenge and a further development of our diversification approach [9] we focused on the evaluation measure itself. One of the problems of ILD is that it relies on having access to the meta-data (categories and attributes) description of each recommender item. This also implies that the accuracy of this method depends on the level and complexity of meta-data available – the more data available the more accurate the measure. Last but not least this also opens the question of using this measure in collaborative systems that i.) do not need metadata for the recommendation procedure and ii.) can contain items of different types (movies and music for example) which can be described with different metadata ontologies.

The aim of this study is therefore to determine whether we can use alternative item features such as Matrix Factorization (MF) latent feature vectors [10] to evaluate the diversity of recommended item lists. For this purpose we revised our previous research [9] and re-evaluated our results using ILD and our proposed latent feature based measure.

II. MATERIALS AND METHODS

In this section we describe the dataset, the baseline RS used to generate recommendations, the diversification methods and the compared evaluation methods.

A. Dataset

For the purposes of our research we used the Context Movie Dataset (LDOS-CoMoDa) [11], that we have acquired in our previous work. The dataset was collected using an on-line application for rating movies (www.ldos.si/recommender.html) that enabled the users to track the movies they have watched and to obtain recommendations from several RS algorithms. In addition, the application features a questionnaire whose purpose is the collection of the contextual data describing the situation during the item consumption.

The dataset currently consists of 4237 ratings given by 184 users to 1782 items. Each rating is also annotated with associated contextual variables. Each user is described with basic demographic data (age, sex, location) provided on a voluntary basis. Each item is described with several attributes: genre, director, actor, language, country, budget and release year.

The on-line application is still available and in use. Additional information about LDOS-CoMoDa can be found in and.

B. Recommender systems

For this paper we used two different RSs: a hybrid RS to generate the diversified recommendation lists and a Matrix Factorization RS to generate the latent feature vectors for each item in the dataset.

Hybrid RS [12]: The hybrid RS used for this experiment was developed as part of our previous research [12]. It is a collaborative RS that selects nearest neighbors based on genre preferences instead of their ratings. Each preference indicates the user's interest for one specific genre (25 in total). By using these preferences we are able to select nearest neighbors who perfectly match the active user in preferences without having a single overlapping item (i.e. item rated by both users). This increases the recommendation pool and the overall quality of the RS.

The hybrid RS generates recommendations for each user by performing the following steps: (i) Calculate genre preferences for the user based on his/her previous ratings, (ii) Find 20 users whose preferences are the most similar to the active user, (iii) Create a pool of potential recommendations from all of the items rated by these users, (iv) Calculate the predicted rating for each item using the Bayesian estimator, (v) Present the user with the top 20 items.

Matrix Factorization RS [10]: The MF RS was used in this experiment was based on the algorithm that was developed and deployed in [10]. For this experiment the algorithm was used solely for the purpose of calculating latent features vectors for each existing item in the dataset. We plan to use the diversification methods described below on the MF RS as well, but decided to leave this for future work as the recalculation of latent features proved to be too time consuming.

The MF RS calculated 7 latent features for each existing items and stored them in a vector. A set of all item vectors was then used as input for our diversification measure (see below).

C. Diversification methods

The three diversification methods used in this research can be implemented in any existing RS without requiring a direct change in the way that RS work. The methods focus on diversifying the top 20 lists generated by the RS. In our case the diversification process follows the next steps for every user's list of recommended items: (i) prepare ordered (descending) list of recommendations and split it into top 20 recommendations list and the remainder of the set, (ii) find exchange candidates in such manner that the diversity of the top 20 items increases without significant harm to the accuracy of the system, and (iii) exchange the selected items to yield a diversified list of recommended items. The second and the third step can be performed iteratively.

As mentioned above, we used three diversification methods that mainly differ in the way how the exchange candidates are picked.

The first method swaps up to three items in a single step (no iteration). It starts by assessing the worst items in the top 20 list. It calculates the ILD value of the list while excluding one item (exchange candidate) of the list at a time. Effectively this means calculating ILD@19. Higher values of ILD@19 indicate better exchange candidates. Next, it searches for the best replacement candidates from the first 20 items of the remainder of the set. The method calculates the ILD@20 after exchanging every combination of up to three items. Final result is the top 20 list with best ILD@20 score after the exchange.

The second method uses the same approach as the first one to determine exchange candidates in the top 20 list. The best item, which yields highest ILD@19 is then replaced with an item from the first 20 of the remainder of the set. The final exchange is done using the replacement candidates that gives best ILD@20 score. In this case we shuffle only a single item at a time, but repeat the process K-times. It can be expected that increasing value of K would favor list diversity in trade-off to lowering list accuracy.

The third method, just like the second one, replaces single item at a time. The difference is, it considers a joint score in form of (1) instead of a pure ILD value.

$$a \cdot avgPR + b \cdot nILD \qquad (1)$$

$avgPR$ stands for the average prediction rating of the list and $nILD$ for the normalized ILD value of the same list. The $nILD$ is calculated using the ILD@20 measure, described in the next chapter. The sum of both parameters is 1. A combination of $a = 1, b = 0$ therefore means that the system creates non-diversified recommendation lists, while $a = 0$, $b = 1$ creates as diverse recommendation lists as possible with total disregard to the final accuracy of recommendations. The shuffling procedure is repeated until best top 20 list (in terms of joint score) is achieved.

D. Diversity measures

We used two measures to determine the diversity of the generated recommendation lists.

The first was the intra-list diversity [7] (ILD@20) which calculates the diversity value of each top 20 list based on the metadata descriptions of each item on the list, that consist of: genre, director, actor, language and country as shown in equation 2.

$$ILD@20 = \frac{1}{2}\sum_{n=1}^{20}\sum_{k=1}^{20}\left(1 - \frac{match(n,k)}{\sqrt{size(n) * size(k)}}\right) \qquad (2)$$

$match(n,k)$ represents the number of metadata description in which items n and k match. $size(n)$ is the number of medatada used in the description of item n.

The second was our proposed measure "latent feature diversity" (LFD@20) which was calculated using the following equation:

$$LFD@20 = \frac{1}{2}\sum_{n=1}^{20}\sum_{k=1}^{20} ldiv(n,k) \qquad (3)$$

Each item is represented with a feature vector composed of 7 latent features [2]. The diversity ($ldiv(n,k)$ of items n and k is therefore calculated as the Euclidian distance between the two latent feature vectors. The greater the distance the more diverse the two compared items.

E. Evaluation

The evaluation of our proposed diversity measure was performed for two experiments.

In the first experiment we focused on the first diversification method since it does not include any additional parameters (number of iterations, weights etc.). We first collected top 20 lists for each user – the original list and the diversified list. We then calculated the diversity 'before' and 'after' using both diversity measures. The aim of this stage was to determine if both measures show similar behavior (for instance if ILD@20 shows an increase in the list diversity does LFD@20 do the same?).

In the second experiment we focused on the second and third diversification method. The aim was to diversify the top 20 list as much as possible while retaining high recommendation accuracy. To achieve this we measured the accuracy of the system (using the F-measure [13]) at different parameter values (number of iterations for the second diversification method and weight values for the third). We then calculated the accuracy, ILD@20 and LFD@20 for each parameter value. The aim of this stage was to compare the trend of each diversity measure in order to determine whether we could achieve the same results using the proposed diversity measure (LFD@20) as when using the measure that was successfully used in our previous experiments (ILD@20).

III. RESULTS

A. First experiment

In the first experiment we created and diversified the top 20 recommendation lists of 44 different LDOS-CoMoDa users. We then calculated the diversity of each diversified and non-diversified list using both ILD@20 and LFD@20 measures.

Since we wanted to compare the behavior of each measure we then calculated the relative change of each measure during the diversification process as shown in table 1.

TABLE I. TYPE SIZE FOR CAMERA-READY PAPERS

	ILD@20	LFD@20
avg. change	5.6%	16.2%
std. dev.	3.6%	27.8%
min. change	1.3%	-21.8%
max. change	14.6%	99.6%

The values show that both measures show a change however the ILD@20 always shows an increase in the list diversity (which is expected since ILD@20 was the measure used to diversify the list) while LFD@20 sometimes actually shows a decrease in the diversity of the items contained in the 'diversified' recommendation list. The negative value of minimum change LFD@20 shows that there were cases where the ILD@20 diversity

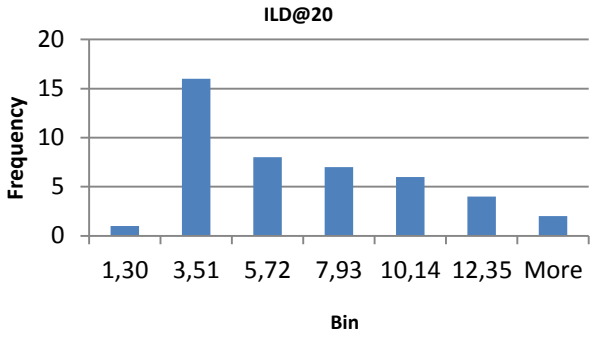


Figure 1. ILD@20 value change in the first experiment.

of the list increased while the LFD@20 value decreased. This was possible as our diversification methods used ILD@20 in the diversification procedure.

Another detail that can be observed from the presented

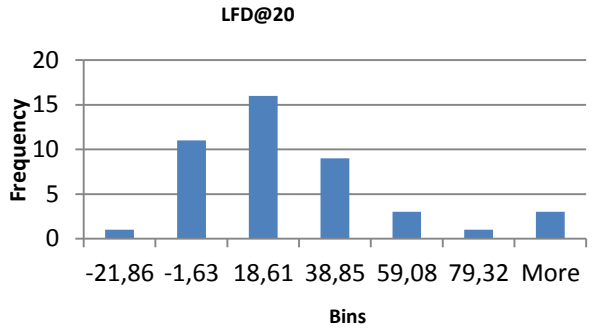


Figure 2. LFD@20 value change in the first experiment.

numbers is the fact that LFD@20 seems to be a lot more dynamic as it's changes range from -21% to as much as 99,6% while the greatest change in ILD@20 value only amounts to 27.8%.

In order to further observe this we performed a data analysis of each set and created a histogram where all 44 values are divided into 7 bins. The results are shown on Fig, 1 and 2. The bin value represents the relative change of the list diversity while the frequency value shows how many of such changes we detected during our experiment.

A comparison of Fig. 1 and Fig. 2 shows that LFD@20 values change in a broader spectrum (from -1.6% to 38.8%) while most of ILD@20 gather around 3.5%.

As we lack the information about how actual users perceive the diversity of these diversified top 20 lists (see further work) we cannot say which of these two measures performs better or is a better reflection of the actual list diversity.

B. Second experiment

In the second experiment we used the recommendation list of 44 different LDOS-CoMoDa users obtained from hybrid RS. We then diversified the list of top 20 items using second and third diversification method (as described in materials and methods section) at different diversification level. At each step we calculated both diversity values, the ILD@20 and the LFD@20. It is important to note that in the diversification procedure, the

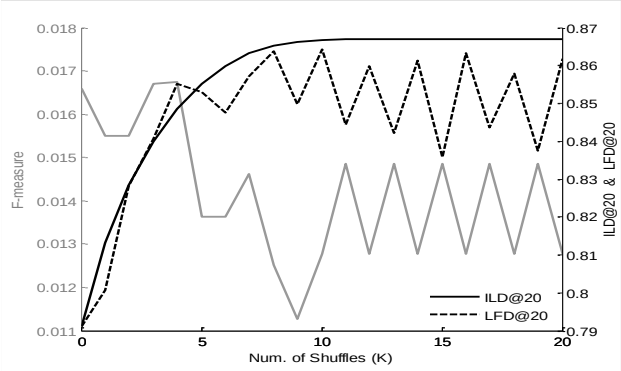


Figure 3. ILD@20, LFD@20 and F-measure values when using the second diversification method.

ILD measure and not the LFD measure was used to find candidates for shuffling step.

Fig. 3 shows the results when using second diversification method (iteratively shuffling single item). It can be seen that as we increase the number of shuffled items, the list accuracy drops, while the diversity, measured as either ILD@20 or LFD@20, increases. Note however, that for the purpose of graph plotting, the LFD values are scaled to the range of ILD values (using ILD@20 at k=0 as the scaling factor). We justify this scaling by our interest in shape of the plot rather than in actual LFD values.

Similarly, the results when using third diversification method (evaluating joint rating/diversity score) are presented on Fig. 4. Again, the LFD values are scaled to the range of ILD values to observe the shape of the plot.

In both cases, we can observe a similarity in the ILD and LFD curves, which in worst case differ by less than 5%. This does not mean that the diversity measures are automatically interchangeable, but they do show similar tendencies.

IV. CONCLUSION AND FURTHER WORK

The aim of this paper was to evaluate whether a latent feature based diversity measure could be used to evaluate the diversity of lists of recommended items. Such a measure would have the advantage of being insensitive to the content type used in the system as well as to the type and complexity of the metadata ontology used to describe the content items stored in the system.

Our results show that LFD@20 shows a different value dynamic when used to evaluate the diversity of each individual user's recommendation list (see experiment one, described above) as compared to ILD@20 which seems to be a more 'conservative' measure showing only

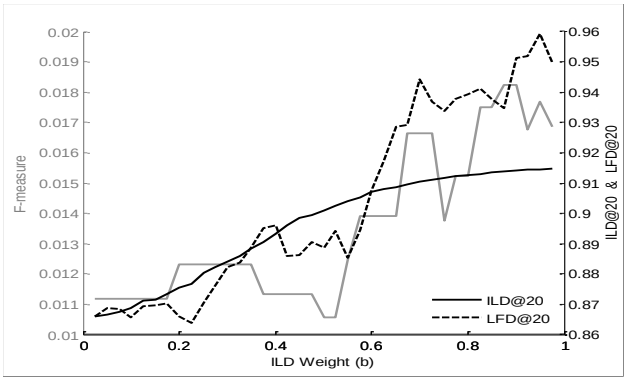


Figure 4. ILD@20, LFD@20 and F-measure values when using the third diversification method.

small relative changes during the diversification process while LFD@20 claims a much larger relative change in the diversity of the same list.

The results of the second experiment on the other hand show that the two measures could be more alike since they both show similar behavior when used to select the best diversification parameter values.

Regardless of this we need to perform additional experiments and evaluations if we wish to confirm that LFD@20 is a viable diversity measure. In order to achieve this we plan to perform a series of experiments:

- The most important experiment that needs to be performed involves a live user trial which would determine which of the evaluated measures best reflects the actual user preferences and would also establish a 'ground truth' about how the users actually perceive recommendation list diversity.
- We should also explore the latent features themselves – in the experiment above we used 7 latent features since this was the number used in our previous experiments. We need to perform a statistical analysis in order to determine whether this number should be adjusted.
- Experiment with different latent diversity measures – for this paper we only used the Euclidian distance. We should also consider other methods such as correlation, p-norm etc.
- Last but not least perform a series of A/B test to determine the user satisfaction with the diversification method.

REFERENCES

[1] A. Dey and G. Abowd. Towards a better understanding of context and context-awareness. pages 304–307, 199.

[2] A. Odic, M. Tkalcic and A. Košir. Managing irrelevant contextual categories in a movie recommender system. Human Decision Making in Recommender Systems (Decisions@ RecSys 13), page 29, 2013.

[3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible

extensions. IEEE Transactions on Knowledge and Data Engineering,17(6):734–749,2005.

[4] B. Smyth and P. McClave. Similarity vs. diversity. In Case-Based Reasoning Research and Development, pages 347–361. Springer, 2001.

[5] D. Jannach, L. Lerche, F. Gedikli and G. Bonnin. What recommenders recommend–an analysis of accuracy, popularity, and sales diversity effects. In User Modeling, Adaptation, and Personalization, pages 25–37. Springer, 2013.

[6] G. Adomavicius and Y. Kwon. Improving aggregate recommendation diversity using ranking-based techniques. Knowledge and Data Engineering, IEEE Transactions on, 24(5):896–911, 2012.

[7] C. Ziegler, S. M. McNee, J. A. Konstan and G. Lausen. Improving recommendation lists through topic diversification. In Proceedings of the 14th international conference on World Wide Web, pages 22–32. ACM, 2005.

[8] M. Kunaver, T. Požrl, Š. Dobravec, U. Droftina, and A. Košir, "Increasing Top-20 Diversity Through Recommendation Post-processing," in SemWebEval 2014, CCIS 475, V. Presutti et al., Eds. Springer International Publishing Switzerland, 2014, pp. 188-192.

[9] M. Kunaver, Š. Dobravec, T. Požrl, and A. Košir, "Increasing Top-20 search results diversity through recommendation post-processing," UMAP 2014 extended proceedings: posters, demos, late-breaking results and workshop proceedings of the 22nd Conference on User Modeling, Adaptation and Personalizaton, I. Cantador, Eds. Alborg, Denmark, pp. 1-6, July 2014.

[10] A. Odič, M. Tkalčič, Jurij F. Tasič, and A. Košir, "Predicting and detecting the relevant contextual information in a movie-recommender system", Interact. Comput. Oxford University Press, vol. 25, no. 1, pp. 74-90, 2013.

[11] A. Košir, A. Odic, M. Kunaver, M. Tkalcic and J. F. Tasic. Database for contextual personalization. Elektrotehniški vestnik, 78(5):270–274, 2011.

[12] M. Kunaver, A. Košir and J. F. Tasič. Hybrid recommender for multimedia item recommendation. Lambert Academic Publishing, 2011.

[13] F. Ricci, L. Rokach, B. Shapira and P.B. Kantor. Recommender Systems Handbook. Springer, 2011.

Efficient Profiling and Distributed Synchronization of Statistical User Models for Content-Targeting and Real-time Applications

Lj. Mercep
Cognizant Technology Solutions, Munich, Germany
Ljubo.mercep@cognizant.de

Abstract – This work presents an efficient approach for managing distributed user profiles in the use-case of the semi-autonomous or fully autonomous connected car. A user profile is a collection about user's personal preferences and statistical data on user's previous interactions with all compatible connected systems outside and inside the car. Managing the user profile practically means dynamically updating the profile based on the interaction with the vehicle's human-machine interfaces and using it to answer complex queries placed by the vehicle's applications. The queries can range between user interface personalization queries (what is the user's favorite radio station), over product-placement related queries (are we going to pass next to a restaurant the user might be interested in) to safety-relevant queries (is the senior driver paying adequate attention to the direction of a detected obstacle). The focus is placed on reducing the computational complexity of user profiling and synchronization of distributed user profiles.

I. INTRODUCTION

Semi-autonomous or autonomous connected cars have the potential to reshape the automotive human-machine interfaces and several business concepts deeply rooted inside the automobile industry. As the automobiles become more autonomous in the upcoming decades, the driver's workspace of a connected car is assumed to change in its purpose and the driver will be able to use the human-machine interface for other non-driving related tasks. The value added through personalized content delivery in the new driver's workspace can be used to offset the increased design and construction costs of the future semi-autonomous connected automobiles and the surrounding intelligent traffic systems. The basis of such business model is efficient management of a large number of user profiles, which can be used for driver workspace personalization and targeted content delivery [1]. User profile is defined in detail later on, but can be summarized as a collection of statistical data on user's preferences generated during hers/his interaction with other systems and a collection of certain known facts about the user.

This work presents a solution to the problem of keeping a large complete user profile in sync with all distributed applications which are using smaller models of the complete user profile to perform complex queries or extensive user profile analysis. The applications are running on remote embedded platforms, such as in-vehicle computing units. It is advantageous to cache the parts of full user profiles on a local embedded platform inside the vehicle, as opposed to performing all the operations remotely on a centralized user profile management platform. Caching is performed indirectly, based on observing the user profile queries of different applications running inside the vehicle. The applications present user-targeted content to the passenger(s) inside the vehicle based on the vehicle location i.e. its vicinity to various product and service profiles on the one side and passengers' user profiles on the other. There are several advantages to distributed handling of user profiles and their local caching in the vehicle - reduced computational complexity and a well-defined interface towards the users of the user profiles and the central user management platform. Once an application inside the connected car requests and retrieves relevant parts of the user profile, it can internally perform custom queries, which might contain business-sensitive data.

The user profiles are generated during everyday interaction with connected handheld devices and other human-machine interfaces and contain static user data (such as gender) as well as learned/observed statistical data about the user's preferences. Bayesian network representation has been chosen as the underlying model of the customer profiles, after performing the analysis in the preceding work of the author on this topic [3]. *Using* the user profile model practically means:

- Updating the statistical user data in the model
- Extending the model structurally by adding new nodes and new dependencies between the new or existing nodes
- Placing questions or queries on the user profile model – this is described as user *profiling*

The first step of the approach is based on an existing improvement of the efficiency of exact querying or exact inference on Bayesian models, which is performed with well-known secondary structures known as junction trees [2]. The user profiles which are stored in this form can be distributed over multiple spatially separated systems and can be queried with different quality-of-service requirements. A query with quality-of-service requirement

The opinions and the evaluations of research topics and business concepts expressed in this work are those of the contributor alone and do not necessarily reflect the views of Cognizant Technology Solutions.

is a pair of query Q and a requirement R where Q can carry the value of "is the driver tired?" and R can carry a set of restraints on the inference procedure of Q. R can, for example, contain the restraint that the query frequency has to be 1 Hz and that the maximal allowed latency in answering the query is 100ms. The impact of quality-of-service requirement, the important quality-of-service dimensions and a further optimization for exact inference with quality-of-service requirements are given in previous work [3].

A proposed metric for evaluating the efficiency of the approach is composed out of two components:

- The computational complexity necessary for the application using the user profile to answer its query and start acting towards the user
- The computational complexity of synchronizing the distributed statistical user models in order to minimize the drift between different copies of the same user profile

This work is organized as follows. In Section II the theoretical background of the approach is present. In Section III the approach is evaluated against the known alternatives. Finally, the work is concluded and the future steps are discussed in Section IV.

II. Proposed Approach to User Profile Management

The first logical step when designing a system for user profile management based on Bayesian representation is designing the underlying Bayesian networks. This work does not evaluate specific Bayesian networks and their effectiveness in practice, apart from using the INSURANCE network to measure computational complexity. Table I. presents the mapping between the user-data-related concepts used in this work and the proposed implementation methods, which are all explained in the following.

The first suggested optimization principle is storing and exchanging the user profile data in the form which is optimized for querying, that means for its practical usage, and not in the more explicit form used during the learning process or during the manual knowledge base design. Secondary knowledge structures such as junction trees, which can be constructed using statistical analysis of queries (through observation) and using the explicit metadata attached to the query (such as query priority or maximal latency allowed for answering the query in the case of strict real-time inference), can drastically reduce the computational complexity of inference. In some cases, meeting the real-time requirements on knowledge updating and parallel inference is not feasible at all without constructing the secondary knowledge structures [7]. This is even more the case if exact inference is required, meaning that the entire knowledge base stored in or available to a particular agent has to be evaluated before answering a query, eliminating the possibility for heuristic approaches [8].

As shown in [3], large computational reduction of exact inference can be achieved after identifying the most commonly used variants of user profile models, meaning the optimal query-optimized junction trees. Identifying these co-called hot-spots also enables high efficiency of information transfer between the central user profile management platform and remote applications, since the remote applications receive a small-scale query-optimized model of the entire user profile. This eliminates the need for intensive remote querying and "internalizes" the machine-machine interface by replacing it with a local user profile model inside the borders of the remote application. To summarize, the distributed applications which need to perform queries on the user profile receive a reduced model of the user profile, optimized for performing exactly the requested queries.

TABLE I. Mapping between Abstract User-Data-Related Concepts Used in this Work and Proposed Implementation Methods

User-Data-Related Concept	*Implementation Method*
Full user profile	Bayesian network stored at the central user profile management platform (CUPMP)
Single property of the user or the minimal non-divisible unit of knowledge about the user	Single node in a Bayesian network
Querying a user profile	Performing exact inference on a set of nodes in Bayesian network
Query-optimized user profile or an optimized non-complete user profile model	Junction tree constructed upon the Bayesian network
Remote application which is querying the user profile	Any application capable of sending a query to CUPMP or performing exact inference
Syncing user profile data after user interaction with the remote application	Updating the Bayesian network in CUPMP and all affected junction trees with the new content of nodes
Syncing the local query-optimized user profile model with the fresh user profile data	Retreiving the content of the nodes inside the appropriate junction tree stored at CUPMP

The internalized user profile model (i.e. the query-tailored junction tree) represents the basic meaning or the utility of the entire user profile, since the optimized models are generated by many real-life queries placed on the complete user profile over a long period of time by real-life remote applications. Exchanging information by the means of directly usable query-optimized knowledge structures makes for a very efficient machine-machine interface, since no time is lost in data encoding or interpretation. Since the user profile information is used for querying very often and since the query profiles change very rarely, the bulk of computational effort is invested into optimizing information consumption and synchronization between all remote applications.

A remote application which uses a user profile has the primary purpose of placing queries on the user profile, reacting to the answers and updating the user profile with new information stemming from the user interaction.

The necessity to construct the secondary knowledge structures around the complete Bayesian network can be positively seen as chance to better understand the system's usage patterns, identify these patterns and properly distribute the computational complexity of querying over all user profile users. Fig. 1 shows the main elements of

the proposed architecture, based around query-optimized user profile models. In the figure, the junction trees are either being used remotely (inference engine used on CUPMP) or locally (inference engine used inside the remote application).

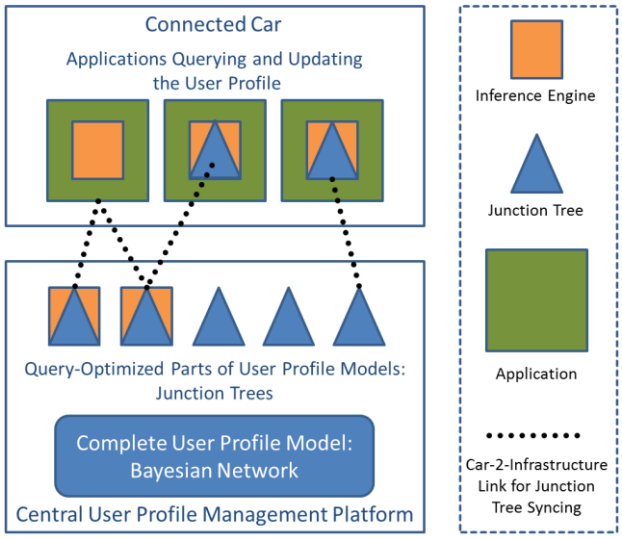


Figure 1. Proposed architecture for user profile management

This work builds upon the following levels of optimization, out of which the first one does not represent a new contribution:

- Queries from users generate query-optimized junction trees to reduce complexity of exact inference
- Similar queries can be answered by the same already generated junction trees, further reducing computational and sync effort on the CUPMP
- Information syncing between the junction trees is perform by write-through the underlying Bayesian network on CUPMP and updating all junction tree bottom-up

It is important to note that all the structural changes on the junction trees have to be controlled by the central user profile management platform (CUPMP), since such changes invalidate all already deployed junction trees for a specific query. A structural change in the underlying Bayesian network would cause invalidation of all distributed junction trees which build upon this network. On the other hand, content changes in junction trees (syncing data between junction trees) can be performed in a completely distributed manner.

Remote applications might be interested in very rapid profiling of the user which has started interacting with it, in order to present her/him with the most appropriate product offer or to customize the driver's workspace. An example is a driver whose vehicle is approaching a shopping center. The application executes a series of important queries, which answers have a high-value utility function. Using the query-optimized secondary knowledge structure is the fastest way to perform wide-range and intensive querying / profiling over the relevant data.

Following procedure for user profile management is suggested:

Applications $a_1,a_2 \ldots a_n$ contained in the set **A** request the profile for the user u_k from the set of users **U**, for the purpose of answering the queries $q_1,q_2 \ldots q_n$ from the set of possible queries **Q** (set contains all nodes inside the Bayesian model representation of user's profile) with the associated quality-of-service requirements $r_1,r_2 \ldots r_n$ from the set of possible requirements **R**.

In the first step, CUPMP checks for existing junction trees for this query profile. If there are none, the service checks how many times this query has already been placed. Based on a fixed threshold, which would be determined in practice, it performs inference directly on the user profile stored in Bayesian network form or it constructs a query-optimized junction tree. The applications receive either the answers to their queries or the generated junction tree.

Alternatively, the applications from set **A** can place a direct request for generating a junction tree for their query-QoS pair (**Q**,**R**). This is a clear statement to the CUPMP that this query-QoS pair will be used for more intensive profiling or more intensive and longer interaction with the user. This type of interaction also provides the application with the freedom to query or profile the user profile without exposing the content of the queries to the central management service. This might be important in the case of business-sensitive queries or performing a passive survey which is relevant to an upcoming, in-the-works, service or product.

After the remote applications have finished using their junction trees and/or after the interaction with user is finished, there are several possible ending steps:

1. The applications delete the junction trees and request a new one when the user interacts with the system next time.
2. The applications cache the junction tree for future use to avoid placing load on the CUPMP. The applications are aware that this will cause a drift between the cached user profile and the actual one, stored on the CUPMP, so they have to re-sync with CUPMP before the next interaction with the same user.
3. The applications wish to sync the possible changes in the junction tree, which happened during the interaction with the user, with the central user profile management service.

The last step, which provides valuable new user data to all applications using the system, has to satisfy at least the following requirements:

1. Applications have to be trusted in order to be allowed to update the central user profile data. Otherwise, one application which injects false user data could compromise all other remote applications.

2. All junction trees which are affected by the change of the synced junction tree should be immediately updated with the new knowledge

A balance between computational and communication complexity and content drift between affected junction trees has to be considered in the last step. In practice, the values of allowed margins for the drift between individual variables in the junction tree representation of the user profile have to be defined based on their impact on most important queries.

III. Evaluating the Approach

One important prerequisite has to be met in order for the following evaluation to hold: A common ontological system or a standard for semantic marking of nodes in the Bayesian representation of user profiles has to be defined. An interesting practical approach is integrating an existing semantic sensor standard, such as Open Geospatial Consortium Software Web Enablement (OGC SWE), with a Bayesian decision making engine [4][5][6]. Please note that the notion of quality-of-service used in the related work does not correspond with the one used in this work. In this paper, quality-of-service is used as an alias for non-functional requirements on a data-centric platform. In the aforementioned publications, quality-of-service is mapped to the quality of data, such as its resolution.

As shown by Thomas and Green in their work on enumerating the junction trees in decomposable graphs, the number of available junction trees is strongly related to the graph structure [11]. A decomposable graph used as an example in their paper, with 23 vertices and 4 components, was shown to have more than 9000 distinct junction forests. They summarize that the conservative upper bound for complexity of enumeration of junction trees (in other words, for calculating the number of distinct junction forests) is $O(n^2)$. The cost for searching for the optimal junction tree in regards to a specific query profile is therefore dependent on the underlying graph's size and structure and on the size and structure of the query. The author is not familiar with any work in this area which attempted to bound the computational complexity of this calculation, based on the properties of the graph and the query. Multiplying the number of the enumerated junction trees with the number of combinations of all vertices which can be meaningfully queried in graph (without taking vertex weighting inside the query into account) would yield an overly pessimistic estimate, since the actual practical combinatorial space is determined by a relatively small number of pairs of nodes which act as data sources and the nodes which are being queried. Low-cost heuristics, such as the average graph distance between the nodes which are members of data sources and the nodes which are members of nodes which are being queried, can be used to eliminate solutions which have a very high probability of being far from optimal. For most practical applications, the optimal solution is not necessary, but rather a good enough or the first solution fitting in the time boundary for both upwards and downwards propagation of the junction tree's message passing protocol.

The first level of optimization in this work is focused on building optimal junction trees based on a specific query profile and is defined in the author's previous work on exact inference on Bayesian networks with quality-of-service requirements. The second and third levels of optimization are the contribution of this work and are based on exchanging entire query-optimized junction trees as opposed to using a central database of junction trees at one location for answering queries. This means that the remote application using the user profile model has to execute the protocol shown in Fig. 2.

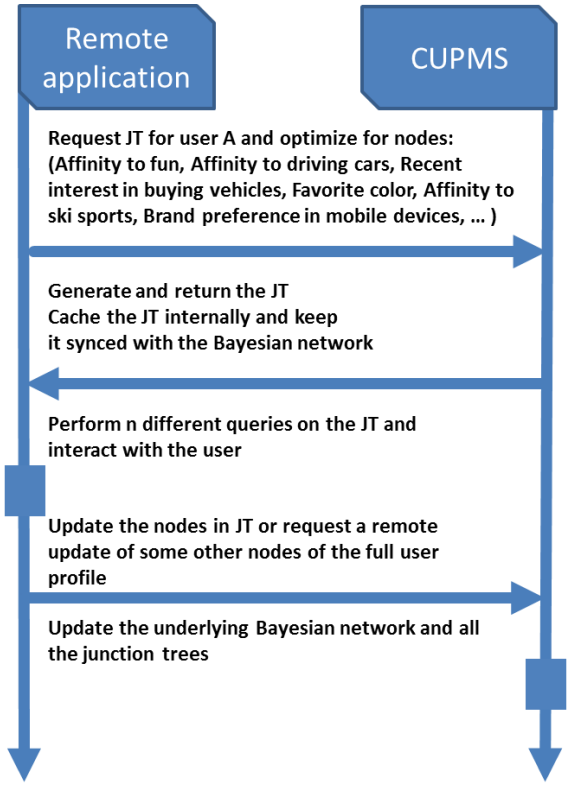


Figure 2. Protocol between a remote application and CUPMP

A. *Computational Complexity of User Profiling*

Important factors to consider:

1. Amount of queries on the junction tree needed to profile the customer
2. End-to-end communication delay with the CUPMP
3. Ratio of high-performance computational power of the CUPMP and the local embedded processing power of the remote applications
4. The computational and delay cost of re-synchronization of distributed user profiles

End-to-end communication delay is not an issue with profiling, since all the queries can be stacked together and sent to CUPMP as a single large query. The delay becomes an issue if a real-time and responsive interaction is necessary or if the application performs user profile updates very often. Even though offloading the entire

computational complexity to the high-performance CUPMP has the advantage of relieving the underpowered embedded platforms, this approach obviously limits the reactivity of the remote applications and exposes the entire user profiling process to the CUPMP.

The hardest element to estimate is the amount of queries on the junction tree which is needed to profile the customer. The ranges of different queries on the junction tree are not only dependent on the range of services or products offered by the data consumer, but also on the desired level of personalization or personalized presentation of information to the customer. Customer profiling can also be related to future products or trends. It could be expected that the depth of customer profiling will increase drastically as the profiling becomes easier to perform.

The well-known INSURANCE [10] Bayesian network is taken as a simple practical example of the complexity of customer profiling. The network has 27 nodes and was designed to calculate the expected claim costs for a car insurance policyholder. Let us assume that the network is expanded into a dynamic Bayesian network, with n history steps where n=3 [9]. Expanding the Bayesian network representation into an "unrolled" dynamic Bayesian network representation enables the application to capture the dynamic of the development of a user profile over the course of time. A customer whose theoretical expected claim costs have been falling over the course of time might be exactly the target audience for a new type of insurance product. Unrolling a Bayesian network in a simplest way means that the number of nodes is increased n times, or 3 times in this example. With a conservative estimate that the company only offers 5 different insurance products whose attractiveness should be evaluated using the current user's data, we assume that 5 different variations of dynamically unrolled INSURANCE network have to be evaluated for this specific user and that each time only one query is performed – the one on the **proposed cost** node. The insurance company however wants to offer a personalized experience to this user and analyzes the models of the car this user is interested in or would possibly want to purchase in future, based on the previous interaction with online car comparison platforms. Nine nodes in the INSURANCE network are directly impacted by the user's current vehicle data. Further assuming that the user was looking into three different vehicles from three different manufacturers in the same price class, it means that each network has to be instantiated with different values of the nine vehicle-dependent nodes at least three times. Varying the vehicle data and the vehicle equipment is where the number of necessary queries starts rapidly increasing, since some variables, such as **mileage**, have no correlation with variables such as **airbag**. Finding the most attractive insurance option for this user includes going through all the variable instantiation combinations and performing inference on **proposed cost**. Let us assume, for the sake of simplicity of this analysis, that the number of relevant variants of each vehicle is not bigger than 5. Total number of queries for this conservative use case is therefore 5 x 3 x 5 = 75. The example for the computational costs of querying the INSURANCE network using junction trees is given in [3], but for the purposes of this analysis it is enough to assume the constant cost **c** which is repeated 75 times. In other words, a single application incurs a computational cost of 75**c** on the central user profile management platform before any interaction with the user took place. If the user, by chance, chooses to select one of the products offered to him on the screen or over another human-machine interface, the application's interest in this specific user grows immensely and it becomes imperative to perform much deeper profiling. User's action of "paying attention" to the first offer can also be seen as permission for deeper profiling. As a side note, since the user's interaction has the potential to create value for the insurance selling company, the company might choose to pay a small fee for this additional or complete user profile. Performing a two-step or n-step profiling based on the level of interaction with the user is very important for such approach. The computational complexity of profiling will increase from the entry-level 75**c**, not only because more user data is now taken into account, but also because wider ranges of variables will be tested. Just like in the simple INSURANCE network, a combinatorial explosion follows from the growing number of variables with impact of user decision which has growing ranges of values.

Compared to the approach where all the profiles are kept on one place and all queries are answered by a massive centralized platform, the advantages of the proposed approach are the following:

- Responsiveness or on-demand real-time reactivity during the interaction with the user, regardless of connectivity issues with the central management platform
- Black-box approach to own profiling, protecting own business-model-related data / data about future products
- Caching local updates in user profile without putting load on the car-to-infrastructure communication

Having all knowledge-update and knowledge-query operations running on a central platform would incur massive performance issues. The proposed approach is highly parallelizable and robust against short-term connectivity losses.

B. COMPUTATIONAL COMPLEXITY OF SYNCHRONIZATION OF DISTRIBUTED USER PROFLE MODELS

Once a junction tree has been retrieved once, it only needs be synced with the currently same junction tree cached at CUPMP. It is important to consider a limit on allowed difference or drift between the information in the previously used and newest junction tree, to reduce the cost of constant minimal-updates or frequent user profile pulling.

The second part of syncing is pushing the new information about the user, acquired during the just-finished interaction, back to CUPMP. The simplest approach is to update the underlying Bayesian network in CUPMP containing the full user profile, and let the CUPMP populate the other junction trees bottom-up. The junction tree construction is now "for free", since it was

already performed once. This provides more predictable timings and forces one central sync point – the Bayesian network – to be constantly up-to-date. The cost of this operation is split over three different phases.

The first phase is updating the underlying Bayesian network, what is linear in the size of the junction tree.

The second phase is updating all other affected junction trees carries the complexity of **(num_nodes * num_trees)**, where num_nodes represents number of average nodes in a junction tree and num_trees represent the total number of junction trees. Lookup tables must exist, keeping track of which nodes are present in which junction tree.

The third phase consists of one pass of upwards and downwards propagation according to the junction tree algorithm. This is bounded by the depth of biggest junction tree.

To put the cost of the three phases into perspective, let us assume 1000000 drivers being on road in a given country – on a daily average. Let us make an estimate on the size of full user profiles in the form of Bayesian networks, containing 100 nodes. Let us assume 1000 different remote applications, with each having own junction tree (no sharing of profiles between applications) with an average of 30 nodes per junction tree. Complexity for syncing remains bounded only by linear complexity of 30x1000.

IV. Conclusion

Exact inference, which is one of the key elements of this work, offers the advantage of perfect reproducibility of the inference results, but at a very high computational cost. Heuristic approaches to Bayes inference, which are going to be the focus of future work on this topic, can result in a drastic reduction of the cost of inference.

One advantage of the proposed approach is a clearly-definable interface towards CUPMP, which enables the consumer to place arbitrary queries freely after receiving the entire junction tree. During the proposed system's operation, it will be possible to indirectly (by observing the usage of various junction trees) valorize different parts of the user profiles based on the frequency of their usage and the impact of their usage on the user behavior.

Driver assistance systems can also tap into the user profile data, at least in the phase of semi-autonomous vehicle operation, and profit from system reactivity and the lack of need for constant connection. Extending the proposed approach from content targeting to safety-critical vehicle systems would further increase the importance of exact inference. The reason is the perfect reproducibility of the results of inference, given the same starting conditions.

One important difference in this approach as opposed to generic user profiling and target content delivery is the fact that the connected car cannot be assumed to have perfect connectivity, but it is required to have perfect reactivity to user interaction.

Junction tree generation is a well-known NP-hard issue which is dependent on Bayesian network size. Cost of the junction tree construction is bounded by the size of the underlying Bayesian network, but this even happens extremely rare and can be performed in parallel to answering other queries. In this sense, the biggest shortcoming of this work is the lack of field trials with evaluation against existing big data, to measure in-the-field computational complexity. The pro and contra argumentation has only been provided by comparing the theoretical computational complexity of the approach.

References

[1] Frank, M., Roehrig, P., Pring, B. "Code Halos: How the Digital Lives of People, Things, and Organizations are Changing the Rules of Business", ISBN: 978-1-118-86207-0, 2014.

[2] Lauritzen, Steffen L., and David J. Spiegelhalter. "Local computations with probabilities on graphical structures and their application to expert systems." Journal of the Royal Statistical Society. Series B (Methodological) (1988): 157-224.

[3] M. Ljubo. "Context-Centric Design of Automotive Human-Machine Interfaces". München, Technische Universität München, Doctoral Dissertation., 2014, 2014.

[4] Yi, Y. and Calmet., Y. "Ontobayes: An ontology-driven uncertainty model." Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on. Vol. 1. IEEE, 2005.

[5] Luján, J. L. P., Yague J. L. P., and Ten., J. E. S. "Middleware with QoS support to control intelligent systems." Advanced Engineering Computing and Applications in Sciences, 2008. ADVCOMP'08. The Second International Conference on. IEEE, 2008.

[6] Yi, Y. and Calmet. J. "From the ontobayes model to a service oriented decision support system." Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on. IEEE, 2006.

[7] Huang, Cecil, and Adnan Darwiche. "Inference in belief networks: A procedural guide." International Journal of Approximate Reasoning 15.3 (1996): 225-263.

[8] Cooper, Gregory F. "The computational complexity of probabilistic inference using Bayesian belief networks." Artificial intelligence 42.2 (1990): 393-405.

[9] Xiangyang, L. and Ji., Q. "Active affective state detection and user assistance with dynamic Bayesian networks." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 35.1 (2005): 93-105.

[10] Binder, John, et al. "Adaptive probabilistic networks with hidden variables." Machine Learning 29.2-3 (1997): 213-244.

[11] T. Alun, and P.J. Green. "Enumerating the junction trees of a decomposable graph." Journal of Computational and Graphical Statistics 18.4 (2009): 930-940.

Text Skew Estimation Based on the Horizontal Entropy Calculation

Darko Brodić*, Milena Jevtić*, Zoran N. Milivojević** and Viša Tasić***
* University of Belgrade, Technical Faculty in Bor, Bor, Serbia
** College of Applied Technical Sciences, Niš, Serbia
*** Mining and Metallurgy Institute, Bor, Serbia
dbrodic@tf.bor.ac.rs

Abstract - This paper proposes the algorithm for text skew estimation based on the entropy calculation. The method consists of three stages. In the first stage, it calculates the horizontal projection profiles. In the second stage, the horizontal entropy for the rough angles is calculated. In the last stage, the horizontal entropy calculation for the smooth angles is performed. The calculated entropy creates the cost function. The position where the cost function has an absolute minimum represents the estimated text skew angle. The experiment on DISEC'13 document database confirms that the method correctly estimates text skew. Besides, it is robust and time efficient.

I. INTRODUCTION

The digitization of documents is typically carried out by a scanner, which usually introduces the text skew [1]. Hence, a some degree of text skew in digitized document image is an unavoidable. Readers are very sensitive to any text skew. They can register the text skew of at least 0.1° [2]. Furthermore, the text skew is undesirable in all stages of the optical character recognition (OCR) system, because it enables that system makes errors. It is valid, because any existence of the residual skew can lead to improper character recognition. If the text skew in document image is not corrected before further recognition processing, then the accuracy of these steps may be significantly reduced. Hence, the text skew detection and correction is indispensable prior to any recognition step in the OCR system. As a consequence, the text skew detection represents a mandatory stage in the OCR system [3].

Many different methods have proposed for the text skew detection in document images [4]. They can be classified into the following groups: (i) methods based on projection profiles, (ii) clustering of nearest neighbor's methods, (iii) cross-correlation methods, (iv) methods based on the Hough transform, (v) Fourier transforms methods, (vi) moment based methods, (vii) straight-line fitting methods, and (viii) other methods.

From all aforementioned methods, we choose to explore the extension of the projection profiles method. A method based on the projection profiles is widely used to detect the text skew angle. It is an easy method which sums up the pixels in the rotated document image creating a histogram at each possible angle. Then a cost function is applied to these histograms. The optimization of the cost function detects the text skew angle. The extension of this method is given in [5] by introducing the entropy concept.

The aim of this paper is to describe the algorithm for the text skew estimation based on the entropy. Furthermore, we suggest the optimization of previously proposed entropy method [5], which leads to speeding up the algorithm by the factor 2 to 4 times. Then, we tested the proposed algorithm on a document database DISEC'13. At the end, the results of the experiment are compared to the well-known Chou's algorithm [6] and validated according to the widely accepted method [7].

The paper is organized as follows. The Section 2 describes the proposed method. Sections 3 defines the experiment. Section 4 shows the experimental results and discusses them. Section 5 makes conclusions.

II. PROPOSED METHOD

The proposed algorithm detects the text skew angle in a document image by calculating entropy of the horizontal projection profiles. The preprocessing stage, which includes cleaning of documents, extracting blocks and denoising images [5] is out of the scope of our approach. Figure 1 shows the flow diagram of the proposed algorithm.

The method is applied to the binary image. Suppose that we have the binary image $B(i, j)$ with M rows and N columns. The horizontal projection profile of the binary image **B** is calculated as:

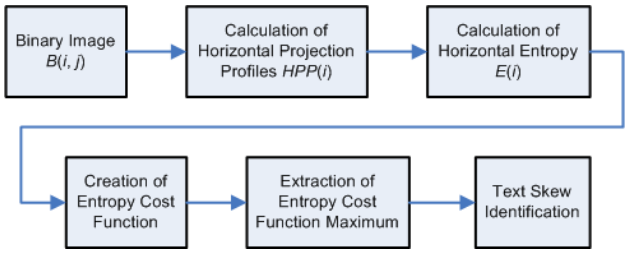
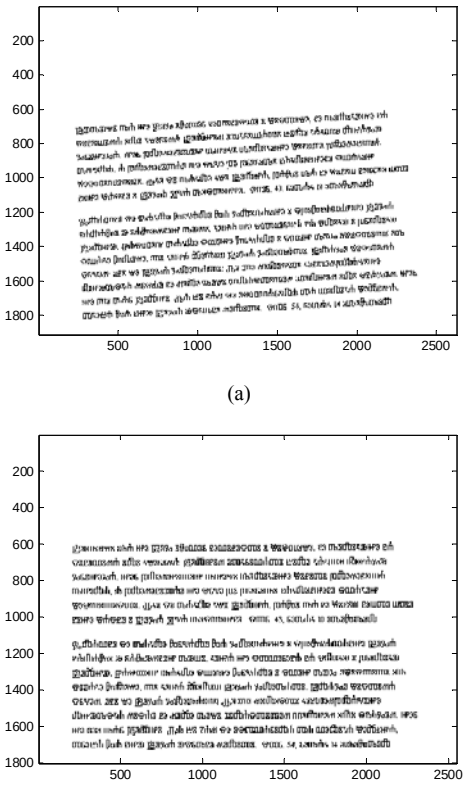
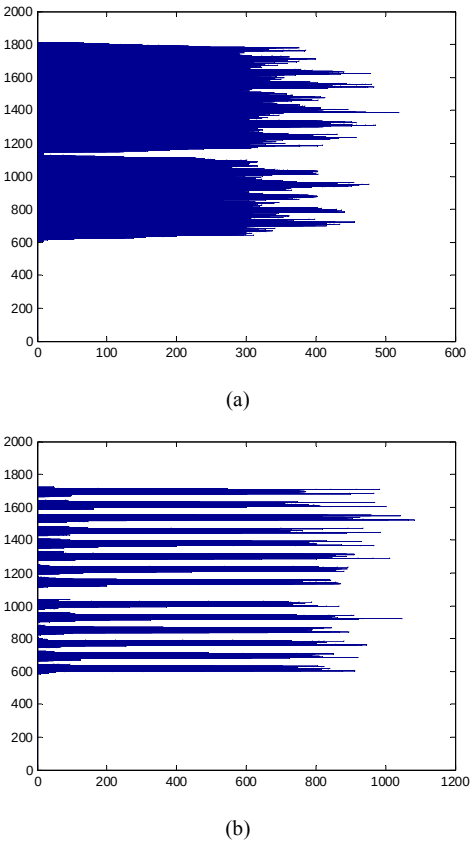


Figure 1. The flow diagram of the proposed algorithm



(a) (b)

Figure 2. Initial binary image: (a) skewed for the angle of 2.6°, (b) unskewed



(a) (b)

Figure 3. Horizontal projection profiles (*HPP*): (a) for skewed image (2.6°), (b) for unskewed image

$$HPP(i) = \sum_{j=1}^{N} B(i,j), \qquad (1)$$

where i represents the row and j represents the column of the image **B**. Each element of the horizontal projection profiles $HPP(i)$ represents the sum of the pixel values in the corresponding rows.

Figure 2 shows the initial binary image **B** ((a) skewed, (b) unskewed). Figure 3 (a) and (b) shows a corresponding horizontal projection profiles according to Figure 2 (a) and (b), respectively. If the number of calculating pixel per row (horizontal projection profiles) is under observation, it is clear that highest value will be received when the text is unskewed. Hence, the procedure of rotating skewed image in the range of angles with certain angle step will establish the *HPP*, which receives its highest value among all rotated images. This angle represents the position where the skew is the closest to 0° [6]. In this way, the text skew can be identified.

Using horizontal projection profiles as a basis, we introduce the concept of entropy. The entropy is determined as a statistical measure of image randomness. The horizontal entropy, i.e. entropy per row is calculated as [5]:

$$E(i) = \sum_{i=1}^{N} -HPP(i) \cdot \log\big(HPP(i)\big). \qquad (2)$$

If the text-line has more randomness in its texture, then $E(i)$ will receive higher values. In the original algorithm [5], the full range of $E(i)$ is used for calculation of the cost function E. However, we introduce an optimized approach, which consists of the two-stage calculation of the entropy function E. In the first stage, the image is rotated from θ_{start} to θ_{end}, where θ_{start} is equal to -15° and θ_{end} is equal to +15°. The step of rotation is $step_{start}$, which is equal to 1°. As a result, the entropy cost function E is obtained. The lowest value of entropy means that it can be considered as the starting (prefinal) estimated text skew angle, i.e. $\theta_{prefinal}$. However, the accuracy is given by the step of rotation equal to 1°, which is too coarse. Hence, some additional justification is needed. In the second stage, the subrange around the prefinal text skew angle $\theta_{prefinal}$ is used for additional testing. Currently, the image is rotated from $\theta_{prefinal} - step_{start}$ to $\theta_{prefinal} + step_{start}$ with the step of rotation $step_{final}$. This step is much finer, i.e. it is equal to 0.1°. The newly obtained result for the minimal entropy represents the final estimated text skew, i.e. θ_{final} with the accuracy of 0.1°.

Figure 4 illustrates the entropy cost function. Figure 4 (a) shows the entropy cost function obtained in the first stage (θ_{start} = -15°, θ_{end} = +15°, $step_{start}$ = 1°), while Figure

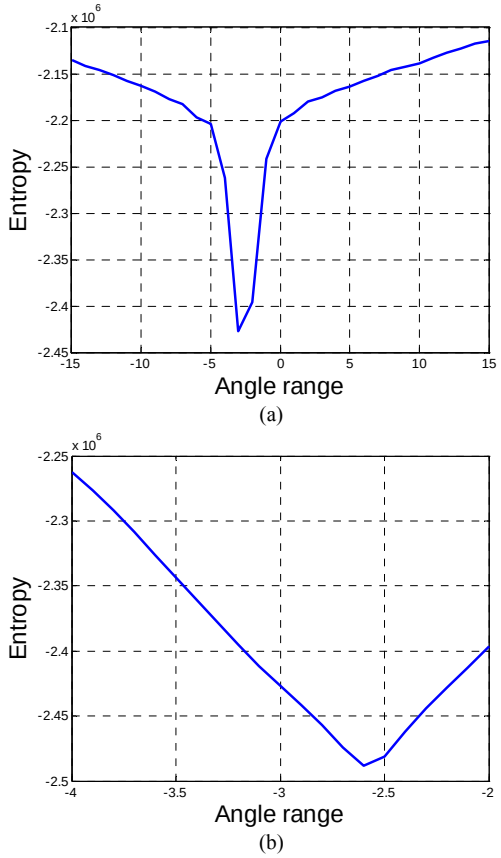


Figure 4. Entropy cost function E for the skewed image (2.6°): (a) established by rotating image using angles from -15° to +15° with the step 1°, (b) established by rotating image using angles from -15° to +15° with the step 0.1°

4 (b) shows the entropy cost function obtained in the second stage ($\theta_{start} = \theta_{prefinal} - step_{start}$, $\theta_{end} = \theta_{prefinal} + step_{start}$, $step_{final} = 0.1°$) of the proposed algorithm. Figure 4 (a) illustrates the so-called coarse detected minimum, while Figure 4 (b) shows the so-called fine detected minimum. The fine detected minimum of the entropy cost function represents the estimated text skew obtained by the proposed algorithm (See Figure 4 (b)).

III. EXPERIMENT

The experiment related to text skew detection needs to evaluate the quality of the algorithm to correctly predict the text skew embeeded into the document image. In this way, the algorithm is applied to the skewed document images from a database. Usually, the document images are randomly rotated at different angles ranging from -15° to +15°. The result of the experiment shows the deviation of the estimated text skew obtained by the tested algorithm and correct text skew value.

As a document database, a part of Document Image Skew Estimation Contest (DISEC'13) [7] is used. It is recognized as a standard tool for unbiased testing of the algorithms for the text skew detection. Document images included in DISEC'13 are extracted from various sources consisting of document images with figures, tables, diagrams, architectural plans, electrical circuits, etc. They are written in different languages like English, Chinese

Figure 5. A few document image samples from DISEC'13 database

and Greek or some multilingual combination. Our part of DISEC'13 consists of 100 images. Figure 5 illustrates a few document images from DISEC'13 database.

To measure a performance of the results obtained from testing algorithm, the following criteria is proposed [7]:

$$AED = \frac{\sum_{n=1}^{N} \Delta(n)}{N}, \quad (3)$$

where $n = 1, ..., N$ is the number of document images from a database, and $\Delta(n)$ is the absolute deviation error (ADE). ADE for document image n is given as:

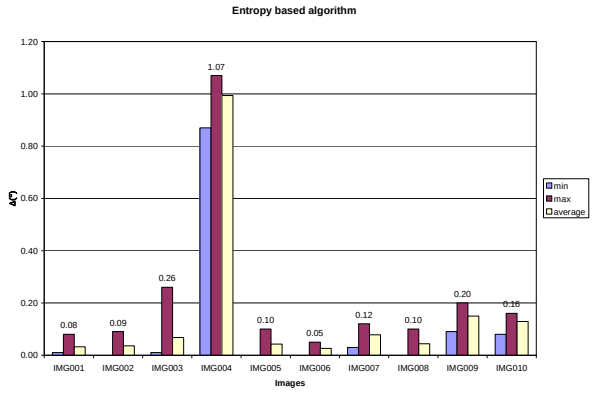
$$\Delta(n) = \left| \theta_{gt}(n) - \theta_{est}(n) \right|, \quad (4)$$

where θ_{gt} is the skew angle of the document image n from a database, commonly called ground-truth, and θ_{est} is detected or commonly called estimated skew angle obtained by testing algorithm.

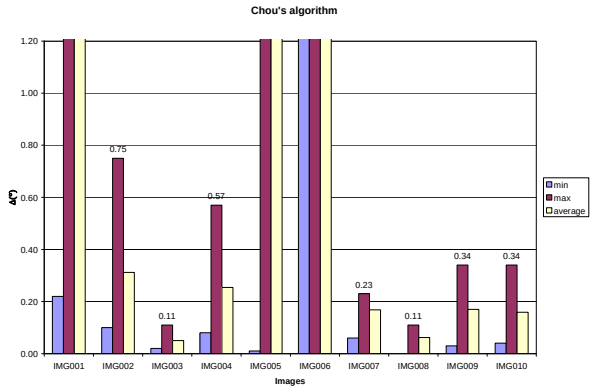
IV. RESULTS AND DISCUSSION

We make a comparison between our algorithm and well-known Chou's algorithm [8] to show the quality of the proposed text skew detection approach. The results of the experiment for the entropy based and Chou's algorithm are shown in Figure 6. Because the text skew of 0.1° may be visible to a human observer, it is chosen as comparison criteria [2].

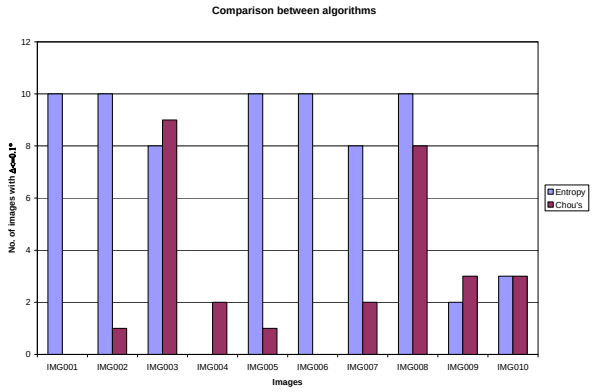
The application of the entropy based algorithm to the DISEC'13 database shows that three out of ten document images are characterized with the highest maximum level of ADE equal to 0.26°, 0.44° and 1.07° with corresponding average ADE level equal to 0.16°, 0.18° and 0.99°. The aforementioned ADE values are quite



(a)



(b)



(c)

Figure 6. (a) ADE for entropy based algorithm, (b) ADE for Chou's algorithm, and (c) Number of images with the skew detection <= 0.1°

moderate. Furthermore, it should be noted that document images with highest ADE level represent so-called multi-column document images.

The application of the Chou's algorithm to the DISEC'13 database shows that three out of ten document images are characterized with the highest maximum level of ADE equal to 9.53°, 20.87° and 22.63° with corresponding average ADE level equal to 2.70°, 10.86° and 8.24°. These values are rather extreme. Two of them represent the comic document images. Obviously, Chou's algorithm is completely unsuitable for text skew detection of such type of document images.

Finally, the comparison between two algorithms is made by counting the number of images with ADE less or equal to 0.1°. The proposed algorithm has a success rate of 71/100, while the Chou's algorithm has a success rate of only 29/100. From all above it is clear that the proposed algorithm has the real potential and it shows the advantages compared to well-known Chou's algorithm for text skew detection.

V. CONCLUSION

The paper proposed the entropy based algorithm for the text skew estimation. The basic entropy method which was previously proposed in [5], we optimized by incorporating two stages with so-called rough and fine step rotation of the analyzed image. In this way, the proposed algorithm was speeding up compared to the basic version at no cost on its accuracy. The algorithm was tested on a part of DISEC'13 database, which contains a variety of document images written in different scripts, including different elements like images, comics and so on. At the final stage, the proposed algorithm was compared to well-known text skew estimation method proposed by Chou et al. [8]. Our method proved to be robust and accurate. It showed much better results than Chou's method in terms of accuracy. Future research will be toward improving the algorithm estimation of the text skew in the multi-column document images.

ACKNOWLEDGMENT

This work is supported by a grant from the Ministry of Education, Science and Technological Development of the Republic of Serbia, as a part of the project No. TR33037.

REFERENCES

[1] Sarfraz M., Rasheed Z. Skew estimation and correction of text using bounding box. Proceedings of 5th international conference on computer graphics, imaging and visualization, (CGIV '08), pp. 259-264, 2008.

[2] Document Analysis Systems II. J.J. Hull, S.L.Taylor, Eds., World Scientific, pp. 40-64, 1998.

[3] Amin A., Fischer S. A document skew detection method using the Hough Transform. Pattern analysis and applications, pp. 243-253, 2003.

[4] O'Gorman L., Kasturi, R. Executive briefing: document image analysis, IEEE Computer Society Press, pp.1-127, 1997.

[5] Arvind K.R., Kumar J., Ramakrishnan, A.G. Entropy Based Skew Correction of Document Images. Proceedings of Pattern Recognition and Machine Intelligence (PReMI 2007), LNCS 4815, pp. 495-502, 2007.

[6] Postl W. Detection of Linear Oblique Structures and Skew Scan in Digitized Documents. Proceedings of 8th International Conference on Pattern Recognition, pp. 687-689, 1986.

[7] Papandreou, A., Gatos, B., Louloudis, G., Stamatopouos, N.: ``ICDAR2013 Document Image Skew Estimation Contest (DISEC'13)"; Proceedings of 12th International Conference on Document Analysis and Recognition, (ICDAR 2013), (2012). 1476-1480.

[8] Chou, C. H., Chu S. Y.; Chang, F.: ``Estimation of skew angles for scanned documents based on piecewise covering byparallelograms"; Pattern Recognition, 40, 2 (2007), 443-455.

Optimization of Ring-Star Transmission Problem in Telecommunication Systems Based on Ant Colony Algorithms

Vedad Letić[1, 2], Zikrija Avdagic[1] and Dušanka Bošković[1]
[1] Faculty of Electrical Engineering/ Department for Computing and Informatics, Sarajevo, Bosnia and Herzegovina
[2] BH Telecom d.d. Sarajevo, Bosnia and Herzegovina
vedad.letic @ bhtelecom.ba, {zikrija.avdagic, dusanka.boskovic}@etf.unsa.ba

Abstract - Solving complex combinatorial optimization problems using classical algorithms is not efficient related to resources and time. To overcome the problem, we used optimal parameters selection based on Ant Colony Optimization (ACO) algorithms. In this paper, we present algorithm for solving telecommunication network using ACO for searching optimal ring star network topology. We analyzed ant's optimization ability based on shortest path between the nest and food location. In our research we used: Ant System, Elitist Ant System, Rank-Based Ant System, and MAX-MIN Ant System. The program is developed using GNU C++ to prove the algorithm theoretical convergence through simulation on variety of topologies regarding to node numbers. The algorithm was adapted to solve design of telecommunication network, which connects terminals to concentrators using point-to-point connections. The algorithm's output is a star topology showing connections of concentrators in a ring creating Digital Data Service. Algorithm uses seventeen parameters, with thirteen metrics to evaluate configurations. Program validation is performed using three different network node configurations for all four ACO algorithms, only changing two control parameters: speed of pheromone evaporation and existence of local search. The best path was evaluated based on: total time, number of iterations, ring size, and value of topology.

Key words: ant colony, optimization, ring-star problem

I. INTRODUCTION

Worldwide, telecommunication companies spend billions of dollars to replace copper lines with new transmission medium - optical lines, and for reasons of increasing the flow of data and information, as well as customizing communication networks for simultaneous transmission of multiple types of these data, including video signals.

Rapid growth of design paradigms for large optical networks is evident, including connecting the clusters of nodes, the synchronized optical network (Synchronous Optical Networking – SONET [1]) in large rings, as building blocks for design of communication systems. Ring is a popular topology for these systems and attracts a lot of research attention.

Optical links are installed along with the rest of the infrastructure of telecommunications equipment and facilities, and using the latest technology they are integrated in an intelligent and transparent transmission network, characterized with high capacity, automatically monitored and controlled.

The ring is separated as a kernel of the network, to which, in accordance with the needs, other parts of the system are connected. Such a solution is transparent for all telecommunication nodes, regardless of whether they are located in the local, regional or some remote telecommunication level.

Prevailing technical standard in the USA is the SONET. Within the SONET ring, the most common node is a telephone exchange, with optical connection into the ring. The topology enables every node to send, receive and forward data. The nodes are available for every edge of the SONET ring thorough the Add-Drop Multiplexer (ADM). The SONET standard has an important role in optical transmission because it helps to prolong survival of legacy networks and their faster transformation into contemporary and functional optical telecommunication networks.

Node symmetry of the ring ensures that each node in the ring has the same role with the all necessary equipment provided. The circular connection of the node protects it from unwanted traffic disruptions. An important problem of telecommunication network optimization is to find a balance when minimizing the sum of routing and connection costs in the network.

The objective of the ring star problem is to find a simple cycle, referred to as a ring, connecting a subset of nodes of a graph. The problem includes two minimization objectives: of a cost associated to the ring connecting the subset of nodes and the minimization of a cost associated to the arcs connecting out-of-the-ring nodes with the nodes within the ring. The problem was introduced by Labbé et al. [2]. The ring-star problem is used in designing telecommunication networks, and in planning fast transport systems. The problem is considered as extension of the classic location-allocation problem from the early 60s, including several comparative versions studied by several authors. One of the recent studies proposes solution based on evolutionary methods [3], and there are methods that employed meta-heuristic and treated the ring star problem as a bi-objective problem [4]. For solving the problem we used optimal parameters selection based on Ant Colony Optimization (ACO) algorithms, a metaheuristic approach proposed by Dorigo et al. [5]. To

our knowledge this approach was not used before for solving the ring star problem.

Typical telecommunication network traffics content from various sources with dramatic increases in system capacity [6]. In order to answer the increasing size of networks, many telecommunication companies support hierarchical structure of the network, built out of different interconnected layers. The various services (voice, video and data) are structured in several layers, allowing telecommunication companies to solve network design problems focused only on a part of the network.

Solution to the network design problem is to connect terminals and concentrators. Design solution presents a set of nodes forming a ring and a set of direct connections from each non-included terminal to a node included in the ring. The solution should also meet the network requirements at the minimum cost.

II. DATA AND METHODS

The Ant Colony Optimization is used in a way to let Ant Agents visit the graph nodes, until they return to the starting node. Upon their return on the starting position we calculated cost of the cycle as a sum of the ring route cost and the assignment costs. Depending on the quality of the cycle pheromone is laid on the branches belonging to the ring.

We had chosen this approach for pheromone trailing, because we did not have *a priori* knowledge about the number of nodes belonging to the ring, and the number of assignment nodes. This approach provided that ant colony had decided about the ring size and which nodes were part of the ring and which not.

Due to a single path nature of the assignments, their costs are 4 to 5 times lower compared with the ring connection costs. This is the reason for the importance of the resulting topology, and significance of the selection which nodes to include in the ring, and which to leave out of the ring and assign separately to respective node in the ring. In case that assignment cost was higher than the cost of the connection within the ring, algorithm would have tried to connect all the nodes into the ring and the problem would be reduced to the classical Travelling Salesman Problem (TSP). Accordingly the Ring Star problem can be considered as a generalization of the TSP.

The program was tested using files Bosnia62.tsp, att532.tsp and d1291.tsp. Files att532.tsp and d1291.tsp are part of TSPLIB95 [7]. File Bosnia62.tsp was created using geo coordinates from 62 municipalities in Bosnia and Herzegovina, and in compliance with the rules for the library TSPLIB95 (in Appendix).

The program comprised four ACO algorithms: Ant System, Elitist Ant System, Rank-Based Ant System and MAX-MIN Ant System. The program was executed using different parameter values and results achieved were captured.

In addition, we have used to algorithms for local search and the following pseudo-random-proportional rule for the next node selection:

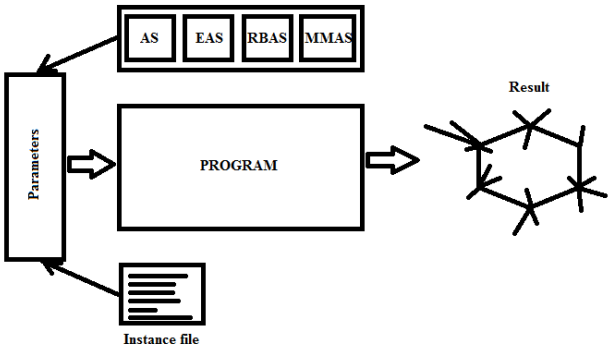


Figure 1. System block diagram

$$p_{ij}^{k} = \frac{\tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}}{\sum_{l \in N_i^k} \tau_{il}^{\alpha} \cdot \eta_{il}^{\beta}}, \quad (1)$$

III. SYSTEM DESCRIPTION

The system we built to utilize the four selected ACO algorithms is shown in Fig. 1. The system was capable to execute the program without user intervention, using the default values for the parameters, otherwise user needed to change the parameter values. After parameters initiation, program calculated or read data about the nodes and their respective distances.

After that initiation phase, the system was searching for solutions. The first solution was saved as a candidate for the result. Every time the system had found a better solution it was saved for reviewing the solution in process and to monitor advancement in a search for a better one.

The system utilizes modified standard ACO algorithms described in detail in [8]. In the following section we shall describe only differences with respect to the original algorithms. In respect to the TSP, pheromone evaporation rate calculation is changed. The evaporation in solving the TSP starts when all ants return to the starting point. In this work, the evaporation is started after the Ant makes more steps than the double maximal ring size. The evaporation implementation is the same for all four ACO implemented in the system.

A. *Ant System (AS)*

1. The Ant System algorithm starts with positioning the Ant Agent at the random node.
2. The Ant Agent moves along the path selecting the next node to visit using the pseudo-random-proportional selection rule (1).
3. If parameter defining the maximum number of nodes in the cycle is equal total number of nodes, the Ant Agent is obliged to visit the starting node again.
4. If the starting and the destination node are the same, the Ant changes the move direction, and traverses the cycle backwards.

5. The Ant checks the quality of the cycle and, if needed, improves it using the local optimization.

6. On its way back, Ant lays pheromone, and the quantity of the pheromone is inversely proportional to the cost of the cycle.

7. The cost of the cycle is calculated as a sum of the ring route cost and assignment costs.

8. After laying pheromone on the branches traversed, the Ant repeats the algorithm.

9. After every few steps, program trigger evaporation of pheromone trail.

B. *Elitist Ant System (EAS)*

Elitist Ant System algorithm differs from Ant System in a way how pheromone quantity is calculated.

The algorithm in accordance with the description in [8], before step 6. (in A. Ant System), assigns extra reword for the best cycle. This is accomplished using the specific parameter that multiplies the pheromone trail.

C. *Rank-Based Ant System (RBAS)*

Rank-Based version of the ACO algorithm for the Ring Star problem has a significant difference in implementation, compared with the previous algorithms. Step 6. (in A. Ant System) is replaced with three new steps.

6a. Whenever the Ant Agent constructs a ring of nodes, that results is captured in the temporary array, until the all Ants complete the task.

6b. After that, the system selects the r best rings and assign to each ring respective rank starting with the r for the best towards 1 for the last ring in the selection.

6c. Pheromone trail is updated for these rings, and quantity of the pheromone is multiplied with the rank of the ring.

D. *MAX-MIN Ant System (MMAS)*

In this algorithm the Ants also construct the path, again by selecting the next node using the pseudo-random-proportional rule (1). The difference is regarding the pheromone update (Step 6. in A. Ant System).

6. The MAX-MIN algorithm limits the upper and lower boundary for the pheromone quantity. After the Ant constructs the cycle, the pheromone is updated according to the calculated boundaries, and the Ant starts its task again from the beginning.

IV. RESULTS

The system was tested using the workstation with core2duo processor and 2GB RAM, with the MS Windows XP operating system.

In the beginning the program was executed to establish input parameter: evaporation rate. The parameter was changed thorough several iterations, and final values for the parameter is presented in the Table I.

TABLE I. INPUT VALUES FOR EVAPORATION

Algorithm	***Evaporation parameter***
Ant Colony System	0,03
Elitist Ant System	0,03
Rank-Based Ant System	0,05
MAX-MIN Ant System	0,018

Each of the four algorithms was tested in duration of 20 minutes, with previously presented parameters and for the four data files. The results are presented in Fig. 2 for data from the files att532.tsp and dl291.tsp [7], and Fig.3 for data from the file Bosnia62.tsp.

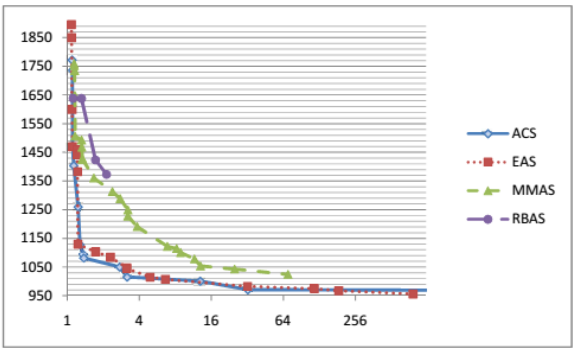


Figure 2. Results for the file: Bosnia62.tsp

Presented results were obtained with system using the following parameters: Number of Ant Agents: 32, Minimal percentage of nodes to be included in the ring: 30%, Maximal percentage of nodes to be included in the ring: 70%, Maximal time interval for program execution: 1200s, Number of nodes in neighborhood: 10, Assignment cost was 20% of the ring route cost.

During construction phase of algorithm, while Ant construct the ring it is not allowed to close ring if he did not include, due to Minimal percentage of nodes, minimal number of nodes in his path. Similarly, if he reach Maximal percentage of nodes, next node which he will chose is his starting node.

Analyzing the results obtained, the MAX-MIN Ant System (MMAS) is constructing the best solution, what is in compliance with [9]. However, it can be noted that in respect to convergence, other algorithms after additional time period construct the better solution.

V. CONLUSION

Presented results bring into attention several issues worth the further investigation. As a next step we would like to build a solution employing a hybrid of the two algorithms: ACO and the Neural Networks algorithms, such as backpropagation .

We expect that this hybrid approach combining the advantages of the ACO with the ability to search for the

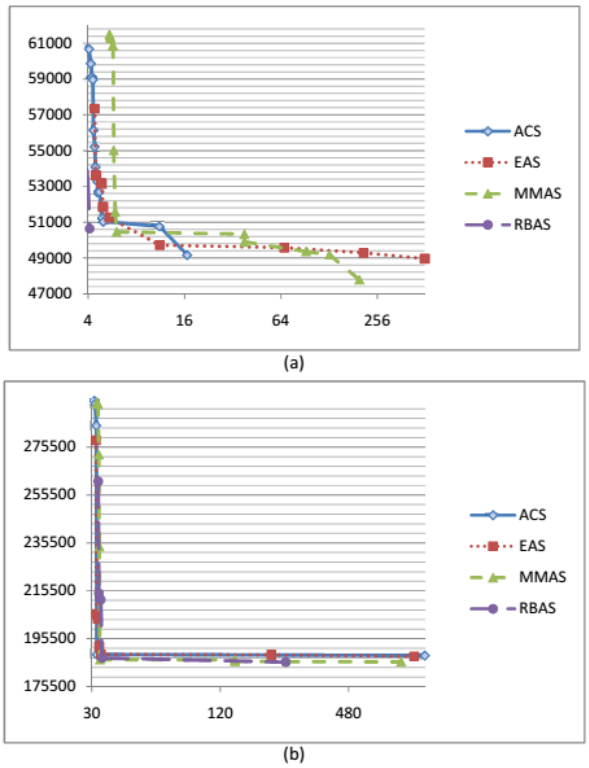


Figure 3. Results for the files: (a) att532.tsp (b) d1291.tsp

global optimal solution, and the backpropagation algorithm with the ability of convergence to the local optima, can be successfully applied for solving the Ring Star problem.

VI. Appendix

Problem instance "Bosnia62.tsp":

```
NAME: Bosnia62
TYPE: TSP
COMMENT: 62 city in Bosnia And Herzegovina (Vedad)
DIMENSION: 62
EDGE_WEIGHT_TYPE: GEO
DISPLAY_DATA_TYPE: COORD_DISPLAY
NODE_COORD_SECTION
1 44.44 17.28
2 44.47 16.16
3 44.45 19.12
4 45.11 16.48
5 45.08 18.00
6 45.09 17.15
7 44.52 16.09
8 44.33 16.22
9 44.52 18.48
10 44.03 17.26
11 44.58 15.56
12 43.06 17.42
13 44.58 17.54
14 44.43 18.05
15 44.08 17.24
16 43.30 18.46
17 43.40 18.59
18 43.56 17.35
19 44.41 18.18
20 44.52 18.25
21 43.49 18.18
22 43.49 18.21
23 44.20 17.15
24 43.39 17.45
25 44.08 18.04
26 44.26 18.54
27 43.41 17.39
28 44.54 17.18
29 43.49 17.00
30 43.11 17.32
31 44.31 18.31
32 44.57 18.18
33 43.20 17.48
34 43.15 18.06
35 42.55 17.37
36 45.05 18.19
37 44.10 17.38
38 44.37 18.21
39 43.28 17.19
40 44.58 16.42
41 44.45 16.39
42 43.51 18.24
43 44.42 18.29
44 44.06 19.17
45 43.22 17.35
46 44.36 17.51
47 44.36 17.59
48 43.43 17.13
49 44.13 17.39
50 42.42 18.20
51 44.14 17.35
52 44.32 18.40
53 45.11 15.48
54 43.59 18.10
55 44.09 17.47
56 44.10 18.56
57 43.54 18.20
58 44.26 18.09
59 44.12 17.54
60 44.23 19.06
61 44.27 18.39
62 44.25 18.02
EOF
```

References

[1] American National Standards Institute, "Synchronous optical network (SONET) — Basic description including multiplex structure, rates, and formats," ANSI Standard T1.105-2001, May 2001

[2] I. Rodriguez Martin M. Labbé, G. Laporte and J. J. Salazar Gonzalez, "The ring star problem: polyhedral analysis and exact algorithm," *Networks*, 43(3):177–189, 2004

[3] H. I. Calvete, C. Galé , J.A. Iranzo, "An efficient evolutionary algorithm for the ring star problem," *European Journal of Operational Research*, 11/2013; 231(1):22-33.

[4] A. Liefooghe , L. Jourdan , M. Basseur , E. Talbi , E. K. Burke, "Metaheuristics for the bi-objective ring star problem,"

Proceedings of the 8th European conference on Evolutionary computation in combinatorial optimization, March 26-28, 2008, Naples, Italy, pp.206-217

[5] M. Dorigo, G. Di Caro, "The ant colony optimization meta-heuristic," in: D. Corne, M. Dorigo, F. Glover, D. Dasgupta, P. Moscato, R. Poli, K. Price (Eds.), *New Ideas in Optimization*, McGraw-Hill, 1999, pp. 11–32.

[6] R.-J. Essiambre and R. W. Tkach "Capacity trends and limits of optical communication networks", *Proc. IEEE*, vol. 100, no. 5, pp.1035 -1055, 2012

[7] G. Reinelth, TSPLIB, Symmetric traveling salesman problem (TSP), University Heidelberg, Institute for Informatics, 2008 [online] ttp://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/ (Accessed: 7 February 2015)

[8] Marco Dorigo and Thomas Stützle, *Ant Colony Optimization*, Published by MIT Press, 2004

[9] T. Stützle, H.H. Hoos, „MAX-MIN Ant system," *Future Generation Computer Systems*, v.16 n.9, 2000, pp.889-914

Using String Similarity Metrics for Automated Grading of SQL Statements

I. Štajduhar* and G. Mauša*
* University of Rijeka, Faculty of Engineering, Department of Computer Engineering, Rijeka, Croatia
istajduh@riteh.hr

Abstract - Manual grading of structured query language (SQL) statements after an exam can be tedious and time consuming for the teaching assistant. Additionally, it can also be subjective to her current state of mind and, thus, prone to errors. In this paper we propose an automated method for grading individual SQL statements. The method uses several common and simple string similarity metrics for comparing the student devised statements against the reference statements. These are then used, along with the manually assigned grades, for building the predictive logistic regression model. The proposed method was evaluated on a dataset consisting of 314 pairs of student-reference statements, along with the discretized average grade assigned by three independent evaluators. The model achieved the expected classification accuracy of 78% on a binary class, thus exhibiting its potential for real-life application. The model can be used as is with the suggested calculated features and reported learnt parameters, or adapted to other examiners' evaluation criteria, presuming their willingness to build manually graded datasets of their own.

I. INTRODUCTION

Sitting in a dark and cold office, working late hours, performing a mind-numbing choir, one cannot help oneself but not to ask himself: "What have I done to deserve this?". The authors of this paper are no strangers to this scenario during midterm and final examinations at their home university, when they are overwhelmed by a large number of SQL exams that need to be manually graded. Here, an SQL exam usually consists of several assignments, custom made for a distinct toy database (one not too simple, or too difficult to grasp), each of which requires an answer in a form of an SQL statement. Although the answer statement might not necessarily be unique for a distinct question, it does need to generate the exact relation, as was prescribed by the assignment, to be correct. Verifying the equivalence of the answer statement output relation to the reference statement output relation is trivial, if done using information technology. Two years ago we developed an SQL examination system in the form of a web service, allowing us to evaluate our students' understanding of SQL in a more grading-friendly and fair way, than it was when we conducted the examinations solely on paper. Additional benefit of such a system is that the students now gained the ability to check their progress during this problem-solving process and correct syntactic errors on the fly, which is a desirable property of a teaching system [1]. This system would automatically test for equivalence between the student relation and the reference relation and automatically grade such cases as correct. Moreover, it would discard syntactically incorrect statements, grading them as incorrect. Yet one problem remained. What was to be done with all the syntactically-correct "close, but no cigar" answers, ones that were automatically graded as incorrect?

Automated grading of SQL statements has been in the focus of the scientific community in the last decade or so. There exist several interesting interactive SQL tutoring and/or grading systems [2,3,4,5,6,7]. The automated tutoring systems, such as [2,3], are focused on improving student learning, and usually provide on-line automated feedback to the student, with guidance based on a personalized assessment. Other systems [4,5,6,7] also include the (partially-) automated grading component, but with some limitations. For example, *AsseSQL* [4] provides query grading, but only on rather simple queries. *SQLator* [5] allegedly uses a heuristic grading engine, but no details are specified in the proceedings paper. *SQLify* [6] introduces other manual grading mechanisms, like peer review, for coping with the problem of continuous grading. The most thorough automated grading system was recently presented in [7], including also the statement cost and style grading, but, again, lacking the option of evaluating partially correct answers.

Recent work done, towards developing a system for fully automated grading of SQL exams, mostly involved mechanisms for thorough semantics parsing. These approaches exploited the SQL semantic specifics to formulate simple rules, to be applied for automated grading. They are, however, mostly limited to simpler statement formulations, which is understandable, since the language semantics are extremely difficult to reasonably bind in a fair SQL understanding-evaluation-system model. For more complex statements, manual grading is usually necessary.

In this paper we demonstrate that simple string similarity metrics can be successfully used for comparing the student-devised SQL statement against the reference statement, without specific knowledge or understanding of the underlying semantic intricacies of SQL. We believe that these could be used to further improve the accuracy of existing, semantic aware, automated grading systems, or as an add-in to the systems that lack soft grading entirely. The prerequisites for this are using a live database management system (DBMS), a nonempty toy database

The work presented in this paper is partially supported by the University of Rijeka research grant 13.09.2.2.16.

and a list of assignments with corresponding solutions in the form of SQL statements.

Structure of this paper is as follows. In the following section, we describe the data used for evaluating the model. Being aware of the structure of the data used is necessary for understanding the workings of the feature transformations and the learning setup, described in section III. In section IV, we report and interpret the results obtained through empirical evaluation of the proposed model, using good machine-learning practice. Our findings are summarized in section V, additionally giving some insight into our probable future efforts in this field.

II. DATA

Exam data was gathered during the summer semester in the academic year 2013/2014, at the course Database Systems, held at the Faculty of Engineering, University of Rijeka. Part of the exam used to assess students' knowledge of SQL was performed on computers, using a live database-interactive system devised here. This system produces an individual exam for each student, generating it by randomly selecting a total of 5 assignments from 3 pools of questions of varying difficulty, tied to an exam database: 2 easy, 2 moderate and 1 question of hard difficulty. For the purpose of this study, we used the "Library" example database, consisting of 13 assignments of varying difficulty, each of which requires an answer in the form of a SELECT statement. The entity-relationship model for this toy database is presented in Fig. 1. During the midterm and term exams in the semester, a total of 393 student answers were gathered for this database. A summary of gathered data is presented in Table I.

A. Old grading process

Each of the 393 student answers was first automatically graded by the system, using a rigid evaluation principle, and then manually regraded by the teaching assistant, to somewhat soften the criteria. If the student-devised statement gave the same resulting relation as the reference statement, it was then marked as completely correct and graded 100%. This equality was compared automatically and reported back by our system using the following expression:

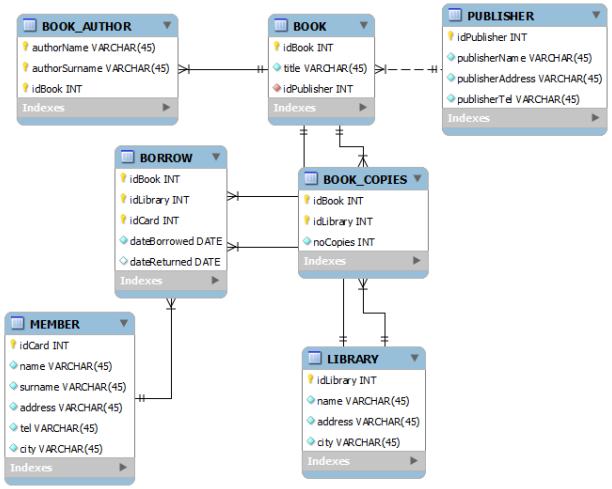


Figure 1. ER model of the "Library" toy database

$$(R_S \text{ UNION } R_R) \text{ MINUS } (R_S \text{ INTERSECT } R_R), \quad (1)$$

where R_S represents the relation returned by the DBMS when running the student-devised statement, and R_R represents the relation returned by the DBMS when running the reference statement. If the expression returned an empty relation, then we could assume that R_S and R_R were equal. The actual implementation of the equality checking function required manual implementation of binary set operations MINUS and INTERSECT, because the MySQL DBMS does not support them directly. If, on the other hand, the student statement was noninterpretable, or far from the correct answer, it was marked as completely incorrect and graded 0%. This was also done automatically.

All of the interpretable statements were manually inspected afterwards for softening the grading principle and detecting cheating attempts. In some cases, when the student-devised statement was semantically close to the reference statement, but not entirely correct, the answer was manually labeled as partially correct and graded 50%. If not for this step, it would remain graded 0%. Cheating would occur when a student would use several statements to come to intermediate results and then use their output manually to achieve the desired resulting relation, instead of writing it as a single, standalone, statement. Detected cheating attempts were then graded 0% (instead of previously automatically assigned 100%).

B. New grading process

To allow a more flexible grading method and to account for the possible teaching assistant bias during the assignment grading, three independent expert reviewers were engaged in a triple blind, fully-manual, grading of the entire answer dataset from scratch. Each of the reviewers was given a spreadsheet, consisting of all of the 393 student-reference statement pairs, along with the assignment difficulty label, access to the live database and several grading guidelines for correctly assessing the knowledge. These were as follows: if the statement cannot be interpreted by the DBMS, i.e. reports an error parsing the statement, label it **N** (noninterpretable); otherwise, if the statement is interpretable, then use the appropriate exclusive scenario:

- If the relation returned by the student statement, R_S, equals the relation returned by the reference statement, R_R, then grade it 100% (1). If the student statement is regarded a cheating attempt, label it **C** (cheating).
- If the student statement is completely meaningless (e.g. SELECT * FROM tblname), label it **M** (meaningless).
- If the student statement is not entirely correct, i.e. $R_S \neq R_R$, then try to assess the amount of knowledge needed for devising such a statement, in regard to the question asked, and grade the answer accordingly in a continuous scale [0%, 100%]. Be strict when grading the answers in the easy category, and be lenient when grading the answers in the hard category.

This last scenario is, obviously, the hardest and most time consuming part of the entire grading process. Separate labels for special cases of noninterpretable, meaningless and cheating attempts were introduced (and used instead of grades) in order to allow later use of various record filtering methods for tidying up the data, e.g. for possible outlier exclusion. The reviewers were advised against sharing the grades they assigned and subjective policies they practiced between them, to retain the correctness of the experiment.

C. Tidying up the data

After the grades have been assigned for every student answer, by each reviewer independently, a mean of those estimated grades was taken as ground truth. Special cases detected during the grading process (i.e. records labeled **N**, **C**, or **M**) by at least one of the reviewers were handled in the following manner. After the grading assignment process, 76 pairs were excluded from the dataset due to their noninterpretability (**N**) by the DBMS. These would (and should) be automatically graded with 0% in our live system, and are, thus, not interesting for the prediction problem. Additional 3 pairs were excluded because they were labeled as cheating (**C**). They were omitted purposefully because they are outliers in the data. If they were more numerous, however, they could be automatically detected by the predictive model learnt and then excluded. Pairs labeled meaningless (**M**) were only graded 0%, and were not excluded from the dataset. This yielded a new dataset of reduced size, consisting of 314 instances. A clearly bimodal grades distribution for this dataset is presented in Fig. 2. Discrepancies between reviewers' grade and label assignments were not taken into account for outlier detection.

Estimated grades distribution on the used dataset clearly suggests that the prediction problem is not solvable via regression. The obvious imbalance of occurrence frequencies of grades, ranging from 10% to 90%, makes it almost impossible to build a good regression prediction model, i.e. one that would estimate the exact grade percentage, ranging from 0% to 100%, based on the given answer statement. This is, clearly, an unfortunate consequence of the rather small exam database used, consisting of a majority of assignments too simple or too difficult for the average student (not justifying the obvious lack of knowledge and skill herein). Summary of the average assigned grades statistics for each question are presented in Table I.

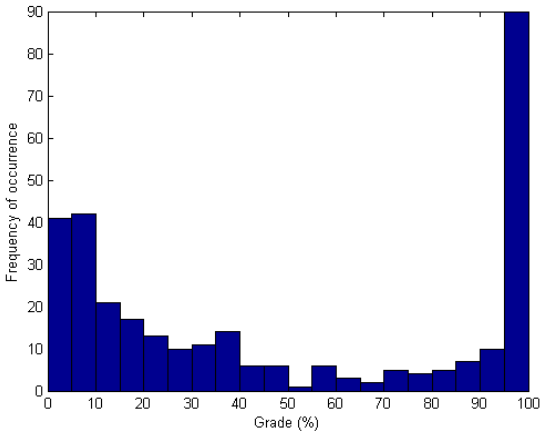


Figure 2. The average grade distribution for the dataset

TABLE I. FINAL EXAM DATASET AND GRADING SUMMARY

Question difficulty	Question ID	Number of answers	Average assigned grade summary			
			Minimum grade	Maximum grade	Median grade	Binary grade distribution 0/1 (%)
Easy	1	24	0	100	15	75 / 25
	2	24	0	100	93.33	33.33 / 66.67
	3	37	10	100	83.33	16.22 / 83.78
	4	29	0	100	96.67	27.59 / 72.41
	5	20	0	100	56.57	50 / 50
	Subtotal (134 answers)				86.67	37.31 / 62.69
Moderate	6	17	0	23.33	13.33	100 / 0
	7	28	0	100	21.67	71.43 / 28.57
	8	28	0	100	45	53.57 / 46.43
	9	28	0	100	21.67	85.71 / 14.29
	10	18	0	100	23.33	61.11 / 38.89
	Subtotal (119 answers)				18.33	73.11 / 26.89
Hard	11	19	0	100	45	57.89 / 42.11
	12	21	0	100	35	57.14 / 42.86
	13	21	0	100	18.33	90.48 / 9.52
	Subtotal (61 answers)				28.33	68.85 / 31.15
Total					35.83	57.01 / 42.99

That being said, we decided to use a binary grading system instead, based on the 50% threshold: answers graded 50% or higher were assigned label 1 (correct) and label 0 (incorrect) was assigned otherwise. This way, all the answer statements that were estimated as being closer to the true solution, were labeled as correct. Unfortunately, this grading system does not include the third, intermediate, grade, because the frequency of the grades around 50% is rather small, thus not providing enough data points for this third label. This binary system is not entirely fair for real-life application, but in this paper serves its purpose for demonstrating the general idea.

Finally, the statement dataset consisted of 314 records, each of them described with the following features: question difficulty, question ID, reference SQL statement, student-devised SQL statement and binary label (0/1). Before these could be used for building a predictive model for automated grading, additional steps were needed to ensure the provisioning of suitable metrics based on SQL statements comparison. These are described in the following section, along with the details on the predictive model used.

III. MODELS AND METHODS

In this section we first describe the data preprocessing techniques that were applied to the SQL statements in the dataset, in a stemming-like manner. The goal was to get rid of the noisy constructs in the statements, thus preparing them for the input to several string similarity metrics functions. Finally, we present the methodology used for learning the predictive model from data.

A. SQL statement preprocessing

SQL statements normally consist of a variety of different constructs and keywords, such as SELECT, FROM, WHERE, GROUP BY, HAVING, variations of JOIN, arithmetic operators, punctuation characters, brackets, table and column names, aliases, etc. Subsets of these form an integral part of every SELECT statement. Some of these constructs and keywords were removed from the statements, in order to insure greater generalization performance, e.g. to lower the noise in the data.

In each SQL statement string, both in the answer from student and in the reference statement, the following characters were omitted: ",", ";", "(", ")", "=", "<", ">", "!", "+", "-", "*". Additionally, newline characters were replaced with whitespace, and all leading and trailing whitespace was removed. Finally, all SQL aliases preceding the table column names with a dot were removed from the statements. This preprocessing step is the only one partially aware of the SQL semantic intricacies.

The choice of activities to be performed during the SQL statement preprocessing heavily depends on the type and content of statements that can be expected. The activities chosen here are far from optimal, but were more than adequate for the preparation of the used dataset.

B. Calculating learning features

Building an accurate prediction model for grading student SQL statements required developing a number of various metrics for measuring the similarity of the student-reference statement pairs. These metrics were adopted and adapted from several well-known information theory concepts [8]. We used the following functions for computing the features (S_S stands for the student-devised statement and S_R stands for the reference statement; function names in the equations are self-explanatory):

- **Normalized absolute length difference (ALD)** is the absolute difference of string lengths between the statements, divided by the length of the reference statement,

$$ALD = \frac{|\,length(S_R) - length(S_S)\,|}{length(S_R)}. \quad (2)$$

- **Normalized character Levenshtein distance (CLD)** is the smallest number of single-character edits (i.e. character insertions, deletions or substitutions) required to change one string into another string (also known as edit distance [8]), divided by the length of the reference statement,

$$CLD = \frac{LevDist(S_R,S_S)}{length(S_R)}. \quad (3)$$

- **Normalized word Levenshtein distance (WLD)** is the smallest number of single-word edits (i.e. word insertions, deletions or substitutions) required to change one string into another string, divided by the number of words in the reference statement,

$$WLD = \frac{wordLevDist(S_R,S_S)}{noWords(S_R)}. \quad (4)$$

  Whitespace is the only token used for retrieving the ordered word list from a statement.

- **Euclidean word frequency distance (EWFD)** is a distance between statement word frequency vector points in the Euclidean space, in regard to the unique words occurring in the reference statement,

$$EWFD = \sqrt{\sum_i (f_i(S_R) - f_i(S_S))^2}. \quad (5)$$

  $f_i(S_R)$ stands for the frequency (number of occurrences) of the i[th] unique word $w_i \in S_R$ and $f_i(S_S)$ stands for the frequency of the word w_i in S_s.

Only the values of the features, normalized to their respective reference value, were used, because experimental results suggested that better model generalization properties were achieved using them, as opposed to using nonnormalized values. Asymptotic runtime of the described feature calculating functions is rather small, varying from O(n) to O(n^2), where n can be the length of the longer string or the larger number of words in two strings. As the last step, values for each of the calculated metrics (separately) were linearly scaled to fit the interval [0,1], whole dataset wise, enabling us later better interpretability of the model parameters learnt.

C. Learning the predictive model from features

After the SQL statement preprocessing step, following the calculation of learning features, the adapted statement dataset consisted of a total of 314 instances, described by 6 features (question difficulty, question ID, ALD, CLD, WLD and EWFD) and a binary class, representing the estimated correctness of the student-devised statement. First two features were handled as categorical, and the remaining four as numeric (values ranging in the [0,1] interval).

We chose to use a binomial logistic regression model to illustrate the applicability of the string similarity metrics for correctly predicting the class value. This was a logical step following the abandoned multiple linear regression idea, highly capable of utilizing the linearity in the calculated string similarity metrics. Inferring the class y from a logistic regression model is done using a simple expression:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}, \quad (6)$$

where x represents the input feature vector, θ represents the model parameters vector and $h_\theta(x)$ is the probability that the inferred estimate equals 1, given x and θ. In the inference phase, classification is done simply by choosing the more probable class, i.e. if $h_\theta(x) \geq 0.5$, then $y = 1$, otherwise $y = 0$. Model parameters are learnt from data by maximizing the log likelihood,

$$L(\theta) = \sum_i (y_i \log h_\theta(x_i) + (1 - y_i) \log(1 - h_\theta(x_i))) + \\ + \lambda \sum_j {\theta_j}^2, \quad (7)$$

in respect to the actual class value of the i^{th} instance, y_i, considering the ridge regression regularization penalty λ to prevent overfitting. The actual implementation of the algorithm used a quasi-Newton method for parameter estimation, with $\lambda = 1$ [9].

IV. EMPIRICAL EVALUATION

To explore the predictive power of calculated features, we performed several experiments on their subsets, bearing in mind their usefulness for future real-life application. The experiments were conducted using stratified 10-fold cross-validation, to test the models generalization properties. Data preprocessing was performed using MATLAB, and model learning and testing using Weka [10]. The learnt models predictive efficiency was estimated calculating two well-known model evaluation measures: classification accuracy and area under receiver operating characteristic (ROC) curve. Classification accuracy is calculated as the percentage of correct predictions against the total number of predictions tested. Estimated classification accuracy that is lower or equal to the percentage of the majority (more numerous) class means that the learnt model is useless. Everything higher than that is pointing out to a possibly useful model. ROC curve is a graphical plot, commonly used for illustrating the models predictive power under various discriminative threshold values [11]. It represents the relationship between the true-positive rate, against the false-positive rate, at a given threshold. Area under the curve (AUC) is commonly used to evaluate the robustness of the model learnt. In regard to the ROC threshold value, classification accuracy is calculated for the discrimination at the threshold value of 0.5. Experimental results obtained through various feature subset setups are presented in Table II and Fig. 3.

In the first experiment, all 6 calculated features were used for grade prediction. Highest classification accuracy was achieved using this exact setup, due to differences in samples belonging to each distinct assignment (question). Good generalization properties are, however, not to be trusted here, because the class distribution balance varies significantly in the data (deviates from its mean), thus contributing to the overall accuracy with trivial models. For example, the percentage of incorrect answers for question ID 13 is 90% (compared to the mean number of incorrect answers, which is 57%). Of course, this also means that for this exact question databank, inclusion of the question ID feature is useful. Anyhow, it would make far more sense to train separate models for each question ID separately, if the data were more abundant.

TABLE II. EXPERIMENTAL RESULTS OBTAINED THROUGH 10-FOLD CROSS-VALIDATION FOR THREE DIFFERENT FEATURE SUBSETS USED

Features used	Classification accuracy (%)	Area under ROC
question difficulty, question ID, ALD, CLD, WLD, EWFD	78.66	0.853
question difficulty, ALD, CLD, WLD, EWFD	71.97	0.735
ALD, CLD, WLD, EWFD	62.10	0.664

In the second experiment, all calculated features, except question ID, were used. The exclusion of the feature led to a significant drop in calculated classification accuracy, as was expected. The overall model usefulness was retained, however, because the class distribution balance has also settled nearer to the dataset mean. The model learnt by using this feature subset is the first one that could be used on other (yet unseen) question databanks, if one was prepared to assign prescribed difficulty levels (easy, moderate or hard) to the questions contained in them, in the same manner as we did. Regression coefficients estimated from the entire dataset are presented in Table III. It is interesting to note that, although individual feature contribution implies the estimation of negative thetas (smaller values should mean higher probability of answer correctness), the model learnt suggests otherwise.

In the third experiment, which was also the last, only string similarity metrics were utilized for model learning. If one is not able to ascertain the correct level of difficulty for questions contained in the databank, this is the way to go. The additional, even more significant, drop in the estimated accuracy suggests that this model is not by far ready to be exploited in this form. We made a comment on this in the following section.

V. CONCLUSION AND FUTURE WORK

Manual grading of SQL exams is a tedious and mind-numbing ordeal, usually done by teaching assistants (those poor sods). Although some effort was made by the scientific community, towards eliminating the need for manual grading of SQL exams, none of them completely succeeded in this task. Unlike these attempts, which are highly aware of the underlying SQL semantics, we

TABLE III. ESTIMATED LOGISTIC REGRESSION COEFFICIENTS FOR THE SECOND EXPERIMENT. INPUT DATA WAS SCALED TO [0,1] INTERVAL, PRIOR TO MODEL LEARNING.

Feature	θ
question difficulty = easy	-0.9984
question difficulty = moderate	0.7057
question difficulty = easy	0.4993
ALD	1.6415
CLD	3.8415
WLD	-0.8138
EWFD	0.3033

decided to use a different approach.

In this paper we present a way of using simple string similarity metrics for comparing the student-devised SQL statement against the reference statement, for the purpose of engineering cloddish, yet useful features, which are then used for supervised learning of a predictive model from data. This model could be used for automated grading of SQL exams. String metrics used in this paper include the difference in character length of statements, the statement Levenshtein distance in terms of characters and words and, finally, the Euclidean word distance between statements. These were used for building a logistic regression model, along with two other, dataset specific features. Empirical evaluation of the proposed procedure was performed on a dataset consisting of 314 manually graded statement pairs. Obtained experimental results suggest, that the procedure shows promise, seeing that the metrics used are able to contribute to the overall predictive accuracy of the model, but is not yet usable (in this form, at least).

What could be done to further improve the predictive efficiency? For one, we should start by building a larger exam databank. Using only a handful of student solutions, for a handful of question assignments (in turn), for only one database sandbox (toy dataset) is not nearly enough for building a good predictive model, nor for the said database, nor for any other, yet unseen, database. A larger exam databank would allow us to exclude the instances, highly-disagreed upon by the reviewers during the manual grading process (these would be the outliers in the data), thus eliminating a significant amount of noise. Additionally, a larger exam databank would probably consist of a more uniformly distributed grades in percentages, or at least a larger number of grades concentrated around 50%, probably allowing us to learn a multiple linear regression model for the continuous grading scale, further increasing the usefulness of such a predictor in real-life application.

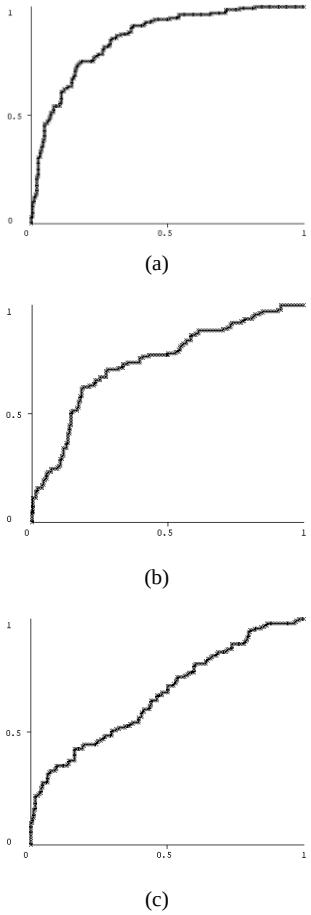


Figure 3. ROC curves plotted for the three feature subset experimental setups used: (a) question difficulty, question ID, ALD, CLD, WLD, EWFD; (b) question difficulty, ALD, CLD, WLD, EWFD; (c) ALD, CLD, WLD, EWFD. Horizontal axis represents the false positive rate, whereas the vertical axis represents the true positive rate.

Utilization of other string metrics for engineering of additional features would most likely lead to further improvements in the model's predictive accuracy. Finally, comparing the ungraded student-devised statements against those student-devised statements that are recognized as completely correct by the system in the early processing stage, in a nearest-neighbor manner, could also improve the model performance. These are the steps that we intend to follow in the future.

ACKNOWLEDGMENT

We would like to thank I. Todorović for his work on developing the original SQL exam system. A very special thanks goes to S. Jugo, F. Hržić and M. Burul, for their invaluable help with the exam statement grading.

REFERENCES

[1] D. A. Schön, "Educating the reflective practitioner: Toward a new design for teaching and learning in the professions," San Francisco, CA, Jossey-Bass, 1987.

[2] Kenny, Claire, and Claus Pahl. "Automated tutoring for a database skills training environment," in Proceedings of the 36th SIGCSE technical symposium on Computer science education, pp. 58-62, 2005.

[3] A. Mitrovic, "An intelligent SQL tutor on the web," International Journal of Artificial Intelligence in Education, vol. 13.2, pp. 173-197, 2003.

[4] J. C. Prior, and R. Lister. "The backwash effect on SQL skills grading," in Proceedings of the 9th annual SIGCSE conference on Innovation and technology in computer science education, pp. 32-36, 2004.

[5] S. Sadiq, M. Orlowska, W. Sadiq, and J. Lin, "SQLator: an online SQL learning workbench," in Proceedings of the 9th annual SIGCSE conference on Innovation and technology in computer science education, pp. 223-227, 2004.

[6] S. Dekeyser, M. de Raadt, and T. Y. Lee, "Computer assisted assessment of SQL query skills," in Proceedings of the 18th conference on Australasian database, vol. 63, pp. 53-62, 2007.

[7] C. Kleiner, C. Tebbe and F. Heine, "Automated grading and tutoring of SQL statements to improve student learning," in Proceedings of the 13th Koli Calling International Conference on Computing Education Research, pp. 157-164, 2013.

[8] G. Navarro, "A guided tour to approximate string matching," ACM computing surveys, vol. 33.1, pp. 31-88, 2001.

[9] S. Le Cessie, and J. C. Van Houwelingen, "Ridge estimators in logistic regression," Applied statistics, vol. 41.1, pp. 191-201, 1992.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I, H. Witten; "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11.1, 2009.

[11] J. A. Hanley, and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143.1, pp. 29-36, 1982.

Multi-agent Modeling Methods for Massivley Multi-Player On-Line Role-Playing Games

Markus Schatten, Igor Tomičić, Bogdan Okreša Đurić
Artificial Intelligence Laboratory
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
Email: {markus.schatten, igor.tomicic, dokresa}@foi.hr

***Abstract*—Massively Multi-Player On-Line Role-Playing Games (MMORPG) give us the opportunity to study two important aspects of computing: (1) large-scale virtual social interaction of people (players) and (2) the design, development and coordination of large-scale distributed artificial intelligence (AI). A common denominator for both aspects are the methods used to study them: social interaction can be described and simulated using agent-based models (ABM) whilst distributed AI is commonly modeled in terms of multi-agent systems (MAS). The important question to ask in both perspectives is how do agents organize in order to perform their tasks and reach their objectives? In the following paper we will present and overview of agent-based approaches to modeling MMORPG's agents including human players, artificial players (bots) and non-playing characters (NPCs). We will put a special accent on useful agent behaviors, coordination and consensus mechanisms as well as organization and social network structure of agent societies.**

***Keywords*—MMORPG, multi-agent systems, modeling, npc, bot, agent behaviors.**

I. Introduction

Role-playing computer games (commonly referred to as role-playing games or RPGs) are a computer game genre in which the player controls the actions of a protagonist (or several party members) in some world which is well defined [1]. A massively multi-player on-line game (MMOG) is a computer game that supports a great number of players playing the game on-line simultaneously often causing interaction among them [2]. Massively multi-player on-line role playing games are thus a mixture of the RPG genre with MMO allowing players to control the action of their avatar (protagonist) by interacting with a potentially large user-base on-line [3].

MMORPGs have become enormously popular during the last two decades. According to Wikipedia [4] numbers of players on some popular MMORPGs are given below:

- Dofus – 10,000,000
- World of Warcraft – 7,600,000 (on average 34,000 simultaneously on-line)
- MapleStory – 5,000,000
- Knight Online – 4,250,000
- Start Trek Online – 3,200,000
- Guild Wars 2 – 3,000,000
- Final Fantasy XIV – 2,500,000 (avg. 30,000-45,000)

The global MMO market [5], [6] is growing continuously with 2011 $\approx$ 8.5 billion €, 2012 $\approx$ 10.2 Bn€, 2013 $\approx$ 11.7 Bn€ and 2014 $approx$ 15.0 Bn€.

While the importance of MMORPGs in economic terms is obvious, another aspect is of equal importance: it allows us to study two aspects of large-scale computing - social interaction of (large numbers of) players through a computing platform as well as the design and implementation large-scale distributed artificial intelligence (in form of non-player characters – NPCs, mobs – various monsters to be fought, as well as artificial players – bots). Both of this aspects can and should be studied using agent-based techniques the former by ABM (social science perspective of agents) and the latter by MAS (computer science perspective).

In the following work in progress paper we will present an overview of agent based modeling techniques which can be applied to study MMORPGs some of which have already been applied in the ModelMMORPG project[1]. Basically we will describe three sets of methods that allow us (1) to model NPCs (finite state machines), (2) user behavior (agent based modeling of social network structures), and (3) bots (agent behaviors, automated planning techniques, consensus and coordination techniques). Our objective in future research is to study user behavior using social network analysis and natural language processing in order to establish agent based models of their behavior that will be mapped into an ontology of organizational design of large-scale MAS (developed partially in [7], [8], [9], [10]) to be used in the implementation of bots to foster organizational behavior among them.

MMORPGs have various subgenres, but a usual setting is that a protagonist is placed into a world in which he interacts with various NPCs and mobs which give out quests that have to be resolved in order to proceed to higher levels, buy better equipment or learn new skills like magic and similar. For the sake of the ModelMMORPG project the MMORPG The Mana World (TMW)[2] has been selected in order to conduct our research. It was selected due to a number of reasons: (a) it is open source (GPL licensed)) allowing us to modify code and add additional functionality, (b) it has a supportive community, (c) it supports a number of interaction techniques which can be studies (e.g. trade among players, IRC based chat, organizing teams called parties, social network functions e.g. friends, enemies, parties etc.), (d) it is a (more or less) finished game featuring lots of quests that can be analyzed.

[1] See http://ai.foi.hr/modelmmorpg for details.
[2] See http://themanaworld.org for details.

In order to study organizational behavior of players, apart from existing quests, we have partially developed a quest which fosters players to organize called "The quest for the dragon egg". The basic idea of the quest is that players have to establish parties of at least three players to transport a dragon egg to their castle to hatch it. Additionally, they have to gather a number of ingredients scattered across the Mana World to create a hatching potion. The catch is that time limits exist for each of the tasks, so they cannot be performed by a single player by himself. An egg hatches on a random location every 24 ± 1 hours, the egg has to be hatched in an hour or it goes bad. The hatching potion has to be available at the moment the egg is transported to the castle. Ingredients can be acquired by fighting mobs or by trading with other parties. Since all of these constraints are in fact (more or less) simple to implement, we use finite state machines to model NPCs needed for the quest.

The rest of the paper is organized as follows: in section II we give an overview of related work. In section III we show how NPCs can be modeled using finite state machines in TMW. In section IV we give an introduction on how to model social network behavior of players using an agent based technique. In section V we provide AI agent formalisms for modeling artificial players - bots. In the end VI we draw our conclusion and give an outline to future work.

II. Related Work

Since MMORPGs are a relatively new phenomena, there are consequently only few studies that investigate MAS and ABM approaches in MMORPGs. Most studies deal with the design of individual AI agents (NPCs). For example Eckschlager et al. [11] try to model emotional behavior through artificial neural networks and a Five-Factor Model of Personality. Tychsen & Hitchens [12] deal with the limited abilities of MMORPGs in story-telling and identify agent based story-telling [13] as a possible solution. Eladhari & Mateas [14] describe a technical framework for modeling personality and emotion for both players and NPCs called the Mind Module. They describe a semi-autonomous agent architecture built to be used in a multi-player environment, employed as a part of the player's avatar. Brom & Lukavsky [15] focus on the usage of full episodic memory in contrast to ad-hoc, or special purpose, memory, and how it can be contributed to the agent's believability.

Some investigations deal with a MAS oriented design of MMORPGs. Robert et al. [16] clarify why classifier systems (CS) are good candidates for action selection mechanisms of NPCs. CSs are used as control architectures for simulated animals or robots in order to decide what to do at each time. The authors describe different classifier systems and introduce a new CS architecture, acting in a MAS environment, which is adapted to the specific constraints of the MMORPGs. Fairclough & Cunningham [17] deal with the area of computer-mediated storytelling, and describe the "development of an expert case-based character director system which dynamically generates and controls a story, which is played out in a multi-player networked game world". The proposed system includes a story director system which utilizes the case based planning paradigm, and facilitates multi-player stories. Stories are modeled as cases, and their planning and scheduling is the primary activity of the system. This work is story-centered, and the notion of AI for the most part refers to the story director agent. Modeling of the NPCs is performed in a layered structure, from low-level behaviors to higher level targeted goals (low level such as collision detection, followed by social simulation, idle behaviors, targeted behaviors, attitudes etc.). Berger et al. [18] describe "an integrated, game-like e-Business environment that follows the role model of (...) MMORPGs", and essentially is a 3D virtual world based on game engine technology. This environment provides a platform for conducting business, and is based on MAS ideas, composed of many autonomous individuals - both human and software agents, which constitute avatars in the 3D virtual world. The goal of this environment is to support the complex interaction patterns between the members in an e-Business setting, and provide grounds for a lively and sustainable on-line community composed of providers and consumers.

Other studies employ ABM techniques to study player behavior. For example Lehdonvirta [19] is trying to apply concepts and techniques from economics to study the complicated interactions inside game worlds and is asking the question if techniques and theories from real economics be applied to virtual economics? Basic motivation for Lofgren & Fefferman [20] was a software update released into the MMORPG World of Warcraft, which created a full-blown epidemic. Players of this game encountered an extremely virulent and highly contagious virtual disease, which caused high rates of mortality and social chaos within gaming population that comes from a large-scale outbreak of a deadly disease. The authors conclude that the virtual outbreak in WoW game was both a missed opportunity to study epidemics, and a new direction for future epidemiological research. Ang & Zaphiris [21] have proposed a new method for sociability design and research through the use of simulation, and have developed an ABM to simulate and study the formation of social networks in a MMORPG guild community. Fishwick [22], on the other hand, introduces personal aspects of agents through first person perspective, inside an open-source, multi-user, virtual environment system called OpenSimulator.

While these studied do analyze the social structure and dynamics of players (trough social network analysis for example) they do not particularly target the question of how players organize in order to perform their tasks.

From this reasoning follows that there are two important aspects which aren't covered in contemporary MMORPG studies: (1) organizational techniques of artificial player's societies development, (2) study of organizational behavior of players.

III. Modeling NPCs

An NPC is an in-game character which is, most usually, used for communication with the player, serving the purpose of delivering all kinds of information to the player, e.g. basic information for new players, introduction to new areas, simple server-wide or local news service, information about a certain quest, etc. An NPC can range from a simple reactive agent to a completely interactive, environment and history immersed agent, based on the purpose it serves. Here we will consider simplest and an agent of average complexity, modeled as a finite automaton.

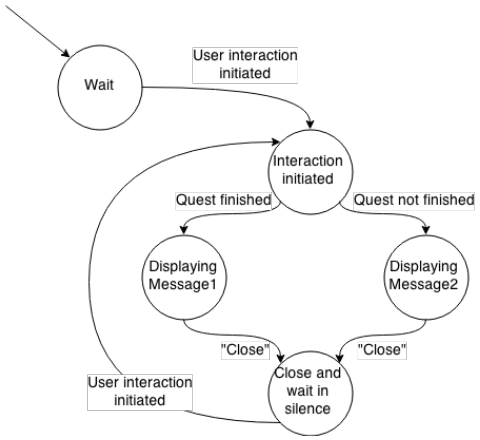


Fig. 1. Graph representation of the described agent

Fig. 2. In-game representation of the two simple interaction examples, dependent on the state of the quest

This finite state-machine needed for our goal is defined by six elements $(\Sigma, \Gamma, S, s_0, \delta, \omega)$, as follows: Σ is a non-empty symbol set called the in-alphabet, Γ is a non-empty symbol set called the out-alphabet, S is a finite non-empty set of states, s_0 is the initial state, δ is transition function defined as $\delta : S \times \Sigma \rightarrow S$, and ω is an exit function defined as $\omega : S \times \Sigma \rightarrow \Gamma$. This more general definition, when specialized for defining agents, is adapted as follows: in-alphabet is a set of possible experiences, out-alphabet is a set of possible actions, set of states is a set of agent's inner-self states, the initial state is the initial inner-self state, transition function is the next internal state function of the agent, whilst the exit is the action function that will be performed by the agent.

The less complicated version is a simple reactive agent which reacts to user-initiated interaction. Based on the status of the quest the user started with another NPC, the NPC delivers simple, static, information about the state of the world. As defined, the NPC has several defined states, activated by user interaction. Figure 1 shows an example of possible states and the transitions needed to reach them.

The in-alphabet of this agent consists of possible variable values, out-alphabet, considering it is a set of actions, gathers all the actions the NPC can take, i.d. send one of the messages to the player, set of inner NPC states are the states of this NPC, as visible in Figure 1, etc.

User interaction is started by clicking on the NPC or other

Fig. 3. In-game representation of the choices a player can make

ways of conversation initiation. The graph on Figure 1 depicts the possible state transitions, e.g. from *Interaction Initiated* state, in order to move to the *Displaying Message1* state, the quest status has to be checked. Based on this action, the agent changes its state, showing either *Message1* or *Message2*, as seen in Figure 2. Similarly, player has to click the Close button, so as to finish interaction with the NPC, and to get it ready for interaction with the next player, NPC or mob.

A more complex example is an NPC which gives out a quest, and is the sole authority on the state of the quest, depending on user interaction and a more dynamic environment, considering it interacts with the player on a more active level, and depends on some environment events, such as a timer. TMW NPCs are implemented using an domain specific scripting language, where quest process is tracked using a variable, and several code segments to be defined. These segments are accessed based on player decisions or status of variables or special elements of the environment. For example, an NPC, the Knight of Ni, after sending a default greetings message to the player, checks if this particular player is a part of a player party, and proceeds accordingly, as is shown in the code listing below.

```
if (getcharid(1)==0) goto L_Deny;
if (getcharid(1)) goto L_Allow;

L_Deny:
    (...)
L_Allow:
    (...)
```

The whole process of interaction between the user and the NPC is manipulated by using conditional expressions, such as in the following piece of code.

```
if (countitem(663) > 0 &&
    $@KnightNiShrubberyQuest == 3)
        set $@KnightNiShrubberyQuest, 4;
if ($@KnightNiShrubberyQuest == 0)
    goto L_Start;
if ($@KnightNiShrubberyQuest == 1)
    goto L_QuestState1;
```

In the code listing above, the first line checks if the player concerned has the needed item, defined by ID 663 (named *Black Rose* in the game), and if the quest state variable has value 3. If those are fulfilled, then the engine will set the value of variable *$@KnightNiShrubberyQuest* to 4. Other lines of code work in a similar manner, since the value of the quest state variable is checked, and a certain action is taken, e.g. the engine is instructed to go to the label *L_Start*, where some other code follows.

These mentioned labels are markers throughout the NPC script, which allow for non-linear script execution, and interactive communication with the user. For example, label *L_Start* contains the code which will try to attract the player to accept a quest. Although in a rather guided fashion, a player can choose the path they want to take, as shown in the code below.

```
L_Start:
    mes "We are
        Keepers of the sacred words.";
    next;
    mes "Ni!";
    next;
    mes "I shall say Ni! again if you do
        not appease me.\"";
    menu
        "You are a what?!",
            L_Ni,
        "All right!  What do you want?",
            L_Describe;
```

Figure 3 depicts the example communication done in the game environment, at the moment when the player is offered to choose their move. Direct interaction with player data, such as player inventory content, is something this game genre cannot do without. The example NPC will proceed to direct player data interaction only if the quest state variable has a certain value, which is, in turn, set by checking the player's inventory. In case of a successful check, the NPC will proceed, taking an item from the player's inventory (*delitem*) and giving them another item (*getitem*). More specifically, when certain conditions for an action are met, the NPC is going to proceed to this defined action, hence producing an effect affecting the player, as shown in code listing below. This example represents a prime example of the finite state automaton, since there is an input, a condition for an action, the action and a certain effect this action has on the environment (i.e. the player).

```
L_Next4:
    mes "[Knight of Ni]";
    delitem 663, 1;
    mes "\"Very good,
         that is a nice shrubbery.";
    next;
    mes "You shall be awarded
        with this mighty gift!";
    getitem 742, 1;
    mes "Now go!\"";
    set $@KnightNiShrubberyQuest, 5;
    close;
```

Non-player interaction with NPC is enhanced by using NPC timers and special code which activates when this timer reaches a defined point in time. In this example, the NPC is going to change value of the quest variable exactly sixty minutes after the second part of the quest is started, as seen in code below.

```
L_Next2:
    (...)
    initnpctimer;
    close;

OnTimer3600:
    set $@KnightNiShrubberyQuest, 11;
    end;
```

This particular action allows for more unpredictable and non-linear NPC behavior, when observed by the player. Introducing a random number generating function would further increase unexpected, yet completely foreseen NPC behavior.

IV. Modeling Player Behavior

In order to model the behavior of players in an MMORPG one needs to acquire statistical data about them using the game. Such logging data in TMW is available in form of chat logs, user (social) behavior logs, user inventory logs, trade logs and other logs that can be collected and analyzed. The collected data can be analyzed using various methods, but two approaches are of particular interest: social network analysis (SNA) and natural language processing (NLP) with possible help of sentiment analysis (SA) [23], [24]. SNA allows us to study the social structure of the players (interactions including chat, trade, fight etc; friendships; antagonisms; parties and similar [25]) as well as processes on networks (the dynamics of network creation [26], spreading of information [27], mixing patterns [28], [29] etc.). While SNA is very insightful in the construction and analysis of player networks, the actual meaning (semantics) of the interactions as well as their intensity can more efficiently be analyzed using NLP and SA techniques. In this way a comprehensive study of behavior can emerge that can be the input to the construction of agent based models of player behavior (see [30] for detailed examples).

A very interesting method to be used here is the so called recipeWorld brought forward by Fontana & Terna which is "an agent-based model that simulates the emergence of networks out of a decentralized autonomous interaction" [31]. The main of their approach is that modeled agents are given so called recipes that represent variable numbers of steps, some of which might run in parallel, which ought to be taken in order to achieve a given result. In this way the social network of agents emerges as a side effect of agent behavior which use recipes to achieve attended objectives. The model of recipeWorld is founded on four distinct sets: A the actual world populated by entities and their actual network; B an ABM with agents that base their behavior on the objectives and recipes derived from A; C the social network generated by B, D the data exchanged on the network. Fontana & Terna proposed that it is possible to populate sets A, B and C, by knowing D with use of inference and an approach similar to reverse engineering.

From our perspective recipes can be patterns of interactions (e.g. message passing, trade transactions and other actions) that lead to well connected subnets over some observed periods

TABLE I. THE STRIPS ALGORITHM

If on top is	Then
A complex or atomic goal which is satisfied	Remove it
A complex goal which is not satisfied	1. Leave it on top 2. Put its subgoals on top in a different order
An atomic goal which is not satisfied	Find a rule which adding list includes the goal 1. Exchange the goal with the rule 2. Put the rule's preconditions on top If no such rule exists put the preconditions on top in a different order If no such order exists find an alternative rule. If no such rule exists the, goal is not satisfiable.
A rule	1. Remove the rule 2. Update the state using the rule 3. Write down the rule (solution)
Nothing	Stop

of time. These patterns of interaction could be extracted using adequate SNA and NLP techniques and then used for simulation experiments to construct artificial agent networks. Such simulated networks could then be analyzed and compared to empirical data to see if they actually conform to real processes by using various network metrics for example. In this way automated organization techniques for agents could be established.

V. MODELING BOTS

The previous contemplation brings us to the most complex part of the project which deals with implementation of artificial agent players. In order for a an artificial agent to play a MMORPG a number of MAS and AI techniques have to be employed. The agent has to have the ability to connect to the game server, learn about its environment, understand instructions from NPCs and mobs, solve quests and interact with other (possibly artificial) players and organize to solve harder (social) quests. For the first part, we have partially implemented a low level client to the TMW server in the Python scripting language that allows us to connect an agent to the server, move around, collect items and interact with other players, NPCs and mobs.

In order to implement intelligent agents we will use the Smart Python Agent Development Environment (SPADE) [32] that provides a FIPA (Foundation for Intelligent Physical Agents) compliant development environment for Python and features a number of predefined agent behaviors, advanced knowledge bases and agent organization facilities like group chat and organizational units. We will use SPADE to build an agent layer above the low level interface that will allow agents to understand their surroundings (by using some of the provided advanced knowledge bases), solve quests (by implementing a belief, desire, intention – BDI agent model and an automated planning algorithm e.g. STRIPS [33] see table I) and interact with their environment (using the integrated communication and organization facilities).

Individual agents are usually defined by a number of behaviors which might be [34]:

- **role factory** (a role added at runtime and then enacted by the agent);
- **itinerary** (allows mobile agents to travel across various locations and perform tasks);
- **periodic** (looped behavior possibly with a given period of time intervals between iterations);
- **observer** (an agents awaits an event in order to perform its actions);
- **listener** (a special type of observer in which an agent awaits a special message of some other agent);
- **client/server** (resembles the client-server model);
- **one-shot behavior or task** (represents a simple task or activity);
- **finite state machine** (resembles a finite state machine in which every node is an activity to be performed);
- **sequential behavior** (a sequence of other behaviors);
- **parallel** (various behaviors are run in parallel).

Most of these behaviors will have to be used to implement a working bot (except maybe for itinerary, since our agents will not have to travel across the network). Role factories come to use in organizational behavior of agents. For example when an agent joins a party he might be assigned a role to act as the carrier of some item (e.g. the dragon egg) and has to enact this role (e.g. has to keep line of sight with two other party members). The periodic behavior can be used in multiple different occasions for example to harvest a given resource mobs of a given type have to be fought. The agent will loop this behavior until it has enough of a given resource. An observer behavior is likewise important – an agent will observe its environment and in case of an event (e.g. an attack of some mob) react accordingly. The listener behavior is similar – if a bot receives a message from another player it might choose to react. Client/server behavior are especially important for trading – (server) agents will offer some items and wait for other (client) agents to pay for them in order to deliver the good. Task behaviors are ubiquitous – every action (moving, attacking, talking etc.) are simple tasks to be performed. FSM behaviors on the other hand might come in handy to deal with various states of quests – e.g. when the player solves one part of the quest (by for example bringing some item to some NPC) it moves to another state (e.g. finding some other item). Sequential behaviors are important for the planning system – all plans will be sequences of (simpler) actions. Parallel behaviors on the other hand are obvious – while solving some task an agent has to be alert about events in the environment, thus the task behavior will run in parallel with some kind of observing behavior.

VI. CONCLUSION

In this work in progress paper we have shown some important agent based modeling techniques that allow us to study MMORPGs with a special accent on techniques to be employed in the ModelMMORPG project in which the TMW MMORPG has been chosen for study. We have firstly shown how to implement NPCs using finite state machines and how they can be implemented in the scripting language of TMW. Later we have provided some recommendations on how to collect, analyze and model behavioral and social player data using SNA, NLP, SA and ABM techniques. Based on the proposed ABM technique to model the social and organizational behavior of players, we have contemplated that artificial agents might implement their consensus and coordination techniques

in the same way by using the recipeWorld approach. In the end we have shown some important modeling techniques for implementing artificial players – bots – by providing important agent behaviors as well as examples of their usage.

Out future research is aimed towards implementing the (dragon egg) quest that will foster organizational behavior of players. Afterwards we will conduct a large scale study with the developed quest by collecting server logs that will be analyzed in the described way. The analysis will yield agent based models of organizational behavior among players that we will embed into an ontology of organizational design techniques for the development of large-scale MAS. In the end we shall implement artificial agent organizations and test their performance against human players on the TMW platform.

ACKNOWLEDGMENT

This work has been supported in full by the Croatian Science Foundation under the project number 8537.

REFERENCES

[1] Wikipedia Contributors, "Role-playing video game — wikipedia, the free encyclopedia," 2015, [Online; accessed 9-February-2015]. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Role-playing_video_game&oldid=642919862

[2] ——, "Massively multiplayer online game — wikipedia, the free encyclopedia," 2015, [Online; accessed 9-February-2015]. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Massively_multiplayer_online_game&oldid=646153954

[3] ——, "Massively multiplayer online role-playing game — wikipedia, the free encyclopedia," 2015, [Online; accessed 9-February-2015]. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Massively_multiplayer_online_role-playing_game&oldid=642752380

[4] ——. (2014) Comparison of massively multiplayer online role-playing games. [Online]. Available: http://en.wikipedia.org/wiki/Comparison_of_massively_multiplayer_online_role-playing_games

[5] Global Collect. (2014) The global mmo games market: Payments, intelligence and trends. [Online]. Available: http://www.globalcollect.com/the-global-mmo-games-market

[6] NewZoo. (2012) The global mmo market; sizing and seizing opportunities. [Online]. Available: http://www.newzoo.com/infographics/the-global-mmo-market-sizing-and-seizing-opportunities/

[7] M. Schatten, "Active graph rewriting rules for modeling multi-agent organizational dynamics," in *Proceedings of the IBC 2012, 1st International Internet & Business Conference*, M. Ivković, M. Pejić Bach, and V. Šimičević, Eds. Rovinj: BIT Society, 2012, pp. 180–185.

[8] ——, "Reorganization in multi-agent architectures: An active graph grammar approach," *Business Systems Research*, vol. 34, no. 1, pp. 14–20, 2013.

[9] M. Schatten, P. Grd, M. Konecki, and R. Kudelić, "Towards a formal conceptualization of organizational design techniques for large scale multi agent systems," *Procedia Technology*, vol. 15, pp. 577–586, 2014.

[10] M. Schatten, "Organizational architectures for large-scale multi-agent systems' development: An initial ontology," *Advances in Intelligent Systems and Computing*, vol. 290, pp. 261–268, 2014.

[11] M. Eckschlager, R. Bernhaupt, and M. Tscheligi, "Nemesys: neural emotion eliciting system," in *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2005, pp. 1347–1350.

[12] A. Tychsen and M. Hitchens, "Ghost worlds–time and consequence in mmorpgs," in *Technologies for Interactive Digital Storytelling and Entertainment*. Springer, 2006, pp. 300–311.

[13] R. Aylett, S. Louchart, J. Dias, A. Paiva, M. Vala, S. Woods, and L. Hall, "Unscripted narrative for affectively driven characters," *Computer Graphics and Applications, IEEE*, vol. 26, no. 3, pp. 42–52, 2006.

[14] M. P. Eladhari and M. Mateas, "Semi-autonomous avatars in world of minds: A case study of ai-based game design," in *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*. ACM, 2008, pp. 201–208.

[15] C. Brom and J. Lukavsky, "Towards virtual characters with a full episodic memory ii: The episodic memory strikes back," in *Proc. empathic agents, AAMAS workshop*, 2009, pp. 1–9.

[16] G. Robert, P. Portier, and A. Guillot, "Classifier systems as' animat'architectures for action selection in mmorpg," in *3rd International Conference on Intelligent Games and Simulation, at London*, 2002.

[17] C. Fairclough and P. Cunningham, "A multiplayer case based story engine," Trinity College Dublin, Department of Computer Science, Tech. Rep., 2003.

[18] H. Berger, M. Dittenbach, D. Merkl, A. Bogdanovych, S. Simoff, and C. Sierra, "Playing the e-business game in 3d virtual worlds," in *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*. ACM, 2006, pp. 333–336.

[19] V. Lehdonvirta, "Virtual economics: applying economics to the study of game worlds," in *Proceedings of the 2005 Conference on Future Play (Future Play 2005), Lansing, MI*, 2005.

[20] E. T. Lofgren and N. H. Fefferman, "The untapped potential of virtual game worlds to shed light on real world epidemics," *The Lancet infectious diseases*, vol. 7, no. 9, pp. 625–629, 2007.

[21] C. S. Ang and P. Zaphiris, "Simulating social networks of online communities: simulation as a method for sociability design," in *Human-Computer Interaction–INTERACT 2009*. Springer, 2009, pp. 443–456.

[22] P. A. Fishwick, "An introduction to opensimulator and virtual environment agent-based m&s applications," in *Simulation conference (WSC), proceedings of the 2009 winter*. IEEE, 2009, pp. 177–183.

[23] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[24] M. Schatten, "Opinion mining on news portal comments – towards a Croatian sentiment database," in *Seminar za metodologijo in informatiko*. Novo Mesto: Fakulteta za informacijske študije, 2012.

[25] S. Wasserman and K. Faust, *Social Network Analysis ; Methods and Applications*, ser. Structural analysis in the social sciences. Cambridge University Press, 1994.

[26] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[27] Y. Moreno, M. Nekovee, and A. F. Pacheco, "Dynamics of rumor spreading in complex networks," *Physical Review E*, vol. 69, no. 6, p. 066130, 2004.

[28] M. E. J. Newman, "Mixing patterns in networks," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 67, no. 2, pp. 1–13, 2003.

[29] R. Fabac, M. Schatten, and T. Đuričin, "Social network mixing patterns in mergers & acquisitions-a simulation experiment," *Business Systems Research*, vol. 2, no. 1, pp. 36–44, 2011.

[30] M. Schatten, J. Ševa, and B. Okreša Đurić, "An introduction to social semantic web mining & big data analytics for political attitudes and mentalities research," *European Quarterly of Political Attitudes and Mentalities EQPAM*, vol. 4, no. 1, pp. 40–62, 2015.

[31] M. Fontana and P. Terna, "From agent-based models to network analysis (and return): the policy-making perspective," in *SwarmFest*. University of Notre Dame, 2014.

[32] M. E. Gregori, J. P. Cámara, and G. A. Bada, "A jabber-based multi-agent system platform," in *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM, 2006, pp. 1282–1284.

[33] R. E. Fikes and N. J. Nilsson, "Strips: A new approach to the application of theorem proving to problem solving," *Artificial intelligence*, vol. 2, no. 3, pp. 189–208, 1972.

[34] T. Marian, B. Dumitriu, M. Dinsoreanu, and I. Salomie, "A framework of reusable structures for mobile agent development," in *Proceedings of IEEE International Conference on Intelligent Engineering Systems (INES2004)*, Cluj-Napoca, Romania, 2004, pp. 279–284.

Temporal Dynamics of Changes in Group User's Preferences in Recommender Systems

Bakir Karahodža, Dženana Đonko, Haris Šupić
University in Sarajevo, Faculty of Electrical Engineering
Bosnia and Herzegovina
bakir.k@gmail.com, dzenana.donko@etf.unsa.ba, haris.supic@etf.unsa.ba

***Abstract*—Using contextual information in recommender systems is a subject of continuous improvement of rating prediction accuracy. Among others, information on temporal rating dynamics contain valuable data that establish foundation for discovering changes in both individual and group user's preferences. Such changes can be caused by multiple factors such as changes of individual user interests, changes in item popularity or other hidden patterns or events. In this paper an improved user-based collaborative filtering algorithm is presented that utilizes changes of group user's preferences over time. We also investigate temporal dynamics of changes in user's preferences within different item categories and propose time weight function that improves prediction accuracy of recommender systems.**

***Index Terms*—Recommender systems, collaborative filtering, contextual information, temporal dynamic, user preferences, hybrid model**

I. Introduction

The general task of recommender systems is defined as finding item i_j that maximizes a given user's utility or optimizing the lifetime of user's utility [1]. The recommendation systems differentiate themselves by the type of knowledge they exploit, e.g. product knowledge or some other domain expert's heuristics or by different strategies the apply. The common strategies in recommender systems are [1]: *Content-based* (CB) recommendation, that recommends items similar to those preferred by the user in the past; *Collaborative filtering* (CF), which recommends items preferred in the past by similar users, using approaches with item-based or user-based filtering; and *Hybrid recommenders*, that combine both CB and CF in order to overcome particular limitations of each individual strategy.

Most of the traditional recommender systems exploit only user and item profiles associated to known ratings in order to predict ratings of unseen items [1], and they consider data collection as static. In such scenario, the ratings produced in the past used to predict future preferences are weighted equally. The fact that user preferences are naturally changing over time is not taken into consideration, leading to a reduced accuracy of recommender systems. The occurrence of changes in user preferences is known as the concept drift or concept shift [2].

Over the recent years, various approaches that cope with this kind of problems have been developed. The most common approach is discounting old rating data with a certain weight depending on the rating time. In collaborative recommender systems, the change of individual user preferences also affects preferences of the neighbor's users used to predict the unseen ratings of individual user, and vice versa. They affect each other indirectly due to fact that the group of similar users in CF is selected based on the user's individual ratings. By adding the item content or item preferences into the consideration, it's also possible to trace and exploit temporal dynamics of individual and group user's preferences over different item content, e.g. item categories. The item content adds an additional dimension to CF approach, and moves the model from pure CF to a hybrid recommender.

In this work is presented an improved user-based collaborative filtering recommender system using temporal dynamics combined with hybrid model that traces the ratings over item categories, thus embeds the item content into the user-based collaborative filtering model.

The rest of the paper is organized as follows: in Section 2 is provided background on the temporal dynamics problem in CF and hybrid recommender systems and its solutions. In Section 3 are described details of the proposed model for improvement. Section 4 explains experimental settings, evaluation metrics and results of different experiments conducted in order to evaluate accuracy of the proposed model. The results are evaluated on a movie data set using movie genres as item categories. The Section 5 presents conclusion remarks with a discussion and directions for future work.

II. Related work

Examples of previous work identify repetitive patterns over time using periodic or discrete time information [3]. There are also works that do not predict ratings based on the future recommendation time. Instead of considering the future prediction time, they simply adopt some of the parameters according to changes in rating data over time.

The most common approaches for adoption of concept drift in CF models generally penalize older ratings presuming them as less relevant at the recommendation time. The set of different models is used, based either on discrete time windows [4], or as a weight function in [3] and [5].

The weight function is defined in [5] as a monotonic decreasing time-function $f(t) = exp(\lambda t)$ in the recommendation step, thus giving more weight to recent item ratings. In function f, t is the time of the rating and λ is a parameter that controls the decay relative to rating time in item-based CF.

Other approaches deliver a series of user ratings, establishing current user interest for individual items or group of items. In [6], ratings are grouped using an interest measure that takes into account item similarity, leading to a unique time series for each user. The users' rating history is divided into several periods and then quantizied by each user's interest in [7], setting a time-window to find users' recent interest. A time-driven filtering strategy is presented in [8] that links items and their features to time functions corrected by temporal curves built from individual user consumption histories. In [9] is proposed weight function dependent on changes in group user's preferences in user-based CF approach.

On the other side, hybrid approaches bring item content to CF methods. In order to select different neighbors for each different target item, a new similarity function for item-based CF is proposed in [10], giving a weight depending on the similarity between the item and the target item. Another hybrid approach is provided in [11] using matrix factorization technique in implementation of item-based collaborative filtering algorithm. In [12] is also proposed item-based collaborative filtering algorithm by combining the information of items categories with item-item interestingness.

However, the most of the related work and models that include temporal information are based on item-based CF methods and do not consider concept drift within a group user's preferences.

III. Proposed Recommender System

Our proposed recommender system considers temporal dynamics in user-based CF recommender system within item categories and proposes model that improves prediction accuracy. The basic idea includes the following assumptions:

a) similar users follow the same concept shift in an item ratings,

b) similar users also follow the same concept shift within the item categories,

c) the recent ratings are more relevant.

Suppose all ratings on an item i during a time period are presented in Fig.1, where the x-asis is the rating time in days, and y-asis represents the given ratings. From all of the ratings, we can mark out the neighbors' ratings N for an user u.

We can see that all neighbor's ratings follow a certain rating trend over time until τ_0, that is a recommendation time in an ideal case. Since an item can belong to one or more item categories, each neighbor's ratings can be divided into corresponding set of ratings over item categories. In Fig.1, from all of the neighbor's ratings, we recognize subsets of ratings $N(C1)$ and $N(C2)$ for categories $C1$ and $C2$ of the item i.

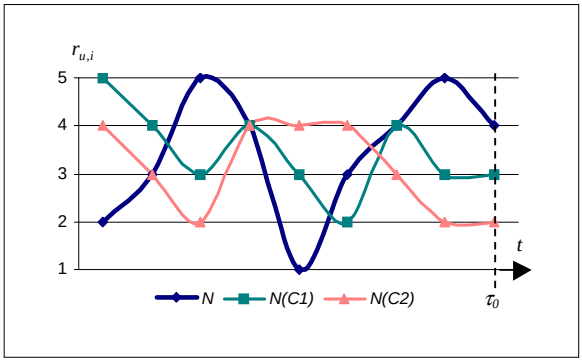


Fig. 1. Trend in neighbors' ratings over item categories

We can see that they also follow a certain trend over time until recommendation time τ_0, but obviously different than the ratings N.

Our goal is to recommend items based on the recorded ratings until the recommendation time τ_0. Based on the assumptions above, we propose an improved user-based collaborative recommender system built from following the steps.

A. Selecting a set of neighbors within item categories

We first find a set of neighbors for each user calculating similarity between users. The common approach for finding similarity P between two users a and u in standard user-based CF is calculated using Pearson correlation coefficient [13]:

$$P_{a,u} = \frac{\sum_{i=1}^{m}\left(r_{a,i} - r_a\right)\cdot\left(r_{u,i} - r_u\right)}{\sqrt{\sum_{i=1}^{m}\left(r_{a,i} - r_a\right)^2 \cdot \sum_{i=1}^{m}\left(r_{u,i} - r_u\right)^2}} \quad (1)$$

where $r_{a,i}$ is rating given to an item i by user a; $r_{u,i}$ is rating given to an item i by user u; r_a is mean rating given by user a; r_u is mean rating given by user u; m is a total number of items.

To accomplish our assumptions a) and b), instead of calculating similarity between users using the common user-item rating matrix, we deliver a set of particular similarity measures for each item category:

- $P_{a,u}(C1)$, $P_{a,u}(C2)$, $P_{a,u}(C3)$, ..., $P_{a,u}(Ck)$

where $C1$, $C2$, ..., Ck are the item categories. The particular similarity measures are used in the following step.

B. Modifying prediction function

The prediction function in traditional user-based CF recommender system is [13]:

$$p_{a,i} = r_a + \frac{\sum_{u=1}^{n}\left(r_{u,i} - r_u\right)\cdot P_{a,u}}{\sum_{u=1}^{n} P_{a,u}} \quad (2)$$

where $p_{a,i}$ is rating prediction for active user a for item i; r_a is the mean rating given by active user a on all items; r_u is the mean rating given by neighbor user u on all items; $r_{u,i}$ is rating

given to an item *i* by neighbor user *u*; $P_{a,u}$ is similarity between users *a* and *u*, and *n* is a total number of neighbors.

To extend our assumptions a) and b) on the prediction step, we use the similarity measures between users from the previous step. Instead of using a global similarity measure between users $P_{a,u}$ in the prediction step, we use a linear combination of calculated similarity measures depending on item categories to which an item *i* belongs:

$$P_{a,u}(i) = \sum_{j=1}^{k} a_j \cdot P_{a,u}(C_j) \tag{3}$$

where C_j are categories of the item *i*.

In case all item categories are equally relevant and there is no declared relationship between them or item taxonomy, the coefficients a_j are equal to 1.

For example, if categories of an item *i* belong to a set *{C1, C2, C3}*, and similarities between user *a* and the neighbor's user *u* over categories C1, C2 and C3 are:

- $P_{a,u}(C1)$ = 0.23,
- $P_{a,u}(C2)$ = 0.08,
- $P_{a,u}(C3)$ = -0.11,

respectfully, then the corresponding similarity is:

- $P_{a,u}(i)$ = 0.23 + 0.09 - 0.11 = 0.21.

The other values from the Equation 2, such as the mean rating r_a and r_u are taken over all available item ratings, without considering the item categories nor other item content. The prediction function from Equation 2 is then modified as:

$$p_{a,i} = r_a + \frac{\sum_{u=1}^{n} \left(r_{u,i} - r_u\right) \cdot P_{a,u}(i)}{\sum_{u=1}^{n} P_{a,u}(i)} \tag{4}$$

C. Adding time weight function

As previously stated in our assumption c), we assume that the recent ratings are more relevant then those given in the past. Consequently, we conclude that the most recent ratings from neighbor's users are closest to a future preference of an active user a. As proposed in [9], we also add time-dependent weight to each neighbor's rating in the preference prediction step as a function of neighbor's rating time. The Equation 4 is then modified as:

$$p_{a,i} = r_a + \frac{\sum_{u=1}^{n} \left(r_{u,i} - r_u\right) \cdot P_{a,u}(i) \cdot W(t(r_{u,i}))}{\sum_{u=1}^{n} P_{a,u}(i)} \tag{5}$$

where $W(t(r_{u,i}))$ is the time-dependent weight function. The function is calculated on the base of difference (in days) between recommendation time τ_0 and the time of actual rating of neighbor *u* on an item *i*. Consequently, the weigh function is expressed as:

$$W(t(r_{u,i})) = \exp[-(1/\lambda) \cdot (\tau_0 - t(r_{u,i}))] \tag{6}$$

where $t(r_{u,i})$ is the time of actual rating of neighbor *u* on an item *i*, and λ is a parameter for accurate performance tuning. As in [9], where the weight function is a monotonic time decreasing function based on the neighbor's rating time, we introduce the same function, but here dependant on the recommendation time, and also combined with the item content. The weight function takes values from the range [0,1]. The more the neighbor's rating time deviate from the recommendation time, the lower is the weight of the rating. The λ parameter and its impact on the weight function and prediction accuracy are discussed in the further section.

IV. Experiments and Results

Following creation of the proposed recommender model as described in previous section, a set of experiments is performed in order to examine accuracy of the system. The main goal in experiments is to compare accuracy of the proposed model with standard user-based CF model.

A. Experimental Setup

In this section experimental design and settings are described for validating the proposed recommender model with temporal dynamics over item categories, including the overview of the used data, selected variables, and performance measures.

Data. We used data from HetRec2011[1] data set that is broadly used in recommender systems research projects. The HetRec2011 data set contains 855,598 ratings from 2,113 users for 10,197 movies. Each movie is categorized by one or more of 19 genres, e.g. Action, Animation, Documentary, Drama, Romance, Sci-Fi, Thriller, etc. The genres assigned to movies are considered as the item categories in the experiment.

Data adjustment. The data set is prepared in order to adjust to a nature of the model that utilizes temporal dimension. The data is ordered by the time of the rating. The frequency of the ratings across the time is analyzed and then chosen the subset of the entire ratings data. The rating frequency is presented in Fig.2.

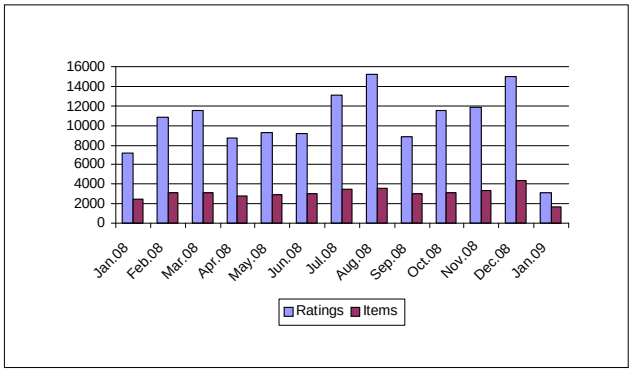


Fig. 2. Rating frequency

From the entire set of the ratings, a newest subset is selected for experiment with 135,555 ratings for period between 20-Jan-2008 and 05-Jan-2009. The subset is

[1] http://grouplens.org/datasets/hetrec-2011/

partitioned into the training and testing data. The training data used to build correlations between users contains 90% of the ratings in order to provide as much closer recommendation time to the rating time. The testing data used in the prediction step contains remaining 10% of ratings. The full description of the data set is presented in Table I.

TABLE I. DATA SET ADJUSTMENT

	Period		Ratings	Users
	From	*To*		
Training – Correlations	20-Jan-08	12-Dec-08	122,297	1,141
Testing – Predictions	13-Dec-08	05-Jan-09	13,258	443

Variables. For each experiment, the following variables are tested to ensure the most accurate predictions in the preference prediction step:

- Neighborhood size,
- Minimum number of contributing neighbors which affects the coverage of the predictions, and
- λ parameter which controls the slope of the weight function.

Performance measures. The performance measures used in experiments are based on two common evaluation metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The MAE is defined as the average absolute difference between predicted ratings and their true user-specified values [13]:

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N} \tag{7}$$

where p_i represent the predicted rating of an item i; q_i represents the user-specified rating values; N represents the number of user's ratings.

The RMSE metrics that puts more emphasis on large absolute errors [13]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (p_i - q_i)^2}{N}} \tag{8}$$

The prediction results are compared with standard user-based CF recommender system in two steps. In first step, only the time-dependant weight function from Section III.C is added to the proposed model (CF-Td) and compared to CF. Thereafter, the linear combination of similarity measures over item categories is added and tested as the final proposed model (CF-Td-IC).

B. Experimental Results

Based on the experimental setup, a set of experiments is performed across different values of input parameters in order to provide the best performing model. These resulted with 30 comparisons for the given data set. The first set of experiments is performed across different values of λ parameter for CF-Td model and compared to traditional CF model.

The second step of experiments is performed on CF-Td-IC model with a subset of experiments testing different values of neighborhood size. The adjustment of different values resulted with conclusion that accuracy increases with the size, as presented in Fig. 3, which is opposed to most of the previous work [13].

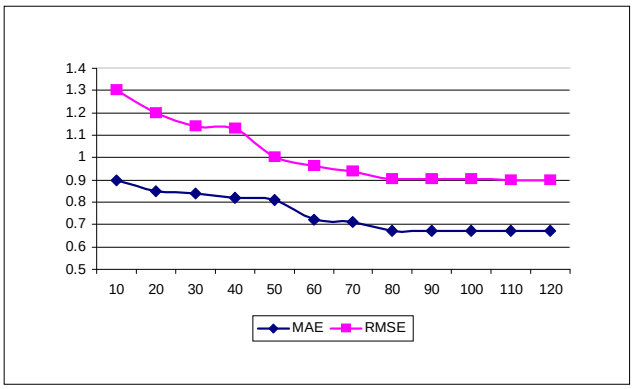


Fig. 3. Testing accuracy on neighborhood size for CF-Td-IC

Having the satisfied results with the given neighborhood size presented in Fig. 3, for all final experiments this variable is set to 100.

The minimum number of contributing neighbors is also tested with different values that affects the prediction accuracy and the coverage of the predictions. For each user-item pair that can not find the given minimum number of contributing neighbors (MinCN) due to sparse rating matrix, the prediction is rejected. Although the condition increases accuracy, the prediction coverage consequently decreases. For final experiments, the variable is tested with two values: 5 and 10, to present changes in MAE, RMSE, and recommendation rating coverage.

Although the λ parameter has less impact on the performance, it's also tested on different values and finally set to 40.

The results presented in Table II show that the CF-Td model outperforms CF model with the given MinCN values. The coverage is same in both models, but becomes poor with MinCN = 10.

TABLE II. EXPERIMENTAL RESULTS

Model	MinCN	MAE	RMSE	Coverage
CF	5	0.8017	1.0677	72%
	10	0.7524	0.9921	60%
CF-Td	5	0.7185	0.9598	72%
	10	0.6887	0.9226	60%
CF-Td-IC	5	0.6993	0.9320	52%
	10	**0.6702**	**0.9049**	43%

The CF-Td-IC model also outperforms CF and CF-Td models in prediction accuracy, especially for value of MinCN = 10, as presented in Fig. 4.

Although the prediction coverage is less than good if compare with CF and CF-Td, the prediction shows very good result with MAE = 0.6702 and RMSE = 0.9049.

Fig. 4. Results for MinCN = 10

V. Conclusions and Future work

In this work is presented an improved user-based CF recommender system that exploits changes of group user's preferences over time. It also utilizes temporal dynamics of changes within different item categories and proposes time weight function for prediction of ratings for unseen items.

The main contribution in this work is in utilizing changes of group user's preferences over time. The proposed model has shown improved results over previous work.

The future challenge of our work is to deal with the future prediction time by utilizing different item content and different recommendation context.

VI. References

[1] Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. ACM Transactions on Information Systems 23(1):103-145. 2005.

[2] Koren Y., Collaborative Filtering with Temporal Dynamics, Communications of the ACM, Vol. 53 No. 4, Pages 89-97, 2010.

[3] Baltrunas L. and Amatriain X. Towards time-dependant recommendation based on implicit feedback. In CARS 2009, New York, NY, USA, 2009.

[4] Nasraoui, O., Cerwinske, J., Rojas, C., Gonzalez, F.A.: Performance of recommendation systems in dynamic streaming environments. In: Proceedings of the Seventh SIAM International Conference on Data Mining, Minneapolis, Minnesota, USA. SIAM, 2007.

[5] Ding Y., Li X., Orlowska, M.E. Recency-based Collaborative Filtering. In: Proceedings of the 17th Australasian Database Conference - Volume 49 (ADC'06), pp. 99-107, Australian Computer Society, Inc., Darlinghurst, Australia. 2006.

[6] Cao, H., Chen, E., Yang, J., Xiong, H. Enhancing Recommender Systems Under Volatile User Interest Drifts. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09), pp. 1257-1266, ACM, New York, NY, USA. 2009.

[7] Yuchuan Zhang, Yuzhao Liu, A Collaborative Filtering Algorithm Based on Time Period Partition, Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on, 777 - 780, 2010.

[8] Y. Blanco, M. López-Nores, J. Pazos-Arias, J. García-Duque, An improvement for semantics-based recommender systems grounded on attaching temporal information to ontologies and user profiles, Engineering Applications of Artificial Intelligence. Volume 24, Issue 8, pages 1385-1397, 2011

[9] Karahodža B, Šupic H., Đonko Dž., An Approach to design of time-aware recommender system based on changes in group user's preferences. In: X International Symposium on Telecommunications - BIHTEL 2014, Sarajevo, Bosnia and Herzegovina, 2014.

[10] K. Choi, Y. Suh, A new similarity function for selecting neighbors for each target item in collaborative filtering, Knowledge-Based Systems, Volume 37, pp. 146-153 , 2013

[11] A.B. Barragáns-Martínez, E Costa-Montenegro, J.C. Burguillo, M.Rey-López, F.A. Mikic-Fonte, A.Peleteiro, A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition, Information Sciences, Volume 180, Issue 22, pp. 4290-4311, 2010

[12] S. Wei; N. Ye; S. Zhang; X. Huang; J. Zhu, Item-Based Collaborative Filtering Recommendation Algorithm Combining Item Category with Interestingness Measure, Computer Science & Service System (CSSS), 2012 International Conference on, pp.2038,2041, 2012.

[13] J. Herlocker, J. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 230-237, 1999.

Predicting sports results using latent features: a case study

Stefan Dobravec
University of Ljubljana, Faculty of electrical engineering, Ljubljana, Slovenia
stefan.dobravec@fe.uni-lj.si

Abstract - Predicting sports results is normally a challenging task, even more in case of a sport that shows a highly stochastic nature. In football, for example, numerous features are tracked and combined with expert knowledge, yielding various predicting algorithms. Our work however, is based on a case where there is no expert knowledge available and the only data comes from previous match results. We built a goal score prediction model that uses latent features obtained from matrix factorization process. We also added a Naive Bayes Classifier to be able to predict outcome of the match. The algorithm has been tested on results of the FIFA World Cup 2014. We also built a match result predictor based on the betting quotas. As these are derived from a complex algorithms that encompass also the expert knowledge, our algorithm can be used to estimate accuracy of an expert knowledge-based system. This case study shows that there is no significant difference between the two algorithms that we tested and that the latent features may provide a valid substitute for real features, when the later ones are not available.

I. INTRODUCTION

Sport result prediction can be viewed as one of the objectives of sport analytics, which aims at helping decision makers (e.g. team managers, odds traders, etc.) to gain competitive advantage. The difficulty of this task depends on many factors, like the availability of data for the past events, the ability to gather data for future events, the knowledge needed to interpret gathered data, and others. In some cases, like in football, the task becomes even more challenging due to the highly stochastic nature of the game. At the same time, however, strategies of a team can be approximated by crisp logic rules [1].

Various techniques for modelling a football match exist that yield different result prediction algorithms. According to [2] the modelling can be put under the four generic categories: (i) empirical models, (ii) dynamic systems, (iii) statistical techniques, and (iv) artificial intelligence (including expert systems). The most basic approach in the category of statistical techniques is to use Poisson distribution for goal-based data analysis [3]. While Poisson distribution is used to predict the number of goals conceded and scored, the other statistical models restrict their prediction to match outcome. However, it has been shown in [4], that both approaches yield similar performance. In the artificial intelligence category there are several approaches that focus on Bayesian network modelling [1] [5] [6]. In most cases these approaches are highly complex, based on numerous assumptions, and require large statistical data [7]. On the other hand, they enable inclusion of expert knowledge, thus making them in general more accurate [5]. The remaining two categories are poorly represented in recent years.

A comparison of the described approaches is hindered due to the fact that they do not use a common dataset. In fact, neither a standard nor an agreed dataset exists. An analysis of football match parameters is performed in [8] pointing out features that influence the game, however various approaches use various sets of parameters. This is at least partly due to the high dynamics of a football match, which makes it hard to systematically gather statistical data. Also, different competition formats (league, tournament) emphasize different parameters or their combinations.

Our research, in contrast, is based on latent features of a matrix factorization model instead of relying on a real (observed) features. We assume that it is better to use latent features that in best way possible reflect the variability of the observed facts, than to rely on a poor set of real features that can't be properly assessed. The proposed method is thus based solely on results of past matches (the number of goals scored by each team), which represent the fundamental data that is recorded in all cases. No other statistical data or expert knowledge is used.

The Matrix factorization technique became very popular in the field of multimedia content recommender systems where it showed good scalability and predictive accuracy [9]. The idea behind using the latent features in our case is to be able to build a successful model even when expert knowledge is not available or when large statistical data is not available.

The proposed method has been tested on the case of FIFA World Cup 2014 tournament. Since different approaches use different techniques to measure the success, direct comparison with our method would be misleading. Instead, we used cross-folding to statistically evaluate the prediction success of our method. We also compared the proposed method to a match result classifier that is based on betting quotas, which indirectly encompass the expert knowledge of odds traders.

In the following sections first the methodology for predicting match results is explained. Following is the explanation of the validation procedure. Next, the results are presented, followed by discussion and conclusions.

II. MATERIALS AND METHODS

A. FIFA World Cup 2014 Dataset

Using World Cup tournament results as the dataset allowed us to observe the short-term performance of competing teams. Thus we were able to exclude time factor from our model, in contrast to a league-type competition where teams compete over long period, which increases the importance of the time factor. We assume that teams enter the tournament in peak condition, with a fixed set of players, and team strategies planned ahead. Therefore major changes in team's performance are not expected.

A major drawback of the dataset is its size, which is quite small. The dataset could be increased by including results of previous World Cup tournaments, however the list of competing teams varies through tournaments. Thus only the number of teams would increase and the number of matches per team would increase only for those teams that qualify regularly. Additionally, the time factor would play an important role as the players, coach, and strategy may change significantly in-between the two tournaments. A sophisticated strategy for handling such dataset, requiring significant research on its own, would be required.

In the World Cup teams compete in the tournament format. In the group stage 32 teams are put in groups of four (altogether 8 groups). Each team plays three matches (rounds) against their group adversaries, which accounts for 48 matches overall in the group stage. Best two teams from each group proceed to elimination stage, where 16 matches are played (8 in the eight of the final, 4 in the quarter-final, 2 in the semi-final, and 2 in the final). The tournament therefore consists of 64 matches.

Additionally, 56 matches played in the preparation period of one month before the World Cup tournament were taken into account.

The data was obtained from the official FIFA World Cup page (http://www.fifa.com/worldcup/). In the elimination stage the match can't end as a draw, therefore overtimes (and optionally penalties) are required to determine the winner. As this rule has huge impact on the final result of the match, we ignored the final score and considered only the score after the regular period, which is also a standard practice with 3-way bets (home win, draw, away win).

Betting quotas for the World Cup matches were obtained on the day of the match from online betting service bwin (https://www.bwin.com/). Betting quotas are in form of 3-way bets and reflect the probability of the three possible outcomes: home team wins, draw, away team wins. Lower quota (approximately) corresponds to higher probability of the given outcome.

B. Matrix Factorization Model

In the field of multimedia recommender systems the matrix factorization model showed robustness and accuracy when predicting user experience [9]. The interaction between users and items is modelled on the base of known ratings that were given by the users to the specific items and that correspond to the user's satisfaction. The model is then used to predict the user's satisfaction with an unknown (unrated) content.

In our case the model is based on the observation of number of goals scored by a given team (T) to its opponent (O) in previous matches - $g_{T,O}$. Every observed match yields two goal score observation:

$$\text{team } A : \text{team } B - \text{scored } G_A : \text{scored } G_B$$

$$first\ observation:\ g_{A,B} = G_A$$

$$second\ observation:\ g_{B,A} = G_B$$

The observations can be arranged in a table (see Table 1), displaying the number of goals scored by a team (T) against its opponent (O). Initially, such table is highly sparse. New values are added as the tournament progresses, however even in the final stages the table remains sparse.

TABLE 1. THE TABLE REPRESENTATION OF GOAL SCORE OBSERVATIONS

$T \setminus O$	A	B	…
A	■	$g_{A,B}$	
B	$g_{B,A}$	■	
…			■

The model is then used to predict the number of goals scored in a match to come $-\hat{g}_{T,O}$. As in the case of observing past matches and match observations as above, in every match there are two goal score predictions, one for each of the two paired teams.

Additional parameters, besides the goals scored by both teams, may be included in the model in the similar way as [10] included the time component. However, this is not straightforward, as the parameters may not be directly correlated to goal score predictions, and an expert knowledge is therefore required. For example, if key player is injured this should influence the match outcome, but you need an expert to point out key players. Since the time factor was excluded in our case due to a short-termed data set, and we avoid using any expert knowledge, we use the matrix factorization model with biases:

$$\hat{g}_{T,O} = \mu + b_T + b_O + \mathrm{q}_O^T \mathrm{p}_T. \quad (1)$$

In the model the team and its opponent are represented using latent feature vectors p_T and q_O and their interactions are modeled as inner products of the corresponding vectors in the latent feature space. By implementing biases b_T and b_O we model the team's inclination towards scoring/conceding goals, while μ represents an overall 'goals scored' average. Bias b_T is calculated so that $\mu + b_T$ represents the average of goals scored by the team. Similarly, the bias b_O is calculated so that $\mu + b_O$ represents the average of goals conceded by the opposing team.

In the learning phase the model is fitted to the known match results, or more precisely to the number of goals scored by one team to its opponent (note again that every match provides two observations). The stochastic gradient

descent algorithm is used to loop through the training set (κ_T), where for each given training case $g_{T,O}$, the prediction $\hat{g}_{T,O}$ is calculated and the model parameters are modified to minimize squared error function:

$$\min \sum_{\kappa_T} \left(g_{T,O} - \hat{g}_{T,O}\right)^2 + \lambda\left(\|p_T\|^2 + \|q_O\|^2 + {b_T}^2 + {b_O}^2\right), \quad (2)$$

where the second part of the sum serves for regularization and it is controlled by the parameter λ.

C. Match Outcome Prediction

The described model predicts the number of goals scored by a team against a given opponent. For a match outcome prediction another step is required. First we need to pair the two predictions, one for home team ($\hat{g}_{H,A}$) and one for the away team ($\hat{g}_{A,H}$), and compare them to predict the outcome. There are three possible outcomes, commonly noted as 1-home team wins, 0-draw, 2-away team wins:

$$\hat{g}_{H,A}, \hat{g}_{A,H} \overset{\mathcal{C}}{\rightarrow} \{1,0,2\}\,. \quad (3)$$

The easiest way to predict the outcome using the pair of goal-scoring predictions is to simply round the predictions to nearest integer to get predicted match score:

$$\mathcal{C}: c = \begin{cases} 1; & round\left(\hat{g}_{H,A}\right) > round\left(\hat{g}_{A,H}\right) \\ 0; & round\left(\hat{g}_{H,A}\right) = round\left(\hat{g}_{A,H}\right) \\ 2; & round\left(\hat{g}_{H,A}\right) < round\left(\hat{g}_{A,H}\right) \end{cases} \quad (4)$$

Such approach, however, proved to be unfavorable to such pairs of predictions, where the two are close together but on different side of rounding threshold. Better results were expected when deploying a standard classifier, trained on a set of prediction pairs. We deiced to use the Naive Bayes classifier, which proved to be robust and performs well in most cases, despite the fact that it builds on the assumption that the input parameters are uncorrelated [11]. As the input parameters for the classifier, the $\hat{g}_{H,A}, \hat{g}_{A,H}$ pairs were used, and the match outcome was used as a label.

While in the first case no training is needed and the prediction is possible 'on the fly', this is not the case with the Naive Bayes classifier. Due to the small size of the data set (64 matches altogether), the classification has been done 'post festum', using the cross-folding technique to train and to test the performance of the classification.

D. Evaluation

Overall classification success rate, which gives the percentage of correctly predicted (classified) samples, is insufficient, as in most cases we need more insight into the classification performance. Thus a confusion matrix is commonly used (see Figure 2). In our case there are three possible classes (match outcomes), predicted classes are marked with 'hat' sign (^), actual (achieved) classes without it. Values p_{ij} represent shares of samples originating from class *j* and classified as *i*.

Several performance measures were developed on basis of the confusion matrix that vary over different research areas. However, in most cases simple and easy to understand measures, like the overall success rate, precision and recall, should be used [12]. Overall success rate (OSR) is used to measure the percentage of correctly classified samples. In our case it is calculated as:

	1	0	2
$\hat{1}$	p_{11}	p_{12}	p_{13}
$\hat{0}$	p_{21}	p_{22}	p_{23}
$\hat{2}$	p_{31}	p_{32}	p_{33}

Figure 1. Confusion Matrix for 3-class classifier

$$\mathrm{OSR} = p_{11} + p_{22} + p_{33}. \quad (5)$$

To get more insight into how the classifier is performing according to a specific class, the precision and recall are used. Both are used to measure percentage of correctly classified samples, however while precision (P) is used for a given prediction class, recall (R) is used for a given actual class:

$$P_i = \frac{p_{ii}}{p_{i1} + p_{i2} + p_{i3}}, \quad R_i = \frac{p_{ii}}{p_{1i} + p_{2i} + p_{3i}}. \quad (6)$$

Fawcett [13] proposes usage of ROC (Receiver Operating Characteristics) graphs as a performance graphing method as they are especially useful for domains with skewed class distribution and unequal classification error costs. ROC graphs show relationship between two ratios: (i) the rate of correctly classified samples from the given actual class (same as recall), named true positive rate (tp rate), and (ii) the rate of incorrectly classified samples that do not belong to the given actual class, named false positive rate (fp rate). In case of perfect classification, the tp rate would be 1 and the fp rate would be zero. ROC graphs allow us to find the point on the curve (which corresponds to a certain classification threshold) closest to the 'perfect case'. The diagonal line represents the 'random classification' where the two rates are equal. The same author also proposes the area under ROC curve (AUC) as the measure for a classifier performance. Higher AUC represents better performance. The performance of a random classifier in terms of AUC is 0.5.

In cases with a small dataset (like ours) it is a common approach to use the cross-folding technique. The correct way to calculate AUC is such cases is to obtain the ROC curve for each fold and represent the classifier with the average curve and the corresponding AUC [14].

In their basic form, the ROC graphs are used for two-class problems. With more classes the situation becomes more complex. One method for handling such cases is to produce several different ROC graphs, one for each class, where reference class is treated as positive and all others as negative [13]. For the multi-class AUC, the weighted sum

approach can be used, where the weights $p(c_i)$ correspond to prevalence of the reference class [15]:

$$\mathrm{AUC}_{\mathrm{total}} = \sum_{c\_i \in C} \mathrm{AUC}(c_i) \cdot p(c_i) \qquad (7)$$

III. EXPERIMENT

On Figure 1 flow of the experiment is presented. Known match results are stored in the local database. In the start only the results of friendly matches were known, the results of the tournament matches were added as the tournament progressed. After every tournament round, the known data was used to (re)train matrix factorization model. Betting quotas were obtained on the day of the match and stored for later comparison.

Due to the small size of the data set the latent feature predictor (classifier) performance evaluation was performed 'post festum'. We used cross-folding of the data set, more precisely the 8 bin folding stratified sampling approach. The 8-fold approach was used instead of more commonly used 10-fold, as this allowed the same size of all bins (64 samples divided across 8 bins).

For comparison, a classifier based on betting quotas was developed and evaluated using the same approach as the latent feature match predictor. In this case, the three quotas (one for each possible outcome) were used for the Naive Bayes classifier input, and the match outcome was used as the label. Betting quotas are (for obvious reasons) prepared using sophisticated proprietary algorithms, backed also by an expert knowledge of the odds traders and others. Such comparison of the two classifiers (the latent feature based and the betting quota based) gives a latent feature approach usability estimation.

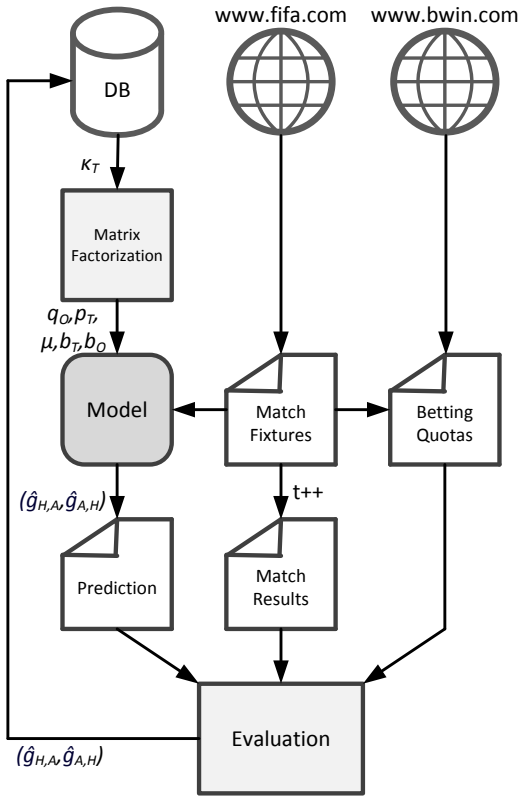


Figure 2. Flow of the experiment

IV. RESULTS

The match prediction performance using the 'rounding' approach (4) yielded 44% overall accuracy (OSR= 0.438). Confusion matrix on Figure 3, as anticipated, shows very poor precision and recall values for the 'draw' class. Using Naive Bayes classifier improved overall accuracy to 50% (OSR= 0.5). Closer look to confusion matrix on Figure 4 shows that prediction value of the 'draw' class improved, as also the recall value of the 'away win' class. There were less candidates for the 'draw' outcome as in the first case, but those were picked better. Consequently we managed to find more 'away wins' and also 'home wins' as in the case of 'rounding' approach.

	1	0	2	P
$\hat{1}$	0.219	0.188	0.109	0.424
$\hat{0}$	0.078	0.047	0.078	0.230
$\hat{2}$	0.078	0.031	0.172	0.611
R	0.583	0.176	0.478	

Figure 3. Confusion Matrix for 'rounding' approach

	1	0	2	P
$\hat{1}$	0.234	0.109	0.109	0.417
$\hat{0}$	0.031	0.031	0.016	0.400
$\hat{2}$	0.109	0.016	0.234	0.652
R	0.625	0.118	0.652	

Figure 4. Confusion Matrix for 'latent features' approach

	1	0	2	P
$\hat{1}$	0.297	0.156	0.141	0.500
$\hat{0}$	0.063	0.047	0.063	0.273
$\hat{2}$	0.016	0.063	0.156	0.667
R	0.792	0.177	0.435	

Figure 5. Confusion Matrix for 'betting quotas' approach

Figure 6 is showing ROC graphs for the classifier that uses matrix factorization model output. The curves represents average over eight folds, used for cross-folding. Shaded areas mark the 95% confidence interval for the averaged curves. In our case the following class prevalence was observed: $p_1 = 0.375, p_0 = 0.266$, and $p_2 = 0.359$. Using (7) we obtained:

$$\mathrm{AUC}_{\mathrm{lf}} = 0.677$$

To estimate the performance of an expert knowledge-based algorithm, we tested Naive Bayes classifier using betting quotas. We achieved the 50% overall accuracy (OSR= 0.5), which is the same as with 'latent features' approach. However by inspecting the confusion matrix (Figure 5) some differences may be observed, for example more candidates for 'draw' outcome at the cost of lower precision, but much higher hit rate (relevance) for the 'home win' outcomes, and also lower hit rate for the 'away win'.

Furthermore, ROC graphs on Figure 7 show superior performance in terms of classifying 'away win' class but at the cost of extremely poor classification of 'draw class'. In this case, we obtained the following AUC value:

$$\mathrm{AUC}_{\mathrm{bq}} = 0.562$$

The AUC that we obtained for latent feature-based classifier is actually higher than the one obtained for betting quotas-based one. However the fact of high measuring uncertainty (shaded areas on Figure 6 and Figure 7) cannot be overlooked. Thus we cannot conclude that the former classifier performs significantly better than the latter.

V. CONLUSION

The prediction algorithm presented in this paper is based only on the base of previous results and does not require any form of match statistics or expert knowledge. We propose to use the described approach in cases when statistics and expert knowledge are not available or when they cannot be imported, interpreted, or evaluated properly.

As football shows highly stochastic nature it is difficult to evaluate any prediction algorithm. The only correct way would be to test all relevant algorithms against each other on the same (reasonably) large dataset. However, it is practically impossible to perform such comparison as various algorithms use various statistical data and knowledge, and a standard dataset for testing purpose containing all required data does not exist yet. The comparison between two classifiers (the latent features-based and the betting quotas-based), performed in this study, can be used only to estimate the possibility to substitute real features with latent ones. This case study showed no significant difference between the two approaches, showing that such substitution is possible.

The major issue with our case study lies in the size of the dataset. We used a short-termed dataset, where 64 matches were played in a period of two months, therefore drastic changes in teams' performance (e.g. due to player injuries) were not expected. It is known that the matrix factorization approach towards latent features extraction

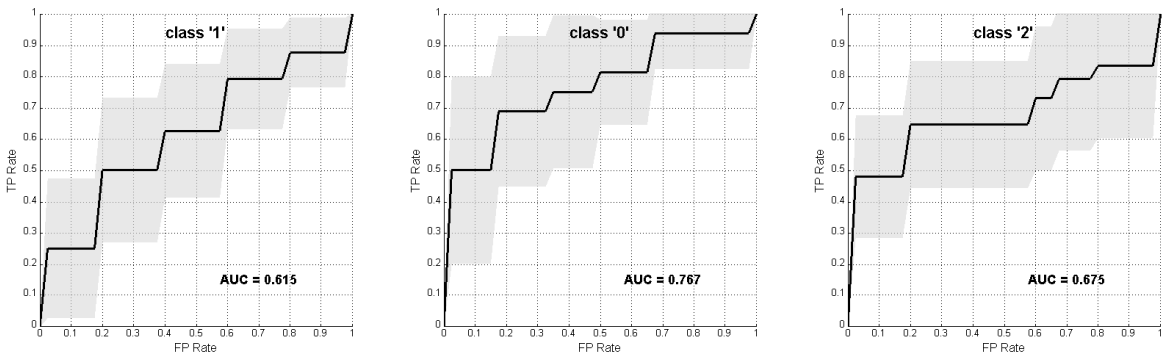


Figure 6. ROC graphs for Naive Bayes classifier using Matrix Factorization model

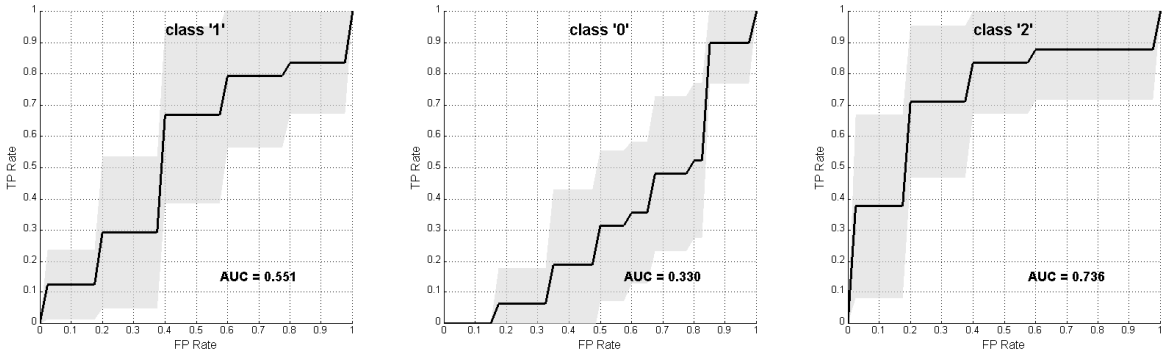


Figure 7. ROC graphs for Naive Bayes classifier using betting quotas

performs significantly better on large datasets. However, larger dataset would require long-term observation bringing (at least) time factor into the equation.

We believe that the prediction process can be further improved by inclusion of expert knowledge into the matrix factorization model training, exposing parameters that do not encompass teams' interaction (e.g. motivation, fatigue…) On the other hand, the parameter selection step would introduce another level of complexity into the matrix factorization process.

Another possibility to improve the prediction success, is to append regression model that would be used to improve latent feature model based predictions. This would allow us to include various 'real' match features into the regression model, where are easier to handle. However, as already pointed out in the paper, the expert knowledge would be required to point out the important features.

REFERENCES

[1] B. Min, J. Kim, C. Choe, H. Eom, R. McKay, "A compound framework for sports results prediction: A football case study," Knowledge-Based Systems, vol. 21, n. 7, pages 551--562, 2008.

[2] M. Hughes, I. Franks, The essentials of performance analysis: an introduction, Routledge, 2007.

[3] D. Karlis, I. Ntzoufras, "Analysis of sports data by using bivariate poisson models," Journal of the Royal Statistical Society: Series D (The Statistician), vol. 52, n. 3, pages 381--393, Wiley Online Library, 2003.

[4] J. Goddard, "Regression models for forecasting goals and match results in association football," International Journal of Forecasting, vol. 21, n. 2, pages 331--340, Elsevier, 2005.

[5] A. Joseph, N. Fenton, M. Neil, "Predicting football results using Bayesian nets and other machine learning techniques," Knowledge-Based Systems, vol. 19, n. 7, pages 544--553, Elsevier, 2006.

[6] A. Constantinou, N. Fenton, M. Neil, "pi-football: A Bayesian network model for forecasting Association Football match outcomes," Knowledge-Based Systems, vol. 36, pages 322--339, 2012.

[7] G. Kumar, Machine Learning for Soccer Analytics, Cambridge University Press, MSc thesis, KU Leuven, 2013.

[8] D. Bunker, R. Thorpe, "A model for teaching games in the secondary school," Bulletin of Physical Education, n. 10, pages 9--16, 1982.

[9] Y. Koren, R. Bell, C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, n. 8, pages 30--37, IEEE 2009.

[10] Y. Koren, "Collaborative Filtering with Temporal Dynamics," Communications of the ACM, vol. 53, n. 4, pages 89--97, 2010.

[11] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, n. 22, pages 41--46, 2001.

[12] V. Labatut, H. Cherifi, "Evaluation of Performance Measures for Classifiers Comparison," Ubiquitous Computing and Communication Journal, vol. 6, pages 21-34, 2011.

[13] T. Fawcett, "An introduction to ROC analysis," Pattern recognition letters, vol. 27, n. 8, pages 861--874, Elsevier, 2006.

[14] F.J. Provost, T. Fawcett, R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," proceedings of ICML-98, vol. 98, pages 445--453, 1998.

[15] F.J. Provost, P. Domingos, "Well-trained PETs: Improving probability estimation trees," Citeseer, 2000.

Role-playing as a knowledge management tool

Luka Brajković
Faculty of Teacher Education
University of Zagreb
10000 Zagreb, Croatia
luka.brajkovic@outlook.com

Winton Afrić
University North
Department of Media design
30155 Koprivnica, Croatia
winton.afric55@gmail.com

Vjekoslav Afrić
Faculty of Philosophy and Social Sciences, University of Zagreb
10000 Zagreb, Croatia
vjekoslav.afric@zg.htnet.hr

Abstract – Role-playing, besides its use as a gaming media, can also be used as a powerful simulation tool. Due to its versatile nature and all the possibilities that it can offer us in building imaginary worlds and situations, we can also apply that technique to simulate real life and realistic behaviour. In project and knowledge management it is possible to use Role playing for optimization and building specific knowledge. Applying the simulation in advance, can greatly help us to recognize, prepare and solve most problems that may occur, reducing overall cost and time of any project.

I. INTRODUCTION

Despite its humble beginnings as a gaming system, Role-playing (RP in further text), has also developed over the years as a simulation system worthy of our attention. In its very core, RP is a system that combines setting, a variety of entities that coexists in it and all interactions among those entities and its surroundings. A setting is a specific environment in which we can create our simulation, making a copy of a real world we would like to simulate, with all its legalities and factors that influence it. Entities, on the other hand, can be all the participants active in the simulation process, events that we wish to study, along with all the tasks and problems presenting goals of our RP simulation. In establishing a simulation by defining its setting, entities and goals, we are able to conduct optimisation processes of real life events, building specific knowledge usable in future projects.

This kind of RP simulation can be used as an optimisation tool in project management. Before actually conducting a project, RP simulation can help us anticipate various problems and situations, explore different phases of our project and optimise them to accomplish better overall results. In every project there are key factors defining successfulness, like the project end time, cost and efficiency. Every step of the way, conducting our simulation, we can explore different solutions to every problem and situation we could come across. Keeping in mind our end goal, exploring different options and solutions and documenting the course of the simulation, we are able to anticipate, and eliminate, a variety of problems that could occur later when a real project starts to develop. With combined knowledge and experience of participants involved in the simulation, and later the real project itself, we can accomplish a high level of optimisation resulting in immense resource benefit. We can even choose the optimisation path making some resources primary and other resources secondary. For example, making time our primary resource, we can reduce overall duration of the project by finishing it prior to the originally planned time. In this example we are making finance one of our secondary resources, not going over the budget but coming to its limits.

II. ROLE-PLAYING SIMULATION IN PROJECT MANAGEMENT

A. Introduction to RP simulation

Project management in its surveillance, optimisation and conduct, treats every project as a system. Each and every project is carefully dissected in order to organise, analyse and optimise all of the components making the system. For every new system that we build, first we need to understand its overall purpose and the way it interacts with its surrounding environment, as every system is part of some already existing larger one. With that in mind, dynamic equilibrium of this new system and control of its parts are essential components on which project the management is focused. In essence we first need to establish the system a structural code of elements and the rules by which the elements correspond to one another so that we may later, when simulating operate the structure and determine specific process codes optimised to our needs. [5]

This step is crucial because once we establish the rules – code of the systems structure the objectives we have in the project and knowledge management in general are process operations of that structural code. The

very structure is the one which will enable us to analyse specific parts of the system, ask different questions, select, make different elements, or attributes pertinent in order to restructure and therefore optimise the system. **The use of RP in simulation in this level is therefore an easy to manage transparent way of producing code (rules) for system optimisation.**

Using RP simulation as an additional tool in project management, we are able to further understand, optimise and achieve that dynamic equilibrium needed in order to successfully build a new system. Applying the simulation it the earliest stage of project development, right after setting up the main idea and general system guidelines, is the best way to use all of its potential. Of course, due its flexibility, we can run the simulation again in every phase of the project and apply different objectives and gain information. If we imagine project management as a surveillance tool when building a new system, we can think of RP simulation as a preparation tool for all our actions during building and implementation phase. Project manager as head figure in project, should choose various relevant field experts which will contribute and be part of simulation process. It is noteworthy that in the case of RP simulation it is desirable to have multiple experts from the same field contribute from various roles. With their combined knowledge and experience put into the simulation, the simulation process creates a knowledge base for solving particular problems. During simulation, participants are looking for different solutions to known problems, and in the end choosing answers that best suit any given occasion. It is necessary to point out that the very RP simulation model preparation should be made in with the problem solving process in mind. The human aspect of the role from which the participants play is crucial and it's where the brilliance of RP simulation lays. Therefore a problem solving approach frequently used in group communications is particularly applicable. [6] There we use a slightly modified classic eight step process for problem solving. The process is modified in the sense that each participant posing the questions is a specific role with a specific perspective. That role is determined in the process of system creation and typically firstly played by a field expert. We say firstly because it is encouraged and welcome that the same role is afterwards played by different experts and non-experts alike as they can provide valuable insight in the whole process of approaching the system from the perspective of playing a specific role. The process is as follows:

1. ***Defining the problems*** – we ask the problems as questions from the perspective of our role in relation to the system structure.
2. ***Identifying the meaning of the problem for the system structure (from our role)*** – what does the problem impact, what can we foresee changing as a result of the problem solving process in general etc.
3. ***Further problem analysis (from the perspective of our role)*** – How long does the problem last? What caused the problem? Does it have any consequences? Is the problem recognised within the system? Does it impact other parts of the system in a known way that is currently managed?
4. ***Identifying the criteria for solution evaluation from the perspective of our specific role***: What would make a solution practical? What are the specific effects a proposed solution could/should have (cost efficiency, time etc.)? If there are multiple solutions what would my criteria be for ranking them?
5. ***Identifying possible solutions*** – this is the phase where the participants should brainstorm. Thus when encountering problems, participants should try to find as many solutions to each specific problem they can, observing them from different angels and exploring additional options. Besides obvious solutions that come from knowledge and experience, simulation itself forces participants to think and look for other unorthodox answers. When brainstorming is concerned the wilder the ideas the better. Each participant should answer these questions from their specific standpoint: what is a solution to our particular problem? How does it impact the system (or my specific part of the system)? How do I view it from my role? Does it create new problems or difficulties for me?
6. ***Solution evaluation*** – each participant should evaluate all solutions according to step 4.
7. ***Selecting the best solutions*** – ones where most participants agree, most relevant criteria are met and the system as a whole reaches the particular goal in the most efficient manner.
8. ***Simulation*** – testing the solutions. From this point on we can return to any of the prior steps.

During simulation, the main point is to explore any relevant solutions recommended by any specific participant/role or defined by an **actant** [7] and research different approaches to encountered problems. Given results are created with all the roles and actants in mind producing different directions in problem solving, ones that are normally difficult to brainstorm. In that way, we can say that **RP simulation itself is in fact an "idea generator".** By documenting problems and given solutions, the RP simulation model provides a base of knowledge, and experience on problem solving, regarding any given situation (project, system), before they occur. Apart from common brainstorming, RP simulation in the end gives us a transparent, precise and well documented workflow on specific problem solving, knowledge base applicable in real situations, time and environment, produced by a group of chosen field experts.

A well-documented simulation workflow gives us a high level of preparedness to almost any situation that could occur during project implementation. Of course, there could always emerge some unpredictable situations, but as soon as those situations appear, we can input them into simulation and develop new steps for solving them. It also helps us that we already have a multitude of other solutions and experience which may relate to this unpredictable one, so the simulation had

already given us general guidelines on how to resolve it, or at least we know the direction to solve this new problem.

B. Simple example of RP simulation

Every project consists of three main components: events or milestones, activities and time. Events are different phases of project that we are trying to achieve through our activities in some estimated time. In order for a project to be successful, each of these components should be done on a predetermined schedule with optimal path (activity) between events. Time needed to finish the project is also divided in four categories. Most likely time, expected time, optimistic and pessimistic time. Most likely time is the best estimate of the time required to accomplish a task, assuming everything proceeds as normal. In reality, we will almost always have some kind of problems and delays to cope with, in the end achieving expected or pessimistic time. Therefore, RP simulation offers us a tool to achieve this most likely time, which is the minimum possible time to finish a project, or maybe even go for optimistic time. By optimising every event and activity in project through simulation, we are able to balance all of our recourses, specially finance, in order to achieve time that we would like. With simulation, we are absolutely trying to avoid pessimistic time and immensely reduce resource consumption.

For example, when building a fifty story skyscraper, one of our events could be building first ten storeys in one month (expected time), with all the activities (physical work, material delivery etc.), in optimal condition to do so. With all of that in mind, a selected group of experts with roles defined and derived from their specific area of expertise should when conducting RP simulation, with their knowledge and experience put into the simulation, be able to optimise the process in order to achieve optimistic time with maximum savings in resource consumption.

On Figure 1 we can see an example of PERT network chart for a seven-month project with five milestones (10 through 50) and six activities (A through F).[1]

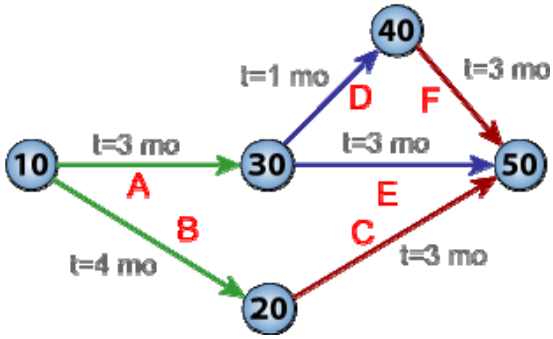


Figure 1: PERT network chart for a seven-month project with five milestones (10 through 50) and six activities (A through F).

In the example above, we know our milestones which are something that generally does not change and activities that up until now were something that we knew had to be done. Using RP simulation, we can take one fragment of the chart, expound one activity and question its parts in order to optimise it. During simulation we are able to in example discover reasons to introduce a new milestone in-between, reorganize the order of some milestones, adjust specific times etc.. By doing so, we guide ourselves through simulation, in order to achieve a better finish time, reduce costs or other resources. In the end, we are trying to find a new or different way of accomplishing our work goals, creating and expanding our own knowledge on the substance. That way, by documenting our newly found knowledge, we are in fact creating a learning system which can be not only applied during project implementation, but also as a teaching method for others.

[1] Source: Wikipedia web page: http://en.wikipedia.org/wiki/Project_management, accessed 30. January 2015.

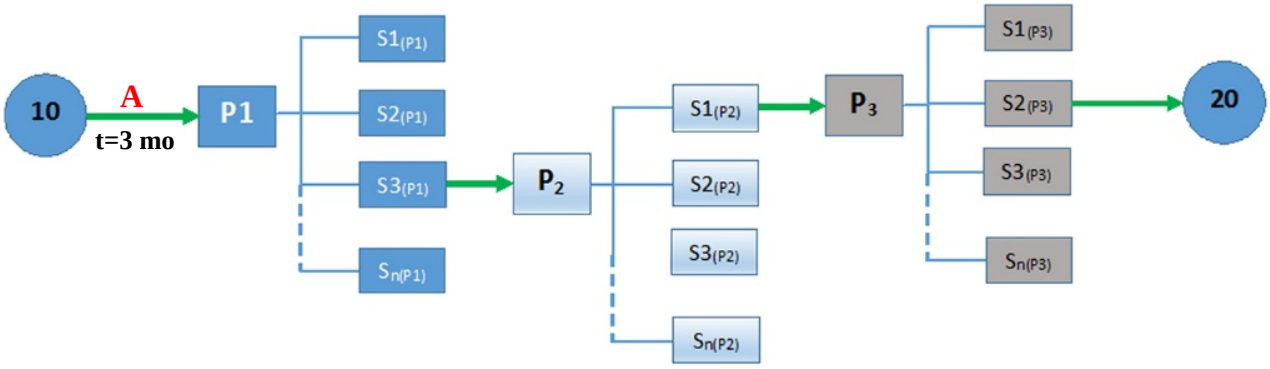


Figure 2: An example of one activity divided by three separate problems with a number of solutions for each

In our second figure, we can see one activity divided by three problems. Problems can also be questions that experts in specific roles ask in order to optimise and improve a course of chosen activity. If the chosen activity is a simpler one, maybe only one question will appear, but with complex activity there could be many. A project manager with his team of experts involved in simulation, using their knowledge and experience, should try to come up with as many questions needed in order to improve given activity (we believe the best approach is using our proposed 8 steps method). Likewise, more answers to each question or problem, gives them more options and paths to direct optimisation process the way they imagined, in the given process, saving resources of their choosing (money, time, material etc.). Picture shows us that on first problem ($P1_{)}$, solution ($S3_{(P1)}$) was chosen. It means that in that particular time in our simulation, one answer seemed most logical and valid in order to be chosen, based on judgement of experts running the simulation. That is of course a specific process code "recommended" by the system and chosen in this particular case but whether or not it is the best solution can only be discovered once the course of the simulation is complete. Let's say that time is the resource we wish to optimise, so in that case answer $S3_{(P1)}$, gives us a course of action with which we will reduce activity time. There are probably other answers that also save time, but may produce other disadvantages, greatly higher costs or similar. The chosen answer leads us to a new problem or question and we repeat the optimisation process.

If we would like to reduce time of our activity, from four months to three months, we should ask relevant questions as part of the 8 step process, and imagine problems that we need to solve from each role in order to achieve our time reduction. In our earlier mentioned example, in order to build the first ten storeys in three rather than in four months, maybe we should apply some new building technique. By introducing that new technique, perhaps we will spend more money on material or mechanisation but reduce manpower needed for construction and achieve our time saving plan. Right there, in our simulation, in that moment we can research all of our options and influence on all later milestones and activities, optimising our overall project plan. Some solutions will open new problems, some will bypass existing ones, but our goal in simulation is to try to imagine all the possibilities we can in order to optimise the project the way we want. That way we also accumulate knowledge and acquire a well developed execution plan for project deployment. In the end, balancing resources but emphasising the one crucial for our overall success, is the right way to approach the process of simulation.

For every project that is being built in a real word environment, we can consider it to be a part of highly dynamic and ever-changing bigger system. After conducting our RP simulation and having it all optimised the way we wanted, taking into consideration all the possibilities that could happen during real life project implementation, since being the part of this bigger dynamic environment, some absolutely unpredictable things could always happen. In that case, we can always go back to our simulation and make necessary changes. This unpredictable problem, in most cases should not represent too big an obstacle, since we already have the optimised project and our resources rather reduced. By introducing that new problem into simulation, the course of affected activity might change and influence all other actions after, but also that might not be the case. It really depends on the scale of that new moment introduced in our project and simulation likewise. Unpredictable problems could cost us a great deal of time and money when encountering them. Best way to reduce their appearance and minimise the damage they could cause, is to have experienced experts running the simulation, but as we stated earlier also include people of different knowledge and background, to have more perspectives on the matter and gain more insight. An architect will know how to solve construction problems, but an economist will promptly adjust the course of an activity in case of a

market fluctuation, while a work safety expert will contribute and likely end up changing the activities attributes (time, finance etc.) as a result.

III. ROLE-PLAYING SIMULATION AS A KNOWLEDGE MANAGEMENT TOOL

Similarly as shown in the example of the project management, RP simulation can be also used as a knowledge management tool. If we imagine a situation or a project that has never been done before, including a field expert research, with RP simulation we are able to explore and develop a knowledge base on the subject. It is very problematic to create a knowledge base and thus develop training or educational means for something that has never been done before. RP simulation is a tool that gives us the needed environment in order to explore our ideas and expand our imagination. Working closely with other experts, it provides us an opportunity to share our ideas and develop them further than we would be able alone. RP simulation gives us the opportunity to expand our ideas even to borderline fiction with various experts as interdisciplinary as possible. In order to create a new knowledge base or learning tool, we usually reach for an already existing situation, review it, optimise and document it. That way others could be trained for that situation and the gathered knowledge can be expanded. The problem occurs when we are faced with a project for which, in order to conduct, necessary knowledge simply does not exist. In that case, RP simulation can provide us means to explore and test our ideas in simulated close to real environment. By using existing knowledge provided by experts, the ones that are related in the field of our research, we are able to build our simulation and put it to test. By doing so, we are creating absolutely new knowledge that could be used later in real-life situations. It is an ideal tool for techno-centric knowledge management perspective, since in its core Role-playing in general is focused at interpersonal interaction in order to share information and ideas and together create new knowledge. Knowledge management is already using storytelling as an instrument for accumulation and sharing information. Since RP simulation is actually an evolved and advanced form of storytelling, it would be wise to further explore it and its use as a valuable tool in knowledge management.

RP simulation is truly unique as it not only educates with the results of its simulation it educates all the participants in the process of role playing. Different experts and non-experts face varying situations from different role perspectives expanding the ways they view and deal with problems. RP is not only interactive it is immediate in the way it affects the participants expanding the way they approach and analyse systems, situations, and problems. When using RP simulation the whole process is a rewarding educating experience which helps participants to develop in various ways within their formal education frame as well as out of it. Not only that but it participants learn to understand different role types tied with different systems and the intricate ways of structural - process relations which stem from interpersonal actor - actor - actant - environment relations. **The usage of RP therefore thrives in longevity and creates a constantly evolving and growing knowledge base which ties knowledge and experience in dealing with knowledge in special roles and situations in ways no other method can.**

IV. CONCLUSION

RP as optimisation tool presents a specific environment in which we can create our simulation, making a copy of a real world we would like to simulate, with all its legalities and factors that influence it. Every step of the way, conducting our simulation, we can explore different solutions to every problem and situations we could come across. The use of RP in simulation is therefore an easy to manage transparent way of producing code (rules) for system optimisation. The human aspect of the role from which the participants play is crucial and it's where the brilliance of RP simulation lays. Therefore a problem solving approach frequently used in group communications is particularly applicable, and we use a slightly modified classic eight step process for problem solving. The process is modified in the sense that each participant posing the questions is a specific role with a specific perspective. RP simulation gives us a transparent, precise and well documented workflow on specific problem solving furthermore it produces a knowledge base applicable in real situations, time and environment, produced by a group of chosen field experts. As we show in the example of the project management, RP simulation can also be used as a knowledge management tool. If we imagine a situation or a project that has never been done before, including a field expert research, with RP simulation we are able to explore and develop a knowledge base on the subject, because a true RP session enables the equivalence of discourse and determines a community of knowledge. RP is a unique tool that establishes the ontology of knowledge for individual projects, defining the context in which we need to think about the problems we face. [8] Therefore RP ties knowledge and experience and can be used as a mighty tool for project management and knowledge production and application of RP to the problems of the project and knowledge management. Therefore RP remains a challenge to maximize the use of human creativity in creating new values, goods and practices.

V. REFERENCES

[1] Afrić, W. (2010): Temeljna načela izgradnje virtualne realnosti u RPG sustavima, Zagreb, p. 151-161.

[2] Šimović, V.; Zovko, V.; Bobera, D. (2011): Projektni menadžment i informacijska potpora, Zaprešić, p. 12-13.

[3] Šimović, V.; Zovko, V.; Bobera, D. (2011): Projektni menadžment i informacijska potpora, Zaprešić, p. 37.

[4] Alavi, Maryam; Leidner, Dorothy E. (1999). "Knowledge management systems: issues, challenges, and benefits". Communications of the AIS 1 (2).

[5] Johannsen, J. Dines; Larsen, S. Eric (2002) “Signs in use”, p.7-24.

[6] DeVito, J. A. (2010); “Essentials of human communication – 7th edition”

[7] Latour B. (1993): “We have never been modern” Harvard University Press.

[8] Afrić, V (2014): Tehnologije e-obrazovanja I njihov društveni utjecaj, edited in “Informacijska tejhnologija u obrazovanju” Zavod za Informacijske studije Zagreb, p.-20.

Nadogradnja inteligentnog sustava aerodromskog održavanja serijskih strujnih krugova

Zdenko Balaž*, Ivo Puhalović**, Goran Bjelić*

*Tehničko veleučilište Zagreb, Elektrotehnički odjel, Zagreb, Hrvatska

**Hrvatska agencija za civilno zrakoplovstvo, Odjel aerodroma, Zagreb, Hrvatska

zbalaz@tvz.hr, ivo.puhalovic@ccaa.hr, gbjelic@gmail.com*

Sažetak -Model inteligentnog sustava koji se nalazi na raspolaganju studentima stručnog specijalističkog studija Elektrotehničkog odjela i polaznici Programa edukacije AERO TVZ za stjecanje licence aerodromskih specijalista, na Tehničkom veleučilištu u Zagrebu, nadograđuje se za potrebe „rada u oblacima". U radu je prikazan koncept nadogradnje inteligentnog sustava koji obuhvaća i prezentaciju baze podataka s rezultatima obrade provedenih istraživanja na aerodromima i ispitivanjima u laboratoriju za potrebe baze znanja. Referentni parametri aerodromskog sustava svjetlosne signalizacije između ostalog se odnose i na stanje otpora izolacije aerodromskih primarnih kabela serijskih strujnih krugova. Redovnim održavanjem, ali i provjerom održavanja inspekcijskim nadzorima utvrđuje se stanje tih vitalnih komponenti sustava koje su u funkciji sigurnosti zračne plovidbe.

Znanjem iz inteligentnog sustava razrađeno je i primijenjeno unapređenje održavanja, koje se kao model usluge dobavljive na tržištu nudi krajnjim korisnicima. Ispitivanje i mjerenje otpora izolacije na serijskim strujnim krugovima aerodromskog sustava svjetlosne signalizacije moguće je povjeriti specijalistima. Uz ekspertni sustav koji služi za edukaciju, postavljena je paralelna agentska platforma koja obuhvaća procedure i uređajem nove generacije koristi se za ispitivanje na terenu. Posebna sofisticiranost ispitnog uređaja uz agentsku platformu je njegova mogućnost da osim sveobuhvatnog ispitivanja, prikupljanja i obrade, relevantne podatke može trenutno, (online), dostavljati na bilo koje zahtijevano mjesto uz punjenje u bazu podataka.

I. UVOD

Sustav aeronautičkih svjetala na zemlji, (*engl. Airfield Lighting System*), pruža pilotu informacije o lokaciji, orijentaciji i poravnanju u nepovoljnim vremenskim uvjetima i noću. U funkciji je sigurnosti zračnog prometa te ovisi o razini opremljenosti zračne luke, (kategoriji prilaza - instrumentalni precizni prilaz CAT I, CAT II-III, A, B, C, slika 1.), može biti kreiran s više podsustava. Ispravnost i pouzdanost cjelokupnog sustava uvelike utječe na sigurno provođenje operacija, (slijetanje, uzlijetanje, taksiranje) zrakoplova na zračnim lukama. Ocjene ispravnosti i pogonske pouzdanosti rezultat su zadovoljavajućih parametara o stanju, koji se nadziru i arhiviraju u baze podataka, a o čemu svjedoče prethodno provedena istraživanja i prezentirani rezultati, [1-2]. Otpor izolacije aerodromskih primarnih kabela serijskih strujnih krugova utječe na izlazne (fotometrijske) karakteristike sustava aeronautičkih svjetala na zemlji, a time i na sigurnost operacija, te ga je potrebno nadzirati i analizirati u svrhu što kvalitetnijeg održavanja sustava.

Slika 1. Aerodromski sustav svjetlosne signalizacije, sustav aeronautičkih svjetala na zemlji

Nastavak istraživanja na aerodromima i ispitivanja u laboratoriju, vezano na stanje otpora izolacije aerodromskih primarnih kabela serijskih strujnih krugova, uz učenje iz inteligentnog sustava potvrdila su mogućnost unapređenja procedura preventivnog i predikatnog održavanja, [3-6]. Nabavljena nova ispitna oprema omogućuje sveobuhvatno ispitivanje, prikupljanje i obradu relevantnih podataka koji se mogu trenutno proslijediti na zahtijevano mjesto, bez prijenosa na računalo. Model inteligentnog sustava koji je na raspolaganju studentima specijalističkog studija Elektrotehničkog odjela i polaznici programa edukacije za stjecanje licence aerodromskih specijalista na Tehničkom veleučilištu u Zagrebu nadograđuje se za mogućnost „rada u oblacima" s navedenim specifičnim zahtjevima, [7-8].

A. Održavanje aerodromskog sustava svjetlosne signalizacije

Za specijalističke aerodromske sustave u funkciji sigurnosti zračne plovidbe, kao što je sustav svjetlosne signalizacije, legislativnim zahtjevima definirani su uvjeti eksploatacije, način održavanja, mjerenja i kontrole stanja otpora izolacije u serijskim strujnim krugovima, [9]. Praćenje parametara kroz duži niz godina ukazalo je na niz važnih spoznaja iz kojih se ocjenjuje pouzdanost ali i strategija održavanja, [10].

Na temelju raspoloživih baza podataka prikupljenih tijekom gradnje fakultetskog ekspertnog aerodromskog sustava, "ExSys AERO TVZ", ali i obradom konkretnih

aerodromskih primjera eksploatacije složeni su algoritmi za koncept odabira održavanja, [11]. Strategije održavanja moraju biti povezane s kreacijama baza znanja kao dijela inteligentnog sustava koje su rezultat praktičnih spoznaja. Ta se unapređenja elaboriraju kroz pripreme specijalističkih programa održavanja i na temelju razmjene iskustava kreiraju nova znanja, [12-13]. Primjer obrade prijašnjih iskustava pretočenih u bazu znanja pokazuje Tablica 1, koja prezentira vrste kvarova na serijskim strujnim krugovima, (ukupno analizirano 166 kvarova na šest zračnih luka, [1]. "Korisnik instalacija, opreme ili uređaja najbolje poznaje njihovo stanje", konstatacija s kojom dvije od pet legislativnih podloga o održavanju započinju poglavlje, kojim se definira plan održavanja, [13]. Kada se radi o planovima održavanja vezanim uz aerodromske specijalističke

TABLICA 1. STATISTIKA KVAROVA IZ BAZA PODATAKA

Red. br.	**Pregled kvarova iz baze podataka**		
	Vrsta kvara		***Vrsta kvara***
1	Kvar na kabelu - oštećenje uslijed atmosferskog pražnjenja, (udar groma)	3/1,8	odstranjen dio oštećenog kabela
2	Kvar na kabelu - mehaničko oštećenje izolacije, starost kabela ili prekid kabela	12/7,3	nadomješten dio oštećenog kabela
3	Kvar na konektoru - nedovoljno čvrst spoja vodiča u spojnici konektora	5/3	zamijenjen konektor
4	Kvar na konektoru - prekid spoja vodiča kod pina konektora	6/3,6	zamijenjen konektor
5	Kvar na konektoru - oštećenje izolacije konektora ili neispravno izrađen konektor	95/57,2	zamijenjen konektor
6	Kvar na konektoru - nečistoća unutar kontaktnih površina oba kontakta konektora	23/13,8	saniran konektor
7	Kvar na transformatoru - proboj unutar namota transformatora	5/3	zamijenjen transformator
8	Kvar na transformatoru ili primarnom konektoru transformatora izvana	12/7,3	saniran konektor
9	Kvar na transformatoru - uzrokovan probojem na strani sekundara transformatora	2/1,2	zamijenjen transformator
10	Kvar na transformatoru i kabelu zbog prodora vode iz zdenca	3/1,8	ispumpana voda i osušen transformator
	UKUPNO	166/100	

sustave, korisno je napomenuti kako su već više od tri desetljeća radne stanice - dio sustava svjetlosne signalizacije u službi za održavanje, opremljene dijelovima ekspertnog sustava - bazama podataka, koje su u početku nazivane "Arhive". One su u početku isključivo služile za "arhiviranje podataka", i značilo je da se utvrđeni podaci mogu arhivirati, pohraniti odnosno spremiti. Danas, u novije vrijeme, razvojem ekspertnih sustava uz baze podataka korisnici uređuju i grade vlastite baze znanja pa se s podacima može zaključivati i stjecati nova znanja, kao što su primjeri provedenih radionica uz edukacijske programe AERO TVZ, [12-13], u skladu s aktualnom legislativom, [14].

Prilikom obrade navedenih primjera za koje su korišteni arhivirani podaci službe održavanja s određenih aerodroma, eliminirani su slučajevi u kojima su greške i kvarovi na serijskim strujnim krugovima bili posljedica greške u proizvodnji kod određenog proizvođača. Također su eliminirani kvarovi i greške za koje je nakon reklamacije proizvođač priznao svoj propust i eliminirani su kvarovi nekih prototipova opreme koji su bili tretirani kao razvojni projekti, i kvarovi koji su bili posljedica nestručnog postupanja - neposjedovanje radne rutine i rad needuciranog osoblja za određene radne postupke, kao i propusti koji su kasnije bili pripisani montažerima opreme. Suradnja s proizvođačima opreme rezultirala je kod korisnika ostvarivanje bonusa i značajne uštede, a tijekom daljnjih eksploatacijskih spoznaja neki od proizvođača su promijenili i svoj pristup u proizvodnji nekih tipova opreme i izdanim uputama za ugradnju, montažu i održavanje.

Navedeni kvarovi obrađeni u Tablici 1. direktno ili indirektno su utjecali na stanje otpora izolacije serijskih strujnih krugova, odnosno utjecali su na ispravnost i pogonsku raspoloživost sustava svjetlosne signalizacije. Utvrđeno stanja otpora izolacije, kroz procedure održavanje, spoznaje su i pomoć aerodromskim stručnim službama. Naime u postupcima održavanja sa stečenim spoznajama moguće je kvalitetnije i kompetentnije održavanje odnosno raspoloživost kontinuirane funkcionalnosti što je direktni zahtjev vezan uz sigurnost zračne plovidbe. Iz opisa dijagnosticiranih kvarova prema Tablici 1. utvrđeni su uzroci na kojima su stečena znanja za buduću eliminaciju, preventivno djelovanje i otklanjanje kvarova.

Tako na primjer zbog udara groma i prolaska prenaponskog vala na instalaciju serijskog strujnog kruga uslijedile su poboljšane verzije prenaponskih zaštita. Najviše mogućih uzroka, (a mogući su i na svim drugim komponentama serijskog strujnog kruga) kada se eliminiraju gore navedeni u obradama su:

- greške i propusti završnih ispitivanja
- hodanje po kabelima postavljenim u i na metalne police u objektima ili zdencima prilikom montaže
- neprimjereno uvlačenje i navlačenje kabela tijekom izvođenja radova na kabelskoj kanalizaciji
- kritičan polumjer savijanja primarnih kabela prilikom uvlačenja i polaganja u zdence
- neadekvatno odabrana vrsta kabela - nemogućnost polaganja za određena projektantska rješenja
- prve tehnologije finožičanih kabela bez ekrana

Velik broj kvarova posljedica je i neadekvatno korištenog alata kojim se preša, (moguća greška u debljini pina konektora), greška dimenzije skinute izolacije, (neuredno uvijen snop žila vodiča, problem promjera primarnih kabela), oštećenje izolacije konektora prilikom nestručne izrade konektora, konektori koji se nakon nekog vremena odvoje zbog nečistoća, zarobljenog zraka i dimenzije dijelova konektora. Sve navedene spoznaje raspoložive su u bazama podataka ExSys AERO TVZ.

II. INTELIGENTNI SUSTAVI ODRŽAVANJA

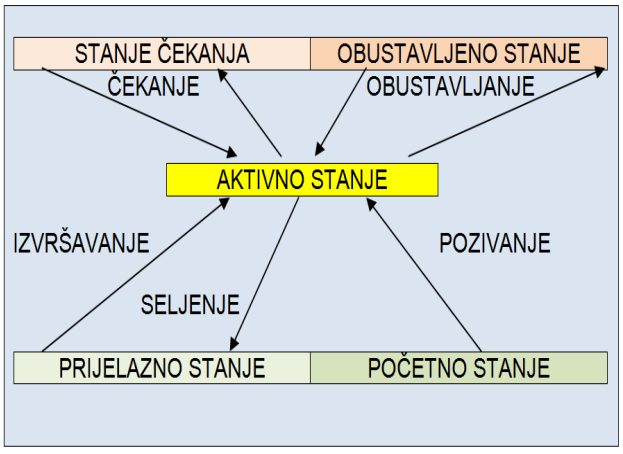
Na tim je osnovama specijalističko održavanje usmjereno na upravljanje kvalitetom, koje podrazumijeva prepoznatljivu i ujednačenu kvalitetu proizvoda i usluga, usmjerenost prema korisnicima, kreativnost i inovativnost, upravljanje procesima, resursima, troškovima i rizicima, komunikaciju i edukaciju. Upravljanje kvalitetom usmjereno je i na stvaranje zadovoljstva zaposlenika i poslovnog entuzijazma, poboljšanje uvjeta i kvalitete rada te razvoj svijesti o kvaliteti pojedinca i cjelokupne organizacije, tvrtke. Uvođenje i praćenje ključnih pokazatelja uspješnosti koja generiraju inteligentni sustavi, koristan je alat za smanjenje troškova održavanja, koje je moguće ostvariti kroz dobru strategiju održavanja, optimalne preventivne ili korektivne aktivnosti, optimalnu raspoloživost doknadnih dijelova, svrsishodnu organizaciju i kvalitetan menadžment. Praćenje ostvarenih ušteda moguće je kroz ekonomske pokazatelje iz baza znanja i baza podataka ekspertnog sustava. Koristan je primjer predstavljenih i standardiziranih ključnih pokazatelja uspješnosti u norma EN 15341:2007, koju je objavio Europski savez nacionalnih društva održavatelja, *(engl. European Federation of National Maintenance Societies)* .

Na tom tragu provedena su i različita dodatna istraživanja na hrvatskim aerodromima, kojima su prikupljeni stvarni podatci o troškovima održavanja i uporabe prema strukturi troškova. Za svaku od navedenih skupina troškova održavanja i uporabe popisane su sve ostvarene stavke spomenutih troškova koje su se javile tijekom promatranog vremenskog razdoblja. Obradom u ekspertnom sustavu došlo se do spoznaja o ostvarenim aktivnostima u održavanja i uporabi kao i implikacije na specifične skupine troškova.

Svakako da je održavanje serijskih strujnih krugova aerodromskih specijalističkih sustava u direktnoj vezi s troškovima, pa tako i postupanje aerodromskih specijalista se veže uz imperativ njihovog racionalnog postupanja. Poznate su razne krivulje troškova održavanja, koji se uglavnom žele svesti na najmanju moguću vrijednost a istovremeno da razina pouzdanosti ostane visoka. Kvalitetnim edukacijama i profesionalnim postupanjem to se i postiže kod aerodromskih specijalista. No u novije vrijeme otvaraju se i nove mogućnosti racionalnog postupanja u pogledu održavanja, naročito kada podaci o kvalitetnom održavanju mogu biti podloga za ostvarivanje određenih bonusa. Iako najčešće policu osiguranja određene opreme zračne luke ugovara ekonomsko-financijski sektor, konačna odluka o mogućem ostvarenju bonusa, (konkretno za dio opreme ALS), u kompetenciji je aerodromske službe održavanja. Primjeri ovih postupanja potvrđeni su u razmjeni iskustava s nekim europskim aerodromskim službama održavanja. Konkretan je primjer investicija s police osiguranja u opremu prenaponske zaštite na temelju koje je ostvarena bonifikacija u razdoblju eksploatacije bez havarijskih posljedica s naslova prenaponskih pražnjenja, a koja je korištena kao model za investicije zamjene primarnih kabela i konektora serijskih strujnih krugova s permanentno lošim otporom izolacije koji je izazivao ispadanje serijskih strujnih krugova. Stoga se može zaključiti kako unutar sofisticiranih održavanja, svi sustavi nadzora i prikupljanja velike količina podataka, omogućuju korisnicima stjecanje novih znanja između ostalog o racionalizaciji održavanja. Za to je vrlo bitno približiti takav sustav korisnicima i omogućiti im lakše snalaženje u toj masi podataka, što je inače i imperativ osnovne, bazne, zrakoplovne legislative, [15], kao i aktualnog unapređenja, [16-17].

A. Koncept nadogradnje inteligentnog sustava za obradu serijskih strujnih krugova

Koncept kreacije rješenja nadogradnje inteligentnog sustava za obradu serijskih strujnih krugova odvija se u fazama. U svakoj fazi inteligentni sustav ne prolazi kroz sva navedena stanja razvoja, već će neka stanja biti vezana sa ciklusima trajanja. Faze stanja mogu se prikazati međusobnim interakcijama, Slika 2.
Globalno shvaćanje kreacije koncepta kroz razvoj temeljeno je na raspoznavanju, sustava, podsustava i



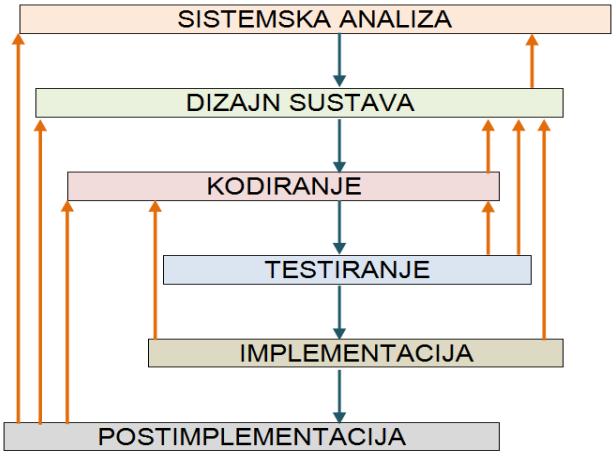
Slika 2. Faze stanja razvoja inteligentnog sustava

platformi. Pa tako u konkretnom slučaju imamo:

- Raspoloživi ekspertni aerodromski sustav iz kojeg učimo - (ExSys AERO TVZ

- Inteligentni sustav u kreaciji za nadogradnju – (InSys AIRPORT TVZ)
- Potrebu za pripadajućim agentskim platformama – (AgPlat AIR TVZ)

Koncept slaganja rješenja nadogradnje inteligentnog sustava aerodromskog elektro održavanja**,** svodi se na slijed koji započinje sistemskom analizom a završava postimplementacijom, Slika 3.



Slika 3. Slijed koncepta nadogradnje inteligentnog sustava

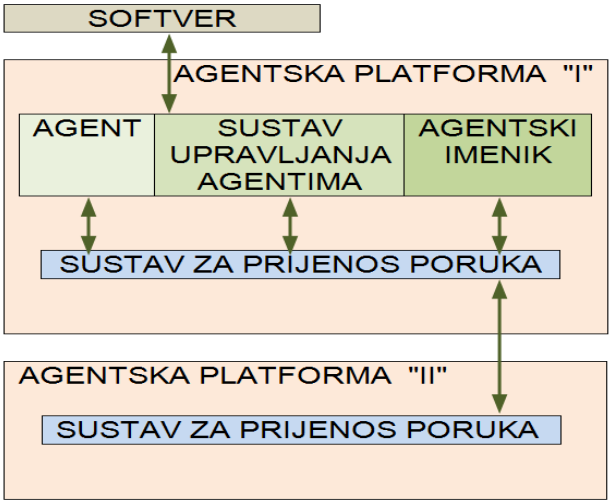
B. Arhitektura agentske platforme za održavanje

Ono, na što će biti stavljen akcent u finalizaciji nadogradnje inteligentnog sustava jest utvrđivanje potrebe za određenim agentskim platformama. U samoj definiciji, Agent, (*lat. agere)* znači voditi, djelovati, raditi. On se razaznaje kao entitet kojim se mogu „opažati" promjene i povratno „djelovati". To je netko tko djeluje ili ima mogućnost djelovanja ili zastupanja nekoga drugoga. Definira se kao sredstvo kojim se djeluje (instrument) ili pozicija s koje se djeluje pa je stoga uvriježen naziv – platforma.

Pojednostavljeno, agentska platforma, (opaža i djeluje), definira se kao nešto što djeluje u ime nekoga kako bi izvršio povjereni mu zadatak. Ona uključuje ljude, virtualne karaktere i računalne komponente, sustava upravljanja i sustava za prijenos podataka, Slika 4.

C. Nadogradnja agentskim entitetom

Svaki inteligentni agent, odnosno agentska platforma, pokazuje elemente inteligentnog ponašanja. On ima sposobnost djelovanja i u situacijama za koje nije unaprijed pripremljen, ima mogućnost učenja i samoispravljanja. Racionalnost agent postiže kao posljedicu svog djelovanja rezultirajući napretkom koji je potrebno vrednovati mjerom uspješnosti *(engl. Performance Measure).* Ta se mjera ponekad naziva i funkcija korisnosti *(engl. Utility Function),* koja



Slika 4. Agentski sustav prijenosa poruka i upravljanja agentima

predstavlja preslikavanje sa skupa ili sekvence stanja na skup realnih brojeva. Cilj je postići dobro ponašanje agenta, njegovu racionalnost (razboritost) o kojoj ovisi konačan uspjeh, što je moguće postići što boljim određivanjem okružja zadatka, i ocjene uspješnosti provođenja, koje definiraju inženjeri znanja, [18].

Prema namjerama ostvarenja unapređenja pogona i racionalizacije potrebno je osim same definicije i razlučiti djelovanje agente.
Agentska platforma se može postaviti kao složeni računalni sustav (budući je sastavljena iz više dijelova), smještena u okolinu iz koje može primati podražaje i koja može fleksibilno, autonomno djelovati u svrhu ostvarivanja vlastitih ciljeva. Dobro ponašanje agenta vezano je uz njegovu pronicljivost, *(engl. Clairvoyant)* i uspješnost, *(engl. Successful)*, pa se od njega očekuje fleksibilnost koja uključuje sljedeća svojstva:

- pristupačnost (prepoznaju svoju okolinu i odgovaraju na promjene u njoj)
- aktivnost (iskazuju i ponašanje usmjereno kroz preuzimanje inicijative)
- društvenost (sposobnost djelovati zajedno s drugim agentima i ljudima u svrhu ostvarivanja svojih ciljeva i u svrhu pomoći drugima u ostvarivanju njihovih ciljeva).

U ciklusima prema slici 2, početno stanje agentske platforme mora biti precizno definirano, odnosno moraju ga karakterizirati:

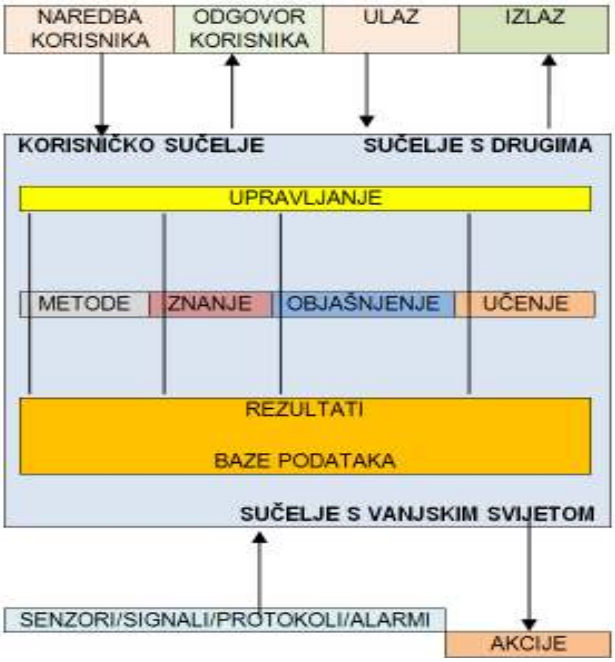
- mjere njegove uspješnosti
- njegovo okruženje
- stvarne akcije i rezultati

Raspoloživi ekspertni demo sustav iz kojeg se uči, dizajniran je i ima već ugrađene svoje senzore i aktuatore te kroz rad u praktikumu moguće je simulirati stvarno stanje serijskih strujnih krugova na aerodromu. Na terenu obradu serijskih strujnih krugova agentska će platforma proći određivanjem trenutnog stanja, koje neće uključiti čekanje, obustavljanje i prijelaz, već raspoloživost parametara utvrđenog stanja. Nadogradnja inteligentnog

sustava za obradu serijskih strujnih krugova obuhvaća samostalno djelovanje agenta. Akcent je stavljen na tehničku napravu, (u našem slučaju sofisticirano mjerilo otpora izolacije) koja samostalno rješava zadatke distribucije utvrđenih parametara u online sustavu. Na tim osnovama agentska platforma je proaktivna u dobro definiranom okružju i uz pomoć inženjera znanja ispunjava svoj zadatka, [19].

Njeno je okružje – operativna površina aerodroma i ona:
- utvrđuje trenutno i stvarno stanje otpora izolacije
- dijagnosticira kritične parametre
- šalje protokol o stanju sustava na temelju utvrđenih parametara na striktno zahtijevano mjesto
- ocjenjuje sigurnost zračne plovidbe na zračnoj luci i
- puni baze podataka

Tako u fazama razvoja, osim prikupljanja parametara broja i vrste modula koji se povezuju važna je interakcija i povezivanje unutar pa i s drugim inteligentnim sustavima ili perifernim jedinicama, kao što će biti nastavak integracije s ekspertnim sustavom u laboratoriju, Slika 5.



Slika 5. Sučelja za komuniciranje agentskih platformi i ekspertnog sustava

D. Kontrolna mjerenja – povjerenje u agenta

Za mjerenje i obradu otpora izolacije serijskih strujnih krugova korištena je agentska platforma oslonjena na ispitni uređaj nove generacije.

Njegove karakteristike su osim razine ispitnog napona koji u potpunosti može udovoljiti zahtjeve legislative, (10.000 V), i mogućnost online dostave izmjerenih podataka. Mogućnost trenutnog slanja podataka znači da se podaci direktno iz samog instrumenta šalju bez posredovanja računala ili bilo kojeg drugog pomagala. To je njegova posebna prednost koja dolazi do izražaja kada su zahtjevi za provjerom stanja vezani uz inspekcijske nadzore jer u tim slučajevima nema mogućnosti manipulacije podacima a niti uključivanje funkcije vremena za koju je dokazano da ima utjecaj na strujne krugove sa starijom i oštećenom izolacijom. To u pogonskim uvjetima znači da je moguće stvarno provjeriti stanje otpora izolacije u nekim ekstremnim klimatskim uvjetima, prilikom puštanja u pogon novih sustava, uključujući i fazne nadogradnje kao i akcidentnim događajima.

Osim slanja, velike količina podataka mogu se prikupljati i obrađivati kroz mehanizme zaključivanja u ekspertnom sustavu. Ta je mogućnost ostvariva pod pretpostavkom da mjerenje i obrada za bilo koji slučaj budu u kompetenciji, inženjera znanja, odnosno da se ima povjerenje u agenta. Elektrotehnički odjel, posjeduje odobrenje Hrvatske agencije za civilno zrakoplovstvo kao licencirana institucija za edukaciju, a odgovornost za dosljednost provedbe u rukama je Voditelja i Instruktora koji su svoje funkcije također potvrdili dobivenom licencom. To bi trebala biti garancija povjerenja uključujući i potvrdu o umjeravanju, [20].

Originalnost i inventivnost sustava potvrđena je razradom agentske platforme, primjenom nove generacije uređaja koji se koriste za provedbu ispitivanja, rezultati kojeg se online protokoliraju i dostavljaju na željenu adresu. Među parametrima moguće je izdvojiti osim stanja otpora izolacije i vrstu, tip, dimenzije, proizvođača kabela serijskih strujnih krugova, način polaganja, (kabelska kanalizacija ili zemlja), stanje zaštitne uzemljivačke trake i drugo. Da bi sve bilo povezano i uzeto u obzir, sudionici u izgradnji IS su osim eksperata, inženjera specijalista, inženjera znanja, i korisnici i osoblje na prikupljanju i obradi podataka. Svi oni sa svojim znanjem pomažu kreiranju algoritama, punjenju baza znanja i baza podataka, i u funkciji su podrške donositelju odluke.

Na tim osnovama, uvođenjem ekspertnih sustava u proces edukacije na TVZ-u pomalo prerasta u vizije rada u oblacima *(engl. Cloud Computing)*. Pokretanjem inicijative i suradnjom s privredom u namjeri konkretizacije ovakvih specijalističkih aktivnosti uključuju se i specijalističke tvrtke. Svojim iskustvom, spoznajama i djelovanjem one također sudjeluju u razvoju ekspertnih sustava ali istovremeno postaju i njegovi korisnici ili pružatelji usluga a vezano za npr. održavanje njegovih specijalističkih sustava. Takav angažman, prema predviđanjima razvoja i primjene ekspertnih sustava, bit će prihodovno orijentiran i prihvaćen, uvođenjem i pružanjem takvih usluga koje još uvijek očekuju višu razinu definirane zaštite intelektualnog vlasništva, [21].

III. ZAKLJUČAK

Sustavu svjetlosne signalizacije na zračnim lukama u funkciji sigurnosti zračnog prometa kroz mjerenje i ispitivanje ocjenjuje se pogonska pouzdanost. Rezultati parametara o stanju, koji se nadziru i arhiviraju u baze

podataka, moraju biti zadovoljavajući u skladu s propisanom legislativom.

Istraživanja na aerodromima i ispitivanja u laboratoriju, vezano na stanje otpora izolacije aerodromskih primarnih kabela serijskih strujnih krugova, uz učenje iz ekspertnog sustava potvrdila su mogućnost unapređenja procedura preventivnog i predikatnog održavanja, ali i ponudu online raspoloživosti izmjerenih podataka. Nabavljena nova ispitna oprema omogućuje sveobuhvatno ispitivanje, prikupljanje i obradu relevantnih podataka koji se mogu trenutno proslijediti na zahtijevano mjesto, bez prijenosa na računalo. Model inteligentnog sustava koji je na raspolaganju studentima specijalističkog studija Elektrotehničkog odjela i polaznici programa edukacije za stjecanje licence aerodromskih specijalista na Tehničkom veleučilištu u Zagrebu nadograđuje se za mogućnost „rada u oblacima“. Specifičnim zahtjevima, bit će moguće udovoljiti kada se ostvare zaštite intelektualnog vlasništva, ali i potvrde povjerenja u agentske platforme koje se u funkciji tih provođenja. Ideja budućnosti takvog rada u virtualnim okruženjima u osnovi podržana ekspertnim sustavima, osim mogućnosti uporabe baza podataka i baza znanja u "bezgraničju" razvijat će se u smjeru pružanja usluga s naslova pouzdanih, provjerenih i zaštićenih podataka.

LITERATURA

[1] Z. Balaž, K. Meštrović, G. Bjelić, "Inteligentni sustav održavanja aerodromskog sustava svjetlosne signalizacije", 37. međunarodni skup za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku (MIPRO 2014) : Hrvatska udruga MIPRO, p.p. 1488-1492, Zagreb, svibanja 2014.

[2] Z. Balaž, “Expertiza specijalističkih sustava u funkciji sigurnosti zračne plovidbe na Zračnoj luci Mostar i prijedlog rješenja za unapriđenje”, Studija br. 10056/01 - Aeroing d.o.o., Zagreb, 2001.

[3] Z. Balaž, “Analiza rezultata ispitivanja stanja otpora izolacije serijskih strujnih krugova Aerodroma Brač, Zračne luke Zagreb i Airport Nafoora”, No: F-68/02, Aeroing d.o.o., Zagreb, 2002.

[4] Z. Balaž, “Upgrading of Airport at Nafoora Field – Arabian Gulf Oil Company”, Benghazi, Lybia, Aeroing d.o.o./SINIT Rovigo, Italy, 2002.

[5] Z. Balaž, „Utvrđivanje stanja otpora izolacije serijskih strujnih krugova Aerodroma Brač“ – Prilog Studiji unapređenja sustava svjetlosne signalizacije operativnih površina, Br. 024/3-14, Aerodrom Brač, listopad 2014.

[6] Z. Balaž i K. Meštrović, „Studija unapređenja sustava svjetlosne signalizacije operativnih površina Aerodroma Brač, iTVZ d.o.o. – Zagreb, Studija, Br. 024/3-14, studeni 2014.

[7] Z. Balaž, “Primjena računala u procesu održavanja manevarskih površina zračne luke – Programi za građevinsko i elektro održavanje“, Stručni skup ISEP 98, Elektrotehnička zveza Slovenije, Ljubljana, 1998.

[8] FAA Department of the Army Maintenanca, FM 3-04.300, Washington, D. C. , 12 August 2008.

[9] Narodne novine RH, “Pravilnik o održavanju i pregledavanju aerodroma te mjerama potrebnim za njegovu sigurnu uporabu”, NN br. 65/05.

[10] Z. Balaž, “Priručnik za osposobljavanje djelatnika koji na zračnim lukama obavljaju specijalistička održavanja“, TVZ, Centar cjeloživotnog obrazovanja, Zagreb, 2013.

[11] Narodne novine RH, “Pravilnik o stručnom osposobljavanju zaposlenika i drugih ugovornih djelatnika aerodroma, pružatelja zemaljskih usluga i korisnika usluga aerodroma koji samostalno obavljaju zemaljske usluge“, NN br. 7/13, Zagreb, 2013.

[12] Z. Balaž, „Radionica/Workshop AERO TVZ-Iso Program, Training for Insulation Resistance Measurements on, Serial Circuits – Airfield Lighting Systems”, Zračna luka Rijeka - Airport Rijeka, Krk, 28.05. 2014.

[13] Z. Balaž, „Radionica/Workshop AERO TVZ-Iso Program, Training for Insulation Resistance Measurements on, Serial Circuits – Airfield Lighting Systems”, Zračna luka Pula- Airport Pula, Pula, 06.10. 2014.

[14] ICAO - Aerodrome Design Manual, Part 5, Electrical systems, Doc 9157-AN/901; " Electrical Circuits for Aerodrome Lighting; Electrical Tests of Series Circuit Equipment; Acceptance Testing of Aerodrome Electrical Circuits;

[15] ICAO -Annex 14, Volume I, “Aerodrome design and operations”, 6th Edition, July, 2013.

[16] Uredba Komisije (EU) br. 139/2014., od 12. Veljače 2014. O utvrđivanju zahtjeva i upravnih postupaka u vezi s aerodromima u skladu s Uredbom (EZ) br. 216/2008 Europskog parlamenta i Vijeća.

[17] Airport Design and Certification, Consolidated version of Aerodromes Implementing Rules and related Acceptable Means of Compliance and Guidance Material, and Certification Specifications and their related Guidance Material, Issued March 2014, ADR, Requirements and administrative procedures related to aerodromes, Commission Regulation (EU) No 139/2014, including AMC and GM to ADR rules, ED Decision 2014/012/R and Certification Specifications (CS) and Guidance Material (GM) for Aerodromes Design CS-ADR-DSN, ED Decision 2014/013/R

[18] Z. Balaž, " Uloga inženjera znanja u održavanju aerodromskih i tunelskih ekspertnih sustava", Zbornik radova Elektrotehničkog odjela, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, God.1 br.1, ISSN 1849-5621, p.p. 101-116, Glavni urednik: Krešimir Meštrović, Zagreb, 2014.

[19] G. Bjelić, "Mjerenje i obrada otpora izolacije aerodromskih serijskih strujnih krugova inteligentnim sustavom", Diplomski rad, Kolegij: Inteligentni sustavi, Mentor, Zdenko Balaž, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, siječanj 2015.

[20] Narodne novine RH, „Pravilnik o mjeriteljskim zahtjevima za mjerila otpora izolacije”, NN br. 81/02, Zagreb 2002.

[21] Z. Balaž, “Inicijativa uređenje zaštite intelektualnog vlasništva stručnih i znanstvenih radova”- Članak pripremljen za Časopis Politechnic & Design, Tehničko veleučilište u Zagrebu, Vol III; No. 2; 2015.

ISS

International Conference on

INFORMATION SYSTEMS SECURITY

Steering Committee

Chairs: Stjepan Groš, University of Zagreb, Croatia
Tonimir Kišasondi, University of Zagreb, Varaždin, Croatia
Željko Hutinski, University of Zagreb, Varaždin, Croatia

PAPERS

TECHNICAL TRACK

Disabling Same Origin Policy for Automatization of Webform Data Entry

Samir Ribić*, Adnan Salihbegović * and Alvin Huseinović**
* University of Sarajevo, Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina
** Enit d.o.o., Sarajevo, Bosna i Hercegovina
samir.ribic@etf.unsa.ba, adnan.salihbegovic@etf.unsa.ba, alvin.huseinovic@enit.ba

Abstract - There is often a need enter data in online documents. Web based interface is popular, but sometimes there is a need to fill the data using scripting and automation. JavaScript may enable automatic data entry, but it requires that script written in JavaScript executes on the same server where the form is being processed, due to protection mechanism of "same origin policy". If needed, some browsers support disabling this policy, or there is alternative method to bypass it. The paper describes this method and includes a sample JavaScript code for automated data confirmation.

I. INTRODUCTION

There are different ways to interact with a computer: command line interface, textual interface, graphic user interface, web interface, gaming interfaces, voice recognition, pen and tablet interface, etc. Inserting large amounts of similar data in online document can easily be automated by using scripting tools and languages developed for command line, although the main disadvantage of using command line is that it is inappropriate for users with lack of experience, and it is considerably prone to errors. Textual and graphical dialogs made the data entry less error prone, but also harder for automatic data entry. It is especially a case when data entry is simple, for example, if one needs to approve large amount of records. The discrepancy between ease of use and possibility to automate the data entry is present also in web interface.

In the second section of the paper will review typical methods for automating data entry in web applications. Web has one ubiquitous scripting language JavaScript, but it has one security property, which makes the automation difficult, called "the same origin policy". The third section of the paper will review popular methods for avoiding this policy. In the fourth section we will illustrate one method of save origin policy avoidance, here is a JavaScript source sample, developed by us, which performs mass data approval on the web site used for the co-ordinated localisation and translation of software. The empirical results for modified workflow will be also given.

II. WEB INTERFACE AUTOMATION

The development of web technologies has created web applications that execute in a web browser. The scripts can execute on the client or server side, depending on the application appliance. Web applications are quite flexible, as they can be quickly changed on the server side without client side interventions. Initially, web interface was designed for entering a minimal amount of data (such as registration forms), but it soon became used for entry of different kind of data. However, when entering a large amount of the similar data, typical data entry process (entering the form, submitting and waiting for server response) might be tedious and slow.

Sometimes this problem is not only related to the design of web application, but also to the application workflow. For example, in a web shop, an officer needs to approve the orders, by clicking check boxes, each of them, separately. After the previous review the officer concludes that all pending orders are correct, and wishes to confirm all of them at once, but the user interface allows only approval one order at the time.

To avoid described situation, web servers offer web services or other kind of API to facilitate automated data entry. An officer who often repeats same queries in web searches may prepare locally hosted web pages which aggregate the queries, as in the following example.

```
<!DOCTYPE html>
<html>
<body>
<iframe width=100% src=
"http://www.bing.com/search?q=weather+paris">
</iframe><br>
<iframe width=100%
src="http://finance.yahoo.com/q?s=TSLA">
</iframe>
</body>
</html>
```

The officer can refresh the page to retrieve the actual data without re-entering the data over and over again. Unfortunately, in web applications, this kind of interface may be unavailable, too complex, and available only for a charge or undocumented. When implementation of such interface on server side is not possible, it is often necessary to try with different approaches to simulate data entry and achieve automatic entry on the web forms. Such approaches might be tried on client side or on proxy.

There are three obvious approaches. The first approach would rely on APIs in the local operating systems. For example, Windows program which sends `WM_MOUSEMOVE` and `WM_KEYPRESS` messages to the web browser window could simulate data entry into the form.

The second approach is based on modified HTTP requests. Instead sending the form data from web browser, it is possible to programmatically prepare a set of HTTP requests with content similar to those generated from the entered data. By using tools (ex. `curl`), one can send the packets.

The third approach would use JavaScript and analyse the remote web page. Document Object Model[1] allows access to each part of the web page, including all elements of web forms. A JavaScript program fills up the web form and simulates required mouse clicks and keyboard actions. For example, one can write this piece of code :

```
var nm=window.document.getElementById("name");
nm.value='John Smith';
```

However, there are obstacles to implement all three approaches. The first approach is quite vulnerable to web page layout and size. It is difficult to predict the exact location on the screen where the button or data entry field will be located. Furthermore, a web browser may intercept system messages or render web pages differently.

Modified HTTP requests have also some obstacles. If they are sent from external software, they need to simulate complete login procedure in a way a browser did it (or to reuse an existing session with browser). Server often checks `Origin` and `Referer` HTTP header as a security measure against Cross site request forgery[2] to be sure that request came from browser, not the external tool. The programmer needs to correctly fill these fields in request. Additionally, the external tool needs to cope with certificates and protocols if HTTPS is used, which makes writing of the external tool more difficult. This is not problem if HTTP request is sent from JavaScript inside browser, using `XMLHttpRequest` object, because SSL library is transparent, but `XMLHttpRequest` object is subjected to Same Origin Policy.

III. Same Origin Policy And Avoidance Methods

A. *Same Origin Policy*

There is one security rule that makes it difficult to use JavaScript as a tool for data entry. It is called the Same Origin Policy. It was introduced early in JavaScript history, with Netscape Navigator 2.0 and it prohibits websites from different domains to interact with each other, except in a limited way [3]. For example, there are popular sites (Amazon, Facebook, PayPal, Ebay), and it is not uncommon that the user opens the tab or window containing session to them. At the same time, this user opens another site in a different window, without knowing that it contains JavaScript code, which reads credit card information from Amazon session or sends an embarrassing message from user's Facebook account. On the other hand, it is desirable that JavaScript initiated from Facebook server is aware of all Facebook windows opened on the client's computer.

Same Origin Policy is established on the client side. It is a purely web browser capability. It allows JavaScript to access the DOM contents of the other web page only if it belongs to the same server, domain, port and protocol, as defined in the standard proposal [4]. Directory and file names need not to be the same. So, `http://www.ex.com/test.js` may access structure from `http://www.ex.com/a/h.htm` but not from `http://www.sam.com/hello.htm` (other domain), `http://w2.ex.com/h.htm` (other host) or `https://www.ex.com/h.htm` (other protocol).

This limitation means that web page influenced by JavaScript and the file with a script needs to be on the same server. This is often not possible to implement.

Two common situations when the same origin policy limits the software usability are JavaScript access to document structure and AJAX request using `XMLHttpRequest` to access web services through JavaScript.

With caution, same origin policy can be avoided if it is necessary. This is not a recommended activity, and should be used with care. This means there is a need to disable same origin policy only during specific activities and re-establish thereafter. While the browser is in insecure environments, it is prohibited to visit the other sites except the ones planed one for data entry scripting.

B. *Macro recording and testing tools*

An approach, which sometimes does not require programming, relies on software that records keys on keyboard and mouse movement and clicks. They can be saved in macro and then replayed, although the use of this method is limited to cases where the same action needs to be repeated many times and if screen layout remains unchanged during the whole web session. Some macro recorders are expanded to full IDE, like Selenium IDE, a Firefox plugin for editing macros in browser, and Selenium WebDriver, a set of language bindings for web browser automation, both of them are available at http://www.seleniumhq.org.

C. *Browser Option*

The easiest method to switch off the same origin policy is to utilize browsers which have this as an option. This also means that one has to avoid these browsers or options in other situations. One example of such a browser is Google Chrome in versions 5 to 20. This browser can be started with an option

```
chrome.exe --disable-web-security
```

This option initially was intended for testing purposes during web development, but here it will be used for automating scripts on sites where automation is not possible by default. In a similar way, in the Mozilla Firefox, it is possible to open `about:config` page and change `security.fileuri.strict_origin_policy` setting to `false`[5].

The disadvantage of this approach is a limitation to a specific web browser version, and need to start it temporarily in non-secure environment. When this is not acceptable, there are other possibilities to avoid the same origin policy.

D. *Change Document Domain*

One approach [6] is designed for the case when the JavaScript program and the web form have the same domain, but not the same host. For example, an university has a common information system, but each department also has its own system. The university system is located at `univ.edu/info.php` while CS department has a system at `cs.univ.edu/start.php`. A program at the CS department (for example data export program) in JavaScript program may contain the line

```
document.domain = "univ.edu";
```

After this statement is executed, the page would pass the origin check with http://univ.edu/info.php. However, it is not allowed to change the domain to otheruniv.edu.

E. *Policy Files for Flash Applications*

The applications in Macromedia Flash can allow cross-domain access using configuration file `crossdomain.xml`[7]. Inside the XML element one can put the tag to allow flash applications to modify data locally, which come from a specific domain.

F. *Proxy Servers*

For frequent web form filling on a remote domain, one can use proxy server. The user can host a proxy server locally to access a web page `http://www.rem.com` using ex. `http://www.local.ba/co/www.rem.com`. A script at `http://www.local.ba/script.js`, from a perspective of web browser will not void the same origin policy.

G. *Cross Origin Resource Sharing*

One recent approach is called Cross Origin Resource Sharing (CORS).[8] It is suitable for AJAX requests. During request, an additional HTTP header may be set as in the following example.

```
Access-Control-Allow-Origin: http://www.srvr.com
```

The browser performing an Ajax request should now ignore the origin.

H. *JSONP*

JSONP stands for JavaScript Object Notation with Padding[9]. Assume that the call to the server `http://www.srvr.com/act.php` results with a JSON document `{"name":'milk',"price":1.8}`. Servers, which return data in JSON format, can also call a callback function when submitted.

On these kind of servers a call to `http://www.srvr.com/act.php?jsonp=Process`, will result with calling the function

```
Process({ "name": 'milk' ,"price":1.8}).
```

This method requires that the server is aware of cross-domain requests.

I. *Use of XdomainRequest and JSONRequest*

In addition to standard `XMLHttpRequest` for Ajax calls Internet Explorer from version 8 has `XDomainRequest` object.[10] This object allows an asynchronous call to HTTP server located on a different domain from the one where the JavaScript program is executed.

```
var obj = new XDomainRequest();
obj.open("post","http://www.serv.com/act.aspx");
obj.send();
```

`JSONRequest` is similar object, installed as additional library but supported by a wider variety of browsers.

J. *Use of GreaseMonkey*

There is a plugin for Firefox called GreaseMonkey [11]. It has `GM_xmlhttpRequest`, a version of the `XMLHttpRequest` which does not use the same-origin policy, object unsafeWindow which provides access to the raw JavaScript window scope of the content page. The unsafeWindow is most commonly used to access JavaScript variables on the page. Usage of this plugin is similar to browser option approach, except for different names of used objects. It is recommended to switch it off when not used.

K. *Use of window.postMessage*

Calling `window.postMessage` method is useful for selectively cross origin scripting [12]. This method causes a `MessageEvent` at the target. The script that receives the data should install the listener using the method `window.addEventListener("message", handler, false);`

In the handler function, it is possible to check for `event.origin` property if it matches allowed domains. The message can be sent using the `postMessage` method over the window opened with `window.open` method.

IV. EXAMPLE

As an example of the situation where disabling the same origin policy might simplify data entry, we will present here our real world JavaScript code. It is used here as an illustration of accessing a form data in JavaScript on remote server, with mouse simulation.

During the localization project of Linux operating system for Bosnian language, we used Launchpad Rosetta[13] system. Typically, such projects might have bigger number of the contributors (translators) than the

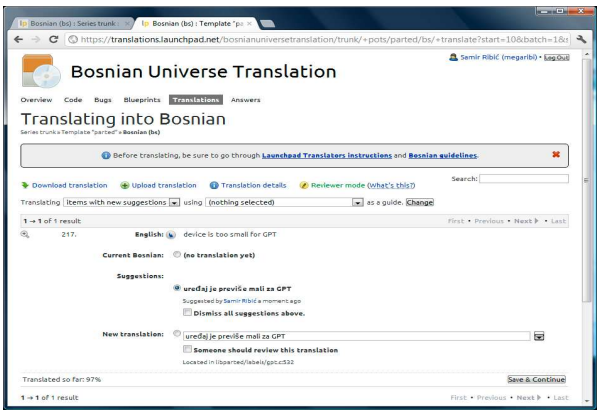


Figure 1. Translation suggestion in Launchpad Rosetta

approvers. In our project, we have had about 70 translators and one approver, who needed to check about 80000 strings. When the approver wants to remark that the contributor works satisfactory, he may start accepting his translation proposals, by clicking radio buttons, and submit buttons. It seems that the job was trivial, just to look at the text and click one radio button, then one submit button (as shown in the screenshot, Fig. 1).

However, due to the fact that the majority of the contributors work just before the deadline, this might result in tedious job for the approver to complete his part in a limited time. Launchpad servers are often overloaded and manual clicking and confirming fails and require resubmitting. Manually, one approver could confirm up to 2000 strings daily. It would be easier to confirm with one script all the proposals from reliable contributors and then to fix only small percent of the strings, which contain translation errors. However, writing such JS string, without accessing the server, is prevented by the Same Origin Policy.

After surveying the different methods for the same origin policy avoidance, we have chosen the browser option method. We decided for it because: it was easiest to implement, it is intended for general JavaScript programming (not only AJAX), it does not require proxy servers, and the JavaScript program can be executed from the local machine. The major drawback was the need for searching older version of Chrome browser. The server side will not recognise it as malicious attack, because the browser will continue to send usual HTTP payload, here we simulate only client side activities.

After the browser Chrome is started with disabled web security, a JavaScript program similar to the following might be used locally.

```
<!DOCTYPE html>
<html>
<head>
<script>
var myWindow,myInterval,finished,poz,tries;
/* chrome.exe --disable-web-security  */
urls=[
"https://translations.launchpad.net/ubuntu/trunk
/+pots/gcc/bs/+translate?show=new_suggestions&ba
tch=1",
"https://translations.launchpad.net/ubuntu/trunk
/+pots/dpkg/bs/+translate?show=new_suggestions&b
atch=1",""
]

function isEmpty( variable ) {
  if ( typeof( variable ) == 'undefined' )
    return true;
  if ( variable  == null )
    return true;
  return false;
}

var dispatchMouseEvent=function(target,var_args)
{
  var e = document.createEvent("MouseEvents");
  e.initEvent.apply(e,
    Array.prototype.slice.call(arguments, 1));
  target.dispatchEvent(e);
};

function afterSubmit() {
  if (isEmpty(myWindow))
    return;
  var u=myWindow.document.URL;
  if (isEmpty(u))
    return;
  window.focus();
  clearInterval(myInterval);
  if (urls[poz]!='') {
    myWindow=window.open(urls[poz], 'scripted');
    myInterval=setInterval(function(){
      doModify() },2000);
  }
  else  {
    myWindow.close();
  }
}

function doModify() {
  if (isEmpty(myWindow)) return;
  if (isEmpty(myWindow.document)) return;
  var all = myWindow.document.
     getElementsByTagName("input");
  for (var i=0, max=all.length; i < max; i++) {
    if (all[i].id.search("bs_suggestion")>0 &&
     all[i].type=="radio") {
       b=all[i];
       b.focus();
       dispatchMouseEvent(b,'click',true,true);
       break;
    }
  }
  var b=myWindow.document.getElementById
   ("save_and_continue_button");
  if (!isEmpty(b)) {
    b.focus();
    dispatchMouseEvent(b,'click',true,true);
  }
  var all = myWindow.document.
    getElementsByTagName("div");
  for (var i=0, max=all.length; i < max; i++) {
    if (all[i].innerHTML.search
      ("There are no messages")>0) {
       poz++;
       break;
    }
  }
  clearInterval(myInterval);
  myWindow.focus();
  myInterval=setInterval
    (function(){afterSubmit()},2000);
}

function startScripting() {
  poz=0;
  if (urls[poz]!='') {
    myWindow=window.open(urls[poz], 'scripted');
    myInterval=
      setInterval(function(){doModify()},2000);
  }
}

function stopScripting(){
  poz=0;
  clearInterval(myInterval);
}

</script>
</head>
<body>
<p>Click the button to trigger a function.</p>
<button
onclick="startScripting()">Start</button>
<button onclick="stopScripting()">Stop</button>
</body>
</html>
```

The array `urls` contains a list of HTML files with translation suggestions. This array can be generated automatically by filtering Launchpad web page, which shows new translation proposals using `grep` and `sed` tools. The approver may remove translation proposal URL

from the array if he concludes that it comes from unreliable translator.

Function `startScripting` fetches the first element from `urls` array, and opens it in another tab. In the Chrome, JavaScript calls are asynchronous and there is the `setInterval` function to repeat calling `doModify` function.

The function `doModify` will check if the page is ready and it will find the HTML element with tag name INPUT and type RADIO, send the mouse event, which will simulate the mouse click. This will confirm the translation suggestion and the next mouse event will click the `save_and_continue_button`. The script will stop with periodical calling of `doModify`, and it will start checking the response by repeated calling the `afterSubmit` function. In the `afterSubmit` function, when response is received, the script will process next `url`. The function is specific to HTML structure of the web page we want to update.

The function `stopScripting` disables the interval calling and therefore stops webpage data populating. Both `startScripting` and `stopScripting` are linked to HTML buttons.

With an automatic confirmation, started from ten client machines, it was possible to confirm 20000 strings in just three hours, the biggest part of this time being attributed to the server delays. Although the translation review is postponed, the list of accepted translations in Launchpad is quickly generated, since it was not overloaded with the GUI part of HTML. The approver finally watches the translations and fixes only those translations that were necessary. The similar script, but intended for situations where edit fields are involved was presented in [14].

V. Conclusion

Web browsers that are in use today, include the Same Origin policy as a security measure for the system by preventing the execution of unwanted scripts on the remote server. This makes the Same origin policy an important part of Web application security. Generally, it is bad practice and dangerous to avoid this policy. However, there are situations that can improve web application usability when this policy is disabled. Examples are the situations when we need to automatically enter big amount of data, but we have only web interface available. In that case the user has to ensure that the application runs in the controlled environment, which cannot in any way execute malicious scripts.

Reference

[1] G.Nicol,, L. Wood, L., M.Champion, S. Byrne,. “Document Object Model (DOM) level 3 core specification”, World Wide Web Consortium, 2001

[2] N. Tandel, K. Patel, “Mitigation of CSRF Attack”, International Journal of Science and Research (IJSR), vol. 3. Issue 6, 2014

[3] C. Jackson, A. Bortz, D. Boneh, and J. C. Mitchell. 2006. “Protecting browser state from web privacy attacks”. In Proceedings of the 15th international conference on World Wide Web (WWW '06). New York, USA

[4] A. Barth, "The Web Origin Concept,"26 Nov. 2010; available http://tools.ietf.org/html/draft-abarth-origin

[5] V. Massol and T. Husted. “JUnit in Action.” Manning Publications Co., first edition, Nov. 2003, pp. 26

[6] D. Flanagan, JavaScript (2nd ed.): the definitive guide, O'Reilly & Associates, Inc., Sebastopol, CA, 1997 , pp. 402

[7] S. Lekies, M. Johns, M., W. Tighzert,: The state of the cross-domain nation. In: Proceedings of the 5th Workshop on Web 2.0 Security and Privacy, W2SP (2011)

[8] A. Van Kesteren, "Cross-origin resource sharing." W3C Working Draft WD-cors-20100727 (2010).

[9] F. Wojcieszyn,. "Cross Domain and Push Communication." ASP. NET Web API 2 Recipes. Apress, 2014.pp 235-236.

[10] J. Lam, J.B. Urlich “Cross Site Request Forgery: What Attackers Don’t Want You to Know A Study of Browser Implementations and Security Mechanisms for XMLHttpRequest and XdomainRequest”, Working Papers in Application Security, Sans september 2014

[11] Pilgrim, Mark. Greasemonkey Hacks: Tips & Tools for Remixing the Web with Firefox. " O'Reilly Media, Inc.", 2005.

[12] A. Barth, C. Jackson, J.C. Mitchell, J. C. Securing frame communication in browsers. Communications of the ACM, 52(6), 83-91. . (2009)

[13] M. Helmke, A. Graner, K. Rankin, B. Hill, J. Bacon, “The Official Ubuntu Book”,. Prentice Hall Press., 2012, pp 258-259

[14] S.Ribić, “Automatizovano popunjavanje web formulara zaobilaženjem pravila istog izvora”, Infoteh .Jahorina, 2012

Towards a Rule-based Access Control Framework for Distributed Information Systems

Sergey Afonin
Institute of Mechanics, Moscow State University
Michurinskij av., 1, Moscow, Russian Federation
Email: serg@msu.ru

***Abstract*—Rule-based access control is a powerful and flexible paradigm for specifying security policies in information systems. Access may be granted or denied depending on object's and subject's properties, or the context of the request, such as current time. Complex access control rules are typically implemented at application level. In this paper we describe a framework targeted to separation of access control rules from the main application logic. It is assumed that access policies are managed in a distributed fashion (discretionary access control), objects are stored in relational database, and requests are independent. Such assumptions correspond to a RESTful web application. We describe a proof of concept implementation and discuss efficiency and decidability issues of access checking and policy verification problems.**

I. Introduction

In many information systems access control rules are hard-coded into main application logic. This is typically so for Web-centric systems. When page request arrives, the server first validates that the requested operation can be executed by the requester. Application level access control rules could be quite complex and depend on various parameters, including object's and subject's properties and context of the request. For example, one can define that a user is allowed to modify document if he belongs to some group (subject's property) and the document was signed (object's property) no more than two days ago (request context). Such rules can be easily implemented in the target programming language (PHP, Python, Ruby or Java in case of web application) and typical application contains code like "return an error message when access conditions are not met, or process the request". The obvious drawback of hard-coded access control rules is that application code should be altered each time the security policy changes. Any modification of the source code, in turn, requires repeated testing, and so on. Moreover, if access control rules are implemented in a Turing-complete language one can not automatically validate any non-trivial property of the implemented security policy.

Classical and the most frequently used in real systems Role-based Access Control (RBAC) model [1] separates application logic form users grouping. It is based on the assumption that roles of an organization are static, while users and their responsibilities might change frequently. RBAC standard provides mechanisms for reducing cost and potential for errors when assigning users to roles within an organization. Static nature of RBAC model makes it less useful for web applications characterized by dynamic access control rules. In basic RBAC model it is difficult to express many natural constraints, such as rule "registered users can modify *their own* objects".

In order to overcome limitations of RBAC model, a number of extensions have been proposed in the literature. These extensions form an actively developing research area frequently referenced to as *attribute-based access control* (ABAC) or *rule-based access control*, allow one to specify security policy by means of *rules* that depend on application level objects' *attributes*. According to [2], [3] such extensions fall into following three classes:

- Dynamic roles. Roles are assigned to user dynamically, depending on user's and/or context's attributes.
- Attribute-centric. Roles are considered as subject's attributes. In contrast to RBAC model this approach does not assume permission to role relation.
- Role-centric. Attributes are used to constrain RBAC roles. Constraint rules can only reduce user's permissions.

Rule-based access control appears in many domains, including information systems, database systems, GRID and Cloud computing, Linux kernel, social networks, XML document-security and many others, see e.g. [4]–[9]. For the moment there exists no commonly adopted model for attribute-based access control. In this paper we discuss a framework motivated by the development of a real system. We believe that methods and tools could be abstracted from the needs of that particular system, yielding a general framework suitable for wide class of information systems. Our goals are: separation of security policy from application logic and ability to perform static analysis of the policy.

It is worth noting that development of a general rule-based security framework requires number of complex issues to be solved. First of all, security policy should be defined in terms of application-level objects, so a kind of object-relational mapping (ORM) have to be supported. In many research, e.g. [5], [10], rules can use fixed number of application level objects' properties, and these properties are computed by dedicated functions or services written in general purpose programming languages. Application developers need to implement appropriate set of functions, so this approach can not be considered as a general solution. The second issue that should be addressed, is the efficiency of an on-line permission checker. As final decision depends on database state, permission checks might produce undesirable database workload, so permission checker's efficiency may be measured in terms of number of database queries. Finally, static validation of specified policy for redundant or conflicting rules, similar to [11], [12], has to be done.

The paper is organized as follows. Next section contains brief description of key properties of information system that triggered this research. We also formulate main goals of the proposed security framework. In Section III we describe rule definition language. Algorithmic issues related to policy enforcement and verification are discussed in Section IV.

II. Motivation and Goals

A. Motivating Example

This research was motivated by the development of a Current Research Information System named ISTINA[1] for Moscow State University. Let us briefly describe its properties and security-related issues.

The system is used to collect information about research and teaching activities (publications, research projects, participation in scientific conferences, membership in program committees, awards, etc.) of university's staff. Some information, such as list of publications in scientific journals, may be imported or verified by accessing external resources, but significant amounts of data are not available from other sources in unified form. As a result, researchers registered in the system are allowed to maintain their own pages. In order to perform more or less reliable scientometric reporting, most critical information is manually verified by authorized users. An authorized user can modify any object authored by anybody from his area of responsibility.

Our access control rules are heavily based on authorship relations, object's ownership information, and organization hierarchical structure. For example, "an article can be edited by any of its authors", or "an authorized user can modify any object created or authored by researchers from his department or its child departments". Some information, such as financial data, may be inaccessible to anybody but authorized users. Role membership, e.g. who are considered as authorized user, may be parameterized by department or other data. For example, membership of a user in *administrative staff* role is limited to department and its descendant departments.

Both permissive (or positive) and restrictive (or negative) rules, defining who are allowed (or not allowed) to perform an operation on specified object are used. Access to individual objects should also be managed. For example, one can grant or revoke editing permissions on specific object to someone who is not in the list of authors, or in the authors list, respectively. Time constrained access is also important, for example, when an administrator wants to be sure that no objects will be modified after the specified deadline.

Another usage scenario of the security framework is related to reporting subsystem [13]. Roughly speaking, a report is a parameterized SQL query. Depending on supplied parameters' values, a user may be granted access to the report, or not. For example, a dean can access data about his department and its parts, but should not access information about other departments. In both cases, the same report is requested, and this problem can not be resolved by standard access control approaches.

In this paper we address only one scenario when all access requests are treated independently and each request is performed on only one object. In other words, there is no interaction between subsequent requests from the same user. This assumption is reasonable for RESTful web applications. If a web page contains information collected from several objects, then this page may be protected as a whole.

It is assumed that each department, or each university that uses the system, can define its own security policy and these policies are managed independently. This distributed nature can produce conflicts in case of joint work. For example, if department A allows anybody to delete any object, and department B does not allow to delete objects at all, is it possible to delete a paper authored by researchers from both A and B? At runtime restrictive rule would take precedence as a safer option, but system's behavior will be surprising for some users, regardless of the decision made.

Finally, some rules of the policy defined at top administrative level, e.g. university, may be redefined on lower levels, e.g. faculties and departments. The list of rules that may be redefined or overwritten by a "subordinate" policy writers should be defined by the top level administrator. From the other hand, policy writers from hierarchically unrelated administrative units could produce any rules. Policy conflicts, similar to conflicts caused by coauthored objects, arise when a user has relation to several administrative domain, for example, as a coworker in two institutions. Such conflicts should be resolved dynamically (at runtime), or statically, during policy validation step.

B. Goals and Requirements

The main goals of the security framework are:

- Separation of access control logic from application's code. Modifications of security policy should not lead to any changes in application's source code.
- Static policy checking. In a distributed environment, when several organizations manage their own policies affecting common database, an automated testing for potential conflicts should be possible.

It is assumed that application developers provide "atomic" permissions that are checked in appropriate points. For example, if a user sends a request for object deletion, then application checks that this user can delete this object by issuing function like

```
has_permission(user, object, "delete").
```

If an object consists of several parts, e.g. paper title and authorship information, access to that parts could be controlled by additional rules and verified by dedicated calls to above mentioned function. For example, a policy could specify that every author of a paper can update title, abstract, and his own authorship information but can not update information about his coauthors. In any case, all checks for specific permissions must be hard-coded, because providing fine-grained access to object's parts can not be realized without programming efforts.

Specific security policy should not be defined or implemented in application's source code. When application calls *has_permission* function providing it with information about

[1] http://istina.msu.ru/

current user, object and requested operation, that function should perform all necessary checks in accordance to specified policy. In order to guarantee policy enforcement, such calls to *has_permission* function may be performed automatically when the request arrives to application server. This can be done by variety of means, including preprocessing of incoming HTTP requests, defining appropriate middleware of application server framework, or using programming language tools, such as Python's function decorators. The actual implementation of policy enforcement in real application depends on available software or hardware tools, and should be as much mandatory as possible.

Another typical problem that arise in Web-centric information system is generation of links available to the user. If user has no rights to perform some operation, e.g. object deletion, then it should be reflected by user interface, for example, by removing or disabling corresponding links. Clearly, generation of a set of active links can be done on top of has_permission function. From the other hand, if we know that several permissions have to be checked in a row a more efficient solution could exists. So, from the programming point of view, the framework should support two functions:

- *check for specified permission*,
- and *list of all operations* that user can perform on given object.

Input parameters for check operation are *subject* (user), *object*, and requested *operation* label (required permission). Second function returns list of valid operations for given *user* and *object*.

We can also specify following additional requirements:

- rules should be defined in a human-friendly formal language (XML-based languages, such as XACML, is not a suitable option);
- both permissive and restrictive rule should be supported;
- if an access is denied, then it should be possible to identify the rule leading to that decision;
- modification of security policy should be controlled by the same access control system.
- administrative domains should be separated in the sense that policy writer from one organization can not specify a policy affecting another organization; all conflicts, caused by existence of coauthored objects or coworkers, should be identified at runtime, or discovered statically.

III. Rule Definition Language

As access rule may refer to object's attributes, rule definition language consists of three parts: definition of objects, roles, and rules itself. First part defines set of entities, attributes and relations that can be used in other parts. Parameterized roles are listed in roles definition. Access control rules defines under what constraint a user has permission to evaluate action on specified object.

Our proof of concept implementation is performed in Common Lisp (due to Lisp's ability of rapid development of domain specific languages), so we use lisp-friendly syntax for all three parts.

A. Definition of Objects and Attributes

Objects definition is a kind of Object-Relational Mapping that reflects database structure. Each statement in this section defines an "object" – an access control entity. In terms of target application, objects correspond to a *classes* or types. In order to simplify descriptions in our proof of concept implementation we associate an object with database table only.

```
(defobject department
  (table department))
```

All other relevant information, including attributes, identifiers and related objects, is extracted from the database dictionary. If a table associated with object A contains a foreign key to a table of object B, then an attribute named "As" (plural form of A's name) is added to B. It is assumed that every database table is equipped with surrogate primary key, so foreign key references are simply integers.

In addition to database table name, a list of possible operations on that object can also be specified:

```
(defobject article
  (table article)
  (operations (read update delete)))
```

This definition means that only that three operations are relevant for *article* objects.

B. Roles

In many applications, user's membership in a role depends on administrative structure of organization, or other parameters. Technically, one can specify the set of roles as a Cartesian product of "real" roles by departments and manage hierarchical department's structure by means of roles inclusion. Clearly, such system is extremely difficult to support and security policy has to be modified every time department's structure changes.

In this work we adopt notion of parameterized roles. User-to-role relation is computed dynamically because the decision on whether or not a user belongs to a role depends on current state of the database. Roles are defined by the following statement.

```
(defrole dean
  (parameters department))
;; membership is limited to department
```

C. Rules

Rules are used to define conditions what should be satisfied in order to perform operation on object. In terms of [2] this models follows role-centric approach: attributes are used to constraint permissions assigned to roles.

Each rule is characterized by:

- access type (allow/deny),
- associated object,
- list of operations,
- grantees (roles or users), and
- access conditions.

Formal grammar for rule definitions is presented in Table I. For the lack of simplicity some "syntactic sugar" constructions are omitted. Let us consider the following example:

```
(rule
  (comment "Delete articles.")
  (access allow) ;; allow/deny
  (object type "article" alias "a")
  (operations ("delete"))
  (grantees
     (roles "admin"
            ("dean" a.authors.department))
     (users "root"))
  (conditions
    (and
      (= a.journal.publisher.name
         "Springer")
      (< 2010 a.year))))
```

Here we define that user "root" or any member of "dean" or "admin" roles are allowed to delete articles that were published in Springer's journals after year 2010. The role "dean" is a parameterized role, so not every dean can delete articles: at least one author of the article should be from the dean's department. When this rule is evaluated during permission checking procedure, object alias "a" is assigned to the object from user's request. Similarly, a predefined name "user" is assigned to the structure representing the requester.

Two points should be noted. First, a rule may be associated with a parameterized role, and values of role's parameters are related to target object's attributes. Second, conditions part of the definition consists of a list of logical expressions that define conditions required for treating this rule as *applicable* to given object. Each atomic expression has form (relational-operator operand operand).

Operands are XPath-like expressions defining access path to desired property of target object. The above example may be rewritten as a oneliner:

```
(= a[year > 2010].journal.publisher.name
   "Springer")
```

Here, filtering condition "[year > 2010]" applied to the object. Filtering condition may contain standard logical connectors and may refer to deeper attributes by means of dot notation.

Possibility to overwrite a rule in subordinate administrative unit is controlled like in [10]: a rule has special boolean property "strict" that states whether or not this particular rule can be redefined. The value of this property itself can be redefined (lowered from True to False) by a subordinate writer. It should be noted that set of objects affected by top-level and low-level rules may differ due to differences in specified

RULE	::=	'(rule' {'(' clause ')'} ')'
clause	::=	'comment' string
		| 'object' 'type' string ['alias' string]
		| 'access' ('allow' | 'deny')
		| 'operations (' operation {operation} ')'
		| 'grantees (' grantee {grantee} ')'
		| 'strict' ['True' | 'False']
		| 'condition (' expr ')'
operation	::=	string
grantee	::=	'(users' user {user} ')'
		| '(roles' [role | '(' role {selector} ')'] ')'
user	::=	string
selector	::=	('object' | alias) [filter] { '.' attribute [filter] }
alias	::=	string
filter	::=	constant relation (attribute [filter] { '.' attribute [filter] })
		| '(' filter ')'
		| filter ['and' | 'or'] filter
expr	::=	'(' relation constant selector ')'
		| '(' ['and' | 'or'] expr {expr} ')'
relation	::=	'>' | '<' | = | '<=' | '>=' | 'like' | 'in'

TABLE I. RULE DEFINITION GRAMMAR.

constraints. For example, top level constraint may by stated as `year > 2010`, while lower level rule may be redefined using `year > 2012`.

More than one rule may be defined for the same object type. In this case access is granted if there exists at least one rule that allows the access, and there are no rules that denies it.

IV. ALGORITHMIC PROBLEMS

In order to make the system useful in real situations, a number of algorithmic issues have to be resolved. First of all, due to highly dynamic nature of access control rules and high rate of such operations, efficient access checking methods should be designed.

A. Checking for Access

In this section we describe a straight forward algorithm for checking access of a user to specified object and discuss possible improvements. The general idea is to find all *applicable rules*, i.e. rules that are assigned to specified object and user, and verify whether an access should be granted, or denied.

Rules are associated to users or roles. In order to compute a set of candidates of applicable rules, i.e. a set containing all correct items and some false positives, membership of the requesting user in roles has to be verified. Membership in a parameterized role depends on current state of the database, so user-to-role relation should be computed dynamically. Given an expression like

```
(grantees
  (roles ("dean" a.authors.department))),
```

where "a" stands for a reference to accessed object (effectively, this is the value of a primary key in corresponding database table), one should verify that current user is a member of any roles "dean x", where x runs over the set of all departments accessed by path expression "a.authors.department". The set of applicable rules candidates consists of all rules associated with such roles.

Once the set of candidates is computed, additional constraints imposed by rule's conditions section will be verified

for every candidate rule (this step is a subject to optimization). Each condition consists of number of expressions of the form (relational-operator value selector), e.g.

```
(<= a[year = 2011 and
      journal.name like "Int.*J.*"].issue
    2)
```

Such expression can be translated into SQL statement naturally. Dots correspond to table joins, and filtering constraints yield conditions in WHERE clause of the SQL query. Consequently, whole constraints section may be represented as an SQL query.

Let us recall that access is granted if and only if there exists at least one positive rule and there are no negative rules for given user, object, and operation. Final decision can be trivially made on basis of applicable rule set. If there exist an applicable rule that denies access, then the finial decision is negative. Otherwise, the decision is positive if an only if the set of applicable rules is not empty.

Our proof of concept implementation supports automatic generation of SQL queries for checking rule's conditions and user's membership in parameterized roles. Checking that a rule matches request parameters is performed in statically defined order. First, matching between request's and rule's object and operation is validated. If this string equality check succeeded, role membership and additional access conditions are validated. This order based on assumption that last check is more complex and time consuming, because corresponding SQL query generated from a potentially complex expression. As long as the number of rules associated with each object and operation pair is small, this simple strategy produces reasonable performance: checking one or two extra rules yields only a couple of extra SQL queries. In general, one can expect that the number of rules will increase over time, so more sophisticated algorithm has to be developed. In the rest of this section we discuss possible improved to the described simple strategy.

One possible approach to performance optimization based on the observation that, if the access was granted, then all candidates for applicable *negative* rules had to be checked. It is worth noting that long lists of negative rules could be inefficient in principle. Positive candidate rules have to be checked only until first match occurrence. Choosing the "right" order of search space exploration can speed up overall performance. Preference may be given to rules that

- do not require to much time to verify, and
- yield positive result with high frequency.

Both values may be estimated using statistics of previous calls to the same rule. Frequency of positive results reflects the quantity of objects satisfying rule's conditions.

Other directions of optimization include:

- choosing the right order of conditions processing, based on the similar statistics on hit rate and efficiency of SQL queries, and
- finding common sub-queries in rules and organize rules hierarchically.

The latter approach is related to view selection techniques in relational database that allow to identify frequently used "patterns" in a set of queries. If conditions, or parts of conditions, of two or more rules are similar, they are probably may by approximated from above (with false positives) by a single query. If this generalized query produces negative result, then none of dependent rules can not be evaluated to positive result. Otherwise, some compensation queries are to be evaluated.

B. Policy Validation

Static validation of security policy is one of the goals of our work. In distributed environment searching for possible conflicts between access rules created by independent administrators is extremely important. Conflicts of two types are possible.

- Explicit conflict: there exists a triplet (user, object, operation) and two applicable rules such that one rule is positive, and the other one is negative.
- Implicit conflict: there exists an object, user, and applicable positive rule defined by a subordinate policy writer, such that there exists no applicable positive rule defined by a higher level administrator.

Both types of conflicts can be detected at runtime. Dynamic detection is, certainly, a useful feature in practice, but static analysis is required for elimination of possible errors in advance.

Methods for static analysis of access control systems were developed for classical RBAC model, e.g. [11], [12]. In [11] authors consider similar problem of collaborative editing of security policy under Administrative RBAC model. The decision problem in this case consists of establishing if a user can be assigned to certain role by a sequence of administrative actions. The problem is reduced to the reachability problem in some infinite transition system.

Decision problems for attribute-based models should be harder because wider range of objects should be considered. Searching for explicit conflicts in some cases may be reduces to a well studied NP-complete problem of checking containment of conjunctive query in a union of conjunctive queries [14]. Indeed, suppose we are checking one particular rule for consistence with a policy. A conflict appears if a conjunctive query corresponding to the rule under consideration is not contained in union of all other queries generated by rules defined for the same object.

In a more general case, policy checking problem may be reduced to a problem similar to [15]. A conflict appears if some counter-example object can be constructed under the specified policy.

V. CONCLUSION

In this paper a very basic role-centric attribute-based access control framework, motivated by the development of a real distributed application, is presented. It is targeted to support RESTful web application where each request is treated independently from all previous requests from the same user.

Attribute-based rules restrict permissions assigned to parameterized roles. Additional constraints, specifying whether or not a rule is applicable to the request, are stated in an XPath-like language that automatically mapped into SQL statements during access checking stage. As access control rules are defined using special language, automated policy validation is possible.

We briefly describe possible research directions in the areas of efficient implementation of decision procedure and static policy validation. Despite principal inefficiency of some operation, such as checking an access in presence of large number of negative rules, we believe that the described framework, which is in the very early stage of development at the moment, may be used in wider class of web-based application.

REFERENCES

[1] D. F. Ferraiolo, J. F. Barkley, and D. R. Kuhn, "A role-based access control model and reference implementation within a corporate intranet," *ACM Trans. Inf. Syst. Secur.*, vol. 2, no. 1, pp. 34–64, Feb. 1999. [Online]. Available: http://doi.acm.org/10.1145/300830.300834

[2] D. R. Kuhn, E. J. Coyne, and T. R. Weil, "Adding attributes to role-based access control," *Computer*, no. 6, pp. 79–81, 2010.

[3] X. Jin, R. Sandhu, and R. Krishnan, "RABAC: Role-centric attribute-based access control," in *Proceedings of the 6th International Conference on Mathematical Methods, Models and Architectures for Computer Network Security: Computer Network Security*, ser. MMM-ACNS'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 84–96. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33704-8_8

[4] X. Jin, R. Krishnan, and R. Sandhu, "A unified attribute-based access control model covering DAC, MAC and RBAC," in *Data and Applications Security and Privacy XXVI*, ser. Lecture Notes in Computer Science, N. Cuppens-Boulahia, F. Cuppens, and J. Garcia-Alfaro, Eds. Springer Berlin Heidelberg, 2012, vol. 7371, pp. 41–55. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31540-4_4

[5] J. Fischer, D. Marino, R. Majumdar, and T. Millstein, "Fine-grained access control with object-sensitive roles," in *ECOOP 2009 Object-Oriented Programming*, ser. Lecture Notes in Computer Science, S. Drossopoulou, Ed. Springer Berlin Heidelberg, 2009, vol. 5653, pp. 173–194. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03013-0_9

[6] B. Carminati, E. Ferrari, and A. Perego, "Rule-based access control for social networks," in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, ser. Lecture Notes in Computer Science, R. Meersman, Z. Tari, and P. Herrero, Eds. Springer Berlin Heidelberg, 2006, vol. 4278, pp. 1734–1744. [Online]. Available: http://dx.doi.org/10.1007/11915072_80

[7] B. Lang, I. Foster, F. Siebenlist, R. Ananthakrishnan, and T. Freeman, "A flexible attribute based access control method for grid computing," *Journal of Grid Computing*, vol. 7, no. 2, pp. 169–180, 2009. [Online]. Available: http://dx.doi.org/10.1007/s10723-008-9112-1

[8] E. Yuan and J. Tong, "Attributed based access control (ABAC) for Web services," in *Web Services, 2005. ICWS 2005. Proceedings. 2005 IEEE International Conference on*, July 2005, pp. –569.

[9] J. Crampton and J. Sellwood, "Path conditions and principal matching: A new approach to access control," in *Proceedings of the 19th ACM Symposium on Access Control Models and Technologies*, ser. SACMAT '14. New York, NY, USA: ACM, 2014, pp. 187–198. [Online]. Available: http://doi.acm.org/10.1145/2613087.2613094

[10] G. Motta and S. Furuie, "A contextual role-based access control authorization model for electronic patient record," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 7, no. 3, pp. 202–207, Sept 2003.

[11] S. Ranise, A. Truong, and A. Armando, "Boosting model checking to analyse large ARBAC policies," in *Security and Trust Management*, ser. Lecture Notes in Computer Science, A. Jsang, P. Samarati, and M. Petrocchi, Eds. Springer Berlin Heidelberg, 2013, vol. 7783, pp. 273–288. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-38004-4_18

[12] M. Pistoia, S. Fink, R. Flynn, and E. Yahav, "When role models have flaws: Static validation of enterprise security policies," in *Software Engineering, 2007. ICSE 2007. 29th International Conference on*, May 2007, pp. 478–488.

[13] S. Afonin, A. Kozitsyn, and I. Astapov, "SQLReports: Yet another relational database reporting system," in *Proceedings of the 9th International Conference on Software Engineering and Applications*, 2014, pp. 529–534.

[14] Y. E. Ioannidis and R. Ramakrishnan, "Containment of conjunctive queries: Beyond relations as sets," *ACM Trans. Database Syst.*, vol. 20, no. 3, pp. 288–324, Sep. 1995. [Online]. Available: http://doi.acm.org/10.1145/211414.211419

[15] N. Moore, "Computational complexity of the problem of tree generation under fine-grained access control policies," *Information and Computation*, vol. 209, no. 3, pp. 548 – 567, 2011, special Issue: 3rd International Conference on Language and Automata Theory and Applications (LATA 2009). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0890540110002051

Improving the Android Smartphone Security against Various Malware Threats

Darko Hrestak and Stjepan Picek
Faculty of Electrical Engineering and Computing,
University of Zagreb
Unska 3, Zagreb, Croatia
Email: darko.hrestak@fer.hr, stjepan@computer.org

Željko Rumenjak
ShoutEm Inc.,
Kalinovica 3, Zagreb, Croatia
Email: zeljko.rumenjak@gmail.com

***Abstract*—Android is one of the most popular operating systems for mobile devices in the world today. One of its greatest advantages, being an open source operating system, represents also one of its major drawbacks. There exists a number of malicious programs that can harm devices running any operating system, including Android. In accordance to that, having a system that is secure presents a goal of a paramount importance. However, the fact that the improvements in the security often come with a penalty in the usability can present a problem. Furthermore, it is difficult to give a good answer to a question when a system is secure enough since that heavily depends on the user's needs.**

In this paper we investigate several ways on how to improve the security of Android devices through various customizations of the operating system or third-party applications.

I. Introduction

When discussing the role of mobile phones in the world today, one could reach the conclusion that they became almost a necessity. Of course, today a mobile phone represents almost a synonym for a smartphone. Although a primary purpose of a phone should be the ability to send and receive voice calls and short messages, it seems that is not enough anymore. Nowadays, characteristics like high resolution screen, high speed multi-core processors or a high quality camera also seem to play a very important role in the choice of a mobile phone. Furthermore, with an increase of Internet speed and availability coupled with a drop in the prices, it is common to be connected to the Internet the entire time. With a number of applications that enable VoIP (Voice over IP), the primary purpose of a mobile phone (calling and receiving voice calls) became almost of a secondary importance. Naturally, this does not mean that people do not call anymore, but only that their phones' GSM modules are becoming less and less important as the primary communication channel.

However, one important question is whether the people think about their phone security. To use a phone, one doesn't need to know its inner workings, but to actually understand various threats one should possess a certain level of knowledge that is less common. The first thing that comes to mind when discussing the security of smartphones is the ability to have a secure, confidential communication (regardless if that is over a voice or a data channel). Such a "need" for secure (encrypted) communication proliferated after the Snowden's revelations [1]. In an effort to fill that need, we are evident of several applications that offer encrypted communication like Skype, Viber or WhatsUp.

However, is a secure communication all that people need to be secure and protected when using smartphones? The answer to this question is a definitive no. Some objectives of importance are also the ability to securely delete data from a phone and to securely store data. Furthermore, the knowledge whether an application has some security weaknesses or does it keep one's location private represents relevant objectives. Out of the aforementioned objectives, securely storing data on the phone is currently the only option readily available in the Android OS, which only requires that a user enables it.

Perhaps the easiest way to achieve all of the aforementioned goals is to use an operating system that is custom tailored to be secure. However, that is often impossible, since such operating systems are usually delivered only with dedicated devices and they are also expensive.

In this paper we concentrate on how to improve the security of smartphones against various malware. Furthermore, we examine only Android OS since it is an open source system that is extremely popular and widespread. When experimenting how to increase the security of a smartphone, we consider two scenarios. The first one deals with the customization of an OS where we do not consider fully custom operating systems, but rather small changes to the official Android OS. We note that such small changes cannot compare with the security of a custom made Android OS, but are much easier to implement. The second scenario concerns with various applications that one can use to improve the security of a mobile device. Since we are only concerned with the security improvements that can help against the malware, this should not be considered a complete study. Indeed, for example one can consider the location data. It is relatively easy to forbid the transmission of data that are sent by the GPS or network. However, recent research showed that even battery consumption can be used to determine phone's location [2]. Similarly, it has been shown that various sensors can be used to identify the device [3] or even to recognize speech [4].

The rest of this paper is organized as follows. First, in Section II, we give a short overview of various malicious

programs and Android OS architecture with a focus on security aspects. Section III deals with the Security-Enhanced Linux for Android as a more general approach to the security of a device/OS. Next, in Section IV, we concentrate on the security of Android applications and how to improve it. Finally, Section V offers a short conclusion.

II. Background

A. Malicious Software

Here, we introduce several classes that various malware can be classified in. We emphasize that this taxonomy can differ from those found in other literature. Here we strive to present various threats, while the exact division of those threats is of secondary importance.

Malware, or malicious software is any program which sole task is to cause destruction of data, steal personal information or gain access to system resources. Therefore, a malware is written with an intention to do harm [5]. Term malware encompasses all classes of malicious software like viruses, trojans, worms, spyware, etc. Note that many malware programs are actually capable of performing various tasks and therefore have elements of several malware classes.

Trojan represents a program that looks like a legitimate one, which after it is activated causes some damage to a user. Examples of Android Trojans are NotCompatible and Trojan-SMS.AndroidOS.Opfake.a.

Worm is a malicious program that makes copies of itself. An example of a worm is Selfmite.b, which spreads with a help of a text message that contains a link to the url address with the malicious code.

Spyware is a malware whose primary purpose is to access and exfiltrate personal or private data from a device [5]. An example of a spyware is ShrewdCKSpy.

Adware is a program in which advertising banners are displayed while different programs are running. An example of an adware is Durak application.

Besides those more common malware classes, there also exists **ransomware** which locks out a device from a user until he pays to unlock his device. An example of a ransomware is ScarePackage.

Chargeware is an application that charges users for content without a clear notification. One example of such an application is SMSCapers.

It is important to emphasize that most of the malicious applications come from small third-party application stores and not from the official Google Play store. For additional information about various malware and threats in general, we refer interested readers to [5], [6].

B. Android Architecture

Android uses a type of UNIX sandboxing method to run its applications [7]. Each application in Android is connected to a separate user and has a unique user ID (UID) and group ID (GID) associated with it. This means that Android security relies primarily on discretionary access control (DAC) provided by the kernel. Under discretionary access control, access to data is controlled at the discretion of the data owner or creator. Therefore, Android application security model is based on application sandboxing and isolation enforced by the Linux kernel through the user and group identities.

An application runs in the context of its own user and has a separate area of storage space reserved for it to store its private data. Besides the sandboxing provided by the Linux kernel, Android also uses a permission model. This model makes sure that only applications with necessary permissions can access system critical resources and resources of other applications. System resources like Camera, Wi-Fi, GPS, etc. are considered system critical resources and an application needs permission to use them. Similarly, if an application wants to access resources of another application, then it must have the necessary permissions required to access them.

The Android architecture defines mechanisms by which an application can communicate with another application outside of the sandbox. This secure cross-application communication is facilitated using inter-process communication (IPC) mechanisms. Primary mechanisms for IPC on Android are Intents and Remote Procedure Calls (RPCs). Intents are messages that components can send or receive. Intents are usually used to send a command to a component. RPC is implemented using binders that are used to directly invoke methods on a remote component.

The user defined permission architecture in the Android permission model enables a developer to define new permissions to certain components of the applications he creates. Furthermore, he can safely export those components to be used outside of the sandbox.

There are four different types of components for an Android application, through which IPC mechanisms are acquired.

- **Activities**. These are the visual elements of an application which users can see. They provide the graphical user interface (GUI) for the application.
- **Services**. Services provide no GUI for the application, but instead they represent the background processes. Usually long running tasks like downloading a file, etc. are implemented through services.
- **Broadcast Receivers**. Responses to system-wide broadcast announcements are accomplished with broadcast receivers. Examples of such responses are when the battery level is low or a picture is captured.
- **Content Providers**. Content provider is an encapsulation over the persistent data storage options available in Android (SQLite) database file system, etc., to facilitate data sharing among applications. Data stored via content providers can be exported and made available for other applications with a proper security channel restricted by the Android permission model.

Intents and Intent-filters are used to perform IPC through activities, services and broadcasts. Content providers use insert, update, query and delete operations to perform their operations. Content providers implemented in an application can be accessed by another application using content resolvers. If one application wants to communicate with another application, it

can accomplish that only through one of the above mentioned components.

One of the main issues with Android permissions, in their current form, is that a user has to approve every permission that a certain application requests before he can install that application. Currently, it is not possible to selectively approve or deny specific permissions to applications, at least not in the official Android builds. Applications also cannot request additional permissions at a later time, so application developers are left with no other choice than to request all permissions during the installation. This can make even legitimate applications look suspicious, because they sometimes require unexpected permissions. This also makes it even more difficult for users to detect which applications could be malicious before installing them. Because applications submitted to Google Play store are not reviewed, ultimately the only way to make sure that applications have no vulnerabilities is to test them.

As an example, a game may require access to user's contacts. That request can be very suspicious to the user that is trying to install that game, but the game developer could use contact information to find which friends of that user are also playing the game, and recommend that they play together. The user cannot know if the game will use this information for a legitimate feature or, for example, share it somewhere on the Internet. A much better approach would be if the game could request only the basic permissions that it needs to function during the installation, and all other permissions could be requested at the time when they are actually needed. In our example this could be when the user wants to invite his friends to play the same game.

For further information about the Android security architecture, we refer readers to [8].

III. Security-Enhanced Linux for Android

Security-Enhanced Linux for Android (SEAndroid) is a project initiated by the US National Security Agency (NSA) in order to identify and address critical gaps in the security of Android [9]. The initial plan of the project was to enable the use of Security-Enhanced Linux (SELinux) in Android, but several additional SELinux extensions specific to Android operating system were added over time. The motivation for the project was the increasing desire to use mobile devices throughout the US government. Android presented itself as a good candidate because of its open source nature that allowed easy customization, and its wide market adoption.

SELinux introduces mandatory access control (MAC) that is based on a system-wide security policy [9]. This security policy is enforced over all processes, even the processes running with root/superuser privileges [9]. The default action in SELinux is denial, which means that everything that is not explicitly allowed is automatically denied. Security rules in the policy are defined by domains or labels. Every process is labeled with a domain, and all processes labeled with the same domain are treated identically by the security policy.

SELinux was included in the Android 4.3 release in a permissive mode, which means that policy denials were only logged and not enforced [10]. In Android 4.4 (KitKat), SELinux was in enforcing mode on a set of crucial domains, and finally, starting with the Android 5.0 (Lollipop) release, SELinux is in full enforcing mode on all domains [10].

One of the most important benefits of SELinux on Android is strong isolation and sandboxing of applications and privileged daemons. SELinux can be used to protect the privileged daemons from misuse, and limit the damage that can be done through them. This means that malicious applications that users can install on their devices can do much less damage to other applications, or to the operating system itself. This, however, does not mean that applications cannot abuse the permissions granted to them by the user during the installation process.

A. AppOps

A step towards allowing users to selectively control application permissions was the AppOps application included in the Android 4.3 release. AppOps was a hidden application that allowed users to deny permissions to installed applications [11]. AppOps application was removed afterwards, but the code used to control the application permissions still exists in the Android open source project.

We display several screenshots of AppOps application in Figure 1. The first screenshot shows the permission that the Facebook application requires, and allows users to selectively deny access to those permissions. The second screenshot lists the applications that recently accessed user's personal information, and the third screenshot shows a list of the applications installed on the device.

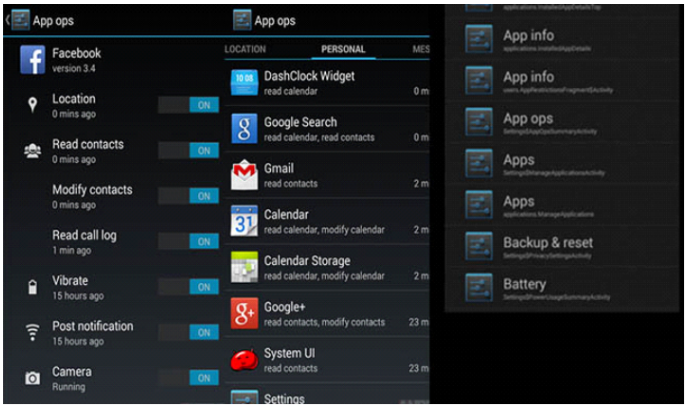


Fig. 1. Screenshot from Appops.

B. EnterpriseOps

An extension to the SELinux called EnterpriseOps (EOps) was developed as a part of the SEAndroid project in order to allow an enterprise like control to application permissions [11]. EOps provides a way to define a policy that can explicitly deny certain permissions to a group of installed applications. This extension relies on the code developed for the AppOps feature, and the primary problem is that the AppOps code is still incomplete [11].

When a permission is denied, an empty data set should be returned to the caller to ensure that applications continue to function correctly. The Android code that handles permissions doesn't currently support those empty data sets, so it completely blocks the operation. Because of this, some applications might crash or behave unexpectedly when some of the permissions they require are blocked. In some cases, the code that checks the policy isn't implemented correctly, and the operation is permitted anyway. Because of those issues, both AppOps and EOps features are not ready for general use. Next, we give an EOps policy example that denies the CAMERA permission [11]:

```
<?xml version="1.0"?>
<app-ops>
    <debug/>
    <seinfo name="default">
        <op name="CAMERA"/>
    </seinfo>
    <seinfo name="system">
        <op name="CAMERA"/>
    </seinfo>
</app-ops>
```

The *seinfo* element of the EOps policy files is used to target applications labeled with a specific name, and the *op* elements inside are used to remove permissions from those applications. The above policy removes the *CAMERA* permission from applications in *default* and *system* domains. Because of that, if a user would start the camera application that comes with the Android OS, the application would launch, but he would not be able to take photographs.

The code and the policy developed by the SEAndroid project has been merged to Android open source project repository as of Android 5.0 release [12]. While this includes the SELinux support, middleware MAC mechanisms, like the EOps extension used for managing application permissions among other things, are still not merged. Developers that want to use or test those features need to create a custom Android build that includes those features.

C. Intent Firewall

Intent firewall is a component of the Android operating system designed to control and enforce intent based inter-process communication in Android. Intent firewall was developed as a part of the Android open source project, and not as a part of the SEAndroid project. Only the tools that simplify the process of deploying intent firewall policies to target devices were developed as a part of the SEAndroid project.

Every intent created in the Android operating system goes through the intent firewall, which means that the intent firewall has the power to allow or deny any intent [13]. Intent firewall can be configured using an XML file with rules that specify how activity, service and broadcast intents will be handled. Rules in this XML file can be used to target one or more intents, and control if those intents will be blocked or allowed. An example of the intent firewall policy that blocks access to settings menu [13] is presented below:

```
<?xml version="1.0"?>
<rules>
  <activity block="true" log="false">
    <component-filter
    name="com.android.settings/.Settings"/>
  </activity>
</rules>
```

The above policy blocks all intents to the activity named *com.android.settings/.Settings*, i.e., the default settings menu activity of the Android operating system.

Intent firewall is implemented in Android, but it is not officially supported or documented [13]. This means that it requires going through the Android source code to determine how to use it properly, and it could also change significantly in the future versions of Android. Intent firewall configuration can currently be modified only by system applications and root users. Because of that, users who wish to enhance the security of their device using intent firewall must first obtain root access on their devices.

D. Customizing Policies

Default SELinux policy is deployed on the device with the operating system build itself, and is loaded during every boot. It is possible to create additional policies and deploy them to devices using tools and the SEAdmin application developed as a part of the SEAndroid project. Those tools will produce a bundle zip file with the policy, suitable for installation using the SEAdmin Android application. SEAdmin application currently looks for policy bundle files on the SD Card device. To deploy a new policy to a device, it is enough to copy the bundle file to the SD Card device, and run the SEAdmin application. After that, the application will detect and install a new policy. The policy bundle file must be signed with the key that was used to build the Android OS. SEAdmin application can also be used to deploy EOps, and intent firewall policies to devices. The procedure to deploy those policies is similar to the procedure to deploy SELinux policies [11]. The main difference is that different tools must be used to create the zip file that will be deployed to the target device.

SELinux for Android introduces a more flexible security model to Android operating system, and increases overall security of the system, provided that the default policies deployed with the OS are good. The main problem for more advanced users who would like to customize the security of their devices is that those policies cannot be easily customized. SELinux policies can be customized by creating a custom Android build and flashing it to a device, or possibly using an exploit to gain root access on the device, if it exists. On the other hand, SEAndroid together with EOps extension may prove to be a good way for organizations to enforce their security policies on a large number of devices using custom Android builds.

IV. Security of Android Applications

In this section, we first deal with the reverse engineering of Android applications. In the second part, we examine several applications that are able to conduct the security analysis of the Android OS or an application.

With Android OS, anyone can create an application and upload it to the Google Play store. Most often, it is necessary to know how to code in order to be able to develop an application. However, tools like App Inventor are available and they make it possible to develop Android applications without any coding knowledge [14]. Inexperienced developers can create applications with substandard code quality and with many vulnerabilities. Such vulnerabilities can become a threat to the user's private data. The most common vulnerabilities are:

- Insecure data storage, which can lead to cross-application data stealing.
- Exported Intents, which can lead to confused deputy problems.
- Insecure network transmission, which can cause user's private data to be sniffed by attackers.

To battle such problems, user can either reverse engineer an application to check what does it do or use some ready-made application to test/change the application of interest.

A. Reverse Engineering APK

Android application installation files (.apk) are quite easy to reverse engineer. Here, we present several tools which make it possible for any amateur to reverse engineer an .apk file in a relatively short time. APK files can be reverse engineered using the following applications:

- apktool [15].
- dex2jar [16].
- Java decompiler tool (e.g. Jd-gui [17], JAD [18]).

Apktool is a tool for reverse engineering third-party, closed, binary Android applications [15]. This tool can decode resources to the nearly original form and rebuild them after making some modifications; it makes it possible to debug Smali (assembler) code step by step. Also it makes working with applications easier because of project-like files structure and automation of some repetitive tasks like building APKs, etc. It transforms binary Dalvik bytecode (classes.dex) into Smali source. Apktool is a multi-platform tool, based on the Apache 2.0 license.

Dex2jar is a multi-platform tool that converts Dalvik bytecode (DEX) to Java bytecode (JAR) [16].

Apktool helps extracting binary resources and allows repacking after which dex2jar converts Dalvik bytecode to Java bytecode.

A better solution are pure Android decompilers, unfortunately there are not as many of them as there are Java decompilers. We mention several popular Android decompilers:

- Open Source: Androguard [19], APK Studio [20].
- Commercial: JEB [21].

Next, we present several publicly available tools one can use to analyze Android applications.

B. TAINTDROID

Enck et al. proposed an analysis system called TaintDroid, which is an extension to the Android mobile phone platform that tracks the flow of privacy sensitive data through the third party applications [22]. The primary goals of this application are to detect when the sensitive data leaves the system via untrusted applications. TaintDroid automatically labels (taints) data from privacy-sensitive sources and transitively applies labels as sensitive data propagates through program variables, files, and inter-process messages. When tainted data is transmitted over the network, or otherwise leaves the system, TaintDroid logs the data's labels, the application responsible for transmitting the data, and the data's destination.

C. SCANDROID

Fuchs et al. developed a tool SCanDroid for the static analysis of Android applications [23]. SCanDroid's analysis is modular in order to allow incremental checking of applications as they are installed on an Android device. It extracts security specifications from manifests that accompany such applications, and checks whether data flows through those applications are consistent with those specifications.

D. X-RAY

Duo Security developed an application, X-Ray, which scans for OS vulnerabilities on the Android device [24]. This application can be directly installed to an Android device. X-Ray scans an Android device to determine whether there are vulnerabilities on it. The X-Ray application presents the user with a list of vulnerabilities that is able to identify and allows him to check for the presence of each vulnerability on their device.

E. APPVIGIL

Appvigil is an advanced security loophole scanner for Android applications [25]. It is a cloud based security scanner that can scan security loopholes in applications. Created by Indian company IIT Bombay, it offers free trial for application scan. Through its audit report, Appvigil provides exact location of loopholes with patch recommendation.

F. ANDROID Application Vulnerability Scanner Tool

Android application vulnerability scanner tool (AAVST) helps to alter applications by disabling some of the permissions they ask and by fixing some of the vulnerabilities that such applications possess [26]. This tool offers significant capabilities whether one is an application developer or a normal user. For an application developer it offers the following functionalities:

- Scanning of Android applications for vulnerabilities.
- Interactively fixing the detected vulnerabilities.
- Generating keystores.
- Building applications.
- Signing application and making it ready to upload to the hosting servers.

As a normal user by using AAVST one has the following options at his disposal:

- Make an “APK" ready for editing (decompile the APK).
- Scan and fix vulnerabilities in the decompiled application.
- Disable undesirable permissions requested by the application.
- Make the application ready to be installed to devices (build and sign the application).
- Generate certificate (important step because only signed applications will run on Android device, when you decompile application it’s signature is removed).

Out of the aforementioned options, one especially interesting is disabling undesirable permissions requested by the application. Many users experience that applications require too many permissions and they would prefer that permissions were more restricted due to the privacy issues. AAVST tool can help the users to deny some of the permissions that an application asks for, and still run the application. Step by step routine to do this is:

1) Run the AAVST tool.
2) Choose the *Decompile application* option where one sets up the output directory for the application source code.
3) Select static analysis with decompiled directory and press the scan button. After that, select *Uses permission* radio button and a choice to disable permissions which one doesn’t want for an application appears. After that, click the update button and check *Re-Build apk* which will build the APK for the modified application.
4) Only signed APK’s will work in an Android device so once the application is rebuilt, one must sign it to make it ready to be deployed to an Android device. In this step one selects *Generate Key* and fills in required parameters to generate a keystore.
5) After that it is necessary to select *Sign application* where one chooses application and a keystore that he generated in the previous step. After the signing step is complete, the application is ready for the installation to an Android device.
6) The final step is installing the permission modified APK to a device.

V. Conclusion

In this paper we discuss the security of Android operating system where we identify several objectives one needs to consider. We disregard the secure communication objective and concentrate on several methods that can be used to improve the security of the Android OS and applications against malware. In order to do that, we present several improvements that are within a reach of an average user. The first set of improvements considers the customization of Android operating system and the second set considers several ready-made applications that can help to improve the security of a system. We note that some of the improvements currently achievable through the third-party applications should become available within Android OS itself in newer releases.

References

[1] S. Landau, “Making Sense from Snowden: What’s Significant in the NSA Surveillance Revelations,” *IEEE Security & Privacy*, vol. 11, no. 4, pp. 54–63, 2013.

[2] Y. Michalevsky, G. Nakibly, A. Schulman, and D. Boneh, “Powerspy: Location tracking using mobile device power analysis,” *arXiv preprint arXiv:1502.03182*, 2015.

[3] H. Bojinov, Y. Michalevsky, G. Nakibly, and D. Boneh, “Mobile device identification via sensor fingerprinting,” *arXiv preprint arXiv:1408.1416*, 2014.

[4] Y. Michalevsky, D. Boneh, and G. Nakibly, “Gyrophone: Recognizing speech from gyroscope signals,” in *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, 2014, pp. 1053–1067. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/michalevsky

[5] S. Gunasekera, *Android Apps Security*, 1st ed. Berkely, CA, USA: Apress, 2012.

[6] “2014 Mobile Threat Report,” https://www.lookout.com/resources/reports/mobile-threat-report, 2014, [Accessed Feb. 20, 2015].

[7] W. Enck, M. Ongtang, and P. McDaniel, “Understanding android security,” *IEEE Security and Privacy*, vol. 7, no. 1, pp. 50–57, Jan. 2009.

[8] N. Elenkov, *Android Security Internals: An In-Depth Guide to Android’s Security Architecture*, 1st ed. San Francisco, CA, USA: No Starch Press, 2014.

[9] S. Smalley and R. Craig, “Security enhanced (SE) android: Bringing flexible MAC to android,” in *20th Annual Network and Distributed System Security Symposium, NDSS 2013, San Diego, California, USA, February 24-27, 2013*, 2013. [Online]. Available: http://internetsociety.org/doc/security-enhanced-se-android-bringing-flexible-mac-android

[10] “Security-Enhanced Linux in Android,” https://source.android.com/devices/tech/security/selinux/index.html, 2015, [Accessed Feb. 20, 2015].

[11] R. Haines, *The SELinux Notebook*, 4th ed. http://freecomputerbooks.com/books/The_SELinux_Notebook-4th_Edition.pdf, 2014, [Accessed Feb. 15, 2015].

[12] “Security Enhancements (SE) for Android,” http://seandroid.bitbucket.org/, 2015, [Accessed Feb. 20, 2015].

[13] C. Yagemann, “IntentFirewal,” http://www.cis.syr.edu/~wedu/android/IntentFirewall/index.html, 2015, [Accessed Feb. 15, 2015].

[14] P. Cashmore, “MIT App Inventor,” http://appinventor.mit.edu/explore/, 2013, [Accessed Feb. 20, 2015].

[15] “A tool for reverse engineering Android apk files,” http://ibotpeaches.github.io/Apktool/, 2015, [Accessed Apr. 15, 2015].

[16] “dex2jar,” https://github.com/pxb1988/dex2jar, 2015, [Accessed Apr. 15, 2015].

[17] “Java Decompiler. Yet another fast Java decompiler,” http://jd.benow.ca/, 2015, [Accessed Apr. 15, 2015].

[18] “JAD Java Decompiler,” http://varaneckas.com/jad/, 2015, [Accessed Apr. 15, 2015].

[19] “Androguard,” https://github.com/androguard/androguard, 2015, [Accessed Apr. 15, 2015].

[20] “APK Studio | Android Reverse-Engineering,” https://apkstudio.codeplex.com/, 2015, [Accessed Apr. 15, 2015].

[21] “JEB - The Interactive Android Decompiler,” https://www.pnfsoftware.com/index.php, 2015, [Accessed Apr. 15, 2015].

[22] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, “Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones,” in *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI’10. Berkeley, CA, USA: USENIX Association, 2010, pp. 1–6.

[23] A. P. Fuchs, A. Chaudhuri, and J. S. Foster, “Scandroid: Automated security certification of android applications,” *Manuscript, Univ. of Maryland, http://www. cs. umd. edu/avik/projects/scandroidascaa*, vol. 2, no. 3, 2009.

[24] “X-Ray for Android,” https://labs.duosecurity.com/xray/, 2015, [Accessed Feb. 15, 2015].

[25] “Appvigil,” http://appvigil.co/, 2015, [Accessed Feb. 15, 2015].

[26] “Android Application Vulnerability Scanner Toolkit,” http://securityresearch.in/index.php/articles/android-application-vulnerability-scanner-toolkit/, 2013, [Accessed Feb. 20, 2015].

Detection of DECT identity spoofing through Radio Frequency fingerprinting

I. Sanchez *, R. Satta†, R. Giuliani‡, G. Baldini§
European Commission – Joint Research Centre (JRC), Institute for the Security and Protection of the Citizen,
Digital Citizen Security, Ispra, Via E. Fermi 2749, 21027, Italy
Emails:* ignacio.sanchez@jrc.ec.europa.eu†riccardo.satta@jrc.ec.europa.eu, ‡raimondo.giuliani@jrc.ec.europa.eu,
§ gianmarco.baldini@jrc.ec.europa.eu

***Abstract*—Digital Enhanced Cordless Telecommunications (DECT) is an European Telecommunications Standards Institute (ETSI) standard for short-range cordless communications with a large worldwide installed customer base, both in residential and enterprise environments. As in other wireless standards, the existence of active attacks against the security and privacy of the communications, involving identity spoofing, is well documented in the literature. Although the detection of spoofing attacks has been extensively investigated in the literature for other wireless protocols, such as Wi-Fi and GSM, very limited research has been conducted on their detection in DECT communications. In this paper, we describe an effective method for the detection of identity spoofing attacks on DECT communications, using a radio frequency fingerprinting technique. Our approach uses intrinsic features of the front end of DECT base stations as device fingerprints and uses them to distinguish between legitimate and spoofing devices. The results of measurement campaigns and the related analysis are presented and discussed.**

I. Introduction

In wireless communication networks, spoofing is a serious attack where a malicious attacker masquerades as a legitimate mobile device and transmits false information. Device identity spoofing is often used as part of larger and more complex active attacks that require sending forged information originated from the identity of a legitimate device of the network.

DECT, Digital Enhanced Cordless Telecommunications [1], is a radio frequency protocol developed by ETSI, widely used in residential and enterprise environments for the transmission of voice in cordless telephony. It is considered by ETSI as its second most successful protocol after GSM, with an estimated install base of almost 820 millions of cordless phones world-wide [2]. DECT is largely widespread in Europe, where it is commonly used by citizens in residential environments as wireless access protocol for fixed telephony. It is also commonly found in enterprise environments, usually integrated into Unified Communication systems.

Despite the security features foreseen in the standard, several attacks against the security and privacy of DECT communications have been described in the literature. In addition to passive attacks against both non-encrypted [3], [4] and encrypted communications [5], [6], a set of active attacks, involving the injection of spoofed DECT packets into the network, has also been described in the literature. In [3], the authors demonstrate how the lack of mutual authentication in some implementations of the standard, can be exploited to perform a Man-In-The-Middle attack able to intercept the voice communications of the victim. In order to do so, the attacker impersonates the identity of the legitimate DECT base station (Fixed Part - FP), spoofing its identity in such a way that the Portable Part (PP) of the victim will eventually connect to it, instead of the legitimate one. In essence, this attack is similar to the one described in the literature [7], [8] for GSM networks, where the identity of the Base Station (BTS) is spoofed by the attacker.

The application of DECT identity spoofing has also been proven to be effective attacking encrypted communications, as part of a more complex active attack. In [9] the authors demonstrated an active attack that sends spoofed DECT packets into the network in order to progressively decrypt an encrypted communication.

The identification of device identity spoofing on DECT networks is an effective way to generically detect a big percentage of active attacks that rely on the injection of spoofed packets into the network. To this end, several approaches can be followed. One approach would be the detection of the spoofed packets based on anomalies in the data and metadata contained in the spoofed packets. This technique could be effective in certain cases where differences between the legitimate and the spoofed packets can be appreciated by looking at their content, as it has been demonstrated in GSM networks for the detection of rogue base stations [10] and in Wi-Fi for the detection of MAC address spoofing on the basis of the analysis of the sequence numbers [11].

In many cases, these detection methods can be circumvented by an attacker, by carefully imitating or jamming the data transmitted by the legitimate station, given the fact that the features used in the detection can be easily modified by the attacker to match the legitimate ones.

In order to tackle this point, the radio frequency (RF) fingerprinting of mobile devices has been proposed as a method for the detection of active attacks in other protocols such as GSM [12] and Wi-Fi [13].

The basic idea behind the RF fingerprinting technique, is that each electronic device has unique features, also called Radio-Frequency Distinctive Native Attributes or (RF-DNA), which can be extracted from its spectral emissions and can be used to uniquely identify the specific model or mobile device. The RF pattern cannot be manipulated at the software level by the attacker, and is therefore hard to forge, even using customisable transceivers, such as those based on Software-defined radio.

In other words, the term RF-DNA is used to embody the coloration of RF emissions (both intentional and unintentional) induced by the intrinsic physical attributes of a unique device and related to components like filters, amplifiers, Integrated Circuits (IC) and so on. While other authors have focused their research efforts in the unintentional emission as in [14], in this paper we focused on the intentional emissions of the DECT base station during transmission and/or signal broadcast.

The use of fingerprinting to detect spoofing has been also proposed in other wireless communication technologies other than DECT. In [15], the authors have applied RF-DNA fingerprints to IEEE 802.16 WiMax-based airport communications by using both Time Domain (TD) and Wavelet Domain (WD) techniques.

Another popular application for fingerprinting is related to Wi-Fi technology. In [16], the authors used unique features extracted from RF-components or from the timing behavior of the MAC-chips to attest the uniqueness of a dedicated wireless station. The authors argue that Wi-Fi Protected Access (WPA) or IEEE 802.11i (WPA2) are not a complete protection against MAC address spoofing because WPA and WPA2 can provide data-frame authentication to prevent clients from being spoofed but unfortunately they do not provide authentication for management frames. The authors applied different approaches for fingerprinting. In one case, Discrete Wavelet Transformation (DWT) was applied to signal phase and signal amplitude to generate features used for the classification process. In another approach, the analysis of the distribution of delay values between 802.11 packets and the corresponding acknowledgment frames was used for fingerprinting as delay values differ in various chipsets.

Still on Wi-Fi technology, Sheng and others [17] applied the analysis of received signal strength (RSS) for fingerprinting on the basis of the RSS patterns (e.g., their Gaussian profile). This approach has the merit to avoid the need for sampling the RF signals, as in [14] and [16], even if this need has become less costly with the latest generations of sampling devices (e.g., the RTL-SDR).

In [18], the authors applied RF-DNA fingerprints to mitigate the risk of impersonation of authorized network devices. The fingerprinting was implemented using Multiple Discriminant Analysis (MDA) for training and authorised device classification and Multivariate Gaussian (MVG) likelihood values to identify rogue devices attempting to gain unauthorized network access by presenting false bit-level credentials. Finally, in a recent paper [19] the authors have applied Support Vector Machine (SVM) to identify GSM phones based on their uplink RF emissions. The analysis in [19] is similar to what done in this paper because the fingerprinting method was based on instant amplitude as in this paper and GSM have some similarities to DECT (e.g., modulation scheme).

As we described above, fingerprinting analysis has been applied to various widespread wireless communication technologies like Wi-Fi, WiMax, GSM and Zigbee but we are not aware of similar analysis for DECT. The purpose of this paper is to cover this gap and provide some preliminary results to mitigate the spoofing of DECT by using a simple, yet effective, RF fingerprinting technique.

The structure of the paper is the following. In section II, we describe the DECT standard. In section III we provide a description of the measurement setup and the definition of the algorithms for classification and identification, which have applied to provide the results in section IV. Finally, V describes the conclusions and potential future developments of this paper.

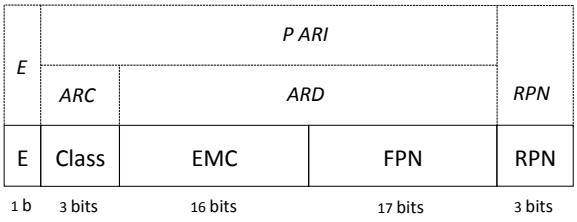


Fig. 1. Radio Fixed Part Identity (RFPI)

II. Digital Enhanced Cordless Telecommunication standard

In Europe, the DECT protocol operates in the range of 1800 Mhz to 1900 Mhz, where a total of 10 physical channels are allocated, using Frequency Division Multiple Access (FDMA), leading to the creation of a total 120 logical channels by dividing each physical channel into 12 logical channels using Time Division Multiple Access (TDMA), with separate time slots for the reception and transmission of data following Time Division Duplexing (TDD). Each DECT device will only transmit in its assigned time slot, producing bursts that will be clearly identifiable in the DECT channel assigned for the transmission.

A typical DECT installation will be composed of one or more cordless phones, denoted as Portable Parts (PP) in the standard, registered to a DECT base, named Fixed Part (FP). In a residential installation, the FP is typically connected to the fixed telephone line and also acts as a battery charger for the portable parts. In enterprise installations, the portable parts are registered to the FP, which is, in many cases, integrated into a Private Automatic Branch eXchange (PABX) system. In recent years, a new generation of residential Unified Communication systems have started to integrate DECT devices for the Smart Home ecosystem. The Generic Access Profile (GAP) of the DECT standard ensures the interoperability of fixed and portable parts from several manufacturers.

Each FP is identified by a RFPI (Radio Fixed Part Identity), which is a 40-bit number mainly composed of a 16-bit Equipment Manufacturer's Code (EMC), a 17-bit Fixed Part Number (FPN) and a 3-bit Radio fixed Part Number (RPN). The first 4 bits of the RFPI, reserved for the E bit and the class type, are set to zeros in residential and small corporate DECT installations (Class A). Figure 1 depicts the composition of the RFPI and the formal relation of each part with the Access Right Class (ARC), Access Right Details (ARD) and Primary Access Rights Identity (PARI) fields.

The EMC code determines the specific manufacturer of the FP. The FPN and FPN codes are assigned by the manufacturer and will uniquely identify a specific FP device. Unless an ECO mode is enforced, a FP will continuously advertise its

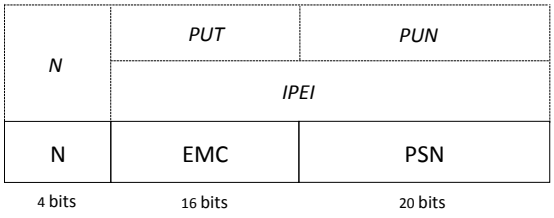


Fig. 2. International Portable User Identity (IPUI)

presence by broadcasting its RFPI, sending 100 DECT frames per second. Due to this fact, a physical channel with a single FP and no active PP communications, will show short bursts in regular intervals of 10 milliseconds. Each burst is a DECT frame that is composed of a preamble, the A field data where the RFPI is located and an optional B field in case there is voice being transmitted in the frame.

On the other end, the PP is uniquely identified by a 40 bit number called IPEI (International Portable Equipment Identity) that is composed of a 4-bit identity type field (type N for residential and small corporate set-ups), a 16-bit EMC and a 20-bit Portable equipment Serial Number (PSN) code. Both EMC and PSN are known as the Portable Part Equipment Identity (IPEI) and uniquely identify the specific PP device, with the EMC uniquely identifying the manufacturer and the PSN the specific device. Figure 2 depicts the composition of the IPEI.

When a PP is registered into a given FP, the RFPI of the FP is known to the PP and it is stored in its permanent memory. Similarly, the IPUI of the PP is registered in the permanent memory of the FP. A PP will monitor the available DECT channels and identify the corresponding FP on the basis of the RFPI code transmitted in the DECT frames.

The DECT standard foresees the usage of authentication, by using the DECT Standard Authentication Algorithm (DSAA), and encryption, by means of the DECT Standard Cipher (DSC). The Generic Access Profile (GAP) part of the standard ensures the interoperability between different DECT models and manufacturers and establishes a pairing process in order to register a DECT handset (PP) into a base station (FP). As part of this process, a long term key, named User Authentication Key (UAK), will be exchanged and permanently stored by both PP and FP.

The DSAA algorithm will use the UAK key for authentication and negotiation of temporal session keys, named Derived Cipher Keys (DSC), which will be used by the DSC algorithm for the encryption of the communications between FP and PP.

It shall be noted that DSC encryption, which protects the confidentiality and up to a certain point the integrity of the data, is only applied against the B-field (voice) and C-channel data (control information). The RFPI and IPUI identities included in the A-field of the FP and PP frames respectively, are not protected.

In that regard, for the purposes of this paper it is sufficient to say that there are several weaknesses on the security of the DECT GAP communications and, as described in section I, various active attacks against DECT communications that involve spoofing of the DECT devices identities have been described in the literature.

III. Fingerprinting analysis

A. Measurement setup

A schema of the measurement setup is provided in Figure 3. The correspondent image of the measurement setup used to collect the radio frequency emissions is provided in Figure 4.

The test-bed is in a shielded semi-anechoic chamber (i.e., absorbers are only partially covering the walls of the chamber). The DECT handset and the base station are separated at a distance of four meters and at a height of one meter from the ground. In the experiments performed, a total of 4 DECT models were analyzed from 2 well known manufacturers of DECT cordless phones.

The receiver used to record the radio frequency samples is an Universal Software Radio Peripheral (USRP) N210 with a configured sampling rate of ten mega samples I/Q per second. The receiver is located in the middle between the DECT handset and the base station (i.e., two meters from each DECT component) and it is also at an height of one meter from the ground. The antenna is a vertical omni-directional full wavelength monopole. The USRP N210 is connected to a desktop computer for the recording and the processing of the samples. To avoid possible biases, the USRP and its antenna inside the shielded semi-anechoic chamber, were the same for all the signal acquisitions.

The experiments were focused on the scenario where the identity of the DECT base station (FP) is spoofed by an attacker, so the recordings of the samples took place with and without PP transmissions. As described in section II, the FP transmits 100 frames per second, even in the absence of a call or PP activity. This was the case for all the DECT models analyzed.

The signals were acquired in 2 rounds for the 4 DECT devices and stored in I/Q format. The signals acquired the first day for all the devices became the training dataset and the set recorded the second day (also for all the devices) was used for testing, as it will be described in the next section.

B. Algorithms for fingerprint verification

Formally, the detection of a spoofing attack can be modeled in the same way of biometric fingerprint *verification* [20], i.e., an one–VS–one comparison of two fingerprints. The first one is the *template* fingerprint of the target DECT base station B, the second one is the fingerprint of the base station who "claims" to be B.

The template fingerprint of a base station is constructed, as follows, from a recorded radio transmission originated from that base station. First, we process in MATLAB the samples recorded by the USRP N210 in I/Q format, in order to identify and process the bursts.

In order to do so, we first calculate the modulus of the I/Q signal. Then we apply a low pass filter to ensure that only the in-band DECT signal is used to fingerprint the DECT device. In our case, we adopted a 2nd order Butterworth filter with a

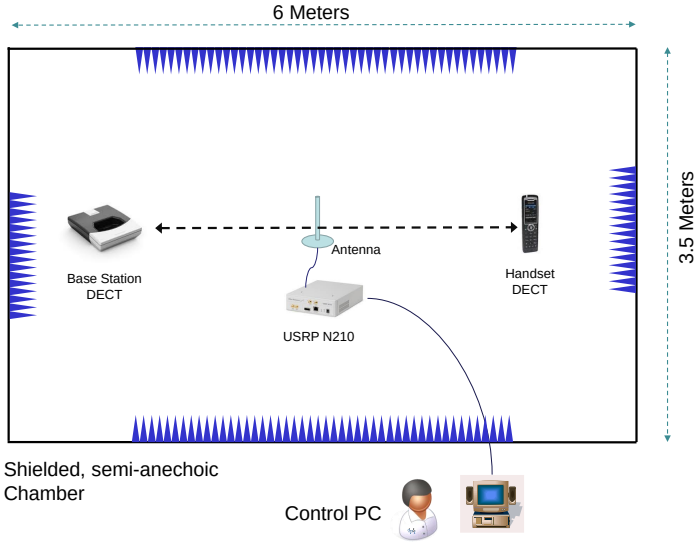


Fig. 3. Schema of the test bed used to collect the radio frequency emissions

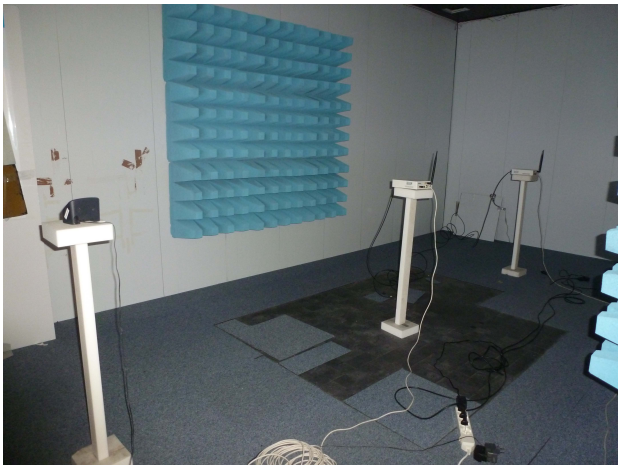


Fig. 4. Image of the test bed used to collect the radio frequency emissions

normalized cutoff frequency of $10e-5\ \pi$ rad per sample in normalized frequency. The Butterworth filter is used because it provides a flat response in the DECT frequency bands. A flat response is more appropriate because we did not know a priori which frequency components of the DECT signal could be more valuable for fingerprinting.

Once the Butterworth filter is applied, the bursts are identified by detecting the ramp-up and ramp-down of the signal. Each burst is normalized to the maximum amplitude of that specific burst and extracted using a fixed thresholding to synchronize it with the others.

We are interested only in the part of the burst which is not related to the content being transmitted (e.g., the voice during the DECT call), in order not to fingerprint content differences rather than the DECT device itself. The content invariant parts in the bursts are the ramp-up and ramp-down sections of a burst plus the preamble. Apart from being content invariant, these sections are appropriate for fingerprinting because they are more related to the physical properties of the front-end of the wireless devices [15].

The ramp-up and ramp-down sections plus the preamble of each burst are then stored in arrays, which are the primary source of data for the fingerprinting algorithms described in the following paragraphs. For simplicity, the combination of ramp-up and ramp-down and preamble are called mini-bursts in the rest of the paper to differentiate them from the initial raw bursts.

The mini-burst of both datasets are normalized into the interval $[0, 1]$, to make their signals comparable. In an initial training phase, a template T_B of the base station B is created by calculating the average of all the mini-bursts stored in the training dataset for that specific DECT device.

Verification of a given mini-bursts b^* of the validation dataset is then performed by computing a *burst similarity score* $S(T_B, b*)$, defined as the Normalized Cross-Correlation between T_B and b^* (both real vectors of the same length):

$$S(T_B, b^*) = \frac{(T_B - \bar{T_B}) \cdot (b^* - \bar{b^*})}{\|T_B - \bar{T_B}\| \cdot \|b^* - \bar{b^*}\|} \quad (1)$$

where $\bar{T_B}$ and $\bar{b^*}$ are the means of T_B and b^*, respectively. The value of $S(T_B, b*)$ measures the likelihood that b^* is a mini-bursts transmitted by the base station B. NCC was chosen as likelihood measure as it evaluates the similarity of the *behaviour* of the two signals (i.e., to which extent the vary in the same way over time) rather than the similarity of their values as it would be done e.g. by the Euclidean distance.

In order to obtain a sharp Yes/No decision, a threshold Th is applied so that b^* is deemed as coming from the base station B only if $S(T_B, b*) \geq Th$. The value of Th impacts on the False Positives (i.e., false alarms triggered by the spoofing detection systems) and of False Negatives (i.e., spoofing attacks not detected). In particular, the higher the threshold, the lesser False Positives, and the more False Negatives. It is worth to point out that False Positives and False Negatives can be reduced by verifying more than one mini-burst, and then decide on a majority vote (provided that the initial number of errors is lower than 50%). In this paper, we measure the accuracy when only one mini-burst is verified.

IV. Results and Discussion

Figure 5 shows the mini-burst templates calculated for each of the four classes, generated by taking the average of all the mini-bursts extracted out of the training data. This is the result of the training phase, where the mini-burst templates of each DECT device are calculated.

Distinct patterns can be appreciated visually in the transient part of the signal, particularly in the ramp-up, possibly generated by front-end components of the DECT device including filters and amplifiers [15].

The verification scenario was carried out with the signals acquired a second day (validation dataset). Figure 6 reports False Positive and False Negative ratios with respect to the selection threshold, as well as the Equal Error Rate (EER), i.e., the point at which the two errors are equal.

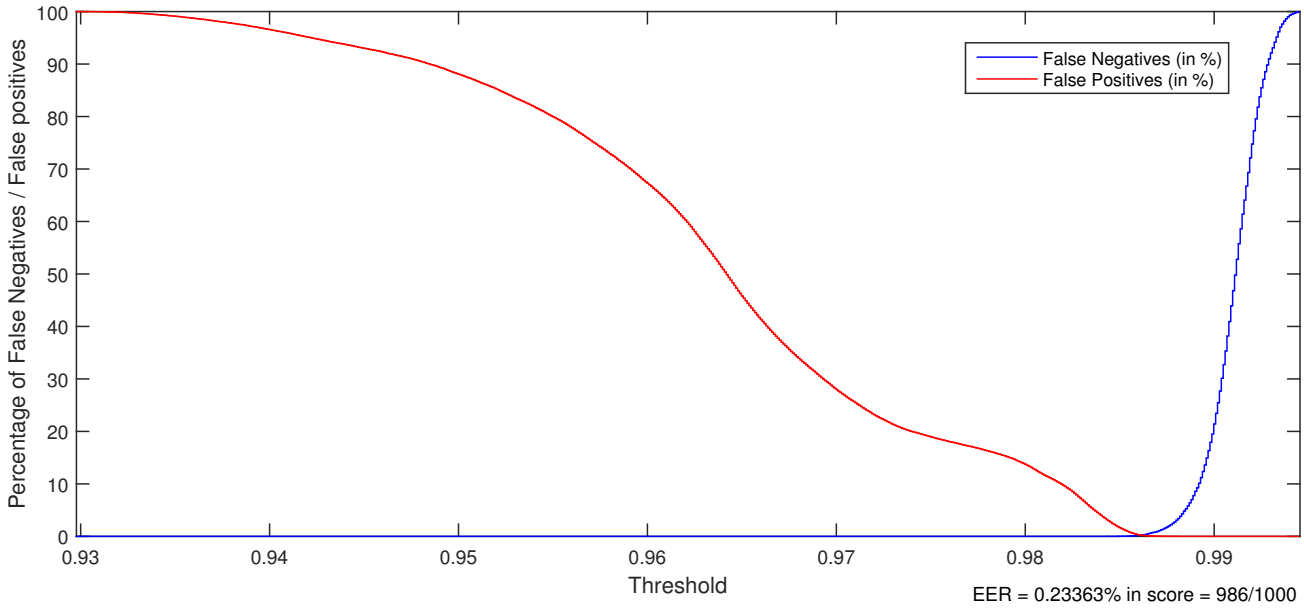


Fig. 6. False Positive and False Negative rates versus decision threshold Th. The Equal Error Rate (EER), i.e., the point where False Positive and False Negative rates are equal, is also reported.

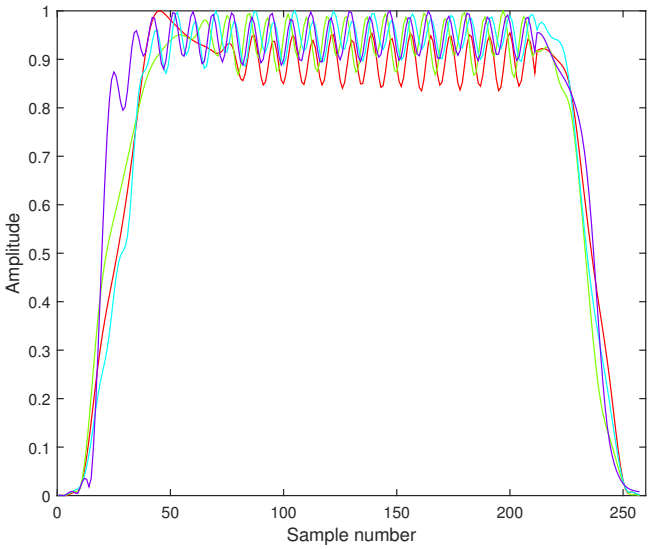


Fig. 5. Averaged Mini-bursts

The results of burst verification clearly show that it is possible to achieve almost 0% Equal Error Rate, i.e. a perfect discrimination between legitimate bursts and attacks. However, it is worth to point out the limited extent of our evaluation, which was limited to a small data set of bursts coming from four base stations. A more fair evaluation should average the results attained with several devices. Nevertheless, as highlighted in III-B if necessary the accuracy can be easily increased by classifying several bursts in place of only one, and then apply a majority voting rule.

In most of the actual spoofing scenarios, the hardware used by the attacker will be sensibly different from the one of the legitimate DECT FP whose identify is being spoofed. This is specially the case where the attacker uses a Software-Defined Radio device as a DECT transmitter and receiver, but also when specific DECT RF chip, completely different than the legitimate FP, is employed to launch the active attack.

The results obtained in the experiments demonstrate that the proposed fingerprinting technique can directly be used to detect such active attacks which involve device identity spoofing, in the probable scenario where the hardware used by the attacker does not perfectly match the one of the spoofed DECT device.

V. Conclusion

In this paper we have described an approach for the detection of device identify spoofing attacks in DECT communications based on the fingerprinting of the radio-frequency signal. We have demonstrated that the proposed RF fingerprinting technique, based on the analysis of the transient parts of the DECT signal, is effective in the detection of active attacks that impersonate the identity of the DECT base station, assuming that the RF hardware use by the attacker is not exactly the one of the legitimate station.

Unlike other techniques usually employed in intrusion detection, the RF signature is intrinsically linked to the RF circuit used by the transmitting device and therefore it is not easily manipulable by an adversary.

On the basis of the results obtained, we think that RF fingerprinting is a promising technique for the detection of active attacks on DECT communications and could be considered, in combination with other approaches, for its practical application by intrusion detection systems specialized in this type of wireless communications.

The proposed approach only takes into account the transient part of the signal and it uses a single burst for the classi-

fication. In order to handle more complex spoofing scenarios, where the same RF hardware than the legitimate DECT station is used to conduct the attack, future work will consider the usage of sequences of bursts in combination with other more complex RF fingerprinting techniques.

REFERENCES

[1] ETSI, Digital Enhanced Cordless Telecommunications (DECT). "Common Interface (CI). Part 1: Overview", ETSI standard EN-300-175-1, 2013.

[2] ETSI, Digital Enhanced Cordless Telecommunications (DECT) official website. Retrieved December 2014. URL: http://www.etsi.org/technologies-clusters/technologies/dect.

[3] S. Lucks, A. Schuler, E. Tews, R.P. Weinmann and M. Wenzel, "Attacks on the DECT authentication mechanisms", *Topics in Cryptology CT-RSA*, San Francisco,CA, USA, 2009, pp. 48-65.

[4] I. Sanchez, G. Baldini, D. Shaw and R. Giuliani , "Experimental passive eavesdropping of Digital Enhanced Cordless Telecommunication voice communications through low cost software defined radios", *Security and Communication Networks*. *8(3)*, 2014, pp. 403-417.

[5] K. Nohl, E. Tews, and R.P. Weinmann, "Cryptanalysis of the DECT Standard Cipher", *Fast Software Encryption*, Springer Berlin Heidelberg, 2010, pp. 1-18.

[6] Coisel, I.; Sanchez, I., "Practical Interception of DECT Encrypted Voice Communication in Unified Communications Environments", *IEEE Joint Conference on Intelligence and Security Informatics Conference JISIC*, The Hague, The Netherlands, 24-26 Sept. 2014, pp. 115-122.

[7] S. Yubo, H. Xili, L. Zhiling, "The GSM/UMTS Phone Number Catcher", *Third International Conference on Multimedia Information Networking and Security (MINES)*, Shanghai, China, 4-6 Nov. 2011, pp. 520-523.

[8] Barkan, Elad, Eli Biham, and Nathan Keller. "Instant ciphertext-only cryptanalysis of GSM encrypted communication", *Advances in Cryptology-CRYPTO 2003*. Santa Barbara, California, USA, 2003. pp. 600-616.

[9] P. McHardy, A. Schuler, and E. Tews, "Interactive decryption of DECT phone calls", *Proceedings of the fourth ACM conference on Wireless network security*, Hamburg, Germany, 2011, pp. 71-78.

[10] K. Nohl. "Mobile self-defense (SnoopSnitch)", *Chaos Computer Congress*, Hamburg, Germany, 2014

[11] F. Guo and T. Chiueh. "Sequence number-based MAC address spoof detection", *Recent Advances in Intrusion Detection*. Springer, 2006. pp. 309-329

[12] A. Chouchane, S. Rekhis and N. Boudriga, "Defending against rogue base station attacks using wavelet based fingerprinting", *IEEE/ACS International Conference on Computer Systems and Applications*, Rabat, Morocco, 2009, pp. 523-530.

[13] Y. Sheng, Tan, G. Chen, D. Kotz, A. Campbell, "Detecting 802.11 MAC Layer Spoofing Using Received Signal Strength", *27th IEEE Conference on Computer Communications INFOCOM 2008*, Phoenix, AZ, April 2008, pp. 13-18.

[14] W.E. Cobb, E.D. Laspe, R.O. Baldwin, M. A. Temple, Y.C. Kim, "Intrinsic Physical-Layer Authentication of Integrated Circuits", *IEEE Transactions on Information Forensics and Security*, 7(1), Feb. 2012, pp. 14-24.

[15] M. D. Williams, S. A. Munns, M. A. Temple, M. J Mendenhall, "RF-DNA fingerprinting for airport WiMax communications security", *Proceedings of 4th International Conference on Network and System Security, NSS 2010*, Melbourne, Australia, 2010, pp. 32-39.

[16] G. Lackner, U. Payer, P. Teufl, "Combating Wireless LAN MAC-layer Address Spoofing with Fingerprinting Methods", *IJ Network Security*, 9(2), 2009, pp. 164-172.

[17] S. Yong, K. Tan, C. Guanling, D. Kotz, A. Campbell, "Detecting 802.11 MAC Layer Spoofing Using Received Signal Strength", *27th Conference on Computer Communications INFOCOM 2008*, Phoenix, AZ, USA, April 2008, pp. 13-18.

[18] C.K. Dubendorfer, B.W. Ramsey, M.A Temple, "An RF-DNA verification process for ZigBee networks", *Military Communications Conference, MILCOM 2012*, Orlando, FL, USA, 2012, pp.1-6.

[19] J. Hasse, T. Gloe and M. Beck, "Forensic identification of GSM mobile phones", *Proceedings of the first ACM workshop on Information hiding and multimedia security*, Montpellier, France, 2013, p. 131-140.

[20] Anil K. Jain, P. Flynn and Arun A. Ross, "Handbook of Biometrics", Springer-Verlag New York, Inc., 2007.

Study of Two Security Constructs on Throughput for Wireless Sensor Multi-Hop Networks

R.D.Sparrow, A.A.Adekunle, R.J.Berry and R.J.Farnish
The Wolfson Centre for Bulk Solids Handling Technology,
University of Greenwich,Chatham Maritime,
Chatham, Kent ME4 4TB, England, UK
{r.d.sparrow, r.j.berry, r.j.farnish} @gre.ac.uk

***Abstract*—With the interconnection of devices becoming more widespread in society (e.g. internet of things), networked devices are used in a range of environments from smart grids to smart buildings. Wireless Sensor Networks (WSN) have commonly been utilised as a method of monitoring a set processes. In control networks WSN have been deployed to perform a variety of tasks (i.e. collate and distribute data from an event to an end device). However, the nature of the wireless broadcast medium enables attackers to conduct active and passive attacks. Cryptography is selected as a countermeasure to overcome these security vulnerabilities; however, a drawback of using cryptography is reduced throughput. This paper investigates the impact of two software authenticated encryption with associated data (AEAD) security constructs on packet throughput of multiple hop WSN, being counter with cipher block chaining and message authentication code (CCM) and TinyAEAD. Experiments were conducted in a simulated environment. A case scenario is also presented in this paper to emphasise the impact in a real world context. Results observed indicate that the security constructs examined in this paper affect the average throughput measurements up to three hops.**

***Index Terms*—Networked Control Systems, Wireless Sensor Networks, AEAD constructs.**

I. Introduction

Wireless Sensor Networks (WSN) have commonly been used in control and instrumentation tasks such as obtaining measurements to transmitting data across multiple wireless devices. The application for WSN have been deployed in control systems to remote monitor; however, due to the nature of the wireless broadcast medium utilised , potential attackers could conduct passive and active attacks to compromise the operation of the system [1].

Many challenges and issues have been documented outlining the current vulnerabilities and constraints associated with WSN, including limited computational power, trust of intermediately devices and the physical medium utilised to broadcast data [2]. The requirement for reliability in WSN is critical with a variety of security challenges that require addressing whilst maintaining the operation of the system [3]. High throughput applications are currently utilised over wireless mediums (e.g. video streaming), with emphasis on quality of service (QoS). WSN can be deployed to relay streaming data to devices outside the range of the broadcasting node and maintain an acceptable quality of experience (QoE) for the end user.

This paper investigates the effect of two software security measures on WSN with focus on Authenticated Encryption with Associated Data (AEAD) concept and how their implementation influences packet throughput of a WSN. Reasons for undertaking this investigation is to record the behaviour of packet throughput dependent applications (i.e throughput measures the successful message deliveries over a communication medium) with AEAD constructs implemented over multiple wireless hops. A mathematical model is proposed to predict packet throughput for multiple hop WSN communications.

The structure of this paper is organised as follows: Section II introduces the terms and phrases discussed in this paper and a relevant case scenario. Section III introduces relevant literature, Section IV introduces the proposed mathematical model. Section V discusses the experimentation methodology. Section VI analyses the results obtained from the simulation conducted with comparison against the proposed mathematical model. Section VII examines the mathematical model as a predictive tool. Section VIII discusses the impact in relation to the case scenario Section IX concludes.

II. Preliminary

This section defines the terms used throughout this paper and presents the context of the case scenario. The Authenticated Encryption with Associated Data (AEAD) concept provides symmetric cryptographic security services to transmitted packetised data. AEAD combines confidentiality and integrity resulting in a secure communication channel. Confidentiality as an encrypting function is thought secure if an adversary is unable to distinguish the ciphertexts from a bit string chosen uniformly at random, from the set of all possible bit strings of a specified length, under a chosen plaintext attack. For the purpose of this paper, an integrity check function is thought secure if it is computationally infeasible to perform an existential forgery under an adaptive chosen ciphertext attack.

Two AEAD paradigms are presented in this paper, they are Counter with cipher block chaining (CCM) and

TinyAEAD. CCM is a National Institute of Standards and Technology (NIST) standardised AEAD construct designed to provide integrity and confidentiality using a 128-bit block cipher [4]. The TinyAEAD construct is designed to provide confidentiality and integrity services using block ciphers of various bit lengths [5]. This construct can be configured to run at a reduced specified number of block cipher iterations in order to enhance the processing speed of the construct. In addition TinyAEAD provides flexibility and adaptability that can be applied to a broad range of contexts.

The case scenario selected focuses on a circuit switched WSN multi-hop communication link viewed as using a linear logical network topology. It is assumed that nodes in the link are 100m due to limited battery power to transmit data over this distance, therefore multiple intermediate nodes are required to relay data between the source and destination nodes [6]. Figure 1 illustrates the case scenario.

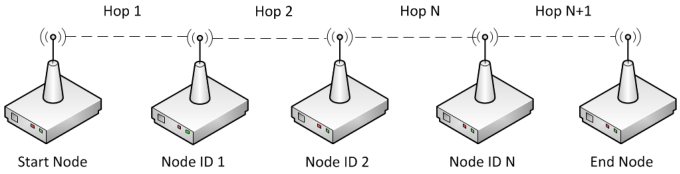


Figure 1. Illustrative concept of a circuit switched linear WSN multi-hop topology

In reference to Figure 1 the Start Node transmits a continuous flow of packets to node ID 1; the data is then transferred over *N* intermediate nodes before reaching the End Node. To secure the data being transmitted between each node, confidentiality, integrity and authentication is applied to prevent adversaries from viewing and modifying data in transit on the communication link; however, a drawback of introducing security is reduced throughput due to the additional processing time of the cryptographic process undertaken at each node during packet propagation.

III. Literature Review

This section introduces relevant literature related to the context of this paper with focus on confidentiality, integrity and authentication for WSN and the current mathematical models utilised to predict packet throughput over multi-hop networks.

Research conducted by Diaz et al [7] examines the attack vectors used against WSN with focus on the impact on power consumption and software execution time. The security vulnerabilities examined by the authors focuses on attacks against the physical and data link layer of the open systems interconnection (OSI) model. Three attack vectors are examined; link-noise, packet injection and direct attacks. Experimentation was conducted using SystemC to model the software execution time. A virtual WSN platform was created to examine collision, interrogation and Sybil attacks to analyse the effect on the power consumption and system performance. Two network topologies were also evaluated in this simulation; linear and mesh networks with each node utilising an ARM 926 processor and 802.15.4 transceiver (Zigbee). Results reported suggest that display power consumption varied depending on the network topology and attack selected.

A journal article presented by Pan et al [8] investigates WSN for healthcare. The main security vulnerability investigated in the article was providing confidentiality and integrity of patients data over a wireless broadcast medium as an attacker can conduct passive and active attacks to monitor and modify confidential data before arriving at its destination. The authors proposed solution uses a Feistel structure with two different methods, the first method incorporates an substitution box (Sbox) and the second method without a Sbox. Each Feistel cipher uses four rounds with a unique 32-bit key called at each round for encryption and decryption. Experimentation of the two Feistel constructs were conducted on an Atmel ATMega128L 8-bit microcontroller operating at 8 MHz crystal frequency. Comparative results against the Data Encryption Standard (DES) cipher suggests that the two Feistel ciphers are more suited to computational constrained WSN as the time taken to operate the encrypt and decrypt functions were faster than DES and memory consumption was also reduced.

An approach undertaken by Vu et al [9] presents a concept for simulation modelling of secure WSN. The focus of the research conducted in this paper examines their own bespoke simulation environment for modelling WSN, with emphasis on secure connectivity between each WSN. The security vulnerability discussed by the authors concentrates on active attack vectors (e.g. man-in-the-middle attacks) to determine what proportion of the WSN links remain secure. Five areas have been configured in the simulator being physical deployment, key establishment, node capture model, network analysis and a graphical user interface. Experimentation conducted by the authors suggests the results obtained for resilience against node capture for model and simulation correlate; however this methodology only take into consideration cryptographic key distribution only.

Research conducted by Marzi and Marzi [10] propose a security model for WSN using trust and reputation models (TRM). The issues highlighted by the authors emphasise the security vulnerabilities incurred by incorporating WSN into a networked environment (e.g. sinkhole attacks) and the computational constraints associated with WSN devices. The proposed solution utilises a Enhanced Bio-inspired Trust and Reputation Model (EBTRM) to calculate which node on the network is the most trustworthy before selecting the node to communicate with. The node also take into consideration if the communication was successful or unsuccessful and factor this into their computational model. Experimentation conducted reviewed the performance of the EBTRM against the Peer Trust System (PTS) and Bio-inspired Trust and Reputation Model (BTRM-WSN) using the TRMSim-WSN simulator. Results indicate that the EBTRM is suited for

selecting the correct trustworthy node in comparison to BTRM-WSN and PTS, even with a high percentage of malicious sensor deployed on the network; however, when compared to the average path length and energy consumption utilised the EBTRM consumed the most energy and took longer wireless paths than the BTRM-WSN approach.

Summary of the literature review undertaken suggests there is a requirement to investigate modern software security constructs suited for WSN as current approaches utilise constructs that are no longer standardised. Secondly, the current models for predicting throughput do not take into consideration the influence of the security constructs selected on the impact of packet throughput; suggesting an approach taking security into consideration is required. This suggests that an investigation is required to examine how modern cryptographic paradigms influences throughput over WSN.

IV. Proposed Mathematical Throughput Model

In this section we propose a mathematical model for predicting the total throughput of packets transmitted across multiple node WSN. The mathematical model presented in this paper utilises the exponential decay function to calculate the change of packet throughput over time (packets per second) for each wireless hop. This model takes into consideration that the wireless nodes would have no knowledge of the previous packet arrival (i.e. the amount of packets arrived at the previous hops), thus making the process memoryless. In addition it is assumed that the packet arrivals do not occur simultaneously, therefore, orderliness has been factored into this model.

The model is calculated by adding the initial packet throughout value for the initial start hop (N_o) divided by the time frame selected in seconds (e.g. per minute) (t) to calculate the packet arrival rate. The inverse of the value is obtained using -1/($\frac{N_o}{t}$) before using the value with the exponential function (e). The value obtained is then multiplied by the total throughput value of the previous wireless hop packet throughput (N_o) to calculate the rate change in packet throughput (r).

$$r = N_o e^{(-1/(\frac{N_o}{t}))} \quad (1)$$

Equation 1: Exponential decay calculation for estimating throughput rate of change per wireless hop

The final calculation is to subtract r from N_o to derive the new packet throughput measurement for N_{o+1} as formulated in Equation 2.

$$N_{o+1} = N_o - r \quad (2)$$

Equation 2: Calculation for estimating throughput per wireless hop

V. Experimentation Methodology

This section discusses the apparatus, metrics and context selected for the experimentation methodology. The programme selected is Proteus ISIS 8 professional with an emulated Microchip PIC18F45K22 selected as the microcontroller. The Serial Peripheral Interface (SPI) is selected as the physical layer (e.g. OSI model) to transmit and receive messages between each microcontroller on the WSN. The AEAD security algorithms used are CCM and TinyAEAD running AES (128-bit key variant). The communication channel for this experiment is simulated.

The experimentation methodology is used for observing packet throughput for ranging multiple nodes with and without software security measures applied. Packets are sent continuously and counted by the last hop device to measure how many packets have arrived at its destination within one minute time sample. All timings and packet throughput counts were taken from the simulator used. It is assumed for this scenario that no noise is present on the wireless channel and consequently all packets transmitted are received without loss or corruption.

Metrics used for the experimentation are seconds for the time frame of the experiment. The packet throughput is counted to measure how many packets arrived in the time frame and number of nodes to state how many intermediate devices were between the start and end node.

Configuration of the components selected are as follows, the crystal frequency selected is 4 MHz to replicate low powered industrial microcontrollers with packet sizes of 36, 52 and 84 bytes. A SPI divisor of 16 is chosen to replicate bandwidth of a wireless link of 250Kbps [11] as calculated using the formula obtained from previous research [12]. It is assumed that a 100m distance is between each wireless node. The experimentation methodology varied the number of intermediate nodes on the linear network, starting from one node to the maximum of five nodes.

VI. Results and Analysis of Experimentation

This section analyses the results obtained from the experimentation methodology discussed in the aforementioned section. Figure 2 illustrates the total throughput for a packet size of 36 bytes. Results presented in Figure 2 suggests that the number of intermediate nodes influences the total amount of successful packets being received within the time frame for a 36 byte packet with one hop showing the largest throughput count and five hops showing the smallest throughput count. The packet size has minimal impact based on the data presented. TinyAEAD at three rounds has a smaller impact on throughput than CCM.

Figure 3 presents data obtained from increasing the packet size to 52 bytes. The data indicates that security has reduced the throughput up to three nodes with CCM having the least throughput recorded. The larger hop counts show a minimal effect on the total throughput readings.

The data in Figure 4 represents results obtained for a 84 byte packet size. The results indicate CCM has a greater impact on

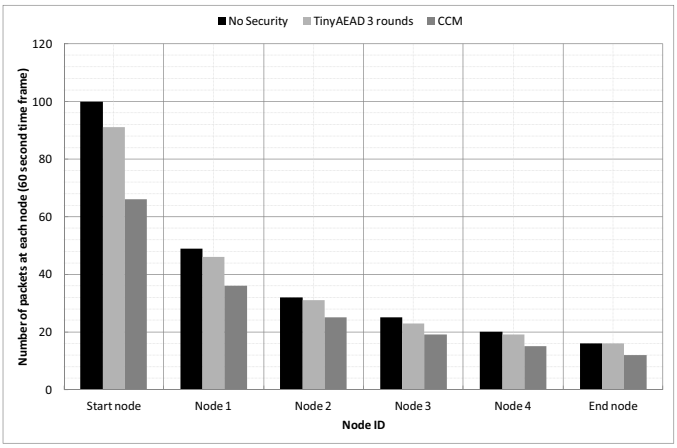


Figure 2. Simulation results for 36 byte packet lengths over multiple intermediate nodes for a sixty second time frame

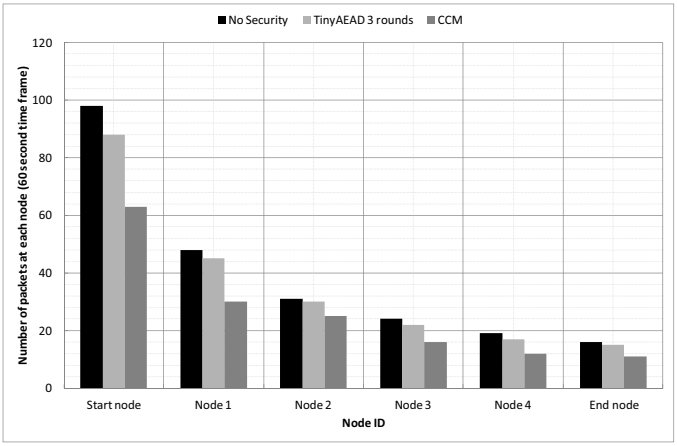


Figure 3. Simulation results for 52 byte packet lengths over multiple intermediate nodes for a sixty second time frame

the total throughput; it is also illustrated that larger packet sizes has ta larger decrease on throughput values in comparison to smaller packet sizes.

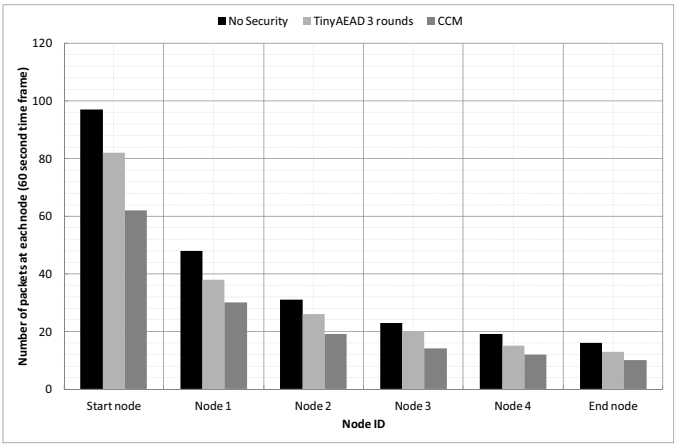


Figure 4. Simulation results for 84 byte packet lengths over multiple intermediate nodes for a sixty second time frame

The number of rounds utilised by the underlying block cipher used by the AEAD algorithms has a direct effect on the throughput values generated as the AEAD algorithms operating at reduced number of rounds outputted more packets in comparison to specified maximum number of rounds.

The quantity of intermediate nodes between the source and destination also has an influence on the total throughput value as the more nodes that were present in the network, the larger the reduction in the packet throughput; however, the AEAD algorithms had a greater impact for nodes closest to the Start Node, whilst nodes further away from the Start Node endured a minimal impact, suggesting that multiple hop wireless propagation over three nodes are suited to operating at stronger security configuration as the degradation of the packet throughput is reduced, whilst lower hop counts below three nodes are suited to lower security parameters as the effect on throughput is more noticeable.

Analysis of the results obtained from the simulation and the results acquired from the mathematical model is presented in Table 1. Table I compares the results for a 36 byte packet size over varying hop counts using a 250Kbps link.

Table I
NO SECURITY THROUGHPUT COMPARISON OF MATHEMATICAL MODEL VS SIMULATION (36 BYTES)

NUMBER OF NODES	SIMULATION RESULTS (PACKETS)	MODEL RESULTS (PACKETS)	CORRELATION COUNT
Start node	97	97	0.9998
1	48	52	
2	31	35	
3	23	28	
4	19	24	
5	16	22	

Data displayed in Table I suggests that the mathematical model correlates with the simulation results obtained up to three nodes. Correlation analysis indicates that the mathematical model and simulation results have a positive linear correlation. Table II shows the comparison data obtained from using TinyAEAD at three rounds.

Table II
TINYAEAD THREE ROUND THROUGHPUT COMPARISON OF MATHEMATICAL MODEL VS SIMULATION (36 BYTES)

NUMBER OF NODES	SIMULATION RESULTS (PACKETS)	MODEL RESULTS (PACKETS)	CORRELATION COUNT
Start node	91	91	0.9989
1	46	47	
2	31	33	
3	23	27	
4	19	24	
5	16	22	

Results in Table II indicates that the mathematical model correlates with the simulation up to three nodes. Correlation analysis suggests a positive linear correlation between the simulation and mathematical model. Table III compares data

obtained from simulation and modelling for CCM.

Table III
CCM THROUGHPUT COMPARISON OF MATHEMATICAL MODEL VS SIMULATION (36 BYTES)

NUMBER OF NODES	SIMULATION RESULTS (PACKETS)	MODEL RESULTS (PACKETS)	CORRELATION COUNT
Start node	66	66	0.9605
1	36	26	
2	25	23	
3	19	21	
4	15	19	
5	12	18	

Data illustrated in Table III indicates that the mathematical model is suited up to three nodes before difference increases. The correlation count between the simulation and mathematical model indicates a positive linear correlation between the two data samples.

Reviewing the comparison between the mathematical model and the simulation results, data obtained indicates that the mathematical model is better suited for calculating the throughput expected up to three nodes, this indicates that the mathematical model error increases as the number of nodes increases. Correlation analysis of the data obtained suggests the model and simulation have a positive correlation with no security and adjustable AEAD algorithms, whilst fixed AEAD algorithms have a lower positive correlation.

VII. PREDICTIVE MODELLING

This section utilises the mathematical model as a prediction tool for the total packet throughput over a select number of nodes over a period of time. Two experiments were conducted, first the prediction of throughput values over six nodes for varying packet sizes and security algorithms. A second experiment assesses the viability of the model as a predictive tool by varying the packet size, time frame, number of hops and rounds used by TinyAEAD. Table IV displays the data obtained for packet throughput values over six hops for a time frame of sixty seconds. The predictions were obtained using the first hop packet throughput value for each packet size and security construct obtained in Figures 2, 3 and 4 as a base reference for the variables used for the calculation.

Table IV
MATHEMATICAL MODEL VS SIMULATION RESULTS FOR THROUGHPUT MEASURED OVER INCREASED MULTIPLE NODES

NUMBER OF NODES	SECURITY CONSTRUCT	PACKET SIZE (BYTES)	PREDICTIVE MODEL RESULTS (PACKETS)	SIMULATION RESULTS (PACKETS)
6	CCM	36	11	16
7	TINYAEAD 3 ROUNDS	52	12	19
8	TINYAEAD 5 ROUNDS	84	7	12
9	TINYAEAD 10 ROUNDS	36	8	14

Results obtained from the predictive model and the simulation results suggest that the results do not correlate, this suggests that the mathematical model is not suited for predicting over five nodes. Table V tabulates data obtained from varying the time frame, rounds used for the underlying block cipher for TinyAEAD, size of the data packets and number of nodes were selected to indicate the viability of the mathematical model as a predictive tool.

Table V
MATHEMATICAL MODEL VS SIMULATION RESULTS FOR THROUGHPUT MEASUREMENTS OVER VARYING PARAMETERS

	TINYAEAD 3 ROUNDS	TINYAEAD 5 ROUNDS	TINYAEAD 10 ROUNDS
NUMBER OF HOPS	2	3	4
PACKET SIZE (BYTES)	36	52	84
SAMPLING TIMES (S)	120	180	240
PREDICTIVE MODEL RESULTS (PACKETS)	93	76	62
SIMULATION RESULTS (PACKETS)	92	80	63

Data presented in Table V indicates that the predictive results obtained from the proposed model correlate with the simulation results.

VIII. DISCUSSION

This section discusses the results obtained from experimentation and how they influence the case scenario of WSN previously discussed in Section II. To clarify how security impacts on packet throughput, Figure 5 graphs the instantaneous throughput measurements for a packet size of 36 bytes.

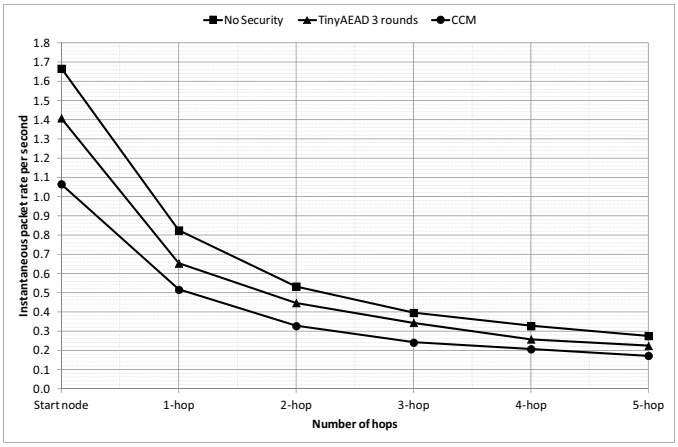


Figure 5. Instantaneous throughput measurements for 36 byte packet lengths over a sixty second time frame

Trends graphed in Figure 5 suggests that implementing security algorithms influences the initial rate of packets received

in the given time frame as no security has the highest throughput rate per second, however, the selection of the security algorithms also influences the instantaneous packet rate as TinyAEAD at three rounds outputs 0.3 packets per second faster than CCM from the initial start hop. The percentage difference between no security and CCM from the start node is 35% whilst the percentage difference between TinyAEAD and no security is 19%. This suggests that the initial impact of security algorithms on throughput is offset from the start node and propagates over multiple hops.

IX. Conclusion

The proposed mathematical model presented in this paper is suitable for calculating the impact of security algorithms using varying packet lengths and security algorithms up to four nodes in length. The amount of intermediate nodes on the network has influence on packet throughput measurements obtained with the number of packet arriving at each hop decaying exponentially; this indicates that the rate of throughput decreases with the increase in intermediary nodes.

Concluding from this investigation undertaken, it has been identified that the use of current standardised fixed AEAD algorithms impact on the throughout of WSN in comparison to adjustable AEAD algorithms, suggesting that current standardised AEAD algorithms using fixed parameters are not as suited in comparison to adjustable AEAD algorithms if packet throughput is a priority.

The mathematical model proposed contributes towards predicting the effect of security algorithms for variable sized WSN for a range of sampling time frames in simulation. Future work will validate the findings obtained on real world platform.

References

[1] A Blilat, A Bouayad, N Chaoui, and M El Ghazi. Wireless sensor network: Security challenges. In *Network Security and Systems (JNS2), 2012 National Days of*, pages 68–72, April 2012.

[2] S. Ahmad Salehi, M.A Razzaque, P. Naraei, and A Farrokhtala. Security in wireless sensor networks: Issues and challanges. In *Space Science and Communication (IconSpace), 2013 IEEE International Conference on*, 2013.

[3] K. Islam, Weiming Shen, and Xianbin Wang. Wireless sensor network reliability and security in factory automation: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):1243–1256, November 2012.

[4] Recommendation for block cipher modes of operation: The ccm mode for authentication and confidentiality.

[5] A. Adekunle and S. Woodhead. An aead cryptographic framework and tinyaead construct for secure wsn communication. In *Wireless Advanced (WiAd)*, 2012.

[6] R Purta, S Nagrecha, and G Madey. Multi-hop communications in a swarm of uavs. In *Proceedings of the Agent-Directed Simulation Symposium*, 2013.

[7] A Diaz, P. Sanchez, J. Sancho, and J. Rico. Design, automation test in europe conference exhibition. In *Wireless sensor network simulation for security and performance analysis*, pages 432–435, 2013.

[8] J. Pan, S. Li, and Z Xu. Security mechanism for a wireless-sensor-network based healthcare monitoring system. *Communications, IET*, 6(18):3274–3280, Dec 2012.

[9] Tuan Manh Vu, Carey Williamson, and Reihaneh Safavi-Naini. Simulation modeling of secure wireless sensor networks. In *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, 2009.

[10] H. Marzi and A Marzi. A security model for wireless sensor networks. In *Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2014 IEEE International Conference on*, 2014.

[11] Texas Instruments. 2.4 ghz ieee 802.15.4 / zigbee-ready rf transceiver, 2014.

[12] R Sparrow, A Adekunle, J Berry, R, and J Farnish, R. Simulating and modelling the impact of security constructs on latency for open loop control. In *Sixth Computer Science and Electronic Engineering Conference 2014 (CEEC'14)*, 2014.

A Practical NFC Relay Attack on Mobile Devices Using Card Emulation Mode

D. Cavdar* and E.Tomur **
* Middle East Technical University, Ankara, Turkey
** İzmir Institute of Technology, Izmir, Turkey
dcavdar@metu.edu.tr, emrahtomur@iyte.edu.tr

Abstract - In this study, a practical card-emulated relay attack is implemented on Near Field Communication (NFC) equipped mobile devices. NFC is a promising communication technology which is also used in smart mobile devices. As an effective and flexible communication technology, NFC is frequently used in innovative solutions nowadays such as payments, access control etc. Because of the nature of these transactions, security is a critical issue that should be considered in system design and development phases. Although inherited from Radio Frequency Identification (RFID) technology, NFC security needs, requirements and solutions differ in terms of its usage areas and solutions. Based on these parameters, security precautions in communication layer of RFID technology do not prevent relay attacks occurred in the application layer NFC solutions. This study is conducted to prove relay attack practicability with using only mobile phones for relaying credentials instead of RFID based smart cards in an access control application. The Host Card Emulation (HCE) mode also eases relay attacks in NFC communication. The study explains the conceptual description of proposed relay attack, development and operating logic of mobile applications working based on card emulation mode and server software and also data communication basics between modules and web services descriptions.

***Keywords*- NFC, Relay Attack, Mobile, Card Emulation**

I. INTRODUCTION

Daily communication devices such as computers, mobile phones etc. have evolved in the last decades in terms of both size and functionality. Although they are small in size, they became very capable devices in terms of performing complex and critical processes. This state is defined as the third era of computing [1] and named "Ubiquitous Computing" or "Pervasive Computing". In this stage, standard computer perception changes both in appearance and logic.

In the last decade, the communication and life style of people have completely changed. People have started to use smart mobile devices for their daily routines such as mailing, dealing with documents, entertaining etc. As a result, production and sales of smart mobile devices, mobile operating systems, mobile marketing and mobile solutions have extremely increased. According to the research conducted by Collins, there are 7.3 billion mobile devices in the world which exceeds world population. These numbers also show the impacts of mobile ecosystem.

Near Field Communication (NFC) technology is one of the communication conveniences that emerge from mobility tendency and device interaction concept. The main function of NFC is to establish connection between two mobile devices or NFC tags and reader. Data exchange or access requests can be easily performed with this concept. NFC enabled smart mobile devices are used in daily applications and create smart solutions. For example; NFC technology is used for contactless payments which are also the most popular NFC applications, loyalty couponing, transport systems (bus, train etc.), data exchange, gaming, location based services and access control systems.

This paper has five sections: After introduction, Section 2 explains Near Field Communication basics and Section 3 includes literature review about the study. Relay attack is described in Section 4 and finally Section 5 concludes the paper with summary and possible future works of this study.

II. NEAR FIELD COMMUNICATION

With the invention of mobile phones, the intent was to enable making voice calls via wireless instead of using traditional wired phones. After Second Generation (2G) GSM technology, mobile phones have been equipped with not only to perform voice but also to send/receive text messages and also to have internet experience capabilities. In addition to standard cellular network connection over base stations, mobile devices have started to establish connection to other electronic devices for data exchange and interaction. The infrared technology was used in first wireless communication trials, then Bluetooth and Wi-Fi wireless technologies have been developed for connection between interactive mobile devices.

1) Basics of NFC

After the studies of hardware developments, Near Field Communication (NFC) technology was introduced in 2002 by Philips and Nokia [3]. In 2004, NFC Forum was established in order to foster cooperation and standardization by Nokia, Microsoft,

This study is supported by the Scientific and Technological Research Council of Turkey under 2211 program.

Sony, and Philips. In early 2010s, NFC Forum has started to introduce NFC standards such as: Logical Link Control Protocol (LLCP) Specification, NFC Data Exchange Format (NDEF) Specification and NFC Record Type Definition (RTD) Specification etc. Then, mobile device manufacturers have been influenced from these developments and started to produce smart NFC products which are NFC enabled mobile devices, tags and readers. Finally, NFC became a standardized short range, easy to use and set up, flexible and stable communication technology.

NFC provides short range (up to a few centimeters) wireless communication between two electronic devices. In order to establish communication, at least one active (energy provider) device should initiate communication process. NFC enabled mobile devices and NFC readers such as contactless payment readers are active devices, whereas NFC tags and NFC posters are passive devices.

NFC operates in 13.56 MHz High Frequency (HF) unlicensed frequency band that is available to all manufacturers with range of up to 4 cm. Data transfer rates of NFC are 106, 216 and up to 424 Kbps which is higher than RFID data transfer rates, however much lower than 3G and Wi-Fi rates. The reason of that is the main purpose of NFC operations is not transfer high volume data to destination.

2) Operating modes of NFC

NFC devices interact via three different modes according to application types and needs. This flexibility creates a customizable environment for developers and manufacturers.

a) Peer-to-Peer Mode

In Peer-to-Peer mode, two active NFC enabled mobile phones interact with each other in two directional data exchange paradigm. Firstly, one mobile device initiates communication set-up process using request-response mechanism, and other mobile device replies its response. The connection is established according to Logical Link Control Protocol Specification (LLCP) created by NFC Forum and data exchange is performed based on Simple NDEF Exchange Protocol (SNEP) protocol basics. The data, picture, URL exchange operations between two mobile devices use this mode of operation.

b) Reader/Writer Mode

Active mobile devices can both read and write passive NFC tags in Reader/Writer mode. Similar to RFID cards, NFC cards respond to requested data sent from initializer mobile device or write demanded data. Passive NFC tags are operated according to NDEF and RTD message format standards. Location based services, information tags or smart posters use this operation mode.

c) Card Emulation Mode

In this mode, two active NFC devices which are mobile device and reader are used. The mobile devices act as smart card based on ISO/IEC 14443 Type A, Type B and FeliCa standards in the application. When mobile phone gets close to reader, similar to RFID applications, reader initiates interaction and reaches mobile phone's NFC module and Secure Element if available. Payment and Access Control applications use this mode for granting access.

III. RELATED WORK

The Near Field Communication (NFC) technology is relatively a new technology which is introduced within the current decade [4]. The first versions of related standards have been released in last five years by NFC Forum and ISO [5], also device manufacturers have started to develop more NFC devices after these developments.

After mobile payment systems have been launched especially after Google Wallet solution, the popularity of NFC has increased. As a result of this popularity, trials of practical attacks on mobile payments were conducted in research [6], [7] and [8]. These studies demonstrate the applicability of security problems in mobile payment domain.

The implementations of NFC relay attack examples were conducted successfully in [6] and [7]. A mobile application installed on mobile device is used in attack; however, no security precautions are discussed in the study [9]. In the studies [6], [8] and [10], ISO/14443 based RFID cards were exposed to relay attack and successfully relayed. As countermeasures, protocols of distance bounding and Frame Waiting Time were offered in [6] and [8], however, they cannot provide sufficient precautions for attacks.

In the ISO/14443 standard, Frame Waiting Time (FWT) is defined for standard smart card and reader communication. FWT variable defines maximum response time after the end of the reader's data. FWT is defined as $(256 \cdot 16/ f_{carrier}) \times 2^{FWI}$, where FWI is a value from 0 (FWT = 300 μs) to 14 (FWT = 5 s) with a default of 4 (FWT = 4.8 ms). The control of this variable was offered in the literature for avoiding relay attack in [6] and [8]. This variable can be modified by attacker or some additional readers may be placed between reader and card in order to overcome this variable. However, the main reason of its unsuitability is that mobile devices are used in the proposed model instead of smart cards and when a relay attack occurred, only NFC based mobile phones

communicate with each other. There is no timing limitation between two NFC based mobile phone communications, because two devices are active not like in the ISO/14443 based passive cards. This problem should be solved with application layer security methods.

Location based control is also offered as another countermeasure in the studies [8] and [11]. Distance-bounding protocols use FWT values for calculating round trip time and finally comment on requester's location. As it is mentioned, using FWT is not a suitable solution for avoiding relay attacks especially where NFC enabled mobile device are used. These studies also show the significance of this research, because, it proves relay attack practicability with using only mobile phones for relaying credentials instead of RFID based smart cards and any other intermediate devices.

IV. Relay Attack

In addition to other security threats such as eavesdropping, data modifications etc. in the short range wireless communication, relay attacks are frequently used for performing interventions.

Attacker captures and forwards relayed message to proxy device and this device delivers the message as if the real owner in a typical scenario. Similar to relay attacks, man in the middle attacks also influence wireless medium, however, the method and also purpose of relay attacks have completely changed. These attacks are frequently used in ISO/IEC14443 based passive smart card communication environment. Since these cards are used in critical payment and access control systems, relay attacks in this environment result in important security threats. For example; in the research [7], ISO/IEC14443 based smart contactless credit card is successfully relayed over proxy devices. In this application, relay mobile phone records debit card data in reader mode and forwards this data to another phone via wireless communication such as Wi-Fi or Bluetooth. Proxy NFC based mobile devices send credentials to reader as the real owner in card emulation mode. If the reader cannot detect the fake sender, relay attack is completed successfully. This activity flow is illustrated in Fig. 1.

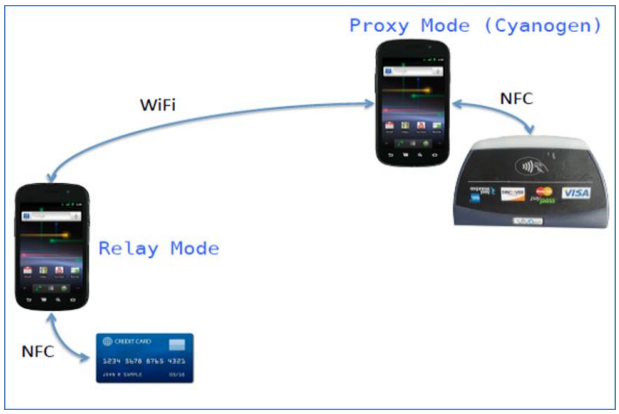


Figure 1. Typical Relay Attack scenario using contactless credit cards.

In the literature, NFC enabled mobile phones are generally used for relaying smart cards, in other words mobile phones read and send the data only. Credit cards, debit cards, and even electronic passports are used in relay attack experiences, however, in the access control systems proposal; mobile devices are used in card emulation mode instead of using smart cards. On the other hand, smart cards are limited and just reply the requested data

A. *Host Card Emulation Mode (HCE)*

NFC devices basically operate in three different implementation modes, Reader/Writer, Peer-to-Peer and Card Emulation mode. Although three modes are specified, developers could use only two of them, Reader/Writer and Peer-to-Peer in Android operating system because of Google policies. Only Google Wallet [12] application could use this operation mode for its transactions. With release of Android 4.4 (API Level 20), developers have started to use Host Card Emulation Mode [13] in their solutions and users have started to use their mobile phones like smart cards. We have also moved to that operation mode in our system because of its functionality and flexibility compared to its other modes.

The use of HCE in solutions also eases the practicability of relay attacks in the applications. Shifting between Reader and Card Emulation modes on same mobile device provides great chance to relay attackers instead of using different intermediate devices. HCE has also been used in practical relay attack scenario of the study.

B. Implementation

The relay attack setup in this study shows differences in transaction methods and relaying medium. As it is illustrated in Fig 3, the high level components of NFC based access control systems are NFC Enabled mobile device, NFC reader, access control panel and central server.

The general activity flow of NFC based access control system begins with controlling whether NFC unit of the mobile device is active or not. If NFC is enabled, user registers to the system with his/her mobile operator number, username etc. using application that is installed on the device. Once user is registered to the system, central server activates his/her account and sends activation key to authenticated mobile device. The activation key is checked based on predefined parameters. After registration process is completed successfully, user can now send access requests with his/her user name, password and expiration time of key parameters.

After getting key from the server using JSON based web services infrastructure, user can now send authentication requests to NFC reader with acquired key. At that time, relay phone can interfere NFC communication between reader and mobile phone. Relay phone can read the key that stored in the real mobile phone by getting close to phone in reader operation mode. In other words, relay phone imitates the NFC reader. If requester cannot identify the real

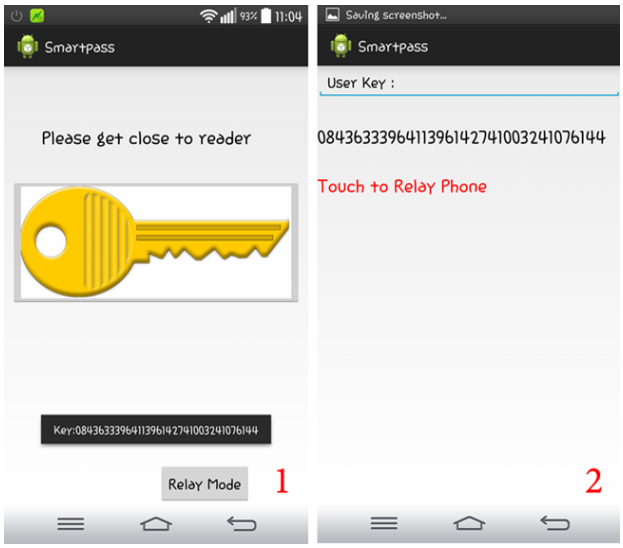


Figure 2. Screenshots of mobile application

NFC reader, relay phone gets the user key from real owner with operating peer-to-peer mode.

In order to complete the attack successively, relay phone should send access request to NFC reader with the key that captured from real phone. In order to do that, relay phone shifts its operating mode to Host Card Emulation Mode, by doing this; it now acts as a smart card and sends requests. These processes are illustrated in Figure 4. After these transactions, NFC reader gets the key and forwards to the server. Because of the validity of the key, unauthorized attempt can access to resource.

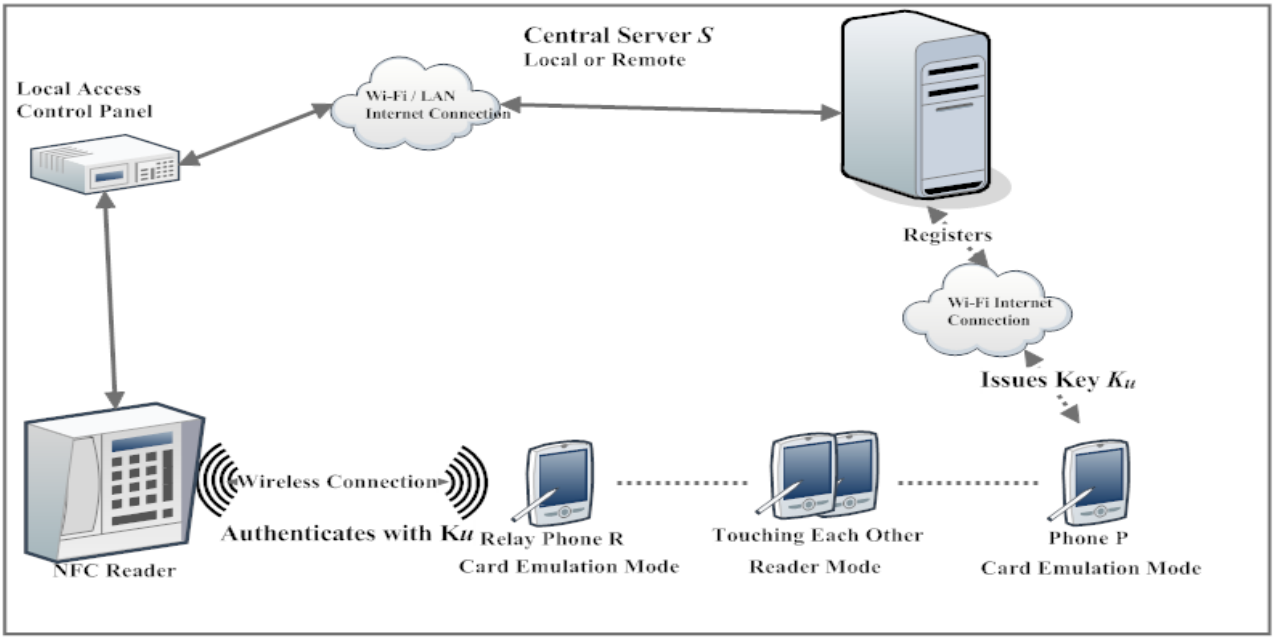


Figure 3. NFC Relay Attack scenario for access control system

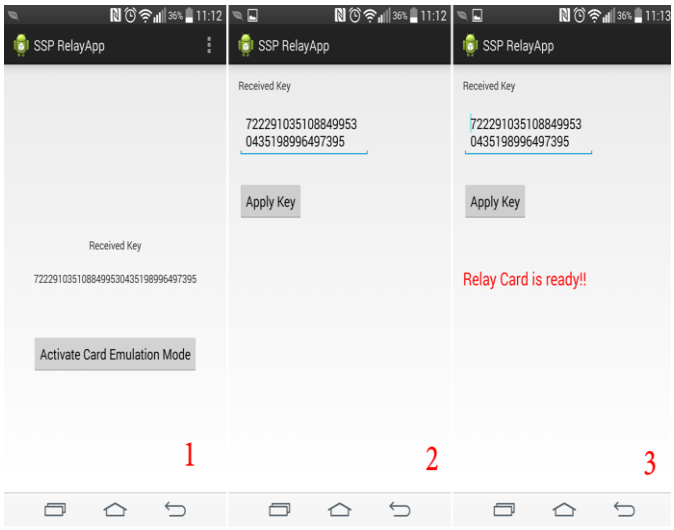


Figure 4. Screenshots of relay application

The whole process is illustrated in Figure 3. In the attack environment, relay phone has no wireless connection to server; therefore any key exchange occurred between server and relay phone. It captured the key from user mobile device using peer-to-peer operating mode and then changed its mode to card emulation mode. After this change, it created a virtual smart card using captured key from user device and requested access. NFC reader, access control panel and central server could not realize the relayed mobile device and attack is successively performed.

In a typical relay attack scenario which is also illustrated in the Figure 1, at least two mobile devices are used for relaying credentials. However, in this study, only one mobile device performs attack with the help of ability of switching operating modes of NFC. This is the main originality of the study. Also this convenience provides flexibility to attackers, because, in this kind of environments, attackers do not need to place between devices physically at the same time. Once he captures the credentials from real phone in reader mode, then he can use them another time after switching NFC mode to card emulation mode. Critical part of this type of relay attacks is whether relayed sender mobile device is detected or not by reader during authentication.

V. CONCLUSION & FUTURE WORK

Practical and successful relay attack scenario on mobile device using card emulation mode and its implementation steps are described in this study. NFC is a relatively new and popular wireless communication technology especially on mobile devices and used in critical solutions frequently. As well as its functionality, it is exposed to the relay attacks. Because of inheritance from RFID standards, the proposed precautions of communication layer such as defining FWI for transactions between receiver and transmitter do not prevent relay attacks in NFC environment. The study proves NFC communication security gap in access control systems using mobile devices. This type of relay attacks occur in the application layer, therefore, the real-time precautions that work on this layer and able to detect real sender should be developed for prevention as a future work.

REFERENCES

[1] Mark Weiser, “The computer for the 21st Century”. Scientific American. Vol. 265, No. 3, 1991, pp.66 -75.

[2] Romen, G. Presentation, 2010, NFC and the NFC Forum

[3] NFC forum, accessed from http://nfc-forum.org/ on January 2, 2014

[4] NFC forum, accessed from http://nfc-forum.org/our-work/specifications-and-application-documents/ on January 2, 2014

[5] NFC Forum. NFC data exchange format (NDEF). Technical specification, version 1.0. 2006

[6] Vounter van Dullink, Wouter, and Pieter Westein. "Remote relay attack on RFID access control systems using NFC enabled devices." 2013. unpublished.

[7] Hancke, Gerhard P. "A practical relay attack on ISO 14443 proximity cards." Technical report, University of Cambridge Computer Laboratory, 2005: 1-13.

[8] Practical Relay Attack on Contactless Transactions by Using NFC Mobile Phones," Lishoy Francis, Gerhard Hancke, Keith Mayes, and Konstantinos Markantonakis. To appear in The 2012 Workshop on RFID and IoT Security (RFIDsec 2012 Asia), November 8 - 9, 2012, Taipei, Taiwan

[9] Zhao Wang, Zhigang Xu, Wei Xin, and Zhong Chen. 2012. Implementation and Analysis of a Practical NFC Relay Attack Example. In Proceedings of the 2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control (IMCCC '12).

[10] Master thesis, Performing Relay Attacks on ISO 14443 Contactless Smart Cards using NFC Mobile Equipment Henning Siitonen Kortvedt, May 17, 2010 Der Technischen Universit At Munchen

[11] Lishoy Francis, Gerhard Hancke, Keith Mayes, and Konstantinos Markantonakis. 2010. Practical NFC peer-to-peer relay attack using mobile phones. In Proceedings of the 6th international conference on Radio frequency identification: security and privacy issues (RFIDSec'10

[12] Google Inc. accessed from https://www.google.com/wallet/ on January 15, 2015

[13] Google Inc. accessed from https://developer.android.com/ guide/topics/ connectivity/nfc/hce.html on January 15, 2015

Honeytokens as active defense

A.B. Robert Petrunić, bacc.ing.comp.,
Algebra university college for applied computer engineering, Zagreb, Croatia
robert.petrunic@racunarstvo.hr

Web applications are one of the mostly attacked platforms today, and because of that new ways to break into the web applications are being invented almost on a daily basis, allowing attackers to steal user's personal data, credit card numbers, and conduct many other frauds related to data and applications hosted on the Internet servers and databases.

Some of the reasons that web applications are constantly attacked is 24/7 availability, mix of technologies used to provide needed functionality, interesting data in the backend databases and easy way to avoid punishment for crimes committed against web sites and website users/owners. There is also an aspect related to cybercrime and cyber warfare that is marching throughout the planet in the last few years, exposing more and more personal data in highly sophisticated and targeted attacks.

This paper will try to summarize few different ways that web application could be written in order to identify, isolate and track the hacker during the attack process. The concept presented in this paper is so called honeytoken – a value the application is using in databases, files, parameters, etc., which should never be changed or touched by the application in normal application lifecycle.

Index terms – Web application security, honeytoken, active defense

I. INTRODUCTION

The reason for writing this paper is because the methods explained inside are only partially being used in the enterprise grade web applications, and because there are not so many papers related to this kind of malicious user protection.

Today, we have to fight against various threats on the Internet that will try to break into vulnerable web site. This threat might come from criminal organizations, trying to infect as much computers a possible in order to build powerful botnets that will be used to steal money, credit card numbers, send spam messages, provide bandwidth for DDoS (Distributed Denial of Service), etc. It might also come from Government sponsored hackers that will use this as Cyber weapon, from hacktivists that will try to use it as a tool to prove they cause, and from other hacker types. It's almost impossible to fight against all of this threats by using traditional methods and tools, and in order to raise the bar a little bit higher for them, one has to write the web application that will be capable of self-protection of some kind. This self-protection is covered throughout the paper discussing the cons and pros for every available method.

In most of the cases, the attack starts with scanning and vulnerability scanning of web application, that consist of testing all of the parameters the application uses, against the database of well-known attack vectors (for example, OWASP project TOP 10 [1]) and vulnerabilities. This is really loud process and it is easily recognizable on the server side and in most of the cases it could be stopped by using WAF (Web Application Firewall [2]). The problem is that WAF could only prevent known attack vectors, and additional protection is mandatory in modern web applications.

The Section "Honeytokens explained" is focusing on basic concepts behind honeytoken and their usage followed by Section "Honeytoken implementation options" that is focusing on different approaches and possible practical honeytoken implementation inside web applications. "Logging and reporting" Section explains the importance of log management, while "Implementation example" is showing the code excerpts one could use to implement proposed functionality.

II. HONEYTOKENS EXPLAINED

In most of the cases, web application is exchanging data with user part of the communication channel by using parameters, and this is the perfect place to use the honeytoken (canary) value in order to fight against attackers and automated tools. The purpose of honeytoken is quite simple: this is the value that should never be used by the application in normal usage scenario. It means that at the same time the value of honeytoken has changed, someone has done something that is not allowed in the current application context, and the session should be disconnected or isolated. Honeytoken could be anything from the web application parameter, through the data inside the database to the file on the file system. Depending on the needs it could be implemented on any or all possible layers of the application and application data. That being said, honeytoken value inside web application parameter will help to identify automated or manual scanning procedures and it could be used as the early warning system that could also be used to prevent the further system compromise. If the system has been compromised (even though the previously explained web application parameter honeytoken has been used) honeytoken inside the database could be used as additional breach identification and possible evidence in forensic investigation. The same logic applies to the files used as honeytokens. Honeytoken implementation is not enough in order to have in depth view on the attack process. In order to track the attacker during and after the attack SIEM (Security Information and Event Management) or centralized log management system had to be used. This system will collect all the related logs (application, web server, operating system, database, etc.).

Conceptual idea of honeytoken based protection inside web application, DBMS (Database Management System) and file on the file system is presented in Figure 1.

Zone 0 is related to firewall and IDS/IPS protection and it is not covered in this paper. This is usually the first line of defense that could be composed of the packet filtering firewall and/or WAF (Web application firewall). This is the zone that could be used to identify classical network attacks (IDS) and prevent them (IPS). If the WAF is being used, it is possible to identify and prevent most of the web application attacks such as SQLi (SQL injection), XSS (Cross Site Scripting), directory traversal etc. OWASP top 10 is ideal place to look for additional information about possible and mostly exploited web application vulnerabilities.

Zone 2 is the zone where web application based honeytokens [3] should be implemented. In this zone it is possible to identify and automatically react against any web application based attack.

Zone 1 and 3 are used to implement file system and database [4] based honeytokens. This is used in case zone 0 and zone 2 are compromised. If the zone 2 has been penetrated, it means that the web application is compromised and collecting as much data about the attacker as possible is mandatory in order to be able to conduct successful forensic investigation.

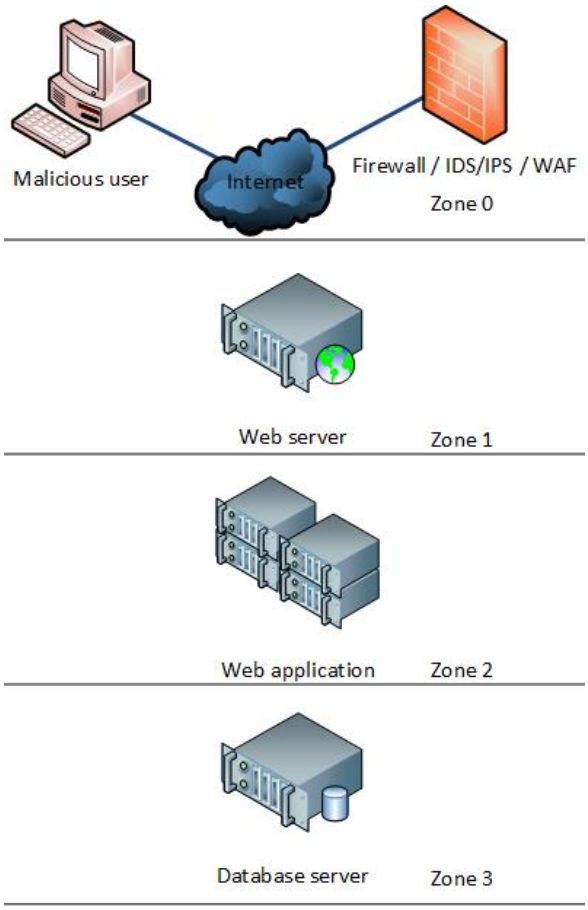


Figure 1: Conceptual honeytoken protection by layers

Logging and centralized log collection should be used in order to correlate the data, identify the attacker, to learn, to support computer forensic investigation, to be able to recognize attack vectors and trends, etc.

Let us define simple honeytoken parameter inside web application. For example, the following URL has one parameter called ID:

https://www.example.com/Site.aspx?ID=1

This link will return the data related to user with ID 1. In order to implement honeytoken value in the previously shown link, we should add additional parameter, and now the example could look like this:

https://www.example.com/Site.aspx?ID=1&Admin=false

Parameter "Admin" is honeytoken, with the value defined as "false". If the value is changed to "true" (or anything else) honeytoken event should be triggered inside the web application that will actively fight against the attacker by disconnecting user session, or allowing further access by placing the user session in the sandbox mode that will partially or fully change the results being sent to the user. This could allow for data collection that might be useful during the forensic investigation and it could be used to collect the crucial evidence that could be used to catch the perpetrator. It can also be used to change the firewall, IDS/IPS, and/or WAF settings in a fully automated manner.

III. HONEYTOKEN IMPLEMENTATION OPTIONS

There are many tools available for web application security scanning [5], and most of those tools (if not all) have spider functionality that is being used as the starting point in application parameter identification. Spider results are later used for vulnerability scanning and identification, by executing known exploits and vulnerability tests against every identified parameter in the application. As an example, SQLi , XSS, parameter tampering and modification, to mention only the few of the possible tests. Some of the tools have additionally modified spider that allows for so called forced browsing that will allow spider to identify existing files and folders even if they are not linked anywhere on any of the existing web pages. By using this functionality it is possible to identify application entry points usually hidden from the end user and one could easily go around the protection of this kind. One of possible options to prevent modified spider is to identify application entry points by using above mentioned functionality is to modify the application logic to response with 200 OK HTTP answer no matter if the web page exists or not. This will slow down the spidering process and it will prolong the process and reliability, thus making it both unreliable and incorrect. If the spidering process is unreliable and incorrect, the whole vulnerability assessment procedure is unreliable, thus it is unusable. As already mentioned, after spider process has identified all the entry points, the tool will try to find all the well-known vulnerabilities and/or ways to breach the system. This portion of the tool's logic could also be tampered

with in a way the results are inconclusive and incorrect, by modifying the application to respond with previously prefabricated answers when the application logic recognize the usage of spider. The proposed methods that could be used in order to actively fight against automated vulnerability scanning tools and hackers are:

Using honeytoken (canary [5]) parameters is one of the easiest way to implement above mentioned logic inside any web application with little or no effort. Developer has to implement parameters with sound and interesting names (in order to fool the hacker also), and those parameters are not supposed to be changed by the client. Should any of the honeytoken parameters be changed, the application logic has to assume that the parameter was modified by malicious user or scanning tool. If this is the case, application will enter highly secure mode that will actively fight against the attacker, no matter if the attacker is automated tool or hacker. The options application developer can use in order to implement this functionality are the following:

I. **Session termination.** This approach has its own advantages and disadvantages. The advantage is that the attacker session will be terminated and it will have to reestablish session again, and disadvantage is that by analyzing the packets that forced the system to terminate the session the attacker could easily identify honeytoken parameters and ignore them during the rest of the scanning or attacking procedures.

II. **Session isolation with 200 OK answers for every possible request.** This method is fairly successful, and it could be easily identified by the attacker if the answer for nonexistent pages is always the same, and it could be much more resilient to detection if the answers are tailored and modified with every single response. The problem is that when the hacker or the tool recognize this pattern it will realize that it is being fooled and in the long run this is just a partial solution to the problem.

III. **Session isolation with 200 OK answers for randomly chosen client requests.** This method is almost the same as the previous one, but in log run it could be more effective, because it will be much more difficult to identify that the server side is fabricating false web pages. Especially if the fabricated results are from prepopulated dictionary that make sense, So it will never respond with the 200 OK for request like this one: http://domain/This_app_does_not_exist_no_way.aspx, and it will respond to the request like this one: http://domain/AdminPage.aspx.

IV. **Session continuation in isolated (sandbox) mode.** This is the option when application does not offer real functionality anymore and attacker will receive preconfigured or dynamically generated fake pages. This could prove to be the most effective way but the problem is that it requires large amount of prefabricated fake results that could prove to be hard to implement.

V. **Throttling the connection between the identified attacker and the server.** Maybe the easiest protection to implement. The attacker will be slowed down to the point when the application will be unusable, thus, it could prove to be the easier and most effective protection that could be implemented. The consequence is: the attacker will need much more time than usual to finish the job and the administrators on the other side will have more time to identify and prevent the attacker.

VI. **Responding with timeouts.** This option is special case of previously explained throttling method, and it could be used as an add-on to this method or as a standalone protection, that will sporadically send timeouts to the hacker, preventing him from finishing the scan, attack, or whatever he is trying to accomplish.

Storing honeytoken values inside databases is method that is being used in order to identify web application protection method breach. Should this happen the decision has to be made in the application whether to terminate the session and prevent further data leakage, or continue the session in order to be able to identify who the attacker is. In order to use this type of honeytokens, the system has to be equipped with the SIEM or centralized log management system that will be utilized in the detection and alarming process.

Storing honeytoken values on the file system. This is the method similar to the one explained before, with the only difference being that the honeytoken is the file on the file system instead a value inside the database table.

Using intentionally generated vulnerabilities. This method is using prefabricated vulnerabilities that will show as the real vulnerabilities (SQLi, XSS …), but they will be isolated and contained. For example, the attacker and/or automated tool will identify this fake vulnerability, and it will try to exploit it further. This could be used to "keep the attacker happy" by giving him fake vulnerabilities, and it could give the system administrator enough time to isolate the attacker and contain his traffic.

Robots.txt entries could also be used as honeytokens. Robots.txt is file used in the root of the website that will inform spiders what should and what should not be indexed. By storing honeytokens (links to the nonexistent web pages) in robots.txt, one could easily isolate misbehaved spiders should the honeytoken link marked in robots.txt as "do not index" be touched by the spider. In this case, the honeytoken is the link itself, because well behaved spiders will obey the entries in the robots.txt, while misbehaved will not.

Active attacks against the attacker. If the attacked web application owner wants to actively fight against the attacker and attack the attacker, this is the option that could be used in order to identify the attacker's real IP address (if the address is masked or hidden by using proxy or proxy chaining software). In order to attack the attacker, web site owner could use XSS vulnerability that will trigger the request from attacker's computer via

direct communication channel, thus circumventing the protection the attacker is using in order to prevent the IP address identification. Of course, the same method could be used to steal the attacker's cookies, hijack his mail account etc.

Above mentioned methods should actively prevent an attack that is utilizing automated tools and/or being additionally watched by malicious users. On a first sight it looks that the database and file system honeytokens are not needed because everything should be stopped on the application layer but the point is that the application layer could always have a vulnerability, even the protection itself, so one have to implement as much of methods as possible in order to be as secure as possible. All the methods provide the administrator, developer, security personnel and others with comprehensive and complete insight into attack itself, methods and techniques being used and it allows for active countermeasures implementation that could be automatically embedded into application or that could be manually started after the attack has been identified.

IV. LOGGING AND REPORTING

In order to be able to identify and prevent the attacker from further attacks it is imperative to have active and correctly configured centralized log management system[7]. This system could be SIEM or "normal" centralized log management with custom made dashboards that will graphically present the active attacks, types of the attacks, attacker's IP addresses, etc. Some additional functionality could be added, like geolocation, routing information, amount of traffic generated, etc. There are many free and commercial tools that could provide some of this functionality out of the box, but usually none of the tool will be able to provide all of the required information without fine-tuning and additional modifications. The most important portion is that all the logs of interest have to be collected and visualized with the tool. Just to mention some of the logs of interest: Operating system logs, Web server logs, Web application logs, database logs, etc.

V. IMPLEMENTATION EXAMPLE

The excerpts of code written as a PoC (Proof of Concept) and as an example how a developer could implement above mentioned protection is written in ASP.NET. Basic ASP.NET knowledge is required to understand this code, though on the conceptual level familiarity with any programing language should be enough. The code presented here is not self-contained protection, it's just an example of where and how the protection could be implemented. In order to track the user sessions the PoC code is using Global.asax file and session_start and session_end functions. "Global.asax file is an optional file that contains code for responding to application-level events raised by ASP.NET or by Http modules" [6]. When new session starts it will be added to a session pool and this is the way of tracking the unique users. When session ends it will be deleted from the sessions pool. Example code of session start follows:

```
protected void Session_Start(object sender,
EventArgs e)
        {
            if
(!Sessions.ContainsKey(HttpContext.Current.Session.S
essionID))
            {

Sessions.Add(HttpContext.Current.Session.SessionID,
new SessionInfo());
                }
```

SessionInfo() contains all relevant and interesting information related to the session, like, whether this user is identified as an attacker, user IP address, UserAgent, Hacking type recognized etc. Anything could be added to this portion of code. This code will create a session, or identify the request with existing session making the user tracking possible. The example function follows:

```
    public SessionInfo()
        {
            //defaults
            IsHacker = false;
            SessionCreated = DateTime.UtcNow;
            LastRequest = DateTime.UtcNow;
            IP = "";
            UserAgent = "";
            RawURL = "";
            HackingType = "Peace"; // Peace - no
attack, CPT - canary Parameter tampering, COSFS - OS
canary tampering, DB canary tampering, SQLI - SQLi,
XSS - XSS//
        }
        public bool IsHacker { get; set; }
        public string Username { get; set; }
        public DateTime SessionCreated { get; set; }
        public DateTime LastRequest { get; set; }
        public string IP { get; set; }
        public string UserAgent { get; set; }
         public string RawURL { get; set; }
        public string HackingType { get; set; }

      }
```

This function is populating variables with the data relevant to the session, like user data, IP address etc.

Function that might be used to recognize hacker by using parameter based honeytoken could look like this one:

```
protected void Page_Load(object sender, EventArgs e)
        {
            if (Request.Params.Keys[1] == "admin")
            {
                if (Request.Params[1] != "false")
                {
Global.Sessions[Session.SessionID].IsHacker = true;


Helpers.MyFunctions.WriteInTheLog(Global.Sessions[Se
ssion.SessionID].Username, Session.SessionID,
Global.Sessions[Session.SessionID].IsHacker,
Global.Sessions[Session.SessionID].IP,
Global.Sessions[Session.SessionID].UserAgent,
Global.Sessions[Session.SessionID].RawURL,
Global.Sessions[Session.SessionID].HackingType);
                    Panel2.Visible = true;
                    Panel1.Visible = false;
                }
              }
```

So, if the parameter admin was tempered with, the user session is identified as malicious one and it will be dealt with appropriately. In this particular example it will only

write the data in the log for later audit, or for direct audit by SIEM, or centralized log management system solution. It could also break the session, isolate the user in specially written sandbox, etc.

An example of function that will write data to the log file could look like this one:

```
public static void WriteInTheLog(string user, string
sessionID, bool Hacker, string IP, string UserAgent,
string RawURL, string HackingType)
        {
            StreamWriter sw = new
StreamWriter("log.txt", true);

sw.WriteLine(DateTime.Now.ToShortDateString().ToStri
ng() + " " +
DateTime.Now.ToLongTimeString().ToString() + " " +
user + " " + sessionID + " " + Hacker + "  " + IP +
" " + UserAgent + " " + RawURL + " " + HackingType);
            sw.Flush();
            sw.Close();
            return;
          }
```

It is not possible to provide the entire code listing in this paper, but the point is that by using few simple functions like the ones shown above, one can start to implement proposed protections.

VI. Conclusion

The future of Internet security is in the hands of developers and security researchers that will add on the security in general, thus raising the bar or preventing the malicious users from compromising the remote systems. It has to be clear to every party included in the web application development that web application will always be vulnerable to some kind of vulnerability, and that the implemented protection could have vulnerabilities of their own. Knowing that should be enough to write the web applications in more responsible way that will make them less exposed to the attackers.

Proposed honeytoken implementation could be used as an additional protection against the above mentioned threats together with firewalls, web application firewalls and various filtering platforms, raising the bar and making it more resilient to intrusions. By using the proposed solution, one could identify, contain and deflect the automated tools and hackers more efficiently than by using the traditional approach only (firewall/IDS/WAF). This is because firewall will not be able to prevent parameter tampering and to recognize the important parts of the application one wants to protect. The same issue is with IDS/IPS, which has to be reconfigured with every new version of application in order to be able to identify what's important. With WAF, which is capable of recognizing variety tampering attacks, there is always an issue of going around the WAF protection itself. The proposed solution will implement the honeytokens inside the application logic, thus making the application resilient no matter if the other protections are implemented or not. Of course, this does not mean that all the other protections are not necessary anymore, it is additional protection with one simple task – to make the web application more resilient to the hacking attempts.

References

[1] The Open Web Application Security Project (12 June 2013), "OWASP Top 10", https://www.owasp.org/index.php/Top10#OWASP_Top_10_for_2013. Retrieved 16 January 2015.

[2] A. J. Hacker, "Importance of Web Application Firewall technology for Protecting Web-based Resources," ICSA Labs an Independent Verizon Business, January 2008. A. J. Hacker, "Importance of Web Application Firewall technology for Protecting Web-based Resources," ICSA Labs an Independent Verizon Business, January 2008.

[3] Craig M. McRaea, Robert Wesley McGrewa & Rayford B. Vaughna (24 February 2007). "Honey Tokens and Web Bugs: Developing Reactive Techniques for Investigating Phishing Scams". Journal of Digital Forensic Practice. Retrived 10 January 2015.

[4] Penny Ross, Amanda Peart (18 July 2013). "The Use of Honeytokens in Database Security". Atiner's conference paper series COM2013-0428. Retrieved 16 January 2015.

[5] E. Fong and V. Okun, "Web Application Scanners: Definitions and Functions," in Proc. the 40th Hawaii International Conference on System Sciences, IEEE, 2007, pp. 280b-280b

[6] Microsoft, "Global.asax file", https://msdn.microsoft.com/en-us/library/1xaas8a2(v=vs.71).aspx. Retrieved 30 March 2015.

[7] R. Meyer, "Detecting Attacks on Web Applications from Log Files," SANS Institute, InfoSec Reading Room, January 2008

POLICY AND RISK TRACK

Artificial Intelligence in Network Intrusion Detection

M. Stampar* and K. Fertalj**
* Information Systems Security Bureau, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
mstampar@zsis.hr, kresimir.fertalj@fer.hr

Abstract – In past, detection of network attacks has been almost solely done by human operators. They anticipated network anomalies in front of consoles, where based on their expert knowledge applied necessary security measures. With the exponential growth of network bandwidth, this task slowly demanded substantial improvements in both speed and accuracy. One proposed way how to achieve this is the usage of artificial intelligence (AI), progressive and promising computer science branch, particularly one of its sub-fields – machine learning (ML) – where main idea is learning from data. In this paper authors will try to give a general overview of AI algorithms, with main focus on their usage for network intrusion detection.

I. INTRODUCTION

Today, lack of network access is unthinkable in both business and personal life. With unrestrained and ever-increasing number of electronic devices connected to the Internet, attackers never had such large attack surface. This brings up the question how to successfully defend from both known and unknown dangers. The answer is not so straightforward, especially having in mind that the number of threats is rising every year [1].

Valuable resources having network access should be permanently protected from all attempts to destroy, expose, alter, disable, steal or gain unauthorized access and/or usage. Resources' confidentiality, integrity and availability have to remain intact. Intrusion detection system (IDS) is a system specially designed to detect such malicious attempts.

As traditional IDS's are mainly signature-based, detecting only known attacks, their biggest problem is the inability to detect new or variant attacks [2]. One topic that intuitively stands out as a potential solution for solving this problem is ML, a type of AI that provides computer the ability to learn without being explicitly programmed for particular purpose.

The rest of this paper is organized as follows. In Section 2, we discuss the basic concepts of AI and ML. In Section 3, we discuss the basic concepts of intrusion detection. In Section 4, we thoroughly discuss the usage of AI for network intrusion detection, including the brief overview of most popular AI algorithms and mutual comparisons found in literature. In Section 5, we present results for conducted research of topic related literature trends, while in Section 6 we offer concluding remarks.

II. ARTIFICIAL INTELLIGENCE

AI is by definition the study of mental faculties through the use of computational models [3]. Main fields of AI research include reasoning, knowledge representation, automated planning and scheduling, ML, natural language processing, computer vision, robotics and general intelligence (strong AI). This paper is mainly focused on ML as it seems to be the most promising AI sub-field for intrusion detection task. Nevertheless, other non-ML methods will be mentioned too.

ML deals with the construction and study of algorithms that can generalize (i.e. learn) from limited sets of data. Such algorithms operate by building models based on input and using those models to make predictions or decisions, rather than following only explicitly programmed instructions. Having such characteristics makes them ideal candidates for intrusion detection tasks.

There are three common problems that ML tries to solve: classification, regression and clustering. Classification means identifying group membership, regression involves estimating or predicting a response, while in clustering a set of inputs is divided into groups, where members have similar characteristics. Depending on output, if the output variable takes class labels, we are talking about classification, if the output variable takes continuous numerical values, we are talking about regression, while if the output values are subsets (i.e. *clusters*) we are talking about clustering. Thus regression is applicable for prediction type of problems as opposed to classification and clustering.

Depending on ways of learning, ML can be further split into three main categories: supervised, unsupervised and reinforcement learning. Supervised learning is commonly used in classification problems where the goal is to get the computer to learn a classification system that we have created. Unsupervised learning seems much harder. The goal is to have the computer learn how to do something that we do not explicitly tell it how to do. It is a powerful tool for identifying structure in unlabeled data, reflecting the statistical properties of the overall collection of input patterns. Reinforcement learning is performed by interacting with an environment, where learning agent learns from the consequences of its actions through trial and error, rather than from being explicitly taught.

There is no single AI algorithm achieving best accuracy for all situations. Hence, one way to improve

results is the usage of combined multiple algorithms to obtain better quality of reasoning than from any single one. There are generally two approaches: ensemble and hybrid. In case of ensemble classifiers, multiple but homogeneous weak models are combined, typically at the level of their individual output, using various merging algorithms (e.g. majority voting) [4]. Hybrid algorithms, in turn, combine completely different, heterogeneous AI approaches (e.g. through cascading).

III. INTRUSION DETECTION

Intrusion detection can be defined as the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource [5]. IDS is a device or software application that gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, including data for both intrusion (i.e. attacks from outside the organization) and misuse (i.e. attacks from within the organization). IDS come in a variety of "flavors" and approach the goal of detecting suspicious behavior in different ways.

Based on a running location, IDS can be network-based (NIDS), inspecting network events at single collecting point, or host-based (HIDS), inspecting system events at individual device(s). NIDS is used to monitor and analyze network traffic, collected using a network tap, span port or hub, to protect a system from network-based threats. HIDS attempts to identify unauthorized, illicit, and anomalous behavior on a specific device. HIDS generally involves an agent installed on each system, monitoring and remotely alerting on local OS and application activity. In this paper we will mainly focus on NIDS systems.

Based on an underlying detection logic, IDS can be signature-based (misuse) or anomaly-based. A signature-based IDS monitors suspicious behavior and compare its characteristics against a database of signatures or attributes from known malicious threats. Main issue is the lag between a new threat and the signature for detecting that threat being applied to the IDS. During that lag period IDS is unable to detect the threat. An anomaly-based IDS monitors behavior and compares its characteristics against an established baseline. The baseline will identify what is "normal" for that subject and alert when anomalous behavior is detected, or significantly different than the baseline. Main issue is the higher false positive rate.

Based on how it reacts to a detected threat, IDS can be passive or reactive. Passive IDS simply detects and alerts, leaving decision to administrator to take action and block the activity or respond in some way. Reactive IDS, also known as intrusion prevention system (IPS), will take predefined proactive actions to respond to the threat. Typically, this means blocking any further network traffic from the source IP address or user.

IV. USING ARTIFICIAL INTELLIGENCE FOR NETWORK INTRUSION DETECTION

A. Datasets

The Cyber Systems and Technology Group (formerly the DARPA Intrusion Detection Evaluation Group) of MIT Lincoln Laboratory, under Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL) sponsorship, has collected and distributed the first standard corpora for evaluation of NIDS's [6]. Such efforts have been carried out in both 1998 and 1999, resulting with DARPA 1998 and DARPA 1999 datasets respectively, with DARPA 1998 being more popular one.

KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge Discovery and Data Mining. The task for the classifier learning contest organized in conjunction with the KDD'99 conference was to learn a predictive model (i.e. classifier) capable of distinguishing between legitimate and illegitimate connections in a computer network [7]. Dataset used during the competition was based on a DARPA 1998 dataset and has become widely accepted as just KDD'99.

Both DARPA and KDD datasets consist of nearly 5 million training connection records (i.e. events) labeled as an intrusion or not an intrusion, and separate testing dataset consisting of seen and unseen attacks. Each connection record consists of 41 different attributes that describe the different features of the corresponding connection, categorized as follows: basic TCP features, content features, time-based traffic features and host-based features. All the forms of attack fell into one of three categories: Remote-to-Local (R2L), User-to-Root (U2R), Denial-of-Service (DOS) or Probing [8].

As already mentioned, KDD'99 is based on the DARPA 1998, which by itself has been strongly criticized by McHugh [9], mainly because of having the characteristics of the synthetic data. As result, some of the existing problems in DARPA 1998 remain in KDD'99 too.

Currently, as there are only few public datasets like DARPA 1998 and KDD'99, much related work considers these for their experiments, while very few studies use non-public or their own datasets. From literature overview results by Tsai et al. [10] it can be concluded that those two datasets are recognized as de-facto standards in intrusion detection field of study.

B. Evaluation

While evaluating IDS, for every possible test value there are two kinds of error: false positive (FP) and false negative (FN). FP occurs when an event is predicted as intrusive but it is in fact normal, while FN occurs when a truly intrusive event occurs without being recognized as one. In the other hand, true positive (TP) measures the proportion of actual positives which are correctly identified as such, while true negative (TN) measures the proportion of negatives which are correctly identified as such. The performance of each classifier can be quantified using the detection rate (DR) and overall accuracy (OA) measures. DR shows the percentage of the true intrusions that have been successfully detected:

$$DR = \frac{TP}{TP + FN} \times 100\% \quad (1)$$

OA is calculated as the total number of correctly classified intrusions divided by the total number of observations:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \times 100\,\% \tag{2}$$

Efficient IDS necessitates high degree of DR and OA, while maintaining low false alarm rates. Accuracy is critical to develop effective IDS, since high FP rate or low DR will make it practically unusable.

C. Neural network

When used without qualification, the terms "neural network" (NN) and "artificial neural network" (ANN) usually refer to a multilayer perceptron (MLP). MLP is a supervised learning algorithm based on the feed-forward neural network with one or more layers between input and output layer. Feed-forward means that data flows in one direction from input to the output layer (i.e. forward). This type of network is trained with the error back-propagation learning algorithm. The true power and advantage of MLP lies in its ability to represent both linear and non-linear relationships and in its ability to learn these relationships directly from the data being modeled. Traditional linear models are simply inadequate when it comes to modeling data that contains non-linear characteristics [11].

Ghosh et al. [12] used MLP for anomaly detection, where the proposed model is a single hidden layer neural network. The performance of this model tested on the DARPA 1998 dataset was a DR of 77% with 2.2% FP. Lippmann et al. [13] used selected generic keywords to detect the attack preparations and actions after the break-in. Results were used to train the MLP using back-propagation algorithm. This approach ensured a DR of 80% when it has been tested on the DARPA 1998 dataset. Debar et al. [14] stated that NNs are very slow in converging, they suffer from dimensioning and stability problems, training takes a lot of time to achieve a reasonable level of performance, and their adaptability is low because a partial retraining can lead to a network that forgets everything it has learned before.

D. Genetic algorithm

Genetic algorithm (GA) is a search algorithm that works similar to the process of natural selection. It begins with a sample set of potential solutions which then evolves toward a set of better solutions. The algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm randomly selects individuals from the current population and uses them as parents to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution. Within the sample set, solutions that are poor tend to die out while better solutions mate and propagate their advantageous traits, thus introducing more solutions into the set that boast greater potential (the total set size remains constant; for each new solution added, an old one is removed). A little random mutation helps guarantee that a set won't stagnate and simply fill up with numerous copies of the same solution.

Li [15] proposed IDS with 57 genes in chromosomes, where each gene represents single connection feature, like: source IP address, destination IP address, source port, destination port, duration, protocol, number of bytes sent by originator, number of bytes sent by responder, etc. Due to the effectiveness of the evaluation function, the succeeding populations are biased toward rules that match intrusive connection. Mukkamala et al. [16] achieved 97.04% OA on DARPA 1998 dataset for their implementation of IDS based on linear genetic programming (LGP). However, Teodoro et al. [17] concluded that main disadvantage of this kind of IDS is the high resource consumption involved.

E. Fuzzy logic

Fuzzy logic (FL) is a form of many-valued logic where one deals with reasoning that is approximate rather than fixed and exact. Compared to traditional binary sets where variables may take on true or false values, FL variables may have a truth value that ranges in degree between 0 and 1. FL proponents claim this generality allows greater flexibility, freedom, accuracy and compactness when representing real world situations. All the usual properties of Boolean algebra can be extended to FL, and probability's degree of belief in a Boolean variable becomes a fuzzy variable's degree of truth. It provides a principled way to encode expert knowledge or heuristics into algorithms that mimic a human's approach to solving challenging problems by weighing different sources of information and making judgments based on a preponderance of the evidence [18].

Chimphlee et al. [19] proposed IDS based on rough sets theory and fuzzy c-means. Proposed system achieved a total of 93.45% OA on KDD'99 dataset, while reduced number of features resulted with enhanced performance. Dickerson et al. [20] developed FIRE IDS which uses simple data mining techniques to process the network input data and expose metrics that are particularly significant to anomaly detection. These metrics are then evaluated as fuzzy set for every observed feature and later used to detect network attacks.

F. K-nearest neighbors

K-nearest neighbors (kNN) is one of the most simple and traditional non-parametric supervised learning algorithms. It stores all available cases and classifies new ones based on a similarity measure (e.g. distance function) calculated between feature vectors. Classification is made by a majority vote of object's k nearest neighbors, where k is typically chosen to be a small positive integer (e.g. 3). For example, if k is chosen to have a value 1, then object is assigned to the class of its nearest neighbor. It has to be noted that if k is chosen to be considerably large, it will result with greater classification time.

Ma et al. [21] proposed kNN IDS based on similarity as a quantitative measure for distance and achieved DR of 90.28% (k=1) on KDD'99 dataset. Liao et al. [22] achieved DR of 91.70% (k=5) on DARPA 1998 dataset with their implementation of IDS based on kNN. They

also concluded that kNN works well in dynamic environments where frequent updates of the training data are required, which makes it attractive for intrusion detection tasks.

G. Naive Bayes

Naive Bayes (NB) is a supervised learning algorithm based on Bayesian theorem with the "naive" assumption of independence between every pair of features. Classification is made by combining prior probabilities and likelihood, to form a posterior probability using the so-called Bayes' rule. Despite its simplicity, NB can often outperform more sophisticated classification algorithms in both speed and accuracy [23]. Also, they are particularly suited when dimensionality of the input is high [24].

Panda et al. [25] proposed NB based IDS and achieved OA of 94.90% on KDD'99 dataset. Amor et al. [26] made a similar proposition and achieved 91.52% OA.

H. Decision tree

Decision tree (DT) is a supervised learning algorithm based on flowchart-like structure in which internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf node represents a class label. Decision is taken after computing all attributes. The paths from root to leaf represent classification rules. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features (i.e. attribute). DT is simple to understand and to visualize, making its explanation for the resulting value easily explainable by Boolean logic.

Abbes et al. [27] compared DR of rule-based Snort and IDS based on DT on a custom dataset based on collected RPC protocol traffic. Snort detected only three, while proposed IDS detected all 46 different forms of attacks. Bouzida et al. [28] made comparison of DT with and without principal component analysis (PCA), a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. They reduced computation time on dataset KDD'99 by a factor of approximately thirty, with slight loss of OA from 92.60% to 92.05%.

I. Support vector machine

Support vector machine (SVM) is a supervised learning algorithm based on concept of decision planes that define decision boundaries. An SVM constructs the hyperplane, or a set of hyperplanes in a high-dimensional space, that separates all data points of one class from those of the other class. The best hyperplane for an SVM is the one with the largest margin between the two classes. It often happens that classes can't be linearly separated. For this reason, the original finite-dimensional space can be mapped into a much higher-dimensional space, using what is called the "kernel trick", presumably making the separation easier [29].

Laskov et al. [30] concluded that the best performance is achieved by the non-linear methods, such as SVM, MLP and rule-based methods. Khan et al. [31] presented results where their solution, utilizing dynamically growing self-organizing tree (DGSOT), enhanced the training time of SVM on DARPA 1998 dataset from 17.34h to 13.18h and improved OA from 57.6% to 69.8%. Chen et al. [32] achieved OA of 86.79% on KDD'99 dataset by using rough set theory (RST) to preprocess data and reduce the number of dimensions before forwarding it to SVM for intrusion detection. Mulay et al. [33] concluded that the integration of DT and SVM model gives better results than the individual models.

J. Random forest

Random forest (RF) is a supervised learning algorithm that is based on a collection (ensemble) of tree predictors, rather than a single classification tree, where to grow each tree a random selection is made from the examples in the training set. Each tree gives a classification and we say that the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest), with basic principle that a group of "weak learners" can come together to form a "strong learner".

Kim et al. [34] concluded that their approach based on RF has been able not only to show high DR, but also to figure out stable output of important features simultaneously. Also, they stated that performance of RF based IDS turns out to be comparable to that of SVM. Zhang et al. [35] achieved average DR of 92.58% on original KDD'99 dataset and 99.86% on balanced dataset, where minority classes have been over-sampled and majority classes down-sampled to increase the DR of minority intrusions.

K. Self-organizing map

Kohonen's self-organizing map (SOM) is an unsupervised learning algorithm producing a low-dimensional (typically two-dimensional), discretized representation of the training samples' input space, simply called a map. It consists of components called nodes or neurons. Each node is associated with a weight vector of the same dimension as the input data vectors and a position in the map space. Nodes in a 2-D layer learn to represent different regions of the input space where input vectors occur. In addition, neighboring neurons learn to respond to similar inputs, thus the layer learns the topology of the presented input space. They provide a way of representing multidimensional data in much lower dimensional spaces, where this process of reducing the dimensionality of vectors is essentially a data compression technique known as the vector quantization [36].

Srinivasan et al. [37] proposed an agent-based SOM IDS called SAPID, designed especially for wireless networks. They achieved an average DR of 97% on an undeclared dataset. Lichodzijewski et al. [38] achieved DR of 94% on DARPA 1998 dataset with conclusion that hierarchically built unsupervised neural network approach is able to produce encouraging results. Rhodes et al. [39] analyzed SOM's potential for intrusion detection. Their experiments shown that even single SOM, when trained on normal data, will detect anomalous behaviors. Also, the ratio by which normal and intrusive packets differed has been greater than an order of magnitude. Their conclusion was that IDS based on SOM should be particularly

powerful because it never needs to be told what intrusive behavior looks like.

L. *Comparisons*

Li et al. [40] compared SVM, NN and kNN based IDS, where SVM had best, while kNN had worst OA on KDD'99 dataset. Chen et al. [41] compared SVM and NN based IDS and concluded that SVM performance is superior to that of NN. Bouzida et al. [42] concluded that while NN is very interesting for generalization and poor for detection of new attacks, DT has proven efficiency in both generalization and new attacks detection.

Amor et al. [43] conducted an empirical investigation on the KDD'99 dataset, comparing NB and DT based IDS's. IDS based on DT obtained a higher OA, while IDS based on NB obtained better DR. Peddabachigari et al. [44] conducted similar study for SVM and DT. IDS based on DT gave better results for both OA and performance.

Wang et al. [45] compared DT, NB and NN based IDS. While OA has been similar for all three cases, with DT having slightly better results, training time of NN based IDS has been by far the worst (couple of orders of magnitude). Chen et al. [46] compared SVM and RF based IDS and concluded that IDS based on RF has better both OA and DR, with slightly worse training time and by far better performance From results presented by Sindhu et al. [47] it can be concluded that NB and DT (having similar values) outperform RF in both DR and OA.

V. LITERATURE TRENDS

For literature trends overview purposes quantitative analysis was done in form of scientific publication search for topic related terms. Each chosen term was inquired in combination with term "intrusion detection" to narrow down results to only the respective field. Google Scholar was used to perform searches through scientific literature, mainly because of its simplicity and openness. Total number of publications (excluding citations and patents) was recorded depending on different search objectives. Main goal of presented results is the outline of current trends provided in a comprehensible format.

In Fig. 1 trends for terms "artificial intelligence" and "machine learning" are shown for period 2010-2014 in form of a number of intrusion detection related publications for each particular year. From given results it is visible that the usage of AI, and ML as its most prominent sub-field, has an increasing tendency. Greater values for ML can be explained by lack of references to

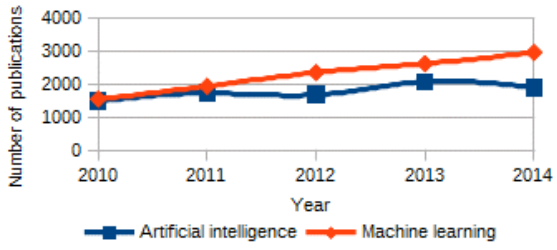


Figure 1. Intrusion detection trends for AI and ML

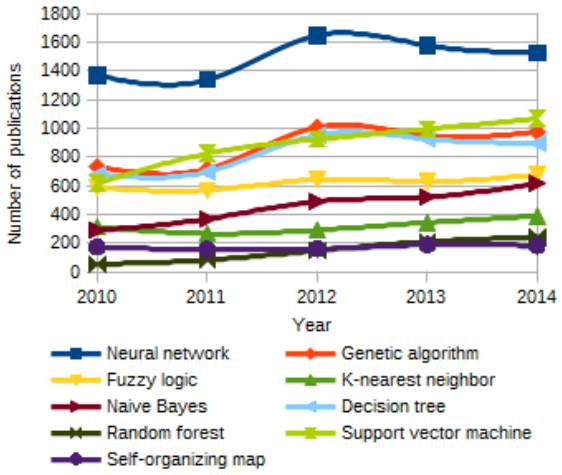


Figure 2. Intrusion detection trends for different AI algorithms

AI in related studies.

In Fig. 2 trends for different AI algorithms are shown for period 2010-2014. It is clear that NN is the most popular AI algorithm used for intrusion detection compared to other algorithms, but with slight decrease in popularity over the last couple of years. Then there are GA and SVM, where SVM is constantly increasing in popularity. Out of the rest, SOM seems to be the least popular, which could be explained by the lack of an intuitive approach to solution of intrusion detection problem. It is clear from observed trends that AI composes a significant part of the intrusion detection study, becoming more and more popular tool of choice.

VI. CONCLUSION

As faster and more effective countermeasures are required to cope with the ever-growing number of network attacks, AI comes as a natural solution to the problem. This paper, albeit briefly, discusses the foundations of AI in intrusion detection field. Most popular AI algorithms are presented along with the review of related studies, showing the maturity of the subject.

From comparisons found in related scientific literature it can be concluded that IDS's based on NB and DT seem to be the most promising, having best overall results for both accuracy and performance when compared to other single-classifier solutions. From presented results for conducted quantitative analysis of scientific literature it can be concluded that AI plays a significant role in the intrusion detection field of study, with NN being most extensively researched.

Further research in this field is necessary as there are very promising results that are obtainable from related algorithms, especially in the area of their combined usage.

REFERENCES

[1] "Application Vulnerability Trends Report: 2014," Cenzic Inc., Campbell, CA, 2014.

[2] L. Wenke, S. J. Stolfo, P. K. Chan, E. Eskin, W. Fan, M. Miller, S. Hershkop, and J. Zhang, "Real time data mining-based intrusion detection," in *Proc. DISCEX'01*, vol. 1, pp. 89-100, IEEE, 2001.

[3] E. Charniak, and D. McDermott, "Introduction to Artificial Intelligence", Addison-Wesley, 1985.

[4] P. Kazienko, E. Lughofer, and B. Trawiński, "Special Issue on Hybrid and Ensemble Methods in Machine Learning," *New Generation Computing 29*, no. 3, pp. 241-244, 2011.

[5] SANS Institute. (2015) Intrusion Detection FAQ. [Online]. Available: https://www.sans.org/security-resources/idfaq/

[6] (2014) DARPA Intrusion Detection Data Sets webpage on The MIT Lincoln Laboratory website. [Online]. Available: http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/

[7] (2014) KDD Cup 1999: Computer network intrusion detection webpage on SIGKDD website. [Online]. Available: http://www.sigkdd.org/kdd-cup-1999-computer-network-intrusion-detection

[8] W. Lee, S. J. Stolfo, and K. W. Mok, "Mining in a data-flow environment: Experience in network intrusion detection," in *Proc. KDD'99*, pp. 114-124, ACM, 1999.

[9] J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory," *ACM transactions on Information and system Security* 3, no. 4, pp. 262-294, 2000.

[10] C. F. Tsai, Y. F. Hsu, C. Y. Lin, and W. Y. Lin, "Intrusion detection by machine learning: A review," *Expert Systems with Applications* 36, no. 10, pp. 11994-12000, 2009.

[11] C. Grosan, and A. Abraham, *Intelligent systems: A modern approach*, vol. 17, Springer, 2011.

[12] A. K. Ghosh, and A. Schwartzbard, "A Study in Using Neural Networks for Anomaly and Misuse Detection," in *Proc. USENIX Security,* 1999.

[13] R. P. Lippmann, and R. K. Cunningham, "Improving intrusion detection performance using keyword selection and neural networks," *Computer Networks* 34, no. 4, pp. 597-603, 2000.

[14] H. Debar, M. Becker, and D. Siboni, "A neural network component for an intrusion detection system," in *Proc. Research in Security and Privacy,* pp. 240-250, IEEE, 1992.

[15] W. Li, "Using genetic algorithm for network intrusion detection," in *Proc. of the United States Department of Energy Cyber Security Group*, pp. 1-8, 2004.

[16] S. Mukkamala, A. H. Sung, and A. Abraham, "Modeling intrusion detection systems using linear genetic programming approach," *Innovations in Applied Artificial Intelligence*, pp. 633-642, Springer Berlin Heidelberg, 2004.

[17] P. G. Teodoro, J. D. Verdejo, G. M. Fernández, and E. Vázquez. "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & security* 28, no. 1, pp. 18-28, 2009.

[18] J. K. Williams, J. Craig, A. Cotter, and J. K. Wolff, "A hybrid machine learning and fuzzy logic approach to CIT diagnostic development," *AMS Fifth Conference on Artificial Intelligence Applications to Environmental Science*, 2007.

[19] W. Chimphlee, A. H. Abdullah, M. N. M. Sap, S. Srinoy, and S. Chimphlee, "Anomaly-based intrusion detection using fuzzy rough clustering," in *Proc. ICHIT'06*, pp. 329-334, IEEE, 2006.

[20] J. E. Dickerson, and J. A. Dickerson, "Fuzzy network profiling for intrusion detection," in *Proc. NAFIPS'00*, pp. 301-306, IEEE, 2000.

[21] Z. Ma, and A. Kaban, "K-Nearest-Neighbours with a novel similarity measure for intrusion detection," in *Proc. UKCI'13,* pp. 266-271, IEEE, 2013.

[22] Y. Liao, and V. R. Vemuri, "Use of k-nearest neighbor classifier for intrusion detection," *Computers & Security*, 21.5, pp: 439-448, 2002.

[23] H. Zhang, "The optimality of naive Bayes," *AA* 1, no. 2, 2004.

[24] S. Caballé, A. Abraham, T. Daradoumis, and A. A. J. Perez. *Computational intelligence for technology enhanced learning*, vol. 273, Springer, 2010.

[25] M. Panda, and M. R. Patra, "Network intrusion detection using naive bayes," *International journal of computer science and network security*, no. 12, pp. 258-263, 2007.

[26] N. B. Amor, S. Benferhat, and Z. Elouedi, "Naive bayesian networks in intrusion detection systems," in *Proc. PGM'03, Cavtat–Dubrovnik, Croatia,* 2003.

[27] T. Abbes, A. Bouhoula, and M. Rusinowitch, "Protocol analysis in intrusion detection using decision tree," in *Proc. ITCC'04,* vol. 1, pp. 404-408, IEEE, 2004.

[28] Y. Bouzida, F. Cuppens, N. C. Boulahia, and S. Gombault, "Efficient intrusion detection using principal component analysis," in *Proc. SAR'04, La Londe, France,* 2004.

[29] B. Scholkopf, "The kernel trick for distances," *Advances in neural information processing systems,* pp. 301-307, 2001.

[30] P. Laskov, P. Düssel, C. Schäfer, and K. Rieck, "Learning intrusion detection: supervised or unsupervised?," in *Proc. ICIAP'05*, pp. 50-57, Springer Berlin Heidelberg, 2005.

[31] L. Khan, M. Awad, and B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering," *The International Journal on Very Large Data Bases* 16, no. 4, pp. 507-521, 2007.

[32] R. C. Chen, K. F. Cheng, Y. H. Chen, and C. F. Hsieh, "Using rough set and support vector machine for network intrusion detection system," in *Proc. ACIIDS'09*, pp. 465-470, IEEE, 2009.

[33] S. A. Mulay, P. R. Devale, and G. V. Garje, "Intrusion detection system using support vector machine and decision tree," *International Journal of Computer Applications*, no. 3, pp. 40-43, 2010.

[34] D. S. Kim, S. M. Lee, and J. S. Park, "Building lightweight intrusion detection system based on random forest," in *Proc. ISNN'06*, pp. 224-230, Springer Berlin Heidelberg, 2006.

[35] J. Zhang, and M. Zulkernine, "Network Intrusion Detection using Random Forests," in *Proc. PST'05*, 2005.

[36] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, "Som pak: The self-organizing map program package," *Report A31, Helsinki University of Technology,* 1996.

[37] T. Srinivasan, V. Vijaykumar, and R. Chandrasekar, "A Self-organized Agent-based architecture for Power-aware Intrusion Detection in wireless ad-hoc networks," in *Proc. ICOCI'06*, pp. 1-6, IEEE, 2006.

[38] P. Lichodzijewski, A. N. Z. Heywood, and M. I. Heywood, "Dynamic intrusion detection using self-organizing maps," in *Proc. CITSS'02*, 2002.

[39] B. C. Rhodes, J. A. Mahaffey, and J. D. Cannady, "Multiple self-organizing maps for intrusion detection," in *Proc. NISSC'00*, pp. 16-19, 2000.

[40] Y. Li, and L. Guo, "An active learning based TCM-KNN algorithm for supervised network intrusion detection," *Computers & security* 26, no. 7, pp. 459-467, 2007.

[41] W. H. Chen, S. H. Hsu, and H. P. Shen, "Application of SVM and ANN for intrusion detection," *Computers & Operations Research* 32.10, pp. 2617-2634, 2005.

[42] Y. Bouzida, and F. Cuppens, "Neural networks vs. decision trees for intrusion detection," in *Proc. MonAM'06, Tuebingen, Germany*, pp. 28-29, 2006.

[43] N. B. Amor, S. Benferhat, and Z. Elouedi, "Naive bayes vs decision trees in intrusion detection systems," in *Proc. of the SAC'04*, ACM, 2004.

[44] S. Peddabachigari, A. Abraham, and J. Thomas, "Intrusion detection systems using decision trees and support vector machines," *International Journal of Applied Science and Computations, USA* 11.3, pp. 118-134, 2004.

[45] G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," *Expert Systems with Applications*, 37.9, pp. 6225-6232, 2010.

[46] Y. Chen, Y. Li, X. Q. Cheng, and L. Guo, "Survey and taxonomy of feature selection algorithms in intrusion detection system," *Information Security and Cryptology*, pp. 153-167, Springer Berlin Heidelberg, 2006.

[47] S. S. S. Sindhu, S. Geetha, and A. Kannan, "Decision tree based light weight intrusion detection using a wrapper approach," *Expert Systems with applications*, 39.1, pp. 129-141, 2012.

Advanced Persistent Threats – Detection and Defense

J. Vukalović, D. Delija
INsig2 Ltd., Buzinska cesta 58, 10010 Zagreb, Croatia
jakob.vukalovic@insig2.eu, damir.delija@insig2.eu

***Abstract* – The term "Advanced Persistent Threat" refers to a well-organized, malicious group of people who launch stealthy attacks against computer systems of specific targets, such as governments, companies or military. The attacks themselves are long-lasting, difficult to expose and often use very advanced hacking techniques. Since they are advanced in nature, prolonged and persistent, the organizations behind them have to possess a high level of knowledge, advanced tools and competent personnel to execute them. The attacks are usually preformed in several phases – reconnaissance, preparation, execution, gaining access, information gathering and connection maintenance. In each of the phases attacks can be detected with different probabilities. There are several ways to increase the level of security of an organization in order to counter these incidents. First and foremost, it is necessary to educate users and system administrators on different attack vectors and provide them with knowledge and protection so that the attacks are unsuccessful. Second, implement strict security policies. That includes access control and restrictions (to information or network), protecting information by encrypting it and installing latest security upgrades. Finally, it is possible to use software IDS tools to detect such anomalies (e.g. Snort, OSSEC, Sguil).**

I. INTRODUCTION

One of the main challenges in information security today is a new type of threat that has become a significant concern for various organizations. Organizations now face advanced groups of attackers who are determined to gain access to their resources. They are not just opportunistic criminals. Instead, they are well organized and possess enough resources to launch prolonged and sophisticated attacks. That is how they got the name – *Advanced Persistent Threat* (APT).

Over the last few years, there has been a significant increase in the amount of APT attacks. FireEye, a US based network security company detected 4,192 attacks associated with APT groups during 2013 [1]. That amounts to 11.48 unique daily attacks. FireEye also reports that APT groups caused 17,995 unique malware infections. Command and control servers that APT groups use to coordinate attacks have been found in 206 countries, most of them in the USA, Germany, South Korea and China. During 2013, they have produced over 22 million control messages. Most often, APT groups target organizations in the USA, South Korea and Canada. Most of these organizations belong to education, finance, government or high technology sector [2].

Exploitation of zero-day vulnerabilities in APT attacks is particularly worrying. During 2013, FireEye reported 11 attacks of that type. Most exploited were Java's zero-day vulnerabilities, followed by Internet browsers and Adobe Reader.

APT attacks cause damage to business operations of organizations affected by them. A study [3] conducted by Ponemon Institute shows that victims of APT attack are most likely to experience downtime of information technologies, loss of sensitive information, and disruption of business processes. Even though all these events cause significant losses, the worst and most expensive consequence is damage to brand and reputation. Ponemon Institute's research shows that, on average, damages to organization's reputation caused by APT attacks amount to $9.4 million.

APT attacks often remain undetected for a long period of time. Some of them, however, gained media recognition. For example, operation Aurora that resulted in loss of Google's intellectual property in 2009, or Stuxnet, a worm discovered in 2010 that infected uranium enrichment plants in Iran.

APTs bring changes in attack methodology. New attack methodologies mandate that organizations change their approach to information security. While there is no such thing as a completely secure system, it is possible to establish countermeasures that will greatly reduce the risk of falling victim to APT attacks.

This paper explains the nature of APTs and provides an overview of appropriate methods of countering attacks. The second chapter gives an overview of APTs, a definition and description of the term, their lifecycle, mode of operation and tools they use to exploit weaknesses in information systems. The third chapter lists some characteristic detection points for APT attacks, particularly outgoing traffic. It also contains a summary of tools that can be used for purposes of detection. The fourth chapter lists basic countermeasures in order to establish proactive defense in an organization.

II. ADVANCED PERSISTENT THREAT

A. What are APTs?

The term *Advanced Persistent Threat* refers to a well-coordinated group of people (typically controlled by an organized entity such as a government, terrorist group, corporate entity or other) with malicious intent that targets

a specific organization, government or company. APTs perform persistent prolonged attacks until they gain access to organization's computer systems, sensitive data or other resources. Individual hackers and opportunistic criminals are not APTs. They either do not possess the resources or abilities to execute complex and prolonged attacks, are rarely well organized, or they are not targeted. Traditional hackers attack for financial gain, because of social and economic issues, or just for fun and to gain recognition. APTs, however, have a very specific target. They are after the target's intellectual property, and they do not just try several methods to get access and then stop trying if they are unsuccessful, but attack continuously. If they are unsuccessful on the first few attempts, they will keep trying until they find a way to compromise the organization.

A description of an advanced persistent threat can be inferred from its name. Traditional attacks lack one or more of these characteristics.

Advanced – the attacks are coordinated by a group of people that possess advanced resources and knowledge, and are often well funded. The attacks themselves do not necessarily have to be highly sophisticated for the breach to be successful. Attackers often use phishing and widely available tools for malware creation [4] [5]. However, if needed, they can also develop their own tools to target specific vulnerabilities (such as zero-day exploits), and use multiple types of attacks in order to breach the target. Therefore, the term *advanced* refers to attackers' capabilities and knowledge which makes it different from traditional attacks.

Persistent – the attackers persist, they repeatedly try to compromise the victim's systems until they gain access to their resources. Then, the attacker maintains access. They create backdoors and, when one connection is severed, another one can be opened and used to continue extracting sensitive information. If the attacker loses access to the victim's system, he will repeat the whole process until he reestablishes access.

Threat – a threat is a possible source of harm or danger. Since APTs possess both the intention and the means to harm their targets, and are not just automated pieces of software, they present a threat.

APTs have a very specific target, and their main objective is to maintain **prolonged access to the target's intellectual property**, sensitive information, or otherwise interesting data that is useful to the attacker. To achieve that, the attack needs to remain concealed.

Even though APT most often attacks target computer systems, attackers can also use other, more traditional and simpler attack methods, such as social engineering or intercepting phone calls. APTs also perform attacks in order to acquire data that may help them execute more attacks in the future.

B. Life Cycle of APT Attacks

APT attacks are complex and consist of several stages. Each step of an attack is carefully adapted to attack the specific victim's vulnerabilities. Although these steps are not a rule, but a general scheme of an APT attack, their existence is vital in maintaining stealthiness and prolonged access to sensitive information, and it is expected in most attacks (**Error! Reference source not found.**).

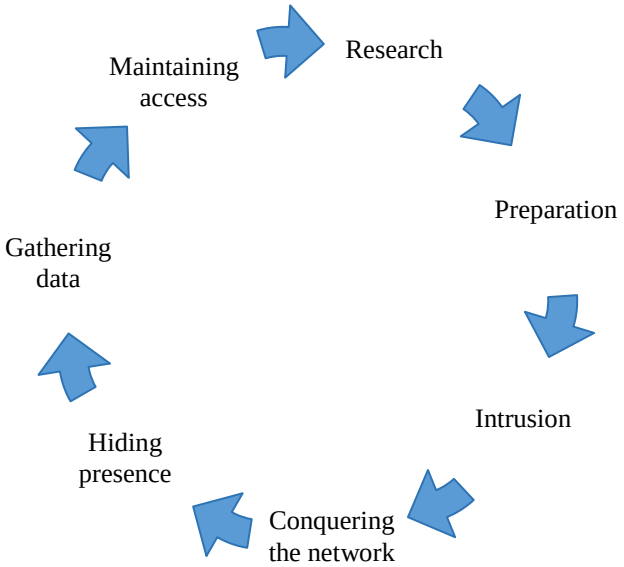


Figure 1 Life cycle of an APT attack

1. **Research.** The attackers gather information about the victim, trying to find weak points to exploit. The information that attackers search for is often publicly available. The attackers research employees' social network profiles and personal information, organization's website, technologies, communications, hierarchy, and premises. APTs differ from traditional attackers at this point. Since traditional attackers (individual hackers, *script kiddies*, inadequately organized hacking groups) do not target a specific organization, or at least do not do so persistently, over a prolonged period of time, they do not care if some of their attacks are unsuccessful, because they know there will be other, successful cases. APTs target a specific organization, therefore they must conduct extensive research if they want to find a vulnerability that can be exploited. Chances of a successful attack increase with the amount of information researched. The attackers will also search for individuals to be targeted (e.g. with phishing messages). At this phase, the attack is almost impossible to detect because it does not involve interaction with the organization's systems. Once the target has been identified, attacker moves on to the next phase.
2. **Preparation.** Using information they acquired in the first step, attackers prepare the attack. In this step they perform active extraction of specific information about the victim. They also prepare and test methods and tools that will be used to exploit found vulnerabilities. Activities in this step include scanning the network, ports, and running services, finding or creating customized malware that will exploit the victim's vulnerabilities, and planning malware

deployment (e.g. phishing messages, removable media). During this stage, APTs also have to prepare the infrastructure that will be used to manage and control attack flow. This includes taking control of servers to use as command and control centers, modifying DNS entries, creating fake domains and email accounts etc. At this phase, the probability of detecting the attack is very low. Once the attack methods and command-and-control infrastructure have been set up, the attacker proceeds to the next step.

3. **Intrusion.** Attackers launch first attacks. Most often, it is done using phishing email messages. These emails contain a message that will convince a gullible employee to open a document or a URL that contains malicious code. The attackers can impersonate a coworker, business partner, or any other person that the employee trusts. Assuming that the attacker gathered enough information in previous steps, and since people are typically the weakest link in the security chain, this approach is probably the easiest way to compromise an organization, and very likely to succeed. Other possible methods are exploiting vulnerabilities in organization's systems, including zero-day vulnerabilities, and spreading malware infections using removable media. Smartphones and tablet computers also present a security risk in this context. Since employees use their personal devices to connect to their workplace network, these devices can be used to spread malware in the organization's network. Absence of security policies and protocols increases the probability of success for this stage of attack. After the attack has been launched, the attacker monitors command and control infrastructure waiting for a sign of successful breach.

4. **Conquering the network.** If the attack has been successful, that means the attackers have gained access to one or more computers in the organization's network. If possible, they will try to obtain administrative (*root*) privileges on victims' systems. After initial breach, they install additional malware and create backdoors (e.g. remote administration tools – RAT) to act as a convenient way to remotely access and control the system. They also try to spread through the network identifying other machines susceptible to compromise, and trying to find valuable data that will be extracted. In this phase the attacker actively controls computers in the network, therefore the intrusion is easiest to detect. Some APTs also execute fake and easily detectable attacks against the organization in order to keep system administrators busy while the real attack takes place quietly.

5. **Hiding presence.** One of the main goals of the attack is to maintain prolonged access to organization's network and, consequently, their data. In order to avoid detection, APTs need to conceal their presence in the system. For traditional attackers this component is not crucial for the same reason it is not crucial for them to research their potential victims – even if they lose access to some of the machines they have infected, they can always find other vulnerable machines. For them, it is a numbers game. On the other hand, APTs have a specific target, which makes it harder to reestablish connection every time they are detected. Because of that, APTs install rootkits, modify event and audit log entries, periodically enter a sleeping state for some time, and delete traces of their own files on the system.

6. **Gathering data.** While exploring the organization's network, attacker searches for data of interest. When found, the data is gathered and, most often, encrypted to conceal its true identity on the way out of the network. Then, the data is masked to resemble legal traffic and slowly extracted from the organization's network. Using the command and control infrastructure that has been set up, it is then routed back to the attacker.

7. **Maintaining access.** In order to extract the maximum possible amount of data from the victim without getting caught, the attacker needs to maintain prolonged access to the victim's network. The attacker needs to make sure the backdoors are working as expected, the infected systems are accessible, and that command and control infrastructure is working, so he periodically checks their status. He could also update existing or install new malware in order to exploit other weaknesses or improve stealthiness, reconfigure the malware, and expand or reconfigure the command and control infrastructure. If, despite everything, the communication is severed, the attacker has to persist and go back to step 1.

C. Attack Toolbox

Attackers employ a variety of tools and methods when trying to access the organization's systems. One of most used techniques is **phishing**. It involves sending a specifically crafted message with an embedded malicious document or link to employees of an organization via email or social networks. In case of **spear phishing**, the attackers send such messages to specific people within an organization, after obtaining their personal information to increase the probability of success. Attackers also spread malware using **infected removable media**, or embedding it into **installation files** of popular programs. **Zero-day** vulnerabilities in commonly used programs or operating systems are targeted as well.

In the second phase, and while spreading through the network, attackers use **network scanning and enumeration tools** to discover computers and other devices connected to the network, their operating systems, open ports, running services and possible vulnerabilities.

When the system has been infiltrated, the attackers create **backdoors**. Backdoors enable the attacker to evade normal security controls and get remote access to the

system. They are created as modifications to the operating system, or they can be installed as a service. Attackers use **rootkits** to hide the presence of unwanted programs in the system. They also use **malware generators** to quickly and easily create customized malware.

In order to facilitate remote access to the machines, the attackers use **remote administration tools** (RATs). While such tools are often used in legal activities, they can also give the attacker complete control of infected systems.

APTs utilize **command and control infrastructure** to communicate with the infected machines, pass control messages, and extract data. Command and control servers are most typically infected servers under attacker's control. The communication is **encrypted** to make it harder to detect and block.

D. Notable Cases

1) Operation Aurora [6]

Operation Aurora was a sequence of multiple hacking attacks in 2009 that targeted Google, Adobe Systems, Rackspace, Juniper Networks and probably many others. The attacks originated from China, and exploited several zero-day vulnerabilities in Internet Explorer. Google's intellectual property was stolen. As a result, Google left the Chinese market.

2) Stuxnet [7]

Stuxnet is a worm, discovered in 2010, that infected industrial control systems, mostly in Iran. The worm propagated through infected USB thumb drives, exploited several zero-day vulnerabilities in Windows OS, and then corrupted PLC control software. The worm destroyed nuclear centrifuges in Iran's uranium enrichment facilities. It is speculated that creation of Stuxnet was a joint effort between Israel and the USA.

3) Operation Shady RAT [2]

A sequence of attacks that started in 2006 and continue to this day. McAfee, a cyber-security company, claims that the attacks targeted more than 70 organizations, most of them in the USA. The attacks utilize RATs and encrypted HTML comments that are used to pass commands to RATs. The attack most likely originates from China.

4) GhostNET [8]

A cyber espionage operation, discovered in 2009, that targets computer systems in more than 100 countries. The attackers use phishing and remote administration tools. The attacks originate from China, and target governments, ministries, and embassies. China's government denies involvement.

5) Darkhotel [9]

An active APT that targets guests using wireless networks in high-end hotels. The APT misuses hotel's check-in data and targets very specific individuals. It offers them software updates for frequently used packages, bundled with trojans and key loggers. It also possess some unusual features. The programs were signed with stolen certificates. Most of the attacks occurred in Japan.

6) RSA [10]

In 2011 it was reported that RSA Security was a victim of an APT attack. The attackers used phishing messages which contained an Excel document which exploited a vulnerability in Adobe Flash Player. The attackers used a RAT tool to control the infected computers and access data on company's servers. This incident caused $66.3 million damage. It is probable that the attackers stole information about RSA SecurID token seeds.

III. DETECTION

Even in the best of circumstances, systems get compromised and computers become infected. Attack vectors are numerous, and covering each option is almost always impossible. If the breach happens, there are several options that can help detect and remove the threat. With APTs, well-timed detection is very important.

Traditional methods of detection rely on examining the signature of unknown files to determine if they correspond to known malware, examining protocols or blocking specific network ports, and intrusion prevention systems (IPS) that can detect a very small subset of threats. APTs, however, rely on masking and encrypting network traffic, port hopping, newly created malware that does not exist in malware databases, malware with mutating code etc. For that reason traditional methods of detection are ineffective. Since malware has developed from infecting individual machines into organized networks of *zombies*, organizations need to implement not just endpoint security mechanisms, but also network security mechanisms.

Since one of the main activities of APTs is to extract data from an organization, analyzing data that leaves the network can help in discovering the command and control messages of an attack. The following characteristics of outgoing traffic can be used to discover malicious activities and should be analyzed:

- **Amount of data and packet quantity.** When normal users establish connections, they typically send out small amounts of data and receive large amounts of data. Since APTs extract data *from* the organization, this is one of key areas to monitor. Infected machines may generate large amounts of outgoing data that looks unusual (encrypted or masked). However, when analyzing outbound data, it is necessary to consider organization's common data flow in normal circumstances. It facilitates adjusting the monitoring engine precisely and decreasing the number of false positives it might produce.
- **Connection duration.** Normal connections are rarely established for longer periods of time. APTs establish connections that last longer when extracting larger amounts of data from an organization. Again, when monitoring connection duration it is necessary to consider types and duration of connections that normal business activities use.
- **Transmission period.** If the machines generate a certain type of outgoing traffic consistently in regular periods of time (a pattern which does not

correspond to normal usage), it could indicate presence of malware that regularly extracts data from the network.

- **Malicious sites.** Network monitoring system should keep a list of sites that have hosted malicious content. Existence of machines that communicate with such sites could indicate the existence of an attack. It will not prevent communication with newly established malware servers, however such communication will stand out from usual and can be detected by other means.
- **Internet Relay Chat (IRC) communication.** Botnets often use IRC to pass control messages between command and control servers and infected machines.
- **Destination IP addresses.** It is possible to determine the country in which the destination IP address resides. If there are outgoing IP packets with destination addresses in countries which have no association with the organization, it is possible that these connections were established by malware.
- **New domains.** In order to create command and control infrastructure, APTs often register new domain names and use dynamic DNS to route traffic between multiple infected machines which makes such traffic hard to track. Frequent communication with a previously unknown new domain could indicate presence of malware.
- **DNS cache.** Normal users typically do not use IP addresses when establishing network connections, they use URLs (uniform resource locators). When the user types a URL or clicks on a hyperlink, their DNS resolver must find the matching IP address for the URL. The domain name and its IP address are then stored in the DNS cache for some time. If outgoing connections do not have a corresponding entry in the DNS cache (which means someone established a pure IP connection), that could indicate malware infection.

Detection of network security breaches can be done using various IDS and SIEM (Security Information and Event Management) tools. It is important to note that these tools do not exempt users from using traditional endpoint security tools, such as personal firewalls and antivirus programs.

Various tools exist which can automate network monitoring tasks, correlate events and help detect breaches in network security:

- **OSSEC [11].** OSSEC is an open source host based intrusion detection system which detects intrusions by examining logs generated by various applications. OSSEC has two work modes, *local* (monitoring only one system) and *agent/server* (collecting and monitoring logs from multiple sources across the network). OSSEC can read and analyze log files of over 40 different programs and devices. It has separate components for collecting, reading, analyzing logs, and sending email alerts. It is also possible to install OSSEC agents on network devices. These agents collect logs on devices and send them to the central server for analysis using UDP. When logs are received, their field values are extracted, crucial information contained within logs is identified, and checked against pre-set or manually created rules. If needed, it sends email alerts. OSSEC works on most operating systems.
- **Snort [12].** Snort is an open source network intrusion detection and prevention tool. It is capable of performing packet analysis and logging on the network. Snort works in one of three modes: sniffer mode, packet logger mode, or network intrusion detection mode. The first two modes are simpler modes used for displaying or storing network packets. The last mode is used for traffic analysis according to a predefined set of rules, and generating alerts. Snort consists of several components which capture packets, decode and preprocess information in packets, analyze extracted information, find patterns within this information, and perform logging, alerting and outputting data. It is also possible to filter generated events in order to reduce false positives (e.g. by rate, number of events, or completely hiding events).
- **Sguil.** Sguil is a set of network monitoring tools that combines an IDS (Snort), network connection security profiling (SANCP), TCP connection analyzer (Tcpflow), stores results from multiple tools in a database (MySQL), and provides a GUI. When deployed, a system consists of a Sguil server and several Sguil sensors which monitor the network and send collected data back to the server. Sguil clients receive and display data from the server. Sguil improves common IDS systems by integrating different data sources and analysis solutions, thereby improving detection rate and reducing workload.
- **Splunk.** Splunk is a tool for processing machine-generated big data [13]. It provides searching, analyzing, indexing, correlating, visualizing and reporting capabilities. Splunk is used in various information technology domains, e.g. web analytics, internet of things, operational analysis, and security [14]. When used for security monitoring, Splunk can collect, index and analyze data from logs, firewalls, IDSs, perform statistical analysis of events within the network and thereby detect threats that would otherwise escape rule-based systems.

IV. DEFENSE

It is possible to undertake actions that can reduce the probability of system security breach. Of course, these efforts should be focused on critical areas – those which

present a high risk of security breach or high cost of repairing the damage. For that reason it is important for organizations to recognize their critical data and potential weaknesses which may present an attack vector.

In order to stop attackers from breaching the system, and discover those who have succeeded, both prevention and detection of attacks is important.

Some basic prevention measures are:

- **User education.** People are often the weakest link in organization's security. Due to the lack of knowledge and understanding of the threat employees become a critical point that attackers exploit. Educating users can greatly improve organization's level of security. It is necessary to educate employees about good password practices, social engineering (particularly phishing messages), malware, and other dangers.
- **Implementing access and usage policies.** Organizations should implement strict access policies and permissions for wired and wireless networks, files and other data. Access should be granted to users at the lowest necessary permission level. When authenticating users, TFA (Two-factor Authentication) should be used and strong passwords should be enforced. Detailed authentication logs should be kept.
- **Controlling external media.** All external storage media (USB drives, optical disks etc.) should be controlled in order to prevent spreading malware infections or losing valuable data.
- **Protecting valuable data.** Valuable data should be encrypted, with limited access. The most critical data should be kept on machines without network connection.
- **Managing endpoint security.** It is important that all operating systems and other programs used by organization's employees have all the latest security updates installed and enabled. It is also important that they have personal firewalls and antivirus programs installed and updated. When detecting or preventing security incidents, endpoint security must not be disregarded, but not exclusively relied upon.
- **Implementing NAC (Network Access Control).** Network Access Control is a method of enforcing network access policies for client devices. Non-compliant devices can be automatically repaired or upgraded when they are first connected. Network access is granted only to devices which satisfy certain security requirements such as security patches or antivirus protection. Other devices are given limited access until they are updated. NAC can also be used to control access based on assigned roles.
- **Blocking high risk applications.** Organizations should block usage of known high risk applications which could be used to spread malware or facilitate data leakage, such as P2P (peer-to-peer) programs, encrypted tunneling applications, and proxies.
- **Blocking known malware servers.** Organizations should keep a list of known malware sites, block access to them, and alert system administrators of any requests to these sites.
- **Analyzing security breaches.** Finally, if an attack does penetrate organization's security, it is necessary to thoroughly analyze how it happened, which parts of the system were compromised, and what data was lost. That way organizations can get a complete picture of the incident and protect themselves from future attacks.

V. CONCLUSION

APTs have become a significant concern for organizations from various domains. This trend is not showing signs of slowdown. Due to their stealthiness and focus on data, APTs present a growing danger that is very hard to prevent, detect and defend against. In the recent years, we have witnessed numerous cases where APTs have caused significant damage to various organizations, not only in form of data theft, but immediate physical damage.

APTs have an advanced approach to attacking organizations, they have the resources and knowledge, they do not stop give up when they are unsuccessful, they conceal their presence, and they target organization's valuable resources. Because of that, new security approaches are needed. Organizations cannot rely on traditional firewalls and malware detection. They need to employ various advanced intrusion detection tools and, based on their activities, develop new methods of detecting anomalies in a specific network, and correlate them in order to discover security breaches.

Even though there is no guarantee that any network is completely secured, organizations can implement tools and security policies that can greatly reduce the risk of an attack. However, doing that requires knowledge, resources and time.

REFERENCES

[1] FireEye Inc., "FireEye Advanced Threat Report: 2013", 2013

[2] D. Alperovich, "Revealed: Operation Shady RAT", 2011

[3] Ponemon Institute, "The State of Advanced Persistent Threats", 2013

[4] FireEye Inc., "Spear Phishing Attacks – Why They are Successful and How to Stop Them", 2014

[5] FireEye Inc., "Poison Ivy: Assessing Damage and Extracting Intelligence", 2014

[6] R. Varma, "McAfee Labs: Combating Aurora", https://kc.mcafee.com/resources/sites/MCAFEE/content/live/CORP_KNOWLEDGEBASE/67000/KB67957/en_US/Combating%20Threats%20-%20Operation%20Aurora.pdf

[7] A. Matrosov, E. Rodionov, D. Harley, J. Malcho, "Stuxnet Under the Microscope", http://www.eset.com/us/resources/white-papers/Stuxnet_Under_the_Microscope.pdf

[8] The SecDev Group Information Warfare Monitor, “Tracking GhostNet: Investigating a Cyber Espionage Network”, 2009

[9] Kaspersky Lab Global Research and Analysis Team, “The Darkhotel APT: A Story of Unusual Hospitality”, 2014

[10] RSA FraudAction Research Labs, “Anatomy of an Attack”, https://blogs.rsa.com/anatomy-of-an-attack/, 2011

[11] D. B. Cid, “Log Analysis Using OSSEC”, 2007

[12] M. Roesch, C. Green, Sourcefire Inc., Cisco Systems Inc., “SNORT User’s Manual: The Snort Project”, http://manual.snort.org/, 2014

[13] Splunk Inc., “Splunk for Security: Supporting a Big Data Approach for Security Intelligence”, http://www.splunk.com/web_assets/pdfs/secure/Splunk_for_Security.pdf, 2014

[14] Splunk Inc., “Advanced Threat Detection and Response: Using Splunk Software to Defend Against Advanced Threats”, http://www.splunk.com/web_assets/pdfs/secure/Splunk_for_APT_Tech_Brief.pdf, 2014

Security Risk Management in Complex Organization

Ivan Sedinić, Tamara Perušić
Croatian Telecom, Cyber & Data Security Section
ivan.sedinic@t.ht.hr, tamara.perusic@t.ht.hr

Abstract: Security Risk Management is foundation and starting point for implementation of security measures in any organization and challenge by itself. But in complex organizations there are additional challenges, how to align IT Security Risk Management with overall Security Risk Management and later with the Company's overall Risk Management. When organization is part of some international corporation, corporative rules also need to be followed in addition to legal and regulation rules. In telecom industry in regular operations also is very important that security assessment could be performed in short timeslot as support for operational decisions. Croatian Telecom as a part of Deutsche Telecom Group is facing all of this issues in addition to ISO 27001 requirements against which the Company is certified. To solve the challenge, the Company developed three methodologies for Information Security Risk Management. All of these methodologies are merged in common Risk Register as well as aligned with the Company's Risk Management. In this paper each Information Security Risk Management methodology will be described including its application area, as well as how recognized security risks are shown in common Risk Register and how they relate to the Company's Risk Management.

I. INTRODUCTION

The goal of this paper is to describe a way to align different Information Security Risk Management methodologies caused by different business requirements. These methodologies are supporting operational decisions process and in the same time are also supporting central risk reporting. Practical way of such approach is described in this paper.

In section I is given introduction to Risk Management and significance of having Risk Management in place. Section II briefly described some of the most popular Risk Management methodologies that are used worldwide. In section III is shown the Company's Risk Management process in the Group as well as three various methodologies that are used for Information Security Risk Management in the Company. Finally, in section IV is described Risk Register that is used in the Company for tracking information security risks.

The Company is Telecommunication Company that provides fixed, mobile, Internet and TV services. It has approximately 5.000 employees and is a part of the International Telecommunication Group. The Group is comprised of European Telecommunication Companies with joint governance and consolidated financial report. At the Group level are defined umbrella policies and guidelines that have to be applied in all group members. The Group and the Company are on different Stock Exchanges and are subject to different Stock Exchange rules.

Careful handling of risks is a fundamental condition for commercial success in today's complex and dynamic environment. Early detection and management of opportunities and risks should be an integral part of any management activity within each company. This also includes instilling awareness of the Risk Management system to all employees as part of the tasks assigned to them. Having well managed Risk Management process provides ongoing support to successful business performance. Focus should be placed on following aspects:

- creating transparency over opportunities and risks,
- early detection of material risks and opportunities,
- precise and careful handling of opportunities and risks,
- risk reporting,
- documentation and management and
- increasing enterprise value in order to recognize risks in early stages and make decisions which enable to focus on activities with the most favorable risk-return profile [1].

IT risk is the potential that a given threat will exploit vulnerabilities of an asset or group of assets and thereby cause harm to the organization. It is measured in terms of a combination of the probability of occurrence of an event and its consequence [2].

II. DIFFERENT RISK MANAGEMENT MODELS

Many risk frameworks have been developed and each of them has its own advantages and disadvantages. All have in common that they require organizational discipline to define assets, list of threats, evaluate controls, and conclude with an estimate of the risk magnitude. Bellow will be shortly described some of the most popular methodologies.

OCTAVE, probably the most well known risk framework, comes in three sizes depending on the organization size (small, medium or large). One of the benefits of the OCTAVE series is that each of the frameworks provides templates for worksheets to document each step in the process. Templates can be used directly or customized for a particular organization [3].

The NIST framework, described in NIST Special Publication 800-30, is a general one that can be applied to any asset. It follows a similar structure like OCTAVE, but uses a little bit different terminology. NIST doesn't provide plenty of templates such as OCTAVE but it is relatively easy to follow. Since it has a focus on concrete components as systems, it can be easily used for organizations that are new to risk assessment. Also because the methodology is defined by NIST, it is approved for the use by US government agencies and organizations that work with them [4].

ISACA's COBIT is IT management framework while the ISO 27xxx are security frameworks that require organizations to have a risk management program. They recommend repeatable methodologies and specify when risk assessments should take place. Risk assessment in COBIT described in RISK IT goes beyond security risks and includes development, business continuity and other types of operational risk in IT [5], whereas ISO 27005 concentrates on security exclusively.

ISO 27005 follows a similar structure to NIST but terms are defined differently. ISO 27005 includes annexes with forms and examples, but like other risk frameworks, it's up to the organization implementing it to evaluate or quantify risk in ways that are relevant to its particular business.

Organizations that do not have a formal risk assessment methodology could firstly review the risk assessment requirements in ISO 27001 and 27002 and consider the 27005 or NIST approach. The ISO standards provide a good justification for formal risk assessments and outline requirements, while the NIST document provides a good introduction to a risk assessment framework.

In this article we will use telecom company listed on Stock Exchange and which is part of international telecom group, as a model.

Described organization's Security Risk Management is based on ISO 27001, 27002 and 27005 frameworks. The Company has decided to certify Information Security Management System according to ISO 27001 standard. Because of it, Security Risk Management System has to be compatible with the ISO 27001 standard.

III. SECURITY RISK MANAGMENENT IN THE COMPANY

Company's Risk Management process

For the purpose of early recognizing and reporting significant risks within the Company, the Company's Risk Management process is established.

This process encourages early identification, analysis and risk assessment, timely communication to decision-makers and implementation of measures and activities that lead to optimal outcomes in relation to objectives of the Company. This process also established additional business transparency for Management and Supervisory Board through reporting process about significant risks.

Management and reporting system of the Company's risks is an integral part of the Risk management and reporting system of the Group. Also, the Company is obligatory to manage and timely report on significant risks according to Capital Market Act [6] and Accounting Act [7].

According to business requirement and obligations the Company's reports have to be prepared quarterly, and every organizational unit in the Company is obliged to ensure continuous monitoring and timely reporting of all risks that may affect the future business as well as work towards promoting awareness of risks at all levels.

The Company reports significant risks:

- *within mandatory periodic financial statement reporting* of the Company for stakeholders,
- *to external auditors* (Audit of financial statements of the Company [8]),
- *to the Group* (Consolidation of risks portfolio for the whole Group),
- *to the Audit Committee* (monitoring efficiency of internal controls of the Company referred to Risk Management system, compliance operations, internal audit, etc.)

The Company's Risk Management process has 6 major sub-processes:

- *identify risk* where risk is recognized,
- *assess* where risk is assessed in relation to its business objectives (plan),
- *correlation and aggregation* where assessment of risk appearance interdependence to other organizational units and processes is done,
- *response to risk situation* where measures for risk management are defined,
- *implementation of action plan* and
- *monitoring* that includes monitoring of residual risk and development of risk circumstances.

All company units have to take care on risks within their jurisdiction and they are following this process when some risk occurs.

This process is shown in Figure 1.

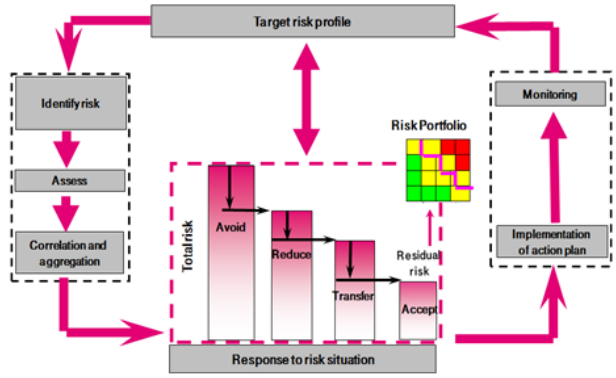


Figure 1: Company's Risk Management process

In Figure 2 is presented reporting line, who reports to whom within Risk Management Process.

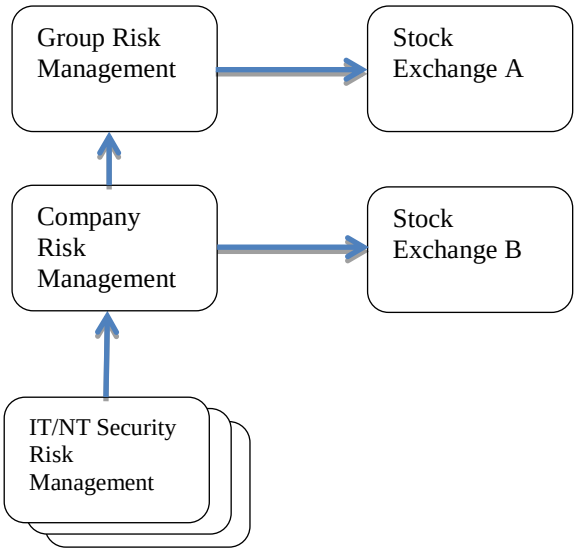


Figure 2: Reporting line

Risks are presented in Risk Portfolio matrix by mapping evaluated impact of recognized risk which is expressed in financial terms and estimated likelihood of occurrence as shown in Figure 3.

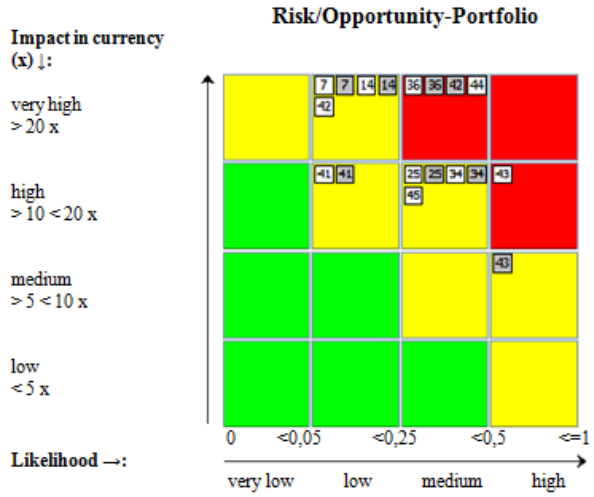


Figure 3: Risk Portfolio matrix

For the purpose of managing information security risks in the Company, three Security Risk Management sub methodologies are developed. All of them are aligned with the Company's Risk Management. Each of them will be described.

A. ISMS Security Risk Management Methodology

This methodology is used in the Company for evaluation of all kind of information security risks. It is developed to support implementation of Information Security Management system (ISMS) according to ISO 27001 standard [9].

ISMS Security Risk Management ensures identification of the critical Company's assets and related risks, and also manages identified risks. In that way, overall security of business processes in the Company is properly considered. Also, by conducting Security Risk Management, continuous increase of security in the Company is ensured.

Methodology that is used in ISMS Security Risk Management process is aligned and very similar to the Company's Risk Management process methodology. Risk matrix is also similar in both processes (4x4) and visually presented in traffic light colors, but scores are calculated slightly different as can be seen in Figure 3 and Figure 5.

ISMS Security Risk Management process consist of the next 5 sub-processes shown in Figure 4.

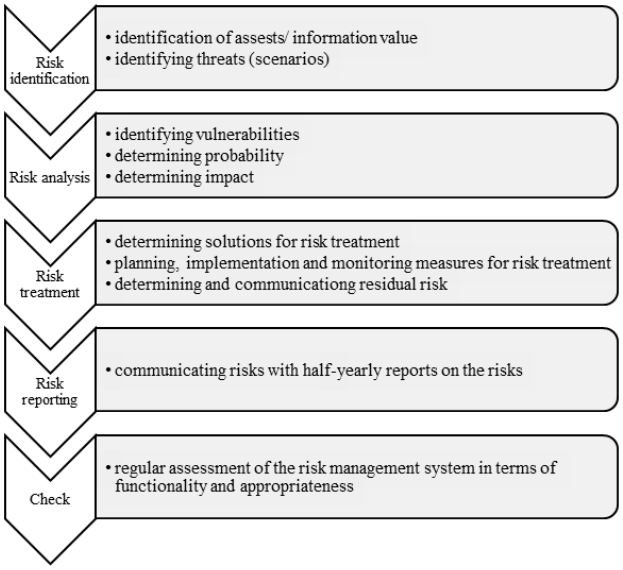


Figure 4: ISMS Security Risk Management sub-processes

Brief description of each ISMS Security Risk Management sub-process is given bellow:

A.1. Risk identification - this sub-process includes identification of assets (people, technology, processes and information) where attention also should be paid to the transmission of information from source to the destination. Scenarios of threats that may have an impact on the identified assets have to be recognized and noticed. Example how it is done is presented in Table 1.

Type	Example of vulnerabilities	Example of threats
Software	Disposal or reuse of storage media without proper erasure	Abuse of rights
	Incorrect parameter set up	Error in use
	Poor password management	Forging of rights
Hardware	Lack of efficient configuration change control	Error in use
	Unprotected storage	Theft of media or documents

Table 1: Examples of some vulnerabilities and threats

A.2. Risk analysis - within risk analysis any threat (scenario) associated with the identified asset is necessary to check from the perspective of certain risk dimensions (availability, integrity, authenticity, confidentiality) and to identify vulnerabilities. The probability of occurrence and impact on asset has to be determined in this step if risk occurs.

In Table 2 are criteria that are used when determining the probability of occurrence.

Category	Probability	Value	Definition
Very low	0-0,05	1	The appearance scenario is very rare (once in the time period of >5 years)
Low	0,05-0,25	2	The scenario appears once within 5 years (based on previous experience)
Medium	0,25-0,5	3	The scenario occurs annually
High	0,5-1	4	The scenario occurs often (several times over a period of one year)

Table 2: Criteria for determining the probability of occurrence

In Table 3 are criteria that are used when assessing the impact.

Category	Value	Impact (Confidentiality, Integrity, Availability)	Financial impact
Very low	1	The impact is rarely felt. The impact is limited to a certain part of the observed process.	No financial impact
Low	2	The impact is limited to the observed process.	Moderate financial impact (<100.000 kn)
Medium	3	The impact can be felt in other processes/ outside the scope of observation.	Significant financial impact (>100.000 kn)
High	4	A major impact within the process and out of processes/ inside and outside the scope of observation; impacts on users	Major financial impact (>1.000.000 kn)

Table 3: Criteria for impact assessment

Risk level is derived from multiplying determined impact value (taking into account financial impact) and probability value (scenario occurrence) and is shown in risk matrix (Figure 5).

Low risk includes risk levels 1 and 2 and is green colored. Middle risk includes risk levels 3, 4, 6, 8 and 9 and is yellow colored. High risk includes levels 12 and 16 and is red colored.

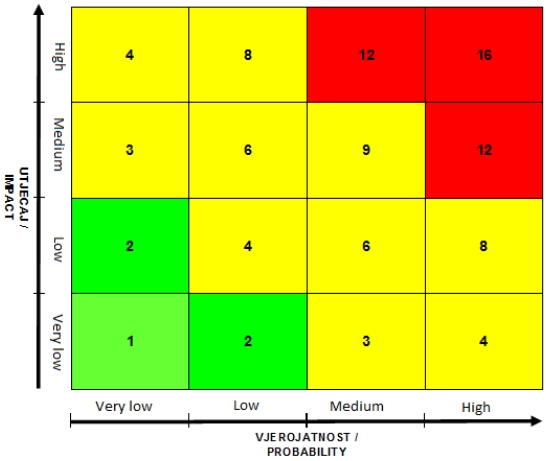


Figure 5: Risk matrix for ISMS Security Risk Management

A.3. Risk treatment - within this sub-process, possible solution for the treatment of risk are determined. After solution is chosen, next step is to plan, implement and monitor measures for risk treatment. Finally residual risk has to be determined and communicated.

The following risk treatments are possible:

- *risk mitigation* in which appropriate controls to reduce the risks are applied,
- *risk acceptance* where organization decides will not to take action and that the risk clearly satisfies the organization's policy and criteria for risk acceptance,

- *risk avoidance* where risks are avoided by not allowing actions that would cause the risks to occur and
- *risk transfer/sharing where risk* are transferred to other parties e.g. insurers or suppliers.

A.4. Risk reporting - risks are communicated half-yearly (internal regulation) within the Report on risks.

A.5. Check - as a final step in ISMS Security Risk Management process, regular assessment of the Risk Management system in terms of functionality and appropriateness has to be done.

B. Basic Security Evaluation (BSE) Methodology

BSE methodology is used for basic security evaluation of all IT/NT systems in the Company against 20 predefined security risks by going through defined check list called BSE questionnaire. Limitation of this methodology is that it is based on a set of basic security measures and therefore it is primarily used for initial security screening of particular system. Advantage is that it gives a solid security insight of IT/NT system as well as possibility to compare levels of IT/NT systems' security risks in the Company.

Following risks are included in BSE methodology:

- inappropriate system location (e.g. in office)
- unsecure access to system from internet,
- administration without Jump-Hosts and/or dedicated admin network,
- possible external access not using approved 3rd party access,
- manual file upload/download,
- missing segregation from office network,
- testing with production data,
- missing central user management,
- missing auditable process for user management,
- missing or insufficient control of given rights,
- missing regular backups incl. recovery tests,
- missing patch process,
- missing continuous scans,
- missing logging and/or log analysis,
- not compliant to near offshore requirements (security rules for offshoring to companies in neighborhood),
- missing agreement for commissioned data processing,
- disaster recovery plan is missing,
- missing system description,
- group accounts used,
- missing data privacy and security concepts

Main goal is to identify security risks in IT/NT systems and create risk repository in short time. Risk repository enables IT/NT management to easily determine systems with the highest risk level from a security point of view.

This methodology is adjusted to needs of the Company and aligned with ISMS Security Risk Management system by using same risk levels and criteria for risk acceptance as well as the Company's Risk Management.

Technical System Owner or his/her representative with assistance of Security Expert has to fill-in BSE questionnaire for particular IT/NT system. Given results are imported in excel tool for security risks assessment and scoring. Risk score is calculated combining impact value and probability value in Risk matrix from ISMS Security Risk Management (Figure 5). Further steps are basically the same as in methodologies shown previously. Results are analyzed and report created about risks found. Given results are presented to management and open issues discussed. Risk treatment plan is created, measures with defined deadlines are agreed and progress of implementation is monitored.

C. IT Security Risk Management

Third methodology is IT Security Risk Management used for evaluation of specific security risks in IT area. Assessment of IT security risks is done within the excel risk assessment tool called RAST. It is an in-house solution where risk estimation is done into nine steps. In each step are entered some data about the treat scenario that is considered. Risk index is calculated by the tool as soon as all required fields are filled-in.

In scope of this methodology are complex IT projects where are evaluated specific IT related assets, processes and solutions. Threats that are considered are:

- threats to confidentiality and integrity of information assets resulting from accidental and deliberate malicious acts done by authorized and unauthorized human threat actors,
- threats to availability of information assets and services provided by technology assets, as long as they are endangered by deliberate malicious acts or human (operative) mistakes. [10].

Risk methodology is very similar to previous ones. Main difference from others is that impact, likelihood of occurrence and risk are expressed as index values on a scale from 1 to 10 (see Figure 6).

The *impact index* is assessed separately for three protection goals – loss of confidentiality (C), loss of integrity (I), and loss of availability (A). Overall impact index is a maximum value of impacts for C, I, A. Impact index has the scale 1-10. The likelihood index is assessed for a described threat scenario and has a scale 1-10.

Risk index = f (impact, likelihood) or in this case
Risk index = sqrt(impact* likelihood)

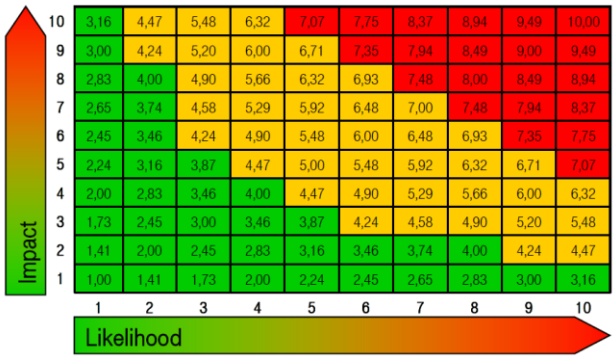


Figure 6: Risk matrix for IT Security Risk Management

This methodology has 5 sub-processes as shown in Figure 7.

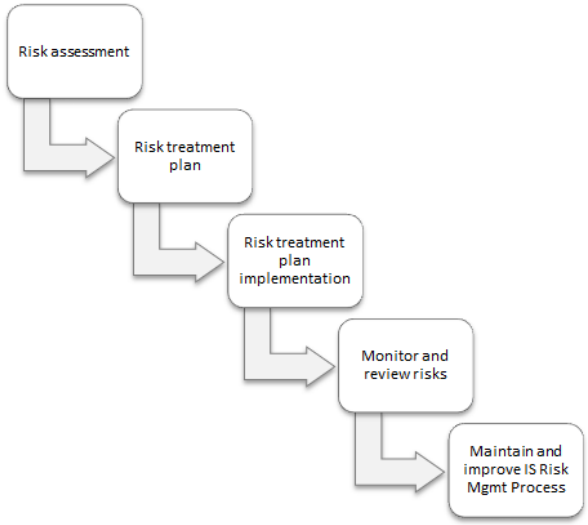


Figure 7: IT Security Risk Management sub-processes

Further is shortly explained each sub-process of this methodology.

C.1. Risk assessment – this sub-process starts with risk identification where vulnerability, threat scenarios and potential impact are documented. Findings have to be confirmed by System Owner. After confirmation, Security Expert estimates the risk index.

C.2. Risk treatment – within this sub-process next options are possible:

- *risk retention* where risk is acceptable according to Risk acceptance criteria and no further action is required,
- *risk reduction* that consists of a list of possible controls to reduce impact or likelihood, cost & benefits, priority of implementation and timeframe for implementation,
- *risk avoidance* in case that risks are too high, or costs of control exceed benefits,
- *risk transfer* where risk is shared with external party and
- *risk acceptance* - where management decision is required for accepting the current level of risk without implementing security measures. This is done in situations where the cost of implementing protective measures exceeds the benefits or risk cannot be decreased with implementation of security controls (it is not possible to decrease threat or vulnerability) nor it is possible to decrease it by propagation or avoiding.

C.3. Risk treatment plan implementation – consists of list of possible controls to reduce impact or likelihood, costs and benefits, priority of implementation and timeframe for implementation.

C.4. Monitor and review risks - All risks in status "open" are monitored in regular intervals.

C.5. Maintaining and improving information security risk management process – regular check of information security risk management process.

IV. RISK REGISTER

In addition to Risk portfolio which is managed within the Company's Risk Management process, Risk register is developed and maintained within IT/NT Security Risk Management process. Its main purpose is to track information security risks within the Company.

Risk register contains all information security risks that are identified in any of 3 above described Security Risk management methodologies, as well as all risks that have been recognized during various internal or external audits and system scans (e.g. penetration scanning, vulnerability scanning, and source code scanning).

In cases where different method for Risk Management scoring is used (like in external audits or system scans), translation tables are designed in order to have comparability of risk levels.

The following information is tracked in Risk register:

- Information asset name – affected assets
- Threats – possible threats to confidentiality, integrity and availability of information assets
- Vulnerability – described weakness in information asset that could expose the system
- Probability/ Likelihood of occurrence – estimation on how easy it is to exploit a threat
- Impact – assessed consequences of risk events if they are realized
- Risk value – value calculated based on assessed probability and impact
- Risk evaluation date – date when risk evaluation is performed
- Recommended measures – proposed risk treatment options for affected information asset

- Deadline – agreed date until risk will be mitigated
- Risk Owner – name of person that is responsible for implementation of measure.

Risk register is updated each time when new risk is identified and assessed or some changes related to risks in evidence occurred.

Risk register is half-yearly reviewed and report on residual risks prepared. Report is discussed with Risk Owners and further activities defined.

V. CONCLUSION

In situation when it was necessary to fulfill requirements and obligations toward different stakeholders in different timeframe and on different detail level, we found that optimal way for us is to design different risk management methodologies which are interconnected at least through same risk levels and risk acceptance criteria. Concept was to have overarching risk management methodology which is capable to cover all area in the Company and plug in results from other, area specific methodologies. On the next level were developed specific risk management methodologies appropriate for dealing with specific problems in information security overall, fast and simple reaction to operational requirements and detail assessment of complex IT projects.

Such approach has numerous benefits as are:

- Practical way to integrate different risk methodologies using bottom up principle
- perform risk assessment in short period when operational requirements dictate it,
- perform risk assessment in very deep and detail level for technologically specific risks,
- unification of results for different area in which risk management tools are used,
- optimization of time spent on risk management.

VI. REFERENCES

[1] DT AG: Manual Group Risk Management, Insurance, Germany 2008 (unavailable to general public)

[2] International Organization for standardization ISO: ISO/IEC "Information technology – Security techniques – Information security risk management", 27005:2008, Switzerland 2008

[3] Carnegie Mellon University, Software Engineering Institute http://www.cert.org/resilience/products-services/octave/index.cfm

[4] National Institute of Standards and Technology: Managing Information Security Risk, Organization, Mission, and Information System View, NIST Special Publication 800-39, Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg, USA, 2011, http://csrc.nist.gov/publications/PubsSPs.html#800-30

[5] ISACA: Risk IT Framework for Management of IT Related Business Risks, USA, http://www.isaca.org/knowledge-center/risk-it-it-risk-management/pages/default.aspx

[6] Capital Market Act, Hanfa – Croatian Financial Services Supervisory Agency, http://www.hanfa.hr/getfile/39633/2_3_0_Capital%20Market%20Act.pdf

[7] Accounting Act, NN 109/07, 54/13, 121/14 http://www.mvep.hr/zakoni/Default2.aspx?pregled=pretraga&pretraga=zakon%20o%20ra%C4%8Dunovodstvu

[8] Accounting tools: Audit of financial statements, http://www.accountingtools.com/questions-and-answers/what-is-a-financial-statement-audit.html

[9] ISO (International Organization for Standardization): ISO/IEC 27001 – Information Security Management, http://www.iso.org/iso/home/standards/management-standards/iso27001.htm

[10] DT AG: IT Security Risk Management Process, Germany, 2011 (unavailable to general public)

PRIVACY TRACK

Violation of User Privacy by IPTV Packet Sniffing in Home Network

Siniša Vukojević, univ. spec. secur. inf.
HAKOM, R. F. Mihanovića 9, 10110 Zagreb
Sinisa.Vukojevic@hakom.hr

Abstract - The aim of this paper is to determine the possibility of learning about and the quantity of collected information about the habits of users of Internet Protocol television (IPTV) in Croatia on the basis of unauthorised monitoring of the IPTV traffic in the users' home network.

The experimental part of the presented work includes collecting IPTV traffic in home networks of IPTV users of the two largest electronic communications operators in Croatia while preforming ordinary user activities (initiating the Set Top Box (STB) devices, switching TV channels, using EPG, program recording, using video on demand etc.). The paper does not explore the possibility and the manner of achieving the unauthorised access to the users' home network itself; instead, it is assumed that it has already been obtained.

Based on the gathered data, each user's action is analysed from the aspect of indicating users' behavioral habits and the violation of their privacy. The results of the conducted analysis presented in the paper provide an overview of gathered information and some concluding remarks regarding the possibility of violation of privacy by IPTV traffic sniffing in the user's home network.

I. INTRODUCTION

Internet Protocol Television (IPTV) is a system which enables delivering the digital television services over the internet using Internet Protocol (IP). The usage of IP technology enabled interactivity, more versatile and higher quality content and the development of numerous new IPTV services.

IPTV services are usually offered via closed networks by electronic communications operators, which control the technical parameters and ensure an adequate Quality of Service (QoS) and Quality of Experience (QoE) to its users. However, apart from international standards and recommendations, there is a lack of publicly available documents which describe the network architecture or its operation, thus leaving several open questions, including the level of users' privacy protection, which is unknown.

Due to that fact, this paper addresses the possibility of IPTV users' privacy violation and what types and amount of user, content or system data can be obtained by an unauthorised IPTV traffic monitoring of user's home network. The aim of the paper is to determine the possibility of IPTV data collection and investigate the quantity of gathered information regarding the behaviour and habits of IPTV users in Croatia based on the IPTV packet sniffing.

The study presented in this paper is performed in home networks of IPTV users of the two largest electronic communications operators in Croatia. The possibility and mode of establishing the unauthorized access have not been explored; it is presumed that sniffing is already established. The protection of protected content and network, as well as IPTV service provider's network security are also not explored, although results of the analysis suggest some aspects of those questions as well.

The experimental part of the study consists of collecting IPTV traffic during usual user actions: initiating the Set Top Box (STB), switching TV channels, using program guide etc. Collected data are analysed with regard to two aspects - user behaviour and the violation of users' privacy.

The paper is organized as follows. In order to provide a better insight into the IPTV functioning, Section II provides a brief overview of the IPTV technology based on the publicly available data. Section III describes the methodology of data collection and Section IV the analysis of IPTV data. Section V analyses gathered information, while Section VI comprises discussion and some concluding remarks.

II. INTERNET PROTOCOL TELEVISION (IPTV)

A. General

IPTV is defined and confirmed by the International Telecommunications Union (ITU) focus group ITU-T FG IPTV as follows [1]:

"IPTV services are multimedia services such as television/video/audio/text/graphics/data delivered over IP based networks managed to support the required level of quality of service (QoS) and quality of experience (QoE), security, interactivity and reliability"

As a term, IPTV was first used in 1995 by the company Percept Software as the title for the application for audio/video traffic transmission via unicast and multicast using IP [2]. In 1998, Precept Software was bought by Cisco, which kept IPTV as a trademark. First IPTV services via Digital Subscriber Line (DSL) ware available in 1999 in UK and it is estimated that today there are approximately 110 million subscribers worldwide [3]. The number of connections in Croatia in 2013 was 393 thousands [4].

B. *Architecture*

Several organisations have been involved in the standardisation of IPTV systems, architecture and services [5]. General IPTV architecture (Fig. 1) consists of four main parts: content provider, IPTV service

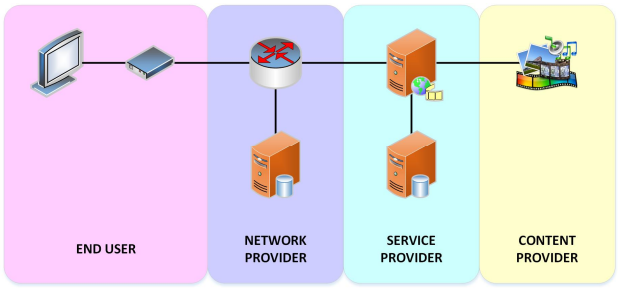


Figure 1. General IPTV architecture [6]

provider, network service provider and the end user. Each operator can be present in more than one entity.

More details on IPTV architecture and its basic elements can be found in the following standardization documents:

- FG IPTV-DOC-0181: IPTV Architecture
- ITU-T Y.1910: IPTV functional architecture
- ETSI TS 182 027: IPTV functions supported by the IMS subsystem
- ETSI TS 182 028: NGN integrated IPTV subsystem Architecture

C. *Protocols*

Providing *end-to-end* IPTV service (from the content provider to the end user) is achieved by the large number of protocols which enable and support different applications and functionalities of IPTV services. ITU document "IPTV related protocols" [7] identifies and categorizes relevant IPTV service protocols (Table I).

In the observed systems the most used protocols are:

- **HTTP** for the communication from the end user STB device to IPTV middleware servers placed in service provider's network
- **RTSP** for the establish and control media sessions (play, pause, rewind and fast forward etc.), acts as a "network remote control"
- **RTP** for the real-time transmission of media content (audio and/or video), and **RTCP** for reporting on the performance and control of the RTP session
- **IGMP** for the TV channels switching

In IPTV networks unicast communication is used for processing unique requests from a single user (e.g. video on demand), while multicast is used for delivering the same content to a large number of end users. Due to the requirements for network bandwidth for every video signal, i.e. TV program, IPTV service would not be possible without using multicast for content transmission.

TABLE I. IPTV PROTOCOLS [7]

TCP/IP LAYERS	*PROTOCOLS*
APPLICATION LAYER	DHCP, DNS, DVB-IPTV, FEC, FLUTE FTP, HTTP, IPDC CDP, LC, MBMS RTP, RTSP, SIP, SNTP, TLS
TRANSPORT LAYER	TCP, UDP
NETWORK LAYER	BGMP, ICMP, IGMP, IP, MLD MSDP, PIM-SM, Anycast-RP, SSM

D. *Services*

IPTV enabled numerous new IPTV services. ITU document "Service scenarios for IPTV" [8] classifies services into content, interactive and communications services (Table II).

TABLE II. IPTV SERVICE LIST [8]

Content Services	Linear TV (audio, video and data)
	Linear Broadcast Audio
	Linear TV with Trick Modes
	Multi-View service
	Pay Per View (PPV)
	Personal Broadcast Service
	PVR service (network or client-based)
	Time-shift TV
	Video on Demand (VoD)
	Near Video on Demand
	Content Push
	Music on Demand (MoD)
	Content download service
	Service Information (EPG, ECG, …)
	3rd party content services
	End-user Originated content
	Regulatory Information services
	Advertising
	Hybrid services
Interactive Service	Interactive TV (iTV)
	Learning services (education, languages, ...)
	Information services (news, weather, traffic, …)
	Entertainment services (photo, games, karaoke, blog)
	Portal services
	Commerce services (security, banking, shopping, …)
	Interactive Advertising
Communication Service	Communication services (e-mail, IM, SMS, Chat, ...)
	Communications Messaging
	Presence services

E. *Market Overview*

At the end of the third quarter 2014 there were around 108.2 million IPTV subscribers, which represents 18.9% annual growth of the number of subscribers, i.e. 17.2 million new subscribers [3]. At the same time, there were 393 104 IPTV subscribers in Croatia [9], i.e. 0.22% lower quarterly rate and 2.06% annual growth, which is significantly lower than 18.9% of global growth (Fig. 2).

IPTV services in Croatia are used in 26.14% households [4], while a major part of 53.07% households uses digital terrestrial television, which is a great potential for the further development of IPTV.

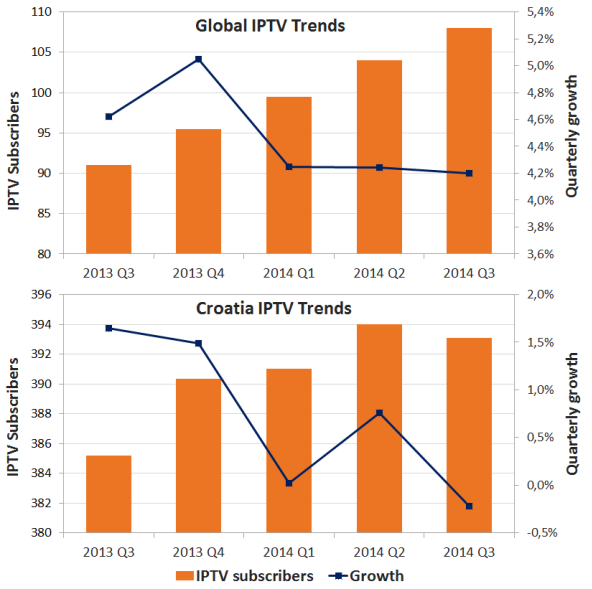


Figure 2. Number of IPTV subscribers and their growth globally [3] and in Croatia [9] for Q3 2014

III. IPTV Data Collection

A. Tools and Method

To determine the amount, type and the content of data which are exchanged when using IPTV service, it is necessary to collect information from the end user's home network. For that purpose the home network has been adjusted: instead of directly connecting ADSL modem and STB, they are connected via switch (Fig. 3). Using Port Mirroring properties of switch, a copy of their communication was forwarded to the computer which collected and stored IPTV packets for the further usage and analysis. For collecting and storing data, as well as for their analysis, a free tool *Wireshark* and Linux operating system *BackTrack 5* have been used. Data were collected in the third quarter 2013. The study included two major operators in Croatia, which are compared in the results of the analysis.

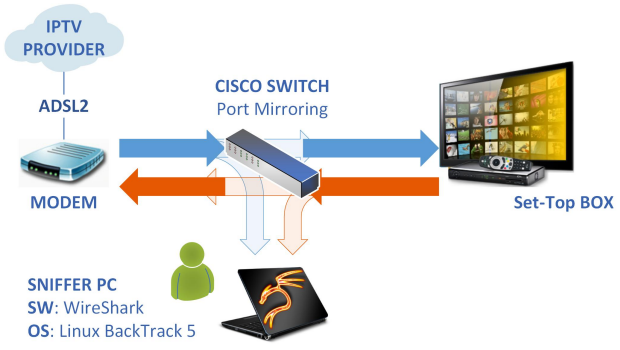


Figure 3. Modified home network for experiment

B. Use Cases

In order to enable faster and better quality collection and later analysis of IPTV packets, the ordinary user activities has been classified into several use cases, which are described and explained in subsequent steps (Table III).

TABLE III. Use Cases

USE CASES	*STEPS*	
Initiating the STB device	1	STB device is not connected to the power, but is connected to the TV and to the ADSL modem
	2	After 30s, power-up the STB device
	3	Monitor the process of initiating the STB device on TV
	4	After image appearing on the TV screen, wait 5 min
Switching TV channels	1*-2	Every 30s user switching TV channels HRT1 → HRT2 → HRT3 → HRT1
Using EPG	1*-2	After 30s, the user open the EPG guide
	3	After 30s, the user presses the button "DOWN" on the remote, and positions to the current show on the channel below HRT1
	4	After 30s, the user presses the button "DOWN" on the remote, and positions to the current show on the channel 2 places below HRT1
	5	After 30s, the user selects "Show description"
	6	After 30s, the user closes the EPG guide
SNIMALICA – recording video content	1*-2	After 30s, the user opens the EPG guide and clicks on next show
	3	After 30s, the user presses "RECORD" and confirms the record TV shows
	4	After 30s, the user closes the EPG guide
SNIMALICA - watching previously recorded content	1*-2	After 30s, the user selects "Snimalica"
	3	After 30s, the user selects one of the pre-recorded shows, starting with watching
	4	After 30s, the user temporarily stops the watch by pressing the "PAUSE"
	5	After 30s, the user continues to watch by pressing the "PLAY"
	6	After 30s, the user interrupts watching by pressing the "STOP" and exits
SNIMALICA - erasing previously recorded content	1*-2	After 30s, the user selects "Snimalica"
	3	After 30s, the user selects one of the pre-recorded shows, and option "DELETE"
	4	Confirms deletion of show and exits
VIDEOTEKA - renting the video content	1*-2	After 30s, the user selects "Videoteka"
	3	After 30s, selects one of the offered programs and clicks the "PURCHASE"
	4	The user enters the PIN and confirms purchase
Youtube	1*-2	After 30s, the user selects "Youtube"
	3	After 30s, selects and starting with watching one of the available video content
	4	After the completion of video playback exits
Facebook	1*-2	After 30s, the user selects "Facebook"
	3	A user logs with username and password (Sinisa.Vukojevic@gmail.com / LoZinka.123)
	4	After 30s, the user selects "FRIENDS"
	5	After 30s, sets the message "Test IPTV" on the wall
	6	After 30s, the user is logged out from "Facebook" and exits
Telefon - Call	1*-2	After 30s, the user receives an external telephone call
	3	After four (4) rings , the user answers the call and hangs up after 10s
Telefon – List of calls	1*-2	After 30s, the user selects "Telefon" and the sub-option "All Calls"
	3	After 30s, the user exits

1* - Start: The user watching TV channel HRT1 (initial step)

C. Description of user IPTV services

IPTV users from operator A and B are subscribed for the following services:

- triple play (Internet, telephone, television)
- Snimalica - TV program recording and watching on demand
- Videoteka - Video on demand

Operator A user is additionally subscribed for:

- Application "Telefon" which allows list of calls overview and current call notification on TV device
- Youtube and Facebook applications via TV

IV. IPTV Traffic Analysis

The IPTV traffic analysis was performed using Wireshark tool. Based on the captured data, each use case is analysed in order to investigate how the observed IPTV system works from the aspect of indicating users' behavioral habits and the violation of their privacy (more detailed technical information are available in [10]).

A. Initiating the STB device

Traffic tracking while booting the user STB provides the most information about the IPTV system, service, content and user under study. Fig. 4 depicts a simplified process of starting user STB.

1) Operator A

After starting, using DHCP protocol, STB device obtains IP address and other network data, as well as the information about multicast UDP flow with IPTV system configuration data - InfoCast. After that, STB connects to InfoCast and downloads configuration data: boot files for its device type, current time and time zone, user interface homepage, IPTV middleware servers. Using boot files configuration data, STB reads files for kernel boot and boot splash, which gives graphical representation of the device start up.

After the kernal boot, follows the download of files necessary for IPTV user interface, based on the homepage *index.svg* whose URL is taken from InfoCast flow. Scalable Vector Graphic (SVG) is an open standard based on XML format and in observed system is used for the graphical representation of the user interface. Besides SVG files, graphical PNG (Portable Network Graphics) files are downloaded, as well as JS script files which consist program code written in JavaScript language.

In parallel with running graphic user interface, based on the JavaScript files content, STB sends multiple HTTP POST requests (26 requests), and receives responses from provider's servers with metadata. Metadata describe user device and GUI interface characteristics, IPTV service and video content. They also contain user, i.e. IPTV subscriber data. Data are written in simple and open text standard format for the exchange and transmission of data, JSON (JavaScript Object Notation).

Initiating the STB device is finalized by registering in the multicast flow containing Promo TV channel video content, i.e. by image appearing on the TV screen. A video stream is transmitted over IP network as MPEG-2 Transport Stream encapsulated into RTP/UDP/IP, and in experiment is observed that is possible to reproduce captured video stream (Wireshark: Telephony – RTP – Stream Analysis – Save payload).

2) Operator B

Using DHCP, STB gets the IP address and other network settings, as well as additional boot files data. Based on the obtained data, STB registers and reads UDP multicast flows with boot files and configuration parameters. By using boot files, STB is restarted, DHCP request repeated and then the current time is set using TIME and NTP request to the current time providers.

Using RTSP DESCRIBE request STB device gets configuration parameters of TV channels present in the IPTV service. After that, it is registered in SAP/SDP flow which contains configuration parameters or IPTV service data, i.e. it redirects to other UDP flows (there are 24 flows in total) with configuration parameters or IPTV service data.

Although UDP flows seemed unreadable, detailed analysis showed that data are sent through UDP in parts and in cycles (UDP Carousel), with the first 8 bytes of UDP load describe its position inside the file. Merging separate parts into one single file, based on the first two bytes (0x1F and 0x8B), showed that it is a GZIP file, which is decompressed into an easily read XML file.

STB device loads configuration parameters first, and then it retrieves device, service, content and user metadata in XML format using several HTTP requests and registering in multicast UDP flows.

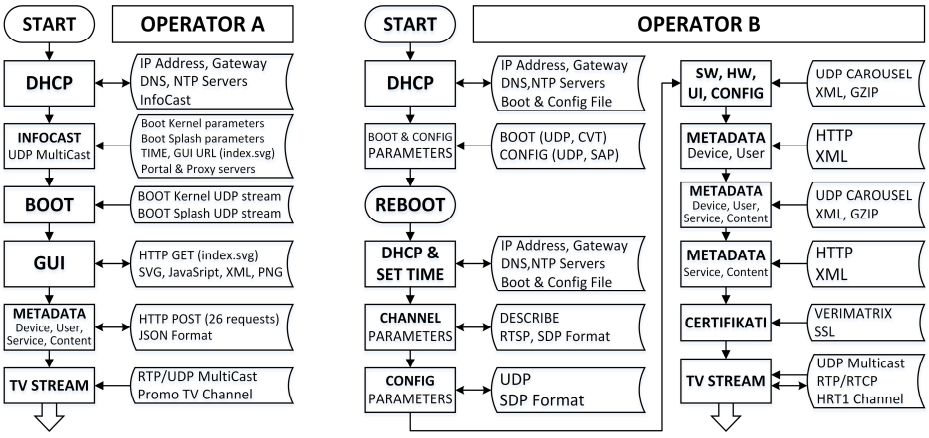


Figure 4. Simplified diagram of initiating the STB device - based on a detailed analysis of traffic in Wireshark

After retrieving metadata, a protected SSL communication occurred for the first time. Typical SSL/TLS handshake and key derivation was observed and the negotiated cipher suite was TLS_DHE_RSA_WITH_AES_256_CBC_SHA. Based on the usage of Verymatrix system, it can be concluded that SSL communication occurred probably due to exchange informations between STB and provider's servers in order to protect video content and digital rights.

Starting the device is finalized by multicast HRT1 TV channel as MPEG-2 Transport Stream encapsulated into RTP/UDP/IP and corresponding RTCP flow, resulting with the image of the channel appearing on the TV device.

3) Analysis of initiating the STB

Analysis of the traffic obtained while initiating the STB showed that it is possible (for both operators) to obtain extensive information about the device, IPTV service, user and the content, such as TV channel list and description of the program content. However, it must be emphasized that it is also possible to obtain the user PIN code (it can be found within the XML file as MD5 hash at Operator A and it can be obtained by hash comparison; at Operator B, it can be found in plain text).

B. Switching TV Channels (Operator A and B)

Switching channels is done using IGMP messages. For each channel switch, the user STB must generate one IGMPv2 Leave Group message and one or more Membership Report messages. Knowing the list of TV channels with the corresponding multicast addresses (it is possible to get them by sniffing traffic when starting STB device) and tracking IGMPv2 messages in the home network, it is possible to track user habits when watching live TV programs.

C. Using EPG

Using EPG (program guide) of Operator A, the information retrieval about the program content is done using HTTP requests, and data are in JSON format.

Operator B does not have an instant communication between the STB and middleware. Information describing program content for the specific time interval is in XML format and they are loaded during device booting, as observed in subsection A: Operator B.

D. Using "Snimalica"

Snimalica is used for recording and later reproduction of video content stored on servers in provider's network (Networked Personal Video Recorder - nPVR). Data collection is done for three use cases: video content recording, watching and erasing previously recorded content.

Both operators use HTTP requests for retrieving all necessary information and starting the actions of recording, erasing or reproduction. Establishing and control of the reproduction of pre-recorded video material is done via RTSP. As arguments to those requests, identifiers of program, i.e. recorded content, are used.

Taking into account that HTTP and RTSP requests are used, it is possible to track habits of using *Snimalica* and subsequently reproducing the content previously displayed to the user.

E. Using "Videoteka" - Video on Demand

Video on demand enables renting multimedia content by using remote control. Data collection is done only for the process of shopping, i.e. renting the video content, omitting the other use cases. Shopping process is verified by user's PIN code at both operators.

HTTP requests with corresponding arguments are used for getting all necessary information and carrying out the renting transaction itself. This enables tracking users' habits and rented content.

By observing details of shopping transactions, we have noticed that Operator B does not provide any safety mechanism from malicious transaction repeating. Operator A has some data from the request cookie which are seemingly safety mechanism, but they can be found in all other HTTP requests, meaning that they are not unique for the transaction itself.

By repeating HTTP requests, with changes of video content identifiers, malicious attacker on user's home network could cause a significant financial damage to the user because the middleware would consider it a legal shopping transaction made by the user.

F. Using application Youtube (Operator A)

Youtube application is used for watching video content on Youtube service. The IPTV user chooses from the given video content or searches the content using keywords. Information about the available content, searching the content and all necessary parameters for carrying out the service, are retrieved via HTTP requests, the same way as the video content itself. Therefore it is possible to keep the track of using Youtube service and later reproduce displayed video content.

G. Using application Facebook (Operator A)

Facebook application within IPTV service enables user to access Facebook profile and most of its functionalities via web browser using the interface adjusted to be displayed on TV device and for using the remote control. Before using the application, it is necessary to sign in using eligible Facebook username and password.

The same as for Youtube application, the data traffic between the application on STB and the rest of IPTV system is done via HTTP protocol and it is fully readable, which enables insight into all exchanged data. Username and password are sent as plain text during the authorization request and can be seen by anyone with the access to the network communication.

Sniffing IPTV traffic enables finding out all exchanged data: personal data, friends' data, photos, messages, Likes etc., including the user's Facebook profile password.

H. Using application "Telefon" (Operator A)

Application "Telefon" enables the possibility of displaying the notification of the current phone call, with the caller's phone number, and checking the phone call lists (list af all, missed, received, outgoing calls) on TV device.

For the current call notification the IPTV system sends STB at TCP 8080 port a file in JSON format which contains the information about the time of the call and caller's number.

For the list of calls overview the communication is performed via HTTP requests. Lists are in JSON format and they contain information about the time and duration of a call.

Sniffing IPTV traffic it is possible to see current phone communication, as well as user's phone call history.

V. GATHERED INFORMATION

The following table summarizes gathered information, relevant for this research, based on the analysis of the data collected via sniffing communication between end user STB and the rest of the IPTV system during the ordinary user activities.

OPERATOR	A	B
STB DEVICE		
Device model	✓	✓
SW – UI version	✓✓	✓✓
IP - MAC address	✓✓	✓✓
Device characteristics	✓	✓
IPTV SERVICES		
Architecture and features	✓	✓
Additional Services	✓	✓
List of TV channels with parameters	✓	✓
TV schedule and show description	✓	✓
OPERATOR	**A**	**B**
USER		
Subscriber ID, name and surname	✓✓	✓✓
User PIN code	✓	✓
User information	✓	✓
IPTV friends	✓	
Information about IPTV packages	✓	✓
USER BEHAVIOUR		
WATCHING TV CHANNELS		
Beginning/End of watching TV channel	✓✓	✓✓
The content displayed to the user	✓	✓*
SNIMALICA		
List of recordings	✓	✓
Recording/Erasing/Reproduction (the content, when)	✓✓✓	✓✓✓
The content displayed to the user	✓	✓*
VIDEOTEKA		
Information of the offered content	✓	✓
Purchase (the content, when, how much money)	✓	✓
YOUTUBE		
Information of the offered and searched content	✓✓	
Watching (the content, when)	✓	
The content displayed to the user	✓	
FACEBOOK		
Facebook username and password	✓✓	
User information / Friends	✓✓	
Photos / Wall messages / Likes	✓✓✓	
TELEFON		
Current call notification	✓	
List of all calls	✓	

✓* - in the case that the collected content that is displayed to the user is not protected

VI. CONCLUSION

The paper provides insight and the analysis of the communication between end user STB and the rest of the IPTV system with an aim to investigate the possibility for the violation of user's privacy and see the amount and type of user, content or system information that can be obtained by sniffing IPTV home network.

Analysis of the collected traffic for both chosen operators has shown that it is possible, and rather easy, to find out user's personal data including their PIN codes, habits of watching television program and video content etc. In case user is subscribed to more advanced IPTV applications, like Facebook or Telefon, it is possible to get additional personal or friends' data, phone call records etc. Besides privacy violation, it is possible to get extensive information regarding the operation of IPTV system, mode of service implementation and available content. Large number of information can be obtained due to the fact that most inter-system communication is performed without the additional protection, using protocols and formats which are easily read (HTTP, RTSP, SAP/SDP protocols, XML, JSON formats). Once known HTTP or RTSP requests can be repeated, with the same or changed arguments, and thus obtain information without waiting for the user's action.

From the network security aspect, the fact that user's PIN code can be easily retrieved at both operators, that at Operator A Facebook username and password are in plain text and open for everyone who monitors the network communication, that shopping/renting transactions are done without the additional protection via simple HTTP requests, is very dangerous for users. Using wireless STB in the future is likely to increase the possibility of abusing described security and privacy issues.

It can be concluded that IPTV service providers should additionally protect delicate communication of user services and assure better user privacy protection, at least at the level of protecting video content, regardless the low possibility of unauthorized access to the user's home network and potential degradation of QoE due to the additional protection measures.

REFERENCES

[1] ITU, IPTV Joint Coordination Activity: "IPTV-JCA Terms of Reference" [online]. Available: http://www.itu.int, January 2015.

[2] RÉMY, J.-G. AND LETAMENDIA, C.: "History" in *Home Area Networks and IPTV*, John Wiley & Son, 2011.

[3] POINT TOPIC: "Global IPTV subscriber numbers – Q3 2014" [online]. Available: http://point-topic.com, January 2015.

[4] HAKOM: "Godišnje izvješće o radu za 2013. godinu" [online]. Available: http://www.hakom.hr, January 2015., pp. 26-27

[5] Boswarthick, David: "ETSI IPTV Standards", IPTV World Forum Asia, December 2008.

[6] ITU, FG IPTV: "DOC-0181: IPTV Architecture", Figure 6-1. [online]. Available: http://www.itu.int, January 2015., pp. 8

[7] ITU, FG IPTV: "DOC-0191: IPTV Related Protocols", [online]. Available: http://www.itu.int, January 2015., pp. 8-10

[8] ITU, FG IPTV: "DOC-0182: Service scenarios for IPTV" [online]. Available: http://www.itu.int, January 2015., pp. 7-8

[9] HAKOM: "Tromjesečni usporedni podaci tržišta za Q3 2014. godinu" [online]. Available: http://www.hakom.hr, January 2015., pp. 4

[10] VUKOJEVIĆ, S.: "Violation of User Privacy by IPTV Packet Sniffing in Home Network", specialist's thesis, FER, 2014

Have we been monetized and become commodity without our consent – Privacy in the time of Big Data Technology

Marinko Žagar, Dubravka Mendeš Poljak
Polytechnic of Zagreb, City of Zagreb

According to Eric Schmidt, Google CEO, the quantity of data which is now generated in only two days, equals the quantity of data which was generated since the beginning of civilization until 2003. Aggregation of technologies which are used for storing and the analysis of huge quantity of data are called Big Data, and are used in many aspects of human activity. Since both structured and non-structured sources of data are used in Big Data technology, there is a question of personal data and privacy protection. Are and in what ways various companies, agencies and services collecting and analyzing personal data from social networks in order to increase their efficiency, make profit or to increase general safety of citizens? Law on Protection of Personal Data regulates the protection of personal data of citizens as well as the supervision of the collection, processing and use of personal data in the Republic of Croatia. Provisions of this Law are applicable to the processing of personal data by state authorities, local authorities, as well as legal entities and private persons who process personal data. Has personal data which is stored and analyzed on daily basis become commodity without our consent?

I. INTRODUCTION

Over the years, the quantity of information available on the Internet has become bigger and bigger. Facebook is an excellent example which illustrates this point. Over 500 TB of data is processed daily in this social network; 300 million photos and 2.7 billion "likes". In the article titled "Big Data Just Beginning to Explode", it is said that the quantity of data expands at the astonishing rate and that the annual growth of data by the year 2020 will amount to 4300%. [1]. Due to continual development of new technologies, as well as trends such as Cloud Computing and Big Data, we are more than ready to store and use all this information for various purposes. Depending on what purpose the information is analysed and used for, ethical issues regarding such use emerge, as well as the need for the protection of customer's privacy.

The term "Big Data"appeared for the first time in the article of NASA investigators M. Cox and D. Ellsworth in 1997. They emphasized the occurance of problems as the result of generating huge quantities of data. At that time, the existing computers with its memory and disc capacities were not able to accept such quantity of data. Nowadays, the technology in use excludes such limitations. The term "Big Data" has been kept until these days, however, it is no longer referring only to marking big quantities of data but active use of tools and procedures for analyzing and obtaining values; that is quality information out of that data

II. BIG DATA TECHNOLOGY

A. *Where is Big Data Technology used?*

World most famous examples of Big Data Technology use are Google, Yahoo, Amazon, Facebook, Netflix and Hadoop.

Google is definitely the leader regarding the quantity of data it processes and is taking number one place as Big Data Technology user. They were one of the pioneers who developed their own method of analysis, known as MapReduce.

By developing its software Hadoop, Yahoo! has brought such analysis closer to a broader circle of users. Hadoop is Java-oriented frame of open code which is used for processing big blocks of data in distributed computer systems.

Another famous global example of using Big Data Technology is Netflix; an American company which provides video content on demand. They have developed the most complex analysis of films and series, based on users' preferences. They have developed algorithms by which they can accurately predict what their users like, what their interests are, what they want to watch and at what time. In this way, Netflix offers their customers exactly what they want. Thanks to such successful model of prediction, they know what a successful film or series should look like and what actors and qualities are expected in such film.

Inspired by Netflix's success, Hrvatski Telekom signed a contract with data analysis company called TVbeat which would monitor viewing statistics of its IPTV platform [2]. By this, HT will get an insight into the viewing habits of its customers. Access to data in limited and entirely anonymous form will be given by HT to third parties (advertisers). The use of this technology will be visible through the definition and adjustment of new or modification of the existing services for the targeted user groups, all based on the conducted analysis.

Apart from the listed examples, there is a range of different areas where Big Data Technology can be applied. However, its biggest application is in companies such as Netflix, which have the need to understand the needs and

preferences of its targeted groups and users. We are talking about marketing agencies with high probability analysis, which give their recommendations to companies regarding the choice of products ideal for a group of customers.

American trading house Target, by using this high probability technology, can know when one of their customers is expecting a baby and can, in accordance with that information, offer them targeted products.

Search engines are another useful source of information. By gathering and analyzing searched terms, one can get very useful data as well as link the person with the inquiries they made.

Smart phones, tablets, smart bracelets and similar devices enable us to monitor and record everything, including the tiniest details of our daily lives such as calories intake, workout time and intensity, sleep rhythm, emotional state etc. When all this data is collected and analysed, one can get very useful recommendations. Simply by observing activities on social networks and from statements like this one: "Today I am feeling terrible, I'm lying in my bed with high temperature", one can predict the outbreak of flu infection in a certain geographical area. In 2009 Google correctly predicted the course of swine flu epidemic spread simply by monitoring swine flu-related searches on the Internet. This technology has been in use in the world of sports for some time now with the purpose of improving sports achievements.

They have come up with a service called MyMagic+ in Disneyland where users effectuate various benefits by wearing smart bracelets. In return, the park monitors all the user's activities, including his/her spending pattern.

In cities, by monitoring traffic, one can predict traffic jams and organize alternative routes, plan major events and more successfully manage the local infrastructure.

Surveillance cameras, private or public, are all around us. This method of monitoring and recording people is the most concerning regarding privacy isssues.Cameras can monitor our movements and if a face recognition algorithm is applied, a person can easily be identified.

The town of Zagreb, by signing a cooperation agreement with Microsoft on Microsoft CityNext project has made their first step towards the use of this technology, with the purpose of fast management of bureaucratic needs and better data management in all its operational segments.

Banks use Big Data technology not only to predict and prevent the abuse of credit cards and other financial malfeasance but to analyse their clients' behavior, their needs and, as consequence, to create specific, personalized offers through their products.

American National Security Agency, NSA, has been collecting huge quantities of metadata from all types of communication and has been using Big Data analytics in order to (among other things) uncover and prevent terrorist attacks and other security threats. Through the Boundless Informant[1] programme 96 billion intelligence units have been collected from servers and telecommunication channels throughout the world in March 2013 alone. [9]

As can be seen from the examples listed above, Big Data technology is more and more in use in a growing number of various systems benefiting individuals directly or indirectly. Crisis management systems have proved to be very efficient and show that the use of personal data without the consent of its owners is crucial in rescuing operations, whereas data owners do not support the use of personal data for commercial and marketing purposes without prior consent.

Analysts predict the future of the Internet closely linked to the term of Internet of things[2]. Internet will simply disappear. There will be many IP addresses, devices, sensors and things which we will carry around and which will be connected, yet we will not even notice that."[3]

The responsibility of each individual when it comes to sharing personal data is often crucial because some individuals ignore EULA[4] which warns the user what software manufacturers will use personal data for. Very often users give their consent without reading the terms of use.

B. Data collection

Information which is used for various Big Data technology analysis is collected from numerous sources. These sources may or may not be private or public, structured or non-structured. The number of users is continually growing, social networks are becoming more and more popular and therefore habits and user activity on the Internet is present in abundance. State and local institutions are becoming more and more aware of the importance of sharing information, partly due to Law on Freedom of Information, partly due to pressures from the European Union, and partly due to the need to provide better services to its citizens. However, all those information, without unification, joint analysis and pattern definition do not have any great value. Picture 1 shows the basic concept of Big Data technology which covers data collection and filtering, storing, sharing and analysis of data, as well as taking actions based on the analysis results of the collected data.

Privacy is becoming an asset based on the laws of supply and demand. Information is a highly valued commodity nowadays, and the information we are willingly sharing on the Internet are being "traded for" certain services. If we want to use some of the services of a webpage we accessed, for example to fill out a questionnaire, play a game or leave a comment, etc., in most cases we are not able to do so unless we accept the terms of use. In this way webpages or applications which potentially collect personal data meet the legal regulation

[1] Boundless Informant – Secret NSA programme which enables collection of information of worldwide communication conducted via phones and computers. A programme called PRISM has been used for this purpose.

[2] Internet of things – IoT – interconnection of uniquely identified devices with the Internet.

[3] Eric Schmidt, the president of Google, Davos 2015.

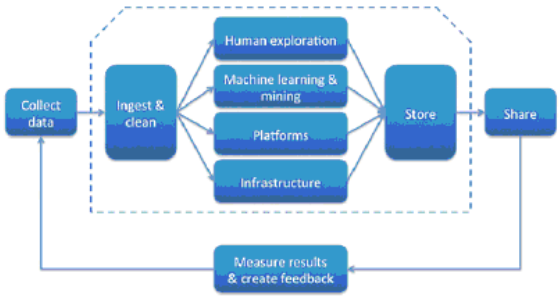
[4] EULA – End-user License Agreement

regarding data collection and processing, as stated in Article 6 and Article 7 of the Law on Protection of Private Data.

"Privacy Policy"is an integral part of every webpage. It defines the terms of use as well as ways in which the service provider will use the data and to whom the data will be available. To sum up, privacy policy is entirely defined by those who offer certain services and the user has a choice whether to accept it or not. Due to the fact that those terms of use are in most cases extensive and elaborate, the users are rarely aware of what they contain and in what ways their personal data is managed, even if they had accepted them. Marketing agencies are very skillful at using this way of "informing" their users. The data used in Big Data technology comes from different sources for which perhaps we gave our consents, but the new set of data emerging from the analysis and obtained through the combination of various sources may not be what we gave our consent to.

Apart from accepting the terms of use, the use of cookies is also regulated. Setting cookies, software which monitors and records computer user's activities when they access webpages is no longer allowed without the consent of the computer user.

Current legislation is not able to entirely protect the privacy rights and therefore individuals should get more involved and assume control over the protection of their data. Individuals should act responsibly during their Internet searches, they should refrain from sharing their data and always read the terms of use of the webpages they visit as well as EULA.



Picture 1. Data flow used in Big Data Technology

Legislation

Thanks to the rapid growth of the Internet users and available services, there is a growing concern regarding the protection of their privacy and personal data. In 90 countries all over the world they have passed laws which regulate personal data access. Nearly all those laws are based on specific measures adopted on the level of individual countries. In Croatia there is a Law on the Protection of Private Data, a Law on Freedom of Information and a Law on Electronic Media. As bylaws in Croatia, the following is in use: Regulation on Keeping Records and the Form of Personal and Regulation on the Procedure for Storage and Special Technical Protection Measures for Special Categories of Personal Data.

The right to protection of personal data is a fundamental human right and, as such, is guaranteed by the Croatian Constitution (Article 37 of the Croatian Constitution). As for the personal data protection, these rights are regulated by law in Croatia. Article 6 of the Law on Protection of Personal Data stipulates that personal data may be collected for the purposes of which the user has been informed on, which have been clearly stated and in accordance with the Law and may further on be processed only for its originally stated purposes. Provision of Article 7 of the same Law provides the legal basis for the collection and processing of personal data. Therefore, personal data may be, among other things, collected and further on processed with the consent of its owners only for the purpose for which the owner gave their consent, in cases specified by law, for the purpose of carrying out legal obligations of the supervisor of collection of personal data and the other enumerated cases. In the case where the individual gave their consent or shared the data themselves, they have the right to withdraw their consent. Croatian Laws conform to the highest European and international norms and standards and fulfill the obligations arising from international conventions that the Republic of Croatia signed and ratified. International law defines data protection: Convention 108 of the Council of Europe - The Law on Ratification of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data and the Additional Protocol to the Convention (MU no. 04/05), Directive 95/46 / EC of the European Parliament and the Council - Directive protection of individuals with regard to the processing of personal data and on the free movement of such data, Directive 2002/58 / EC of the European Parliament and the Council - Directive on the processing of personal data and privacy in the electronic communications sector and the Manual on European legislation on data protection. As the document states, it is very important to understand that the right to data protection is not an absolute right, but it has to be balanced against other rights, national legislation and individual cases. For this reason, for example, publication of photos in some cases may be interpreted as a breach of privacy, while in some cases this may not happen. All these laws share the concept of consent. Consent means "each voluntarily given, special and informed expression of will." In many cases it is the legal basis for the lawful processing of data[5]. According to Article 2 of the Law on Personal Data Protection of the Republic of Croatia, the subject's consent is freely given and it is a specific manifestation of the subject's will by which they express their agreement to the processing of their personal data for specific purposes.[6] This means that the consent must be clear, unambiguous and free statement, given by the individual and regarding the use of data which this individual can withdraw at any time. The individual must also have the right to choose in order to make their consent valid, and the withdrawal of such consent should not invite any type of sanctions.

[5] Directive on Data Protection, Article 2. Act (h).

[6] Directive on Data Protection, Article 2. Act (h).

However, the problem arises when it tries to answer the question of how to regulate the collection, storage and exchange of information on the Internet, especially when all that is spread throughout different countries and continents.

Current research regarding the issues of privacy concerns on the Internet shows that most respondents feel concern about that. According to research conducted by Tech Pro Research, 84% of the total number of respondents confirmed their concerns regarding privacy violations [8]. Similar results were obtained through extensive research conducted by the committee formed by order of the President of the United States Barack Obama on Big Data and privacy, ie. how this technology affects the way we live and work [10].

Picture 2. Most occurred words regarding Big Data technology inquiries [10]

Despite that, the users of smart phones, social networks and the Internet, generally speaking, do not think about these issues as they use various accessible and widespread services and publish their statuses or opinions on the Internet. Awareness about the violation of privacy is only perceived after all those opinions, comments and data typed in various search engines, profiles etc. are displayed back to their originator. There was a case like that in a divorce hearing when an ex-wife was faced with a document containing all her online activities collected in several large files [4]. It contained information on how much time she had spent on the Internet, which webpages she had visited, what opinions and comments she had shared and how she had felt every day. All that data was used as character traits of the person in question. Certainly, this felt like the violation of privacy and it caused great discomfort for her. Regardless of the fact that all such activities are recorded in the vastness of the Internet, the users do not think about who may see or use such data due to the falsely created feeling of anonymity on the Internet. On one hand the Internet promotes the freedom of expression and it has become the largest civil force that has emerged in recent years, and citizens have a voice that can be heard[7], on the other hand we try to protect and restrict all these information. The protection of such information and privacy is still the responsibility of each country. Therefore, this is more or less (depending on the country) regulated by law. EU members are trying to secure the highest level of privacy protection for their citizens. On the other side, the USA is not placing emphasis on the protection of privacy, whereas in countries such as China or India privacy as such does not even exist. However, regardless of laws, interpretations of the right to protection of personal data and privacy are very elastic.

In addition to incomplete definition of the protection of privacy in the Internet environment, virtually no or very small penalties are prescribed for the violation of it. The European Parliament currently adopted only a formal proposal for a new legislation on data protection of citizens which provides much greater sanctions of companies which will be violating this[4].

C. What does Big Data know about us?

It is easy to conclude that, based on the areas where Big Data technologies are used, the collected and processed data can easily identify individual persons. In order to placate concerns regarding the privacy violations, the organizations which use and publish analysis based on Big Data information harvesting claim that all information by which an individual could be identified have been removed. They claim that it is impossible to link records with actual persons who made those records. In the article titled Data Issues, Pete Warden gives several examples to prove the opposite. One of these examples is Netflix, who released an anonymous list of film rentals. This list was correlated with and analyzed along with IMDB reviews, which made it possible to determine the identity of certain individuals and view their entire rental history. Another example is the analysis and the comparison of two different social graphs, the result of which also led to the identification of the persons in question [7]. Information such as user names, e-mail address, webpages visited, inquiries done, comments left, IP addresses, times of action and many, many more are sent through web browsers. We leave traces. Regardless of that, individuals voluntarily give up their privacy in order to participate in all the services provided on the Internet. It is very easy to monitor the location and movement of persons via their smart mobile phones, however such knowledge does not diminish the number of smart mobile phone users.

III. BIG DATA AND PRIVACY

A. Data management in the public sector

Our data can be found in various data collections which are stored and kept at data centres of ministries and government bodies. Government agencies and services have our personal data and use them to improve their services towards us, in order to make government services more efficient. The use of personal data in the times of crisis, natural disasters or terrorist threats is absolutely justified and to citizens quite understandable. The use of personal data in such situations is legally regulated by the Law on the protection of personal data, if the data processing is required for the purposes of protection of life, preventive medicine and the like. The systems of predicative policing, based on Big Data analytics use both structured and non-structured data from various public

[7] Eric Schmidt, Google, The World Economic Forum in Davos 2015.

services. The analytics of such complex and enormous quantities of data is extremely efficient. Crime rate in cities where such systems have been implemented have been decreased by approximately 20% [11]. Apart from the above listed situations, our data is used in situations which are not predicted by laws, nor did the citizens give their consent to it. For example, during the political campaigns, some politicians send personalized messages to their voters and ask for their votes in text messages or via post. Apart from this, seemingly innocent example, every day we witness the use of someone's personal data for the political purposes, and the sources of such data are government agencies or services. We are in particular sensitive to the abuse of our health-related data as well as data of underage persons.

The use of personal data from such databases in combination with the data from social networks and services can be used for various purposes and their latent value is huge.

Data management in the private sector

Private sector is becoming more and more aware of the value of their databases. The implementation of CRM systems has become the condition for the provision of the efficient service to their customers and clients, and company owners and management are aware of this fact. Such systems bring certain benefits to their owners as well as to their clients. Significant benefits to both sides are brought by the use of Big Data technology, which can analytically manage both structured and non-structured data from various sources. The value of data is best illustrated through the example of company Facebook, which is a classic example of a company whose business is entirely based on data. We can see a big discrepancy between the estimated market value of $ 104 billion and the value of company's physical assets estimated at $ 63 billion. This enormous discrepancy in value represents the data the company works with. Doug Laney, the president of company Gartner, was the first one who warned us about the lack of data value in presenting the book-keeping value of Facebook. In his research he pointed out that even though Facebook owns "just data", in a value sense they own 2.1 trillion units of tradable goods, such as likes, uploaded materials and comments, which have been collected in the period between 2009 and 2011. This number is even more impressive nowadays. In the same research, he highlighted the actual value of each such data item at 5 cents, which sheds a totally different light on Facebook, where every user is worth $ 100[12].

Daily use of technology brings many advantages but also reveals many details which we would rather keep private. How to harmonize these apparently opposing facts?

B. How to protect privacy?

Even though continual efforts are invested into solving this issue, measures that have been adopted and in force are failing at protecting the privacy. This is the result of several things: lack of help, poor awareness, and in many cases the lack of interest of the individuals. The individuals are the persons who share and manufacture "products" in big quantities and should therefore get informed about where and how their data will be used. On the other hand, terms of use and consents on websites should be clearer, more concise and concrete regarding management methods of the data they provide. Luckily, data and privacy violations are not used in order to cause certain damage or malicious actions. Regardless of that, the individuals are feeling threatened when they receive personalized e-mail messages for which they know they had not given their consent. People are also unwilling to share their information in order to help third parties to make profit. One of the proposals of privacy protection in this case is that organizations which use data inform their users about where, when and how they would use their personal data [6]. The negative side of this method is that persons might be swamped with a large number of notifications in cases where many organizations would use their personal data and there would be very little they could do about that.

The largest challenge today, in times of Big Data when large sources of data are available, is the prevention of any kind of data leakage and securing that organizations which have access to private information do not use them for commercial purposes.

IV. CONCLUSION

Even though the benefits of using Big Data technology are, by far, great and useful, it is clear that this way of data collection and data analysis violates privacy in all its segments. Laws do exist, however, regulations and provisions are not able to protect privacy and at the same time enable the free use of Big Data technology. Despite the fact that it is the individuals who share their data and who should be more responsible for what they do, they do not want to be the ones who think about such issues and expect legislative control of organizations which collect and process data and therefore it is expected from states to reinforce their activities in this area.

LITERATURA

[1] Big Data Just Beginning to Explode , http://www.csc.com/big_data/flxwd/83638big_data_just_beginning_to_explode_interactive_infographicJ. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] HT i TVbeat uvode IPTV BigDatu, http://www.bug.hr/vijesti/ht-tvbeat-uvode-iptv-bigdatu/133385.aspx

[3] Number of Internet Users (2014), http://www.internetlivestats.com/internet-users/M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[4] Theresa M.Payton & Theodore Claypoole: Privacy in the age of Big Data, Rowman&Littlefield, 2014.

[5] ictbussiness.info, http://www.ictbusiness.info/vijesti/postrozuju-se-pravila-zastite-privatnosti

[6] The Promise and Peril of Big Data http://citpsite.s3.amazonaws.com/events/big-data/Aspen-Big_Data.pdf, David Bollier, Rapporteur (2010)

[7] Big Data Now by O'Reilly Media, 2011.

[8] Research: 84 percent more concerned about security and privacy in 2015., Teena Hammond,

http://www.zdnet.com/article/research-84-percent-more-concerned-about-security-and-privacy-in-2015

[9] Luke Harding; " Snowden: Dosjei", EPH media, February 2014.

[10] The White House; Big Data: Seizing opportunities preserving values; May 2014.

[11] W.L. Perry, B. Mclnnis, C. C. Price, S.C. Smith, J.S. Hollywood; Predictive policing, The Role of Crime Forecasting in Law Enforcement Operations; RAND corporation 2013.

[12] Viktor Mayer-Schönberger and Kenneth Cukier – Big Data A Revolution That Will Transform How We Live, Work and Think, 2014

SECURITY AWARENESS AND EDUCATION TRACK

Impact of Security Education on Password Change

Viktor Taneski*, Marjan Heričko* and Boštjan Brumen*
* University of Maribor, Faculty of electrical engineering and computer science/Institute of Informatics, Maribor, Slovenia
viktor.taneski@um.si, marjan.hericko@um.si, bostjan.brumen@um.si

Abstract - Users and their passwords are the weakest link of security, because they tend to create passwords that are not strong, easy to remember and contain words that are familiar to them. Users' lack of security consciousness and their behavior can be influenced by security education. Our research explores the effect of password security training on the strength of the passwords chosen by the users. The objective of this paper is to explore possible differences in the quality of passwords between students from different faculties with different foreknowledge and education about passwords and security. In this paper we analyze the passwords of the students from the Faculty of Tourism and compare the results to the ones from the Faculty of electrical engineering and computer science at the University of Maribor. We collected the data by means of an online questionnaire, performed among undergraduate and graduate students. Despite our efforts to educate the users about the importance of frequent password change, a large percent of users did not change their passwords following the lectures. The overall conclusion of this preliminary study is that users still lack security knowledge regarding password change and need to be further educated in this direction.

I. INTRODUCTION

The combination of a username and password is the most common method for gaining a legitimate access to a computer system. Alternative authentication methods such as biometrics (based on "what you are"), smart cards (based on "what you have") and two-step verification are becoming more available, but passwords remain the most common method for authentication for computer systems, despite the fact that they suffer from a number of problems. Password-related problems are firstly identified by Morris and Thompson [1]. In their experiments they noticed that users of the system chose passwords that are short, simple and contained only lowercase letters and digits, or appeared in various dictionaries. 20 years later Zviran and Haga [2] concluded that one of the biggest vulnerabilities to a computer system's security is the user. Almost 50% of the users surveyed in their study reported passwords composed of five or fewer characters, 80% used only alphanumeric characters, and 80% never changed their password.

Many approaches and solutions, for increasing the security and usability of textual passwords, were proposed through the years but none have proven widely acceptable [3]. Users and their textual passwords are still considered "the weakest link", because they still tend to choose weak passwords and passwords that can be found in a dictionary [4], use a very small number of passwords and are likely to use words that are familiar to them as their passwords [5]. Users create the easiest-to-remember passwords regardless to any recommendations or instructions and tend to trade security for memorability [6], [7]. These problems can arise as a result of users' lack of security motivation and the fact that users tend to circumvent password restrictions for the sake of convenience [8]. This indicates a deficit in user's consciousness about security and this user behavior can be influenced by information security training [9]. Educating users about password security and assisting them with creating secure passwords can raise the security consciousness of system users and can help achieve greater security [2].

This study, which is research in progress, explores the recommendation of Zviran and Haga [2] by analyzing the effect of password security training on user's practices regarding password creation, frequent password changes and their consciousness about security and the importance of creating strong and hard-to-guess passwords. We performed a pilot study among undergraduate and postgraduate students from the Faculty of Tourism and the Faculty of electrical engineering and computer science at the University of Maribor. We investigated their password practices, attitudes and knowledge about password creation and password security. The survey consisted of two phases over one month period along with several lessons on password security between the two phases. The first phase of the survey was performed among students that had no password security training. The second phase of the survey was performed one month later on the same group of students. For a period of two weeks, between the two phases, the students were educated about the importance of password security and how to manage their passwords. Following the lectures there was a two-week period for learning decay. The contribution of this study in this manner is that we succeeded in educating the participants about some password practices, in contrast to some previous studies [10], [9] that have not proven to be quite successful.

The rest of the paper is organized as follows: In Section II we review and analyze related work, Section III describes the research method, Section IV presents the results of the survey and a discussion, while Section V concludes the paper.

II. Related Work

Research on passwords has been active for almost 35 years [3]. Zviran and Haga [2] in their paper on password security investigated in more details the characteristics of user-selected passwords and the relationships between key password characteristics. They used a questionnaire as an instrument to investigate users' password characteristics. After finding that user-selected passwords are relatively weak and easy to guess, Zviran and Haga propose a further investigation of the effectiveness of educating users about passwords security and raising their security consciousness.

There have been a few previous attempts to research the effect of education on users' password practices. Delber Hart in [10] presents the results of an experiment performed to determine the effect of a brief presentation of password security best practices on users. The methodology used was a survey, which was administered five times among different student groups along with some password education. The results of the study show that the education had no significant effect, even possibly had an opposite effect: students realized they do not want any more information about password security best practices.

The study by Horcher and Tejay [9] follows the recommendation by Zviran and Haga and examines the use of cognitive load theory to design the information security training on password strength. It is a study in progress involving employees in a chemical company with different level of technical expertise and work location. The preliminary results show that after twelve hours of cracking users' passwords, 93% of the passwords were compromised, using the dictionary of pre-computed passwords.

Our approach can be viewed as a combination of the aforementioned approaches. Our questionnaire is an improved version of the one created by Zviran and Haga in [2]. Furthermore, our education was conducted over a longer period of time compared to that of Hart in [10], and in addition to the password strength we also analyzed the general password characteristics and the password change frequency.

III. Method

A. The Survey

Data for this study were collected by means of an online questionnaire. The main objective of the questionnaire was to determine the characteristics of textual passwords used by individuals to log in, mainly to a university account, but also to other accounts, such as Facebook, Twitter, Google, or their own mobile devices and phones. Because asking users about their passwords is a sensitive topic, we did not ask them about their actual passwords, but rather about the characteristics of their passwords for different accounts. We classified the questions into six categories:

- Account information – which services or devices does the student have a password for (desktop computer, notebook computer, mobile phone, tablet, University account, Facebook account, Twitter account, Google account)
- Password information – the characteristics of the selected passwords (first character, length, how did the user choose the passwords etc.)
- Password recall – how often do users forget their passwords, how often do they change them and how often do they write their passwords down
- Frequency of password use – information about how often do users log in by using the password, and whether they use the password to log in to multiple accounts
- Data importance and data sensitivity – how important and how sensitive are the data to users they are protecting
- Demographic data about the user – gender, year of birth, university, faculty, postal code etc.

TABLE I. Number of Participants in the Study

Faculty	Initial results		Filtered results	
	Phase 1	Phase 2	**Phase 1**	**Phase 2**
Tourism	43	37	34	34
EE&CS[a]	33	30	30	30
Total	76	67	**64**	**64**

a. Electrical engineering and computer science

For creating our questionnaire, we drew inspiration from the questionnaire that was performed by Zviran and Hagga [2]. Our questionnaire has additional questions and additional answer options about password characteristics for different accounts for different services or devices, thus allowing us to compare characteristics of different passwords for different accounts. The completion of the questionnaire took between 15 and 20 minutes.

B. Procedure

The survey was conducted in the spring term of 2014 and consisted of two phases. The first phase of the questionnaire was performed among students that did not receive the lessons. After the first phase of the survey, the students attended lectures dedicated specifically to passwords and related issues, like password creation, management and security. The education consisted of topics about the importance of creating strong and secure passwords, how to choose such passwords and how to manage them (frequent changes, no writing down, no reusing one password for more than one different accounts etc.). After the education part followed a two-week period of learning decay, followed by the second phase of the survey. We slightly modified the questionnaire by adding an additional answer option "I prefer not to disclose" to every question in the "Password information" section. The goal is to see if there will be any significant difference in the general password characteristics and the frequency of password change, before and after the students attended the lectures.

TABLE II. SUMMARY OF COMMON PASSWORD CHARACTERISTICS – FACULTY OF TOURISM

Account	Common password characteristics			
	Phase 1	Count	Phase 2	Count
desktop account	lowercase letters only	4 (26.7%)	lowercase letters only	4 (26.7%)
notebook account	mixed case letters and 3 or more digits	3 (20.0%)	lowercase letters only	6 (37.5)%
	mixed case letters and 1-2 digits somewhere	3 (20.0%)	I prefer not to disclose	3 (18.8%)
mobile phone	digits only	11 (64.7%)	digits only	6 (37.5%)
			lowercase letters only	3 (18.8%)
tablet device	digits only	2 (50.0%)	digits only	2 (33.3%)
University account	lowercase letters only	5 (19.2%)	single case letters and 1-2 digits in front	4 (14.8%)
	mixed case letters and 1-2 digits somewhere	4 (15.4%)	I prefer not to disclose	5 (18.5%)
Facebook account	lowercase letters only	7 (22.6%)	mixed case letters, several special characters and several digits	4 (14.8%)
			I prefer not to disclose	5 (18.5%)
Twitter account	mixed case letters and 3 or more digits	1 (33.3%)	long sentence, mixed case	1 (25.0%)
			I prefer not to disclose	1 (25.0%)
Google account	lowercase letters only	7 (25.0%)	mixed case letters, several special characters and several digits	2 (10.0%)
			I prefer not to disclose	4 (20.0%)

C. Participants

The participants were undergraduate and postgraduate students from the Faculty of Tourism and the Faculty of electrical engineering and computer science at the University of Maribor. Table I presents the number of participants in the first and the second phase of the study. As shown in the Table I, a total of 76 students participated and completed the first phase of the web-based survey (43 students from the Faculty of Tourism and 33 from the Faculty of EE&CS). 67 of them also completed the second phase of the study (37 from the Faculty of Tourism and 30 from the Faculty of EE&CS). As we can observe from Table I, the number of participants is not equal in the first and second phase, which means that not all participants completed the two phases of the study. Having the same sample/participants in the two phases of the study is crucial in order to properly analyze the data and determine whether our education had any significant effect. For this matter, before we proceeded with analyzing the data, we excluded those participants that did not complete one of the two phases (either the first or the second phase). Thus, we used a total of 64 records, as legitimate answers for performing our analysis. The last two columns in Table I represent the number of total participants in the study, after we filtered the results. Ages of the participants range from 21 to 51 years for the Faculty of Tourism ($M = 28.50$; $M_d = 24.50$; $SD = 9.046$; a Shapiro-Wilk test of normality indicates a deviation from normality at $p << 0.05$) and from 21 to 26 years for the Faculty of electrical engineering and computer science ($M = 22.93$: $M_d = 23$; $SD = 1.172$; a Shapiro-Wilk test of normality indicates a deviation from normality at $p = 0.015$). Ten participants were male and 24 were female at the Faculty of Tourism, and 26 were male and 4 were female at the Faculty of electrical engineering and computer science.

IV. RESULTS

We were interested in the general password characteristics and password change frequency before and after the lectures. We were also interested in finding out whether there is any significant difference in the general password characteristics and the password change frequency between users' passwords from both faculties.

A. General Password Characteristics

Since asking about the actual users' passwords is a sensitive subject, we asked our participants about their password characteristics and did not inquire their actual passwords. The question that we asked in both phases was "What are the characteristics of your password?". The most common answers to the question for both phases and both faculties are summarized in Table II and Table III for both faculties, respectively. In the tables, the first column denotes the specific account, the second column represents the most common answers about password characteristics for every account in the first phase. The third and the last column represent the most common answers and the percentage of users who chose that specific answer in the second phase.

We were interested to investigate whether there is any significant difference in the general passwords characteristics for the accounts before and after the lectures, first for the Faculty of Tourism, and second for the Faculty of EE&CS. A before and after measurement of password characteristics was taken on a categorized ordinal scale, from 1 = "digits only" to 23 = "long sentence, digits and special characters" and 24 = "I prefer not to disclose". We ordered this scale by calculating the number of possible password combinations for every

TABLE III. SUMMARY OF COMMON PASSWORD CHARACTERISTICS – FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Account	Common password characteristics			
	Phase 1	Count	Phase 2	Count
desktop account	lowercase letters only	4 (28.6%)	mixed case letters and 1-2 digits somewhere	5 (31.3%)
			I prefer not to disclose	2 (12.5%)
notebook account	lowercase letters only	6 (23.1%)	mixed case letters and 1-2 digits somewhere	7 (25.9%)
			I prefer not to disclose	2 (7.4%)
mobile phone	digits only	12 (63.2%)	digits only	11 (57.9%)
tablet device	digits only	2 (66.7%)	digits only	2 (66.7%)
University account	single case letters and 1-2 digits behind	9 (33.3%)	mixed case letters and 1-2 digits somewhere	5 (17.9%)
			I prefer not to disclose	2 (7.1%)
Facebook account	mixed case letters and 3 or more digits	7 (25.0%)	mixed case letters, several special characters and several digits	6 (21.4%)
Twitter account	mixed case letters and 3 or more digits	4 (44.4%)	mixed case letters and 1-2 digits somewhere	3 (33.3%)
			I prefer not to disclose	2 (22.2%)
Google account	mixed case letters and 3 or more digits	7 (28.0%)	mixed case letters and 1-2 digits somewhere	4 (17.4%)
			I prefer not to disclose	2 (8.7%)

password category. For the purposes of this paper, we presumed that the average password length is approximately 9 characters, as stated in [11], a long sentence would have at least four words, and we have the English language and the English alphabet as a dataset (e.g. for a "one letter uppercase, other lowercase" type of password the number of possible combinations would be $26^1 \times 26^8 \times 9$).

The data for the Faculty of Tourism were mostly not normally distributed (Shapiro-Wilk test of normality indicated a deviation from normality for nearly every account password, except in the case of tablet account passwords where the data were normally distributed. Therefore, we performed non-parametric Wilcoxon Signed-rank test for dependent samples to compare the two dependent samples (before and after). In the case of tablet account passwords we performed a Paired samples t-test. The results indicate that there is a significant difference in password characteristics after the lectures for the mobile, University, Facebook and Google accounts (p-values: 0.043; 0.004; 0.013; 0.011 respectively, at significance level 0.05). If we take a look at Table II we can observe some changes in users' passwords for the mobile, University, Facebook and Google accounts, which we find significant. The passwords for the University, Facebook and Google account in the first phase mostly consist of lowercase letters only. The results from the second phase demonstrate that the passwords for these accounts are slightly stronger (according to our ordinal scale): consist either of single case letters and 1-2 digits in front, or mixed case letters, several special characters and several digits. We can also observe changes in the password characteristics for other accounts, but our statistical analysis shows that these changes are not statistically significant and may be due to chance alone.

The data for the Faculty of EE&CS were also not normally distributed (Shapiro-Wilk test of normality indicated a deviation from normality for every account passwords). As previously, we performed a Wilcoxon Signed-rank test for dependent samples to compare the two dependent samples before and after. The analysis shows that there is no significant difference (the significance level is 0.05) between password characteristics before and after the lectures, for none of the accounts. By looking at Table III we can, however, observe certain changes in the characteristics of users' passwords, but the statistical analysis shows that these changes are not significant. According to our ordinal scale of password characteristics, passwords for the desktop, notebook, University and Facebook account are slightly stronger in the second phase. But, this is not the case with the passwords for the Twitter and Google account where we can clearly notice that passwords in the second phase are weaker than the ones in the first phase.

We can provide at least two possible explanations for this unexpected behavior. First, there is a possibility that participants did not quite understand our lectures about what makes a good and strong password. The second explanation may lay in the fact that a large percent of the participants from both faculties, as we can notice in the both Table II and Table III, chose not to disclose any information about their password characteristics. It is most likely that after attending the lectures the awareness of users about the importance of their data and the importance of not disclosing information about the passwords for their accounts, increased; thus participants in the second phase deliberately falsified data about their password characteristics.

TABLE IV. FREQUENCY OF PASSWORD CHANGE – FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Accounts	Frequency of password change	Count
desktop account	never changed it since first use	5 (35.7%)
	less than once a year	5 (35.7%)
notebook account	less than once a year	11 (42.3%)
mobile device	less than once a year	8 (40.0%)
Tablet device	never changed it since first use	1 (33.3%)
	less than once a year	1 (33.3%)
University account	never changed it since first use	17 (63.0%)
Facebook account	less than once a year	11 (39.3%)
Twitter account	never changed it since first use	5 (55.6%)
Google account	less than once a year	7 (28.0%)
	up to three times a year	7 (28.0%)

TABLE V. FREQUENCY OF PASSWORD CHANGE – FACULTY OF TOURISM

Accounts	Frequency of password change	Count
desktop account	less than once a year	11 (64.7%)
notebook account	less than once a year	10 (62.5%)
mobile device	never changed it since first use	6 (35.3%)
	less than once a year	6 (35.3%)
Tablet device	less than once a year	2 (50.0%)
	four to six times a year	2 (50.0%)
University account	never changed it since first use	13 (48.1%)
Facebook account	less than once a year	16 (51.6%)
Twitter account	never changed it since first use	2 (66.7%)
Google account	less than once a year	13 (46.4%)

B. *Password Change Frequency*

We educated the participants about the importance of a frequent password change in order to ensure that a stolen password can not be used to compromise other passwords and accounts of other users in the system. The frequent password change is considered a basic security measure [2]. On the other hand, the rapid growth of the popularity of the Internet and the increased number of online accounts, can lead to users having to maintain many different passwords [11]. Because of this, the forced and too frequent password changes can have negative effect on the users [12]. Users may quickly forget which password restrictions apply to which system, which password is current, and this may lead to users tempting to write their passwords down or to reuse an old one. Despite such claims, the research in literature still supports frequent password changing in order to reduce predictability and suggests that a realistic time for password change should be 90-120 days [13].

We were interested in finding out whether our lectures and recommendations about frequent password changing had any effect on the participants and whether they changed their passwords after attending the lectures. For that matter, in the first phase of the questionnaire we asked our participants "How often do you change your password (when not required by the system?)", and in the second phase, following the lectures, we asked them "Have you changed your password since the last survey or not, and why?" The most common answers to the first question are summarized in both Table IV and Table V for the students from the Faculty of Tourism and the Faculty of EE&CS, respectively. By observing Table IV we can notice that many students at the Faculty of Tourism never change their passwords since the first use, or change them very rarely (less than once a year). We find a statistically significant difference between the answers to this question from the first and the second phase of the study, for the desktop (Wilcoxon Signed-rank, p=0.031), notebook (Wilcoxon Signed-rank, p=0.014), mobile (Wilcoxon Signed-rank, p=0.039), Facebook (Wilcoxon Signed-rank, p=0.023) and Google (Wilcoxon Signed-rank, p=0.012) accounts. This significant difference indicates that there are some changes and effects of the lecturing. When asked whether they changed their passwords since the first phase, some percent of the students answered that they did not change their passwords since the first phase (especially for the University account – 59.3%). On the other hand, overwhelming are the results about the students who changed their passwords after attending the lectures: 43.8% for the desktop computer, 43.8% for the notebook computer, 50.0% for the mobile phone, 33.3% for the tablet device, 44.4% for the Facebook account, 50.0% for the Twitter account, and 42.1% for the Google account. These students stated that they changed their passwords because they realized that the old passwords were very week and could compromise their other passwords and accounts.

Table V summarizes the most common answers to the question about the frequency of password change, for the students from the Faculty of EE&CS. The table shows that most of the students from this faculty change their passwords very rarely (less than once a year) or never change them since the first use, which was also the case with the password change frequency at the Faculty of Tourism. A Wilcoxon Signed-rank test for dependent samples shows no statistical difference (at significance level 0.05), between password change frequency before and after the lectures, for none of the accounts (desktop – p=0.248; notebook – p=0.560; mobile – p=0.850; tablet – p=0.317; University – p=0.431; Facebook – p=0.739; Twitter – p=0.705; Google – p=0.059). By analyzing the answers to the question "Have you changed your password since the last survey or not, and why?" we find that a large percent of the students (desktop 52.9%; notebook 57.1%; mobile 57.9%; tablet 66.7%; University 75.0%; Facebook 50.0%; Twitter 66.7%; Google 52.2%) reported that they have not changed their passwords after attending the lectures. However, we find some promising

results (desktop 35.3%; notebook 32.1%; mobile 26.3%; tablet 33.3%; University 21.4%; Facebook 28.6%; Twitter 22.2%; Google 30.4%) where students decided to change their passwords because they realized that the old one was very weak and could compromise their other passwords and accounts.

V. CONCLUSION

We investigated the effect of password security training on user's practices regarding passwords creation and password change frequency. We performed a pilot study with university students as participants to explore their password characteristics and habits regarding password creating, using, managing and security. We collected our data by means of an online questionnaire. The survey consisted of two phases. Between the two phases, the students received lectures about the importance of creating strong and secure passwords, how to choose such passwords and how to manage them.

We find significant changes and an improvement in the characteristics of the passwords for the University, Facebook and Google account at the Faculty of Tourism. The passwords for these accounts in the first phase mostly consist of lowercase letters only. The results from the second phase demonstrate that the passwords for these accounts are slightly stronger: consist either of single case letters and 1-2 digits in front, or mixed case letters, several special characters and several digits. However, we find no significant changes in the characteristics of users' passwords at the Faculty of EE&CS. We also noticed that a large percent of the participants from both faculties chose not to disclose any information about their password characteristics during the second phase. Our deduction is that the awareness of users about the importance of their data and the importance of not disclosing information about the passwords for their accounts, increased after attending the lectures.

The results regarding the frequent password change from the first phase at the Faculty of Tourism show that many students never change their passwords since the first use, or change them less than once a year. On the other hand, we find an increased percentage of students that changed their passwords after attending the lectures (43.8% for the desktop computer, 43.8% for the notebook computer, 50.0% for the mobile phone, 33.3% for the tablet device, 44.4% for the Facebook account, 50.0% for the Twitter account, and 42.1% for the Google account). These students stated that they changed their passwords because they realized that the old password was very weak and could compromise their other passwords and accounts. Our analisys also shows a simillar behaviour about the frequent password change at the Faculty of EE&CS: students very rarely change their passwords (less than once a year) or never change them since the first use. A large percent of the students reported that the have not changed their passwords after attending the lectures.

The lectures and recommendations had a positive effect regarding users' password characteristics and password change at the Faculty of Tourism, but not quite positive at the Faculty of EE&CS. Despite our efforts to educate the users about the importance of frequent password change, a large percent of users did not change their passwords following the lectures. The overall conclusion of this preliminary study is that students at the Faculty of EE&CS still underestimate security risks and need additional security knowledge. As paradox as it may seem, students studying computer science may have a false feeling that nothing can happen to their accounts. Lax security behavior regarding rare password change and password write-down, still exists, but compared to the results from the research made by Zviran and Haga [2], our the results show some improvements.

For future research we are interested in determining the reasons for the students behavior related to password characteristic (choosing passwords with lower quality in the second phase of the study) and low frequency of password change, even after attending our lectures. This work will serve as a starting point for our further research in this area where we want to determine how students choose their passwords, and what other factors can affect the choice of passwords (data importance, data sensitivity, log on frequency, field of study, etc.).

REFERENCES

[1] R. Morris and K. Thompson, "Password Security: A Case History," *Commun. ACM*, vol. 22, no. 11, pp. 594–597, Nov. 1979.

[2] M. Zviran and W. J. Haga, "Password Security: An Empirical Study," *J. Manag. Inf. Syst.*, vol. 15, no. 4, pp. 161–185, 1999.

[3] V. Taneski, M. Hericko, and B. Brumen, "Password security - No change in 35 years?," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, 2014, pp. 1360–1365.

[4] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley, "Does My Password Go Up to Eleven?: The Impact of Password Meters on Password Selection," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2379–2388.

[5] M. Bishop and D. V Klein, "Improving system security via proactive password checking," *Comput. Secur.*, vol. 14, no. 3, pp. 233–249, 1995.

[6] M. Zviran and W. J. Haga, "Cognitive passwords: The key to easy access control," *Comput. Secur.*, vol. 9, no. 8, pp. 723–736, 1990.

[7] M. Zviran and W. Haga J., "A Comparison of Password Techniques for Multilevel Authentication Mechanisms," vol. 36, no. 3, 1993.

[8] A. Adams, M. A. Sasse, and P. Lunt, "Making Passwords Secure and Usable," in *Proceedings of HCI on People and Computers XII*, 1997, pp. 1–19.

[9] A.-M. Horcher and G. P. Tejay, "Building a better password: The role of cognitive load in information security training," *Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on*. pp. 113–118, 2009.

[10] D. Hart, "Attitudes and Practices of Students Towards Password Security," *J. Comput. Sci. Coll.*, vol. 23, no. 5, pp. 169–174, 2008.

[11] G. Notoatmodjo and C. Thomborson, "Passwords and Perceptions," in *Proceedings of the Seventh Australasian Conference on Information Security - Volume 98*, 2009, pp. 71–78.

[12] M. A. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'Weakest Link' — a Human/Computer Interaction Approach to Usable and Effective Security," *BT Technol. J.*, vol. 19, no. 3, pp. 122–131, 2001.

[13] P. Cisar and S. M. Cisar, "Password – a Form of Authentication." 2007.

Empirical study on ICT system's users' risky behavior and security awareness

K. Solic*, T. Velki** and T. Galba***
*J.J. Strossmayer University, Faculty of Medicine, Osijek, Croatia
**J.J. Strossmayer University, Faculty of Education, Osijek, Croatia
***J.J. Strossmayer University, Faculty of Electrical Engineering, Osijek, Croatia
kresimir.solic@mefos.hr, tena.velki@gmail.com, tomislav.galba@etfos.hr

Abstract - In this study authors gathered information on ICT users from different areas in Croatia with different knowledge, experience, working place, age and gender background in order to examine today's situation in the Republic of Croatia (n=701) regarding ICT users' potentially risky behavior and security awareness. To gather all desired data validated Users' Information Security Awareness Questionnaire (UISAQ) was used.
Analysis outcome represent results of ICT users in Croatia regarding 6 subareas (mean of items): Usual risky behavior (x_1=4.52), Personal computer maintenance (x_2=3.18), Borrowing access data (x_3=4.74), Criticism on security in communications (x_4=3.48), Fear of losing data (x_5=2.06), Rating importance of backup (x_6=4.18). In this work comparison between users regarding demographic variables (age, gender, professional qualification, occupation, managing job position and institution category) is given.
Maybe the most interesting information is percentage of questioned users that have revealed their password for professional e-mail system (28.8%). This information should alert security experts and security managers in enterprises, government institutions and also schools and faculties.
Results of this study should be used to develop solutions and induce actions aiming to increase awareness among Internet users on information security and privacy issues.

I. INTRODUCTION

The importance of ICT system's users' knowledge and awareness about information security issues should be acknowledged when dealing with user's privacy and information security in general [1-3]. Users with potentially risky behavior can significantly affect overall security level of different information and communication systems [4-6].

Generally, main goal of empirical studies is to produce some new knowledge based on gathered data analysis. In that manner aim of this work was to produce some new conclusions about ICT system's users' knowledge, behavior and awareness regarding information security issues. For purpose of collecting data authors used previously validated Users' Information Security Awareness Questionnaire (UISAQ) [7]. Total of 701 participants included in this study were ICT users from different areas in Croatia with different knowledge, experience, working place, age and gender.

Other empirical studies with similar aim in their research examined only certain segments of ICT users' awareness or behavior, mostly focused on password usage and password quality [8-13]. However, by using UISAQ questionnaire when examining ICT users this empirical research covers wide range of awareness, knowledge and user's behavior. Additional quality of empirical research was succeeded by using UISAQ questionnaire as statistically validated measuring instrument.

The UISAQ questionnaire has two main scales and six subscales, each with five or six items. Associated abbreviations are used in further text and tables:

- Potentially Risky Behavior (PRB; k=17)
  - Usual Behavior (UB; k=6)
  - Personal Computer Maintenance (PCM; k=6)
  - Borrowing Accessing Data (BAD; k=5)
- Knowledge and Awareness (KA; k=16)
  - Security in Communications (SC; k=5)
  - Secured Data (SD; k=5)
  - Backup Quality (BQ; k=6)

These subscales describe user's behavior, knowledge and awareness. Participants should evaluate their agreement with the statement on a 5-point Likert-type scale where five means excellent from aspect of information security. At the end of UISAQ questionnaire there were two additional questions about behavioral security of users and one part with demographic data.

For statistical analysis in this work statistical software tool MedCalc 14.12.0 was used. Statistical significance, when comparing differences among groups was defined as $p<0.05$, using nonparametric tests Mann-Whitney U Test and Kruskal-Wallis Test with Bonferroni correction when needed.

II. PARTICIPANTS

In this research a paper version of UISAQ questionnaire for data collection was used. Sample was defined in a way to be as similar as possible to general ICT user in the Republic of Croatia restrictive to adult users, but covering different regions (Dalmatia, Slavonia,

Zagreb area), both rural and urban areas, both government institutions and business organizations and also including students, unemployed and retired users with different background knowledge and experience regarding information security issues.

Participants (n=701) in total were 32.0 ± 11.5 years old (arithmetic mean ± standard deviation), youngest participant was 18 and oldest was 66 years old. Among participants there was 61.6% of female participants and 31.4% of all participants were working at private sector meaning business organizations. Regarding professional qualifications most of the participants were with high education (masters) 36.8%, while there was similar percentage of those with high school (25.2%) and bachelor degree (24.5%). The 28.8% of all participants revealed their password for professional e-mail systems' access by writing it down on the questionnaire.

Analysis outcome of the whole sample represents average results of ICT users in Croatia regarding 6 subareas (mean of items): Usual risky behavior (x_1=4.52), Personal computer maintenance (x_2=3.18), Borrowing access data (x_3=4.74), Criticism on security in communications (x_4=3.48), Fear of losing data (x_5=2.06), Rating importance of backup (x_6=4.18).

III. COMPARISON RESULTS

In order to compare ICT users, authors made groups regarding gender, age, workplace in government or private sector, professional qualification, managing job position and regarding revealing password. Following results are more interesting part of total results, depending on existence of statistical significant difference.

Comparing results regarding gender are showing that female ICT users got generally better results, except in subscale "Usual Behavior" (Table 1). There is no statistically significant gender difference regarding password revealing (p=0.547, Chi-Square Test).

Comparing results regarding age, where four groups were defined, are showing that middle age and older ICT users got generally better results (Table 2). For analysis between each group was used Mann-Whitney U Test with Bonferroni correction (with p<0.0125) Analysis results have shown that youngest group of users is significantly different of all other groups in total and most other subscales (UISAQ, PCM, KA, SC) while significant difference among other three groups of users is found only regarding subscale "Personal Computer Maintenance".

TABLE I. GENDER DIFFERENCES IN USERS' INFORMATION SECURITY AWARENESS

gender	Male (n=269) x±SD	Female (n=432) x±SD	p*
UISAQ	3.68±0.37	3.71±0.32	0.237
PRB	4.17±0.40	4.15±0.34	0.409
UB	3.30±0.96	3.14±0.87	0.018
PCM	4.45±0.48	4.58±0.38	<0.001
BAD	4.75±0.38	4.72±0.39	0.050
KA	3.18±0.54	3.27±0.46	0.013
SC	3.33±0.83	3.57±0.81	<0.001
SD	2.06±0.84	2.05±0.76	0.734
BQ	4.16±0.72	4.20±0.65	0.667

* Mann-Whitney U Test

Regarding workplace of ICT users, working in government institutions or in private sector, results have shown that both groups of ICT users got similar results, except in subscale "Secured Data" where users working at private sector got significantly lower result (p=0.015; Mann-Whitney U Test with Bonferroni correction). However, ICT users that work in private sector significantly more often reveal their password (p<0.001, Chi-Square Test), 48.2% of them.

Results of comparison between groups of ICT users with different professional qualification have shown that participants with masters got total result and two subscales regarding behavior (UISAQ, PRB, UB) significantly better than all other groups (with p<0.001; Mann-Whitney U Test with Bonferroni correction) (Table 3). Most significant differences (with p<0.0125; Mann-Whitney U Test) were found between users with masters and users with high school (UISAQ, PRB, UB and SD) while users who attended gymnasium are more skeptical in securing data than users with high school (SD).

Results of comparison between groups of ICT users regarding managing job position have shown significant difference between top management and the rest of employees and also significant difference between employed and unemployed users (Table 4). Statistical analysis between each group (p<0.0125; Mann-Whitney U Test with Bonferroni correction) has shown that

TABLE II. AGE DIFFERENCES IN USERS' INFORMATION SECURITY AWARENESS

age	18-30 (n=166) x±SD	31-40 (n=206) x±SD	41-50 (n=190) x±SD	51-66 (n=139) x±SD	p*
UISAQ	3.58±0.36	3.73±0.31	3.74±0.35	3.73±0.32	<0.001
PRB	4.07±0.39	4.21±0.37	4.18±0.34	4.14±0.33	0.028
UB	3.25±0.78	3.40±0.90	3.16±0.93	2.90±0.95	0.105
PCM	4.30±0.50	4.48±0.42	4.61±0.33	4.76±0.27	<0.001
BAD	4.66±0.49	4.74±0.37	4.78±0.29	4.77±0.38	<0.001
KA	3.09±0.51	3.26±0.43	3.30±0.55	3.31±0.46	<0.001
SC	3.15±0.89	3.49±0.83	3.66±0.74	3.62±0.74	<0.001
SD	2.04±0.78	2.00±0.69	2.11±0.92	2.07±0.75	0.832
BQ	4.07±0.67	4.29±0.56	4.13±0.80	4.24±0.65	0.009

* Kruskal Wallis Test

TABLE III. DIFFERENCES IN SUBSCALES OF UISAQ REGARDING PROFESSIONAL QUALIFICATION OF USERS

professional qualification	High school (n=177) x±SD	Gymnasium (n=77) x±SD	Bachelor (n=172) x±SD	Masters (n=258) x±SD	p*
UISAQ	3.64±0.36	3.65±0.40	3.66±0.30	3.77±0.32	<0.001
PRB	4.13±0.38	4.07±0.40	4.08±0.35	4.24±0.34	<0.001
UB	3.05±1.02	3.00±0.90	3.03±0.80	3.46±0.83	<0.001
PCM	4.56±0.48	4.49±0.55	4.51±0.42	4.53±0.35	0.103
BAD	4.77±0.38	4.70±0.43	4.70±0.42	4.75±0.36	0.211
KA	3.15±0.49	3.24±0.60	3.24±0.46	3.30±0.48	0.090
SC	3.49±0.83	3.39±0.97	3.49±0.83	3.49±0.78	0.976
SD	1.93±0.81	2.17±0.78	1.99±0.76	2.13±0.81	0.005
BQ	4.03±0.85	4.15±0.73	4.22±0.59	4.27±0.58	0.088

* Kruskal Wallis Test

unemployed users are significantly different from all other groups in total and in several other subscales (UISAQ, PCM, KA, SC, BQ). Significant difference was also found between top and middle management regarding "Borrowing Accessing Data", while there was no difference found between each management group and employed users.

Comparing results between ICT users that did or did not reveal password for accessing professional e-mail system are shown in last table (Table 5). ICT users that revealed their password got significantly lower overall result and results for three subscales that examine "Potentially Risky Behavior" (UB, PCM and BAD). Also there is significant difference regarding age, where younger ICT users significantly more often reveal their password.

However, ICT users with lower level of education significantly more often reveal their password ($p<0.001$, Fisher's Exact Test).

IV. CONCLUSION

Some general conclusions about ICT user's behavior and awareness emerging from analysis results are:

- Female users are generally more careful and more skeptical comparing to their male colleagues;
- Regarding age difference, middle age and older ICT users got better results in total and in most of the subscales;
- ICT users that work in private sector significantly more often reveal their password;
- Comparison of users with different professional qualification has shown that participants with masters got overall result significantly better than other users. Most significant differences were found between users with masters and users with high school;
- Unemployed users got significantly lower results than all other groups in total and in several subscales, both regarding behavior and awareness. Significant difference between three groups of employed users was only regarding borrowing access data;
- Participants who did not reveal their password generally got better results than participants that did reveal their password.

Regarding gender, age and different professional qualification results are expected. However, top management participants achieved surprisingly well results, which is very important as that kind of ICT users is most often target in direct phishing hacker attacks.

Maybe the most interesting information is percentage of users that have revealed their password for professional e-mail system (28.8%) and they are working in private

TABLE IV. DIFFERENCES IN SUBSCALES OF UISAQ REGARDING MANAGING JOB POSITION OF USERS

job position	Top management (n=24) x±SD	Middle management (n=126) x±SD	Employee (n=495) x±SD	Unemployed (n=55) x±SD	p*
UISAQ	3.85±0.35	3.72±0.32	3.70±0.34	3.54±0.35	0.001
PRB	4.26±0.31	4.15±0.36	4.15±0.37	4.12±0.31	0.259
UB	3.47±0.79	3.17±0.93	3.18±0.90	3.37±0.92	0.151
PCM	4.43±0.37	4.58±0.34	4.55±0.43	4.20±0.48	<0.001
BAD	4.89±0.16	4.71±0.36	4.73±0.41	4.77±0.26	0.020
KA	3.44±0.56	3.28±0.44	3.25±0.49	2.96±0.53	<0.001
SC	3.63±0.86	3.65±0.73	3.50±0.82	2.87±0.81	<0.001
SD	2.25±0.75	2.02±0.65	2.06±0.82	2.03±0.87	0.470
BQ	4.46±0.43	4.18±0.65	4.20±0.69	3.98±0.71	0.012

* Kruskal Wallis Test

TABLE V. DIFFERENCES IN SUBSCALES OF UISAQ REGARDING USERS' PASSWORD REVEALING

password revealed	No (n=499) x±SD	Revealed (n=202) x±SD	p*
UISAQ	3.72±0.35	3.64±0.31	0.003
PRB	4.19±0.36	4.06±0.37	<0.001
UB	3.25±0.92	3.07±0.86	0.008
PCM	4.56±0.40	4.46±0.49	0.029
BAD	4.77±0.34	4.65±0.48	0.001
KA	3.25±0.51	3.21±0.46	0.530
SC	3.51±0.83	3.41±0.83	0.259
SD	2.04±0.79	2.08±0.80	0.461
BQ	4.20±0.68	4.14±0.68	0.134
Age	40.34±11.42	37.92±11.49	0.009

* Mann-Whitney U Test

sector significantly more often. This information should alert security experts and security managers in companies, government institutions and also schools and faculties.

There are few possible drawbacks of this study. It was not possible for authors to check out if the revealed password is true and active, and also there would be higher amount of revealed passwords if some employees in some departments were not warned in advance. Other recommendations for future studies would be additional questions in demographic section of UISAQ questionnaire and bigger sample size.

Results of this study should be used to develop solutions and induce actions aiming to increase awareness among Internet users on information security and privacy issues.

REFERENCES

[1] A. Tsohou, S. Kokolakis, M. Karyda and E. Kiountouzis, "Process-variance models in information security awareness research", Information Management & Computer Security, vol. 16, pp. 271-287, July 2008.

[2] S. Williams and S. Akanmu, "Relationship Between Information Security Awareness and Information Security Threats", IJRCM, vol.3, pp. 115-119, August 2013.

[3] P. Tasevski, "Methodological approach to security awareness", CyberSecurity for the Next Generation. (Politechnico di Milano, Italy), 11-12 December 2013.

[4] P. Puhakainen and M. Siponen, "Improving Employees' Compliance Through Information Systems Security Training: An Action Research Study", MIS Quarterly, vol. 34, pp. 757-778, December 2010.

[5] K. Solic, D. Sebo, F. Jovic and V. Ilakovac, "Possible Decrease of Spam in the Email Communication", Proceedings IEEE MIPRO, (Opatia), pp. 170-173, May 2011.

[6] K. Beckers, L. Krautsevich, and A. Yautsiukhin, „Analysis of Social Engineering Threats with Attack Graphs", Proceedings of the 3rd International QASA - Affiliated workshop with ESORICS, (Wroclow, Poland), September 2014.

[7] T. Velki, K. Solic and H. Ocevcic, "Development of Users' Information Security Awareness Questionaire (UISAQ) - Ongoing Work", Proceedings IEEE MIPRO, (Opatia), pp. 1417-1421, May 2014.

[8] A. Keszthelyi, "About Passwords", Acta Polytechnica Hungarica, vol.10, pp. 99-118, September 2013.

[9] K. Solic, H. Ocevcic and D. Blazevic, "Survey on Password Quality and Confidentiality"", Automatika, vol. 56, April 2015. (accepted for publication).

[10] A.G. Voyiatzis, C.A. Fidas, D.N. Serpanos and N.M. Avouris, "An Empirical Study on the Web Password Strength in Greece", 15th Panhellenic Conference on Informatics, (Kastonija Greece), pp. 212-216, September-October 2011.

[11] M. Dell'Amico, P. Michiardi and Y. Roudier, "Password Strength: An Empirical Analysis", Proceedings IEEE INFOCOM, (San Diego, CA) pp. 1-9, March 2010.

[12] Ma Wanli, J. Campbell, D. Tran and D. Kleeman, "Password Entropy and Password Quality", 4th International Conference on Network and System Security, (Melbourne, VIC), pp. 583-587, 1-3, September 2010.

[13] P.G. Kelley, S. Komanduri, M.L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L.F. Cranor and J. Lopez, "Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms", IEEE Symposium on Security and Privacy, (San Francisco, CA), pp. 523-537, May 2012.

Moore's Curse on Textual Passwords

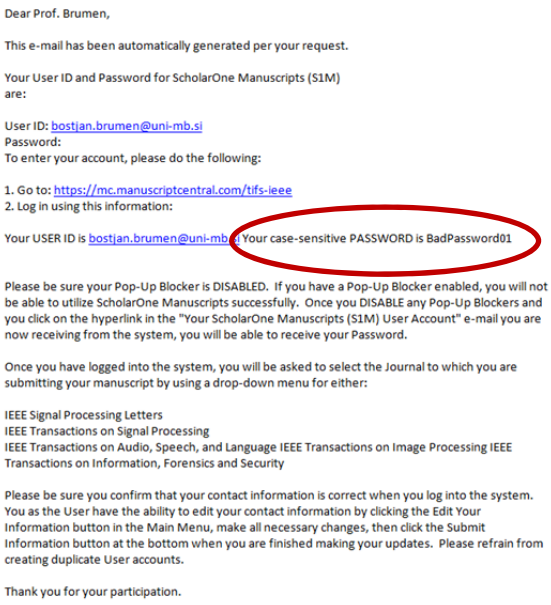
Boštjan Brumen* and Viktor Taneski *
* University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, Si-2000 Maribor, Slovenia
bostjan.brumen@uni-mb.si

Abstract - Background: Passwords are still the predominant way of authentication in information systems, and are mostly at user's responsibility. They conceive, use, re-use, abuse and forget passwords. In absence of strict password policies and at minimum required user training, passwords tend to be short, easy to remember, connected to the user's personal or professional life and consequently easy to break. The additional problem with passwords is their aging: Moore's law is affecting the available computing power to crack passwords and those deemed secure today may easily be broken in the near future. Objective: The aim of this paper is to study various scenarios of the effect the Moore's law is having on passwords and their security. In addition, advancements in other fields, e.g. quantum computing and Internet of Things, are taken into the account. Method: We analyzed various password types and the lengths required to withstand an off-line brute-force attack. The analysis was performed under various scenarios and combinations thereof: the Moore's law will continue to be in the effect for years to come with varying parameters, quantum computing will become feasible, improvements in hash tables computations will speed up the cracking process, and others. Results: The paper shows the minimum password length in characters for each password type under various scenarios. Even the most optimistic scenario shows that the minimum required password length today should be of 11 randomly drawn characters, rendering most of the passwords inappropriate due to their poor memorability. Conclusion: The current textual passwords are cursed by the Moore's law and other advancements in the field. Soon, classical textual passwords will need to be replaced by other mechanisms, which are, fortunately, already emerging.

I. Introduction

The ability to authenticate the identity of a user via passwords continues to be basic and the most relied-upon security mechanism in information systems. Weak passwords have led to very serious breaches, exposing millions of users and/or causing billions in damages. The history of password-related problems pre-dates the seminal paper written Morris and Thompson in 1979 [1] – it goes way back to mid-1960s and to the CTSS operating system exposing all the passwords as a daily welcome message [2]. A comprehensive list of breaches since 2005 can be found at [3].

Sadly, with all the years passing and evidence against bad password practices piling up, such practices continue. For example, the ScholarOne Manuscript system[1], used by many IEEE societies and journals, is storing passwords in plaintext. To upload a manuscript, a user must first create an account by selecting a username and a password. If the password is forgotten, the system is able to re-send it back to the user, meaning the system stores the passwords in original plaintext, or in best case it stores them encrypted but still able to decrypt upon request.



Dear Prof. Brumen,

This e-mail has been automatically generated per your request.

Your User ID and Password for ScholarOne Manuscripts (S1M) are:

User ID: bostjan.brumen@uni-mb.si
Password:
To enter your account, please do the following:

1. Go to: https://mc.manuscriptcentral.com/tifs-ieee
2. Log in using this information:

Your USER ID is bostjan.brumen@uni-mb.si Your case-sensitive PASSWORD is BadPassword01

Please be sure your Pop-Up Blocker is DISABLED. If you have a Pop-Up Blocker enabled, you will not be able to utilize ScholarOne Manuscripts successfully. Once you DISABLE any Pop-Up Blockers and you click on the hyperlink in the "Your ScholarOne Manuscripts (S1M) User Account" e-mail you are now receiving from the system, you will be able to receive your Password.

Once you have logged into the system, you will be asked to select the Journal to which you are submitting your manuscript by using a drop-down menu for either:

IEEE Signal Processing Letters
IEEE Transactions on Signal Processing
IEEE Transactions on Audio, Speech, and Language IEEE Transactions on Image Processing IEEE Transactions on Information, Forensics and Security

Please be sure you confirm that your contact information is correct when you log into the system. You as the User have the ability to edit your contact information by clicking the Edit Your Information button in the Main Menu, make all necessary changes, then click the Submit Information button at the bottom when you are finished making your updates. Please refrain from creating duplicate User accounts.

Thank you for your participation.

Fig. 1: A ScholarOne system message on lost password request

Because the "lost" password is sent by a plaintext e-mail, it becomes vulnerable in the system of origin, in all the passing systems, and in the end-user's e-mail box system and probably also on the user's hard drive. Such a password management system is very vulnerable to all kinds of attacks.

In such cases when a password is used for an on-line system with a lockout policy (e.g. "three times and you are out"), the stored (and later compromised) user credentials (username/email, password) become available offline for attacks without lockouts. The transition from an on-line password to an off-line one is very common [4].

Based on the survey by Brown et al [5], almost all respondents reuse passwords, and about two thirds of password uses are duplications. Passwords have been forgotten by a third of respondents, and over half keep a written record of them. Despite all the warnings and research in the last 50 years users continue to have bad passwords and practices and are thus easiest targets [6-10], while passwords continue to be the predominant way to authenticate the identity of a user.

[1] A product by Thompson Reuters, http://scholarone.com/products/manuscript/

Without the mentioned problems, yet another one is slowly but surely dooming the textual passwords: the effect of Moore's law on computing speed and consequently on password cracking.

A. Aims and Scope

In this paper, our research goal is to find out the length of a randomly generated password to withstand the brute force attacks using future computing power. Namely, passwords may be long enough today to protect valuable information, but what about in 10 years and beyond? The scientific contribution of this paper is to give guidelines for the password selection and evaluation criteria, e.g. for the password meters.

II. Related Work

Typically, a user is authenticated based on one of the three underlying principles (or combinations thereof): "what you know", "what you are" and "what you have" [11, 12]. "What you know" is based on secrecy known only to authorized user, "what you are" is based on user's physical characteristics (e.g. retina image, fingerprint) and is better known by the term biometrics, whereas "what you have" is based on smart cards or other tokens.

The passwords that are generated based on "what you know" principle can be divided into textual passwords and graphical passwords [13, 14]. In our contribution we focus on textual passwords as graphical ones require a different user interface for entering passwords [14] and are thus not directly comparable to our study. In addition they require a redesign of user interface for entering a password [14].

There are several alternatives to (nearly) random long textual password:

- passwords based on personal data or characteristics (e.g. birthdate, names, pets, addresses, etc.) [1, 15, 16].
- cognitive passwords require user to answer a randomly selected set of personal questions which only an authorized user can answer correctly [17]
- pass-sentences and pass-phrases which are composed of long, grammatically correct phrases [18, 19].
- randomly generated but pronounceable passwords made up of concatenated pronounceable syllables [20, 21].
- mnemonic passwords are those where a user chooses a memorable phrase and uses a character – often the first letter – to represent each word in the phrase [17, 22].

Each of these passwords types has several security issues, as discussed in the relevant references.

When protecting the data using passwords, one needs to observe one of the underlying principles of security in computing: "*Principle of Adequate Protection: Computer items must be protected only until they lose their value. They must be protected to a degree consistent with their value*." [12]. Namely, different data have different life spans. For example, stock exchange data have a short life span between their creation and publication date; for medical data this time span easily reaches 60 or more years. These data need to be protected even when processed, as discussed in [23-26].

It is to notice that present paper deals only with randomly generated passwords; further detailed studies are needed to make comparisons between each of password types with respect to issues covered in this paper.

III. Moore's Law and Other Effects on Password Security

The general rule is that passwords strength is proportional to their length and the size of the symbols space – i.e. the pool from where the password characters can be used from. This is true only if the process of the character selection is truly random [11]!

The number of possible combinations in a random password can be expressed by

$$s=b^{le}, \tag{1}$$

where s is the total number of combinations, b is the total number of symbols in a domain and le is the length of the password in number of characters.

The longer the password, the stronger and more resilient it is to the attacks. The speeds for unlocking a plaintext password from a given (unsalted) hash value are today about $v=10^9$ combinations/second [27-30], up to 10^{12} combinations/second [11].

Because the Moore's Law [31] is still doubling the computing power roughly every 18 months, we can expect the speeds to increase in 10 years by a factor of $2^{(10*12/18)} = 2^{6,67} \approx 100$. The increase factor in 60 years is amazing 1.099.511.627.776, or 10^{12}; the computers in 60 years from now are assumed to be one trillion-times faster than those today. Although Moore's law was doomed decades ago [32], recent studies show that it will still be in force for a few coming decades [33].

Other improvements in the field of computer science, besides improvements in hardware, make password cracking easier. The most important improvements are quantum computing [34, 35], improvements in parallel and GPU farming computations [36], improvements in hash computations [37] and applying Markov model preprocessing and filtering [38]. All these improvements have had and will have an impact on the password cracking process, while new improvements are, obviously, on the way.

Quantum computing is already a reality [39]. We can compare its state-of-the-art as being in late 1930s and early 1940s for computers based on Von Neumann's architecture. If today's quantum computers are not yet faster than conventional, we can assume their speed growth will be following the Moore's law too. In our model, we suppose that quantum computing will be generally available in 20 years from now [34] and its leap in password security will be one thousand-fold [34, 35], with standard exponential increase in eighteen months.

Other advancements in the field, e.g. construction of graphical processing units (GPU) farms, have further

increased the cracking speeds. It was shown that GPUs can speed up the cracking by a factor of 100 compared to the conventional systems; in addition, GPUs as electrinic circuits follow the Moore's law. Next, by using the Markov models and filters in modelling the password probability, the speed-up improvements are again substantial. For the purpose of our model we asume that all such improvements in the field (aside the hard computing power increase due to hardware improvements) will cumulatively increase the speed of password cracking by 10 % every year.

A user, when creating a password, needs to observe one the following principle: (1) a password is considered secure if the cost of breaking it exceeds the value of the protected information, or (2) the time required to break the password exceeds the lifetime of the protected data [11]. Today, we cannot rely on the costs because the ratio of bits processed per unit of currency is continuously sinking making the password cracking machine affordable [27, 29, 30]. One can only increase the time required to break a password. Thus let us assume a typical 60-year time span of usefulness of stored personal information [23, 25, 26, 40, 41].

Under the mentioned assumptions we calculated the time needed to crack various passwords using brute-force technique today and in a time line from today on. We assume that data have a useful life span of 60 years [23].

In other words, a safe password today (in 2015) would be made of at least ten characters from upper and lowercase letters, numbers and symbols, but it would not be safe in 10 years from now.

For the purpose of modelling, we asume that users use characters from the symbol space of approximately 50 when no possword creation policy is in place (i.e. uppercase and lowercase letters) and from 80 characters when such a policy is enforced (uppercase & lowercase letters, at least one digit and a symbol). Table I lists the models and their descriptions.

TABLE I. LIST OF MODELS

Model No	*Description*
1	Moore's law, 50 characters in symbol space
2	Moore's law, 80 characters in symbol space
3	Same as model 1 plus improvements in the field add 10 % annually
4	Same as model 2 plus improvements in the field add 10 % annually
5	Same as model 3 plus quantum computing
6	Same as model 4 plus quantum computing

For each model we first calculated the number of combinations using different password lengths and the then the worst time required to crack them using brute-force attack.

For each of the model the equations are as follows:

Model 1: $$t_i = \frac{b_1^{le}}{cs \cdot 2^{\frac{i*12}{18}}} \quad (2)$$

Model 2: $$t_i = \frac{b_2^{le}}{cs \cdot 2^{\frac{i*12}{18}}} \quad (3)$$

Model 3: $$t_i = \frac{b_1^{le}}{cs \cdot 2^{\frac{i*12}{18}} \cdot f^i} \quad (4)$$

Model 4: $$t_i = \frac{b_2^{le}}{cs \cdot 2^{\frac{i*12}{18}} \cdot f^i} \quad (5)$$

Model 5: $$t_i = \frac{b_1^{le}}{cs \cdot 2^{\frac{i*12}{18}} \cdot f^i \cdot q}; q = \begin{cases} 10^3; i \geq 20 \\ 1 \end{cases} \quad (6)$$

Model 6: $$t_i = \frac{b_2^{le}}{cs \cdot 2^{\frac{i*12}{18}} \cdot f^i \cdot q}; q = \begin{cases} 10^3; i \geq 20 \\ 1 \end{cases} \quad (7)$$

where *cs*, the initial cracking speed, equals to 10^9 passwords / second and b_1 and b_2 are set to 50 and 80, respectively. As mentioned, other improvements add 10 % each year (f=1,1). Another assumption is that (general-purpose and publicly available) quantum computing will be a fact in 20 years (parameter q).

IV. RESULTS

We have calculated the time needed to break a password for each of the models. The results of calculations for Model 6 are presented in Table II (for reasons of space other detailed results are not shown).

From the Table II can be seen (see row 1: "year 0" and column 1: "length 10") that a password consisting of 10 randomly drawn characters from a pool of 80 (lower and upper case letters, digits and symbols) is considered safe. With today's computing power it takes approximately 340 years to brute-force attack it. However, in only 5 years from now the computing power will increase (based on the assumptions built in the Model 6) and it will take only 25 more years to break such a password. In other words, if someone waits for 5 years and utilizes then-available computing power, the password cracking will be done in 30 years from now; for 20-year wait it will be done in seconds.

TABLE II. RESULTS OF CALCULATIONS FOR MODEL 6

length / years	10	11	12	13	14	15	16	17	18	19	20	21	22
0	340,5	27238,5	2,2E+06	1,7E+08	1,4E+10	1,1E+12	8,9E+13	7,1E+15	5,7E+17	4,6E+19	3,7E+21	2,9E+23	2,3E+25
5	20,97	1678,0	1,3E+05	1,1E+07	8,6E+08	6,9E+10	5,5E+12	4,4E+14	3,5E+16	2,8E+18	2,3E+20	1,8E+22	1,4E+24
10	1,29	103,4	8269,51	6,6E+05	5,3E+07	4,2E+09	3,4E+11	2,7E+13	2,2E+15	1,7E+17	1,4E+19	1,1E+21	8,9E+22
15	0,08	6,37	509,43	40754,26	3,3E+06	2,6E+08	2,1E+10	1,7E+12	1,3E+14	1,1E+16	8,5E+17	6,8E+19	5,5E+21
20	4,9E-06	3,9E-04	0,03	2,51	200,85	1,6E+04	1,3E+06	1,0E+08	8,2E+09	6,6E+11	5,3E+13	4,2E+15	3,4E+17
25	3,0E-07	2,4E-05	1,9E-03	0,15	12,37	989,83	7,9E+04	6,3E+06	5,1E+08	4,1E+10	3,2E+12	2,6E+14	2,1E+16
30	1,9E-08	1,5E-06	1,2E-04	9,5E-03	0,76	60,98	4878,12	3,9E+05	3,1E+07	2,5E+09	2,0E+11	1,6E+13	1,3E+15
35	1,1E-09	9,2E-08	7,3E-06	5,9E-04	4,7E-02	3,76	300,51	2,4E+04	1,9E+06	1,5E+08	1,2E+10	9,8E+11	7,9E+13
40	7,1E-11	5,6E-09	4,5E-07	3,6E-05	2,9E-03	0,23	18,51	1480,98	1,2E+05	9,5E+06	7,6E+08	6,1E+10	4,9E+12
45	4,4E-12	3,5E-10	2,8E-08	2,2E-06	1,8E-04	0,01	1,14	91,23	7298,65	5,8E+05	4,7E+07	3,7E+09	3,0E+11
50	2,7E-13	2,1E-11	1,7E-09	1,4E-07	1,1E-05	8,8E-04	0,07	5,62	449,62	3,6E+04	2,9E+06	2,3E+08	1,8E+10
55	1,7E-14	1,3E-12	1,1E-10	8,5E-09	6,8E-07	5,4E-05	4,3E-03	0,35	27,70	2215,84	1,8E+05	1,4E+07	1,1E+09
60	1,0E-15	8,1E-14	6,5E-12	5,2E-10	4,2E-08	3,3E-06	2,7E-04	0,02	1,71	136,50	1,1E+04	8,7E+05	7,0E+07

For these reasons passwords of length 10 are not adequate for future use. The same hold true for all passwords shorter than 19 characters (in model 6). With 19 characters and 60 years from now it will still take 136 years to break such a password. "Safe" password lengths for the given time frames are shaded.

It is interesting to notice the quantum leap that is modelled to happen in 20 years from now: it will take 500 years to break a 12-character password with computing power in 15 years from now on, but only roughly two weeks in 20 years from now.

Table III summarizes the calculations for all the models. It can be seen that required password lengths today are 10 to 11 characters, randomly drawn. In 10 years, they will need to be 2 characters longer, in 30 years approximately 4-5 characters longer and in 60 years their required length will need to be double of today's length. Today, a typical password has (only) 8 characters [42].

TABLE III. COMPARISON OF RANDOM PASSWORDS RESILLIENCE TO BRUTE FORCE ATTACKS

Model	*Required length today*	*Required length in 10 years*	*Required length in 30 years*	*Required length in 60 years*
1	11	12	14	18
2	10	11	12	16
3	11	13	14	20
4	10	11	13	18
5	11	13	16	22
6	10	11	14	19

The table shows that what is safe today will not be safe tomorrow [11], this especially holds true for off-line files which will remain unchanged for years, containing interesting medical data about an important person. The breach of privacy of data in such a file has a timelessness property [23] – no matter when in the lifetime of data the breach happens, it poses a problem.

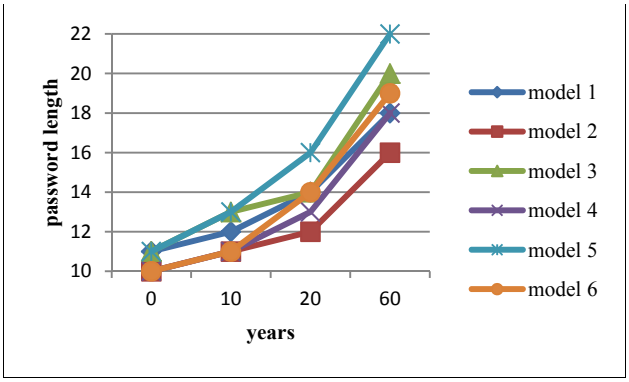


Fig. 2: Required password length with respect to time

The graph in Fig. 1 depicts the dependency of the required password lengths and the time.

V. DISCUSSION AND CONCLUSION

Egelman et al [43] have calculated that 10-character passwords are required today to be safe tomorrow (i.e in 10 years) as well. This is somehow true; however, the value of data has a longer time span. Computing power will continue to increase exponentially and what is safe tomorrow due to unavailable computing power can easily be affordable in 10 years from now. In the most pessimistic scenario regarding the power increase we need to use 16 random characters in password (base 80), and in the most optimistic one 20 or 22 for base 50 and 80, respectively. For example, the following password consists of 20 characters: "Q€Tu!3%7|rZu2$6(oPŁL".

This amount of characters to remember constitutes a much larger corpus than is the capacity of human memory, the well-known 7±2 principle [44]. Human memory in addition is temporally limited (short-term) when it comes to memorizing sequences [45] and has a limited capacity with a fixed final error rate [46-48]. For

this reason good passwords that are consisting of an abundant number of randomly selected characters are doomed: the users will either forget them [49] or write them down, or both [50-52]. In this paper we have shown that already today's password lengths are inadequate and that tomorrow's speed improvements will make security in computing even worse. What are the alternatives?

For one, special cognitive methods should be used to compose passwords. The required length of cognitively-generated passwords exceeds the randomly generated ones for a comparable resilience; however, they are much easier to remember [23]. But, users will have to be educated about proper password generation and management very early, possibly already in the primary school, but certainly not later than in the first year of secondary education.

Secondly, the industry will have to start replacing textual passwords as the predominant way of authentication with other means, such as keys [53], two-phase authentication and biometric. But, each of these replacements has some disadvantages. Common to all is the possibility to disclose the user's real identity [24].

ACKNOWLEDGMENT

The work was partially supported by the Slovenian Research Agency under grant number P2-0057..

REFERENCES

1. Morris R and Thompson K. Password security: A case history. Commun. ACM 1979; 22(11): 594-597.
2. Corbató FJ. On building systems that will fail. Communications of the ACM 1991; 34(9): 72-81. DOI: DOI: 10.1145/114669.114686.
3. PRC. Privacy Rights Clearinghouse. Chronology of Data Breaches. Security Breaches 2005 - Present 2014; Available from: http://www.privacyrights.org/data-breach. Accessed: 2014-10-20. (Archived by WebCite® at http://www.webcitation.org/6AfYySWJK).
4. El Emam K, Moreau K and Jonker E. How Strong are Passwords Used to Protect Personal Health Information in Clinical Trials? J Med Internet Res 2011; 13(1): e18. DOI: 10.2196/jmir.1335.
5. Brown AS, Bracken E, Zoccoli S and Douglas K. Generating and remembering passwords. Applied Cognitive Psychology 2004; 18(6): 641-651. DOI: DOI 10.1002/acp.1014.
6. Adams A and Sasse MA. Users are not the enemy. Communications of the ACM 1999; 42(12): 40-46. DOI: DOI: 10.1145/322796.322806.
7. Adams A, Sasse MA and Lunt P. Making passwords secure and usable. People and Computers XII, chapter 1. . London: Springer London; 1997. p. 1-19. ISBN: 3540761721
8. Notoatmodjo G. Exploring the 'Weakest Link': A Study of Personal Password Security. MSc Thesis. The University of Auckland, New Zealand; 2007.
9. Sasse MA, Brostoff S and Weirich D. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. BT technology journal 2001; 19(3): 122-131. DOI: DOI: 10.1023/A:1011902718709.
10. Tam L, Glassman M and Vandenwauver M. The psychology of password management: a tradeoff between security and convenience. Behaviour & Information Technology 2009; 29(3): 233-244. DOI: DOI 10.1080/01449290903121386.
11. Stallings W. Cryptography and Network Security: Principles and Practices. 4th ed.Upper Saddle River, NJ: Prentice-Hall; 2006. ISBN: 0131873164
12. Pfleeger CP and Pfleeger SL. Security in computing. 3rd ed.Upper Saddle River, NY, USA: Prentice Hall PTR; 2003. ISBN: 0130355488
13. Davis D, Monrose F and Reiter MK. On User Choice in Graphical Password Schemes. USENIX 2004 Security Symposium; 2004.
14. Suo X, Zhu Y and Owen GS. Graphical passwords: A survey. in 21st Annual Computer Security Applications ConferenceTucson, AZ, USA: IEEE; 2005. p. 463-472.
15. Zviran M and Haga WJ. Password security: an empirical study. Journal of Management Information Systems 1999; 15: 161-186.
16. FIPS. PUB 112 Password Usage. National Institute of Standards and Technology; 1985. ISBN:
17. Kuo C, Romanosky S and Cranor LF. Human selection of mnemonic phrase-based passwords. Proceedings of the second symposium on Usable privacy and security, July 12-14; 2006. Carnegie Mellon University, Pittsburgh, PA, USA: ACM.
18. Spector Y and Ginzberg J. Pass-sentence- a new approach to computer code. Computers & Security 1994; 13(2): 145-160. DOI: DOI: 10.1016/0167-4048(94)90064-7.
19. Brostoff AS. Improving Password System Effectiveness. PhD Thesis, Department of Computer Science, University College London. London, UK: University of London; 2004.
20. Gasser M. A Random Word Generator for Pronounceable Passwords, MTR-3006. ESD-TR-75-97, AD-A017676. Bedford, Mass: MITRE Corp.; 1975. ISBN:
21. Ganesan R, Davies C and Atlantic B. A new attack on random pronounceable password generators. Proceedings of the 17th {NIST}-{NCSC} National Computer Security Conference; 1994. Baltimore, MD, USA.
22. Zviran M and Haga WJ. Cognitive passwords: the key to easy access control. Computers & Security 1990; 9(8): 723-736. DOI: DOI: 10.1016/0167-4048(90)90115-A.
23. Brumen B, Heričko M, Rozman I and Hölbl M. Security Analysis and Improvements to the PsychoPass Method. Journal of Medical Internet Research 2013; 15(8): e161. DOI: DOI 10.2196/jmir.2366; PMID: 23942458.
24. Brumen B, Heričko M, Sevčnikar A, Završnik J and Hölbl M. Outsourcing Medical Data Analyses: Can Technology Overcome Legal, Privacy, and Confidentiality Issues? J Med Internet Res 2013; 15(12): e283. DOI: 10.2196/jmir.2471.
25. Brumen B, Welzer T, Družovec M, Golob I and Jaakkola H. Protecting medical data for analyses. Fifteenth IEEE Symposium on Computer-Based Medical Systems; 2002. Maribor, Slovenia; 4 June 2002 through 7 June 2002: IEEE Computer Society Press.
26. Brumen B, Welzer T, Družovec M, Golob I, Jaakkola H, Rozman I and Kubalík J. Protecting medical data

for decision-making analyses. Journal of Medical Systems 2005; 29(1): 65-80.
27. Hunt T. Our password hashing has no clothes. http://www.troyhunt.com/2012/06/our-password-hashing-has-no-clothes.html. Accessed: 2014-10-20. (Archived by WebCite® at http://www.webcitation.org/6AobEryR5) 2012; Available from: http://www.troyhunt.com/2012/06/our-password-hashing-has-no-clothes.html. Accessed: 2014-01-30. (Archived by WebCite® at http://www.webcitation.org/6AobEryR5).
28. InsidePro. Password Recovery Software. http://www.insidepro.com/. Accessed: 2014-08-30. (Archived by WebCite® at http://www.webcitation.org/6AoaUZsx3) 2012; Available from: http://www.insidepro.com/. Accessed: 2014-01-30. (Archived by WebCite® at http://www.webcitation.org/6AoaUZsx3).
29. Kingsley-Hughes A. Cheap GPUs are rendering strong passwords useless. in ZDnet.com, June 1, 2011 Available from http://www.zdnet.com/blog/hardware/cheap-gpus-are-rendering-strong-passwords-useless/13125. Accessed: 2014-08-30. (Archived by WebCite® at http://www.webcitation.org/6Aob1QCot); 2011.
30. Zonenberg A. Distributed hash cracker: A cross-platform gpu-accelerated password recovery system. Rensselaer Polytechnic Institute, Troy, NY 2009: 27.
31. Moore GE. Cramming more components onto integrated circuits Electronics 1965; 38(8): 114-117.
32. Kish LB. End of Moore's law: thermal (noise) death of integration in micro and nano electronics. Physics Letters A 2002; 305(3–4): 144-149. DOI: http://dx.doi.org/10.1016/S0375-9601(02)01365-8.
33. Robison RA. Moore's Law: Predictor and Driver of the Silicon Era. World Neurosurgery 2012; 78(5): 399-403. DOI: http://dx.doi.org/10.1016/j.wneu.2012.08.019.
34. Ladd TD, Jelezko F, Laflamme R, Nakamura Y, Monroe C and O/'Brien JL. Quantum computers. Nature 2010; 464(7285): 45-53. DOI: DOI 10.1038/nature08812. PMID: 20203602.
35. Elmannai W, Elleithy K, Pande V and Geddeda E. Quantum security using property of a quantum wave function. Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island; 2014. DOI: 10.1109/LISAT.2014.6845217.
36. Lee C-C, Liu C-H and Hwang M-S. Guessing Attacks on Strong-Password Authentication Protocol. International Journal of Network Security 2013; 15(1): 64-67.
37. Feldmeier DC and Karn PR. Unix password security-ten years later. Advances in Cryptology—CRYPTO'89 Proceedings; 1990. Springer.
38. Marechal S. Advances in password cracking. Journal in Computer Virology 2008; 4(1): 73-81. DOI: 10.1007/s11416-007-0064-y.
39. Hsu J. D-Wave's year of computing dangerously. Spectrum, IEEE 2013; 50(12): 11-13.
40. Brumen B, Golob I, Welzer T, Rozman I, Družovec M and Jaakkola H. Data protection for outsourced data mining. INFORMATICA (Ljubljana) 2002; 26(2): 205-210.
41. Brumen B, Golob I, Welzer T, Rozman I, Družovec M and Jaakkola H. An Algorithm for Protecting Knowledge Discovery Data. Informatica 2003; 14(3): 227-288.
42. Dell'Amico M, Michiardi P and Roudier Y. Password strength: an empirical analysis. INFOCOM, 2010 Proceedings IEEE; 2010. IEEE.
43. Egelman S, Sotirakopoulos A, Muslukhov I, Beznosov K and Herley C. Does my password go up to eleven?: the impact of password meters on password selection. Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems; 2013. April 27 - May 2, Paris, France: ACM.
44. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review 1956; 63(2).
45. Johnson GJ. A distinctiveness model of serial learning. Psychological review 1991; 98(2): 204-217. DOI: DOI: 10.1037/0033-295X.98.2.204.
46. Brumen B, Golob I, Jaakkola H, Welzer T and Rozman I. Early assessment of classification performance. Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation-Volume 32; 2004. Australian Computer Society, Inc.
47. Brumen B, Hölbl M, Harej Pulko K, Welzer T, Heričko M, Jurič MB and Jaakkola H. Learning Process Termination Criteria. Informatica 2012; 23(4): 521-536.
48. Brumen B, Jurič MB, Welzer T, Rozman I, Jaakkola H and Papadopoulos A. Assessment of classification models with small amounts of data. Informatica (Lithuan.) 2007; 18(3): 343-362.
49. Florencio D and Herley C. A large-scale study of web password habits. Proceedings of the 16th international conference on World Wide Web; 2007. ACM.
50. Yan J, Blackwell A, Anderson R and Grant A. Password Memorability and Security: Empirical Results. IEEE SECURITY & PRIVACY 2004; 2(5): 25-31. DOI: DOI: 10.1109/MSP.2004.81.
51. Zviran M and Haga WJ. A comparison of password techniques for multilevel authentication mechanisms. The Computer Journal 1993; 36(3): 227-237.
52. Yan J, Blackwell A, Anderson R and Grant A. The memorability and security of passwords: some empirical results. in Technical report # UCAM-CL-TR-500.Cambridge, UK: University Of Cambridge; 2000.
53. Grosse E and Upadhyay M. Authentication at Scale. Security & Privacy, IEEE 2013; 11(1): 15-22. DOI: DOI 10.1109/MSP.2012.162.

DIGITAL FORENSICS TRACK

Concepts and Methodology in Mobile Devices Digital Forensics Education and Training

Damir Delija
INsig2 d.o.o, Zagreb, Croatia
damir.delija@insig2.hr

Abstract - This paper presents various issues in digital forensics of mobile devices and how to address these issues in relation to the education and training process. Mobile devices forensics is a new, very fast developing field which lacks standardization, compatibility, tools, methods and skills. All these drawbacks have an impact on the results of the forensic process as well as a deep influence in the training and education process. In this paper real life experience in training is presented, with tools, devices, procedures and organization with the main purpose of improving the process of mobile devices forensics and mobile forensic training and education.

I. INTRODUCTION

Some interesting issues and situation were noticed while providing digital forensic services, especially during training and education. These situations are part of the whole digital forensic process but usually not stressed enough, it is worth mentioning them and analyzing them. Most of these issues are coherent with the current state of the mobile forensics, some are specific while some of the problems already recognized and reported in mobile forensics field "Is Mobile Device Forensics Really Forensics?" [3], others are specific but still rooted in the same basic issues.

Experience comes from the business position such as digital forensics service provider, which covers the full life cycle from defining solution proposal, development, implementation, support, consulting, training, improvements, and upgrades. This is a project oriented life-cycle where training and support holds the biggest share of it with the aim of improving the solution and keeping it operational.

Basically all projects in mobile digital forensics include the introduction of new functions in existing system (law enforcement, business etc) through providing forensic tools and methods. In such setups training and consulting is essential for success. From the client's viewpoint it is often very challenging because the personnel don't usually have the basic training or skills, also there is usually no infrastructure for quickly setting up internal training and skill distribution. In a few words this means you actually fail if you provide only tools and installations without training, since everything will be shelved because no one will be able to efficiently use it.

Experience stresses on the education and training as the most important part of mobile forensics projects. Aim is to improve ability of users to work independently and in a forensically acceptable manner while dealing with various mobile devices. This conceptual goal, with both firm understanding what tools and procedures can and can't do, is essential for any acceptable mobile forensic practice.

The profession of digital forensics requires continuous education, training, and practice based on all above mentioned concepts. It is necessary to define methodology which can fulfill these requirements in the context of projects (law enforcement environment, military, governmental). This must provide training in basic forensic science, basic computer science and engineering, understanding of forensic tools and procedures

Mobile forensics is a very dynamic field in digital forensics which is, by itself, a new field in forensic science. This concept is taking to extremes where it is presented in recent 2014 paper "Is Mobile Device Forensics Really "Forensics"?", [3] which addresses some open questions and gives clear definitions on mobile forensics and digital forensics. Fast development rate, and actually a slave position to development in other fields, puts mobile forensics into a very unpleasant position where lack of standardization stands among other problems [3].

As an excellent overview into tools, skills, knowledge and procedures required for mobile forensics is the presentation "Cell Phone and GPS Forensic, Tool Classification System, (2009 Update)" done by Sam Brothers in 2009 which defines classification of tasks, tools and skills [5].

II. MOBILE FORENSICS SPECIFICS

It is important to stress that mobile forensic is a specific field, it is even hard to find coherent definitions as to what is a mobile device and what is mobile device forensics. There are various definition for mobile devices most of them defining mobile device as „small size device which is mobile and with network connection" [3]. Even if the definition of mobile forensic is more straightforward, it is sill biased enough because due to the ambiguous definition of mobile devices. As early as in definition there are problems with regards to procedures, tools, methods and compatibility with the rest of the digital forensics. Key issues to be solved are with relation to essential principles of digital forensics. The simplest firm relation in mobile device forensics is a „sub-science of digital forensic science" [3], while digital forensics is a computer and engineering science [3]. In precise terms, forensics are the application of scientific knowledge to

legal problems [3]. Still it is possible to argue about this definition based on Vietse Wenema definition of digital forensic "Gathering and analyzing data in a manner as free from distortion or bias as possible to reconstruct data or what has happened in past of the system" [6] which is used by SANS Institute. It is crucial to see it as a scientific and engineering process which is wide a enough concept, in that sense mobile forensics is a digital forensics applied on the mobile devices.

No one actually can count the number of mobile device, let alone the number of different models and types. The only sure thing is constant the arrival of new devices. Mobile devices has short lifespan models soon become extinct, while all of them are on the market as not sufficiently finished devices, as in not always working correctly. To complicate things even more, it is impossible to get data from mobile service providers with regards devices used and connected to their infrastructure, while this data surely exists and it is not confidential or illegal, but it is often under nondisclosure or privacy rules.

Table 1: Mobile device evidence extraction process – mobile device forensic process [2] [5].

Preparation Step	1. Device Intake, device is taken into forensic process 2. Device Isolation from mobile and WIFI network 3. Device Identification, type, model, features 4. Device Preparation, preparation for extraction and manipulation
Analyses Step	1. Processing extracting data 2. Verification of extraction and findings 3. Documentation/ Reporting of findings
Finalization Step	1. Presentation of findings and reports 2. Archiving device and results 3. Tools calibration and maintenance

One very distinctive mobile forensics trait is how commercial vendors approach mobile forensics. We can roughly separate vendors in specialized in mobile forensic and those general purpose digital forensic tool vendors. Among these groups there are no common tools nor standards from the mobile devices viewpoint. What is even more interesting is the separation among vendors based on how they handle mobile devices. Digital forensics vendors usually only support smartphones and provide very limited support for other mobile phones, while mobile forensic vendors almost completely ignore anything other than mobile devices. What is also stunning is the lack of compatibility and standards among mobile vendors and lack of using existing established forensic data formats.

From a practical point of view, of conducting investigations or just using forensic tools and methods, it is important to recognize real work tasks and address them properly in training and education. Basic steps in mobile forensic processes are defined in Table 1, it is the same process as any other digital forensic process, with a few difference among stressing preparatory steps and having very specialized tools, almost on data recovery level. Skill-sets and mobile forensic analyses levels are related to mobile forensic process as presented in Table 2, skills are hard to achieve and requires a lot of practical work [5], [2].

From our current experience gathered trough training and support, data made available, through logical level extraction or even manual level extraction, is sufficient for closing the investigation. Only in a small number of situation, about 10%, requires phone memory dump (physical extraction) or more complex extraction methods. The same happens with a number of mobile devices included in one investigation process; mostly only one or two mobile devices are included (basic communication among two people), often related to one or two computers forensic images. This estimates are hard to confirm since statistics data for ex-Yugoslavia area are not possible to access, law enforcement organization currently does not have such statistics publicly available.

Table 2: Mobile forensics analyses types and skills levels [5].

	Operation	Skill level	Description
1.	Manual Analysis	Introduction / beginner	taking picture of device screenshots, manually accessing reading data
2.	Logical Analysis	Introduction / beginner	extracting data from mobile devices by logical synchronization commands, usually done by forensic tool
3.	Hex Dump (Physical Analysis)	Intermediate	extracting bit by bit copy of mobile device memory, done by forensic tool or specific usage of non forensic tools (flash boxes, jtag)
4.	Chip-Off (Physical Analysis)	Expert /advanced	extracting flash chips from devices and reading chip contest, required dedicated laboratory equipment
5.	Micro Read (Physical Analysis)	Expert /advanced	extracting chips and reading logical gate states by electronic microscope

III. EXPERIENCE IN MOBILE FORENSIC TRAINING AND EDUCATION

There are many possible classifications of digital forensics training processes. In theory it can be vendor specific or vendor independent, academic or professionally oriented, certified or uncertified, practical or theoretical. In reality it is always a compromise especially in the case of mobile forensics. As for any digital forensics education training devices and tools are essential, because of possible varieties of models and scenarios. With mobile forensics it is essential to provide practical methods of configuration for mobile devices in order to provide same set of artifacts and same results during different training steps. Being unable to provide relevant training environments, with up-to-date tools and configurations, render training process ineffective and sometimes counterproductive, since outdated methods usually compromise evidence and results.

Methods used by various vendors are different from academic and no-tool specific trainings, in most situation academy is not capable with keeping up with development, while vendors always get concentrated on their tools lacking general knowledge and skills.

Mobile forensic tools become part of our portfolio (INSIG2) in 2009, with now defunct EnCase Neutrino. Since 2009 many other products have been accepted like Cellebrite UFED [7], MicroSystemation XRY [8], and Oxygen provide with full service. With such a wide product range, vendor independent mobile forensic services and support is provided, based on the requirements and current state of the market. Full mobile forensics training infrastructure was implemented and added to the training curriculum in 2012. Since 2012 each training is evaluated and analyzed for valuable actionable data, but because of contracts data is under nondisclosure clauses, so just general observations can be presented.

Table 3: Number of mobile forensic trainings provided

	Attendees	Region	Type of training
2012	40	Croatia	vendor specific/ law enforcement
2013	30	Croatia	vendor specific and / or vendor refresher / law enforcement military
2014	30	EU	vendor independent / law enforcement / military

To present the volume of activity in mobile forensics training, training events are presented in Table 3. Each training event is fully documented, analyzed and evaluated for lesson learned, but not only from the attendees' viewpoint but also from the trainers and logistic viewpoint. It is crucial to acquire such data to understand the practices of different vendors and how these practices and tools fit into the current environment. Such data is necessary in order to keep the quality and understand what should be improved or modified, also the whole process must formalized to prevent overseeing important factors. Based on this data, and data acquired through vendor products trainings and usage, the process is developed which keeps the necessary skills and ability to teach mobile forensics.

The key concept is understanding that the training event is specific and unique, since version and tools are changing fast. For demonstration we can use data about UFED family of products. Just since 2012 UFED forensics portfolio has substantial changes both in hardware and software, while continuously being upgraded with new supported mobile devices and analyses capabilities. All volume and type of changes have to be incorporated into the training process to keep it efficient. Supported phone models grow in the hundreds per year, not only as a new device model, but also with support for new applications, functionalities and operating systems.

Very important chat tool Skype can be used as example. In 2009 Skype was not supported as an analyses feature in automated analyses on smartphones, while in 2013 it is a standard part of the analyses. Since Skype artifacts are extremely important, that new functionality has to be introduced and supported in the training, with examples, drawbacks, practical issues etc. This is typical for feature and application support in mobile forensics. To show the volume of work required for only one application, in this case Skype analyses, in 2012 extraction was done through cooperation with other forensic tools [4], while in 2013 generic support was introduced. Again, during training, both methods were presented. General methods of using other forensic application, to verify artifacts' findings, is mentioned since it is an example of common practice. Application evolves and it is always possible to find a new version which is not supported on the current tool, but is supported by some other tool like Belkasoft, InternetEvidenceFinder, or by custom developed script [5].

It is the same for all features and applications on mobile devices, especially about encryption support, geolocation information and other new developments. Available forensic methods have to be presented, while optional solutions must also be provided at least at the conceptual level, as shown in Table 10.

Since changes are in all steps, each step has to fully include both the theoretical and practical parts of the training, with appropriate training mobile devices and artifacts on mobile devices. Maintenance and troubleshooting issues are key to making training kits operational, it must also be included. These topics are emphasized since they are usually overlooked in trainings.

Other mobile forensic tools are close to UFED example since mobile phone development forces forensic vendors to keep up with the change. With general purpose forensic tools vendors' situation is different, since the tool has limited mobile forensic capabilities, mostly only smartphone support. Good illustration for general forensic tool is EnCase from GuidanceSoftware. With introduction of EnCase version 7 in 2011 former mobile forensic version of EnCase, Neutrino, was discontinued. Its functionality later was added as special smartphone module into main EnCase v7 product. Since 2011 EnCase v7 got 26 versions and subversions with various upgrades, functionality changes and bug fixes.

IV. PROCEDURES AMD METHODS FOR MOBILE FORENSICS TRAINING AND EDUCATION

Preparation of training for mobile devices and forensic images requires forensically sound approach, method and tools which will guarantee that the training results are reliable and useful. As for specific vendor training, forensic images and artifacts are usually provided by the vendor, the same goes for the training materials. In reality it can happen as it was for early UFED situation where there were only forensic images of mobile devices provided by the vendor, but no officially approved training materials.

Training process varies in duration, usually from 3 days to 5 (24 or 40 hours) depending on the type of training. Duration for the tool specific trainings it is usually up to 3 days with 1/3 time for practical work.

Each skill level from Table 2 usually has its own skill level training, which deeply depends on the previous trainings. No-tool specific and general purpose trainings follow the same idea being more concerted on general skills than specifics tools. Training materials are copyrighted and protected from any kind of distribution or usage out of contract limited trainings that even goes for trainings result analyses data.

Preparing training materials, mobile devices and forensic images is an important and complicated task. It is common for any type of practical computer training, but because if nature of mobile forensic it is crucial part of training. Basically, for each new current or future versions images have to be recreated and reinstalled on the mobile devices. This process is independent from the training and unusually covers three basic steps and some additional logistics steps.

Basic steps

- **Initialization of mobile devices to known state**: Resetting mobile device to factory defaults, then installing the image from backup or other source, depends on mobile device model, operating system etc. After this step the mobile device is ready for the next training event
- **Creating mobile artifacts for each specific mobile device platform:** complex step which requires using application, tools on mobile device in real life scenario. For example for each supported mobile phone real chat sessions were done, emails sent and received, sms's going through town, images taken, video, connecting to WIFI etc. Each action is documented and time stamped so that the artifacts can be compared and verified in preparation for the training. It is a lengthily task requiring a lot of time and resources, usually done by students as part of their internship work.
- **Creating mobile device images:** when mobile devices have all necessary artifacts, forensic images and backups are created as baseline images from which other mobile devices - of the same types - will be cloned. Methods of cloning are different depending on mobile devices models, varies from backup/restore, synchronizations methods or using dedicated cloning tools like Ufed. It is also important to remember creating relevant forensic clones of SIM cards, to prevent mobile devices from changing installed images. Forensic SIM clone is SIM copy of user data but without data required for GSM connection, so mobile device can not connect to network while keeping it configuration and artifacts unchanged. This SIM copy is done with mobile forensic devices like UFED through the specialized SIM cloning function.

Additional logistic steps

- **Preparing and maintaining training documentation:** keeping manuals and documentation up to date, depends on the source of the documentation.
- **Acquiring and maintaining the necessary fleet of mobile devices**: mobile devices are changing and to keep with these changes typical models and functionalities must be obtained. From experience, this means having about 5 devices of the same type in the training kit. At the moment it is about 60 mobile devices, 30 smartphones and 30 of other phone models some of them ageing. From this kit about 5 devices are in various states of degradation, being replaced by new models. Also it is important to keep some broken devices to show techniques and methods available for partially functioning devices.
- **Acquiring and maintaining sets of SIM cards**: SIM cards from various mobile service providers have to be obtained, initialized, used and cloned, in order to provide realistic usage patterns. Each SIM has a limited lifetime so it has to be regularly maintained, subscription renewed and replaced if necessary. For each event it is necessary to have a set of SIMs in specific condition, even locked or damaged to provide realistic training scenarios.
- **Maintaining forensic equipment:** forensic kits for mobile forensics have different elements, but it must be maintained and kept in order as any other tool, basically it is keeping with the vendor updates and changes.
- **Maintaining versions of forensic images and backups and documentation:** each mobile device forensic image, backup or configuration must uniquely named, documented, listed and stored. Without these administrative practices the training process can not be reliable.
- **After each training event forensic image of each used mobile device can be created and compared with baseline image**. Forensic images should be created with available forensic tools and procedures. This is not a mandatory step, however, it helps to keep track of changes and possible bugs or malfunctions in the forensic software or equipment. Also it presents the patterns of how the trainees work with mobile devices as well as the efficiency of the training, since any change in mobile device configuration or content out of planned actions shows failure in training procedures. Up to now such failures were detected in forensic images.

As for the no-vendor trainings and education, it is almost the same situation but with specifics issues. This

type of training and education covers general issues about mobile devices and presents specific tools and tasks to solve them with such tools. In a typical training curriculums, there is a list of topics with expected interval for renewal and change as presented in Table 4.

Each of these changes require a full cycle of preparation and training refreshing both for trainers and specialists who already have attend a previous training. Since each topic has very fast update rate, special type of training refreshers are needed. To keep with this lifecycle and requirements, efficient organization and logistic processes must be established, with well defined procedures for each step in the training process.

Table 4: Topics in mobile training curriculum and its life cycle influences

Training Topic	How often has to be updated	Specific forensic image required or specific device
Introduction to mobile devices, technologies	Yearly	No
Introduction to mobile networks	Yearly	No
SIM Cards and key serial numbers	Yearly	SIM cards and forensic images
Common challenges with Devices	For each version or event	No
Forensics of the mobile devices	For each version or event	example of locally used devices
Seizure of Mobile Devices	Yearly	example of locally used devices
Types of extractions with mobile devices and comparison to regular computer extractions	For each version or event	example of locally used devices
Tools for mobile forensics – Analysis, Searching, Reporting	For each version or event	New devices and tools with updated kits, connection cables, flash boxes etc
Introduction to Smartphones and other "Smart" devices	yearly	example of locally used devices
Introduction to mobile devices operating systems	For each version or event, forensic images have to updated	example of locally used devices and forensic images of devices, forensic images have to updated, same for the mobile devices
Key features challenges with "Smart" devices in forensic sense	yearly	example of locally used devices
recovering deleted data from a wide range of mobile devices, including locked devices	For each version or event	example of locally used devices and forensic images of devices, forensic images have to updated, same for the mobile devices
understand how forensic software extracts and decodes data	For each version or event	, forensic images of devices, forensic images have to updated
understand how you can approach a forensic problem, defining a forensic strategy which may use a combination of tools and techniques to obtain evidence from a mobile device	For each version or event	example of locally used mobile devices, forensic tools
Understand the different challenges in the field of mobile device forensics compared to those in traditional computer forensics	For each version or event	example of locally used devices and forensic images of devices, forensic tools, forensic images have to updated, same for the mobile devices
Understand the different acquisition methods available when examining mobile devices	For each version or event	example of locally used devices and forensic images of devices, forensic tools, forensic images have to updated, same for the mobile devices
Understand how and when to use the different approaches	For each version or event	example of locally used devices and forensic images of devices, forensic tools, forensic images have to updated, same for the mobile devices
Awareness of the limitations of each method forensic method	For each version or event	example of locally used devices and forensic images of devices, forensic tools, forensic images have to updated, same for the mobile devices
Know how to approach defining an acquisition strategy for a new device	For each version or event	example of locally used devices and forensic images of devices, forensic tools, forensic images have to updated, same for the mobile devices

I. LESSON LEARNED AS DIGITAL FORENSIC TRAINING PROVIDER

II. Lessons we gained trough provided trainings and education since 2012 are in fact very disappointing, even on European level and especially for former Yugoslavia countries. Lack of ability for any type of digital forensic is astonishing also lack of investing into required infrastructure, training and education are overwhelming. It is the same for governmental, academic and business communities. At the moment most crucial problem is in law enforcement because it seriously influence ability to protect community. For the long term prospects lack of programs, funding and related knowledge in academic community is also worrying, but

it is not so important because of growing numbers of excellent online academic resources. The most astonishing lesson is that internal organization on client side is key for success, a fact which one will take for granted since organization is sending members to training. These lessons are worthwhile for anyone included in digital forensics training and education.

III. Based on our experience gathered trough delivered training, it can be said that mobile education and training is a key part in keeping forensic skills, especially in law enforcement and military organizations. In such organization there is a dichotomy among needs and ability to implement organizational structure which keeps skills and expertise. This dichotomy must be constantly addressed and remedied with various formal and informal methods of training and education. Such methods as conferences, workshop, and refreshers etc, while are not efficient as full set training, provide at least minimal necessary updates for skills and knowledge.

IV. For organizations sending personnel to training and education, it is important to emphasize the fact that it is actually skill set and internal organization that must be acquired, implemented and maintained. This should be done by establishing internal organizational structure, career path through continuous education and training for organization members [1]. It is often that internal organizational structure that is missing while need for specific skills are recognized, which is a very common scenario in law enforcement and defense organization as well as any other organizations which have a strict legally defined structure. Methods and practical solutions on how to cope with such problems are presented in "Digital Forensic Triage" [1]. Definition of such organizational structure is not part of the training and education, it is important to deliver the message about it's importance, because without it, the skills and knowledge are lost thus nullifying the training results.

V. As for the personnel receiving training, it is important to stress on the necessity of continuous working and keeping with current developments, which again leads back to the internal organizational structure of the trainee's organization. Again, without such organization, personnel will probably get only one minimal introduction training, there will be no upgrades and no career path. Skills and forensic capabilities will soon be lost.

VI. In some organization, informal internal keeping-up events can be organized, but this is sometimes double edge situation which is often maintained for a short period of time and later fails. Another key element for such organizations is implementing an internal communication network among personnel to keep skills and knowledge active. There are many possible models; the right model depends on the structure of organization and legal requirements, since it is not the same for the law enforcement, military or business environment. What is often seen as good model is kind of help desk or internal forum type of organization which works well in most situations. There we have experts helping others and keeping knowledge and skills alive with very simple computer infrastructure for support.

V. CONCLUSIONS

As a general conclusion we can say that training processes must be tailored for the needs of the organization and attendees. To cope with this conclusion very efficient, but resource consuming, processes must be implemented where trainings are kept up to date with various inputs: vendor development, user requirements, current best practice and client abilities (mostly budget restrains). This means that vendor / tool specific trainings and educations should be customised for specific needs and abilities, it is impossible to directly apply it without adaptation. Adaptations must be carefully prepared and examined since for example what is proposed by vendor or it is best practice in another country is not applicable in local context because of various technical or even legal reasons. Major goal for any mobile forensic trainig and education should be:

- to improve expert level of knowledge and skills for a usually small number of expert users and
- to keep growing number of user able to efficiently apply basic mobile forensics.

LITERATURE

[1] Stephen Pearson;Richard Watson: "Digital Triage Forensics", Syngress ,July 13, 2010, ISBN-13: 978-1-59749-596-7

[2] Sam Brothers: "iPhone Tool Classification" http://www.appleexaminer.com/iPhoneiPad/ToolClassification/ToolClassification.html,

[3] Gary C. Kessler:"Is Mobile Device Forensics Really "Forensics"?", NIST Mobile Forensics Workshop, Gaithersburg, MD, June 2014

[4] Paul Henry: „Quick Look - Cellebrite UFED Using Extract Phone Data & File System Dump", SANS Forensic Blog 2010

[5] Sam Brothers :„Cell Phone and GPS Forensic, Tool Classification System (2009 Update)", State of the Market Place as of: May 2009"

[6] SANS Secuirty 508.1 Forensic training, „Forensic and Investigative Essentials", SANS, 2007

[7] UFED, mobile forensic tool, http://www.cellebrite.com/

[8] XRY, mobile foresnci tool, https://www.msab.com/

RAM data significance in Digital Forensics

K. Hausknecht[1], D. Foit[1], J. Burić[1]

[1]INsig2 do.o.o., Zagreb, Croatia
Kresimir.Hausknecht@insig2.eu
Davorka.Foit@insig2.eu
Jerko.Buric@insig2.eu

Abstract **- In present modern times when operating systems require larger amounts of RAM or Random Access Memory, we usually come across computers with 4 GB RAM, but given the price drops, it is quite usual to come across computers with 64 GB of RAM as well. By imaging this part of computer memory and by performing forensics analysis of the data located in RAM, it can be easily concluded that performing RAM imagining and analysis should be one of the essential steps in any forensic investigation.**

This paper will give a short introduction to digital forensics and the role of live data forensics. Furthermore, the mail goal will be to show and explain the importance of forensics of live machines and artefacts which can be found as well as methods and tools which are used for extracting and analyzing data from RAM. In addition, it will be shown that sometimes in forensic investigations, data contained in RAM can contain enough evidence to solve the whole case and actually be everything a digital forensics investigator really need.

Index Terms **- RAM, Live data forensics, Imaging, Volatility, Batch scripts, Windows processes, Network connections, Passwords, Windows registry, Extraction, Malware.**

I. DIGITAL FORENSICS

Digital forensics is a relatively new area which is constantly spreading and progressing quickly. Accordingly, it requires continuous monitoring of new technologies in order to be up to date. The development of technologies such as smart phones, tablets as well as the constantly changing operating and computer file systems requires an exhaustive analysis in order to determine the best way how to get information required in investigations. Additionally, forensic investigation techniques on existing and new technologies are constantly developing.

In such a dynamic working environment a forensic analyst must be aware of all changes and developments of technologies. Many techniques are complex, they require a lot of time, specialized trainings and often special tools. It is exactly due to the frequent changes and developments that certain fields i.e. divisions of investigations emerged within the digital forensics itself such as computer file systems, mobile device forensics, operational systems forensics and the like. It is common that a forensic analyst opts for just one of these areas. One of these areas is the Live Data Forensics [1].

II. WHAT IS AND WHY IS LIVE DATA FORENSICS IMPORTANT?

Live Data Forensics

One of the new tools of a digital forensic analyst is also the Live Data Forensics. It is defined as the procedure of analyzing volatile or partially volatile computer data i.e. data which disappear upon shutting down the computer or data that is not persistent – often changed. In digital forensics, the analysis of such live data is important due to accessing data and evidences which are not reachable otherwise. In reality it often happens that the procedure of acquiring data includes solely the analysis of data-storage media such as hard disks, USB memories, CD/DVD media and other, better known as post-mortem analysis. The analysis of live data includes saving and analyzing volatile data such as Pagefile, Hibernation file, Crash Dump files and most importantly RAM - Random Access Memory.

III. RANDOM ACCESS MEMORY – RAM

In order to understand the importance of analyzing RAM data, it is first necessary to explain where such data are located and what information's does it contain. All data "travelling" through the computer have to pass through RAM at one point. Regardless whether we are referring to network traffic, working with hard disks, copying files from or to a certain medium or simply displaying pixels on the screen, all information are recorded in RAM memory blocks. Information will be present in RAM as well as on the computer hard disk, the only difference being that RAM is telling us what was happening on the computer at a certain period of time i.e. at the moment we recorded RAM while data on the hard disk are telling us what was happening in general. It is important to note that a huge quantity of data and information are never registered on the hard disk but solely in RAM. Common example being internet data [1][9][10].

IV. IMPORTANCE OF RANDOM ACCESS MEMORY ANALYSIS

Volatile data are constantly changing, they often are not structured in the same way as the computer file system and it is considerably more difficult to analyze and interpret the same. However, exactly such data are often the basis for continuing an investigation since certain data can be obtained solely from RAM. Currently, the analysis of volatile data is

still not defined as the analysis of hard disks drives or USB memory. Hard disks have a predefined structure and we can always find certain data at the same place which is not the case when analyzing RAM. Random Access Memory may but doesn't need to be allocated in the same places, depending on the type of memory in question. It is not possible to anticipate where a certain data can be found. It is obvious that the analysis of this data requires higher expertise and that it is a more demanding than post-mortem analysis. What also needs to be emphasized is that collecting such data also requires a certain expertise and work flow. Since the data is constantly changing, the moment and the way in which they are collected can greatly influence the final result. In Table 1 it is shown how RAM data will change (degradation of data) in time while the computer is in sleep mode (turned on but not in use).

TABLE 1 How RAM data changes with time in sleep time – data degradation

Activity	% Changes in RAM	
	256 MB RAM	512 MB RAM
Start	100	100
Sleep mode 1h	90,4	96,7
Sleep mode 2h	79,7	96,1
Memory dump	76,9	89,8
Sleep mode 15h	74,8	85,6
Memory dump	67,2	69,4

Information in table one was gathered by performing 2 tests in a lab environment. An older PC machines was used with single core AMD processor, Win XP SP3 and one 256 MB DDR 1 RAM module. After boot process, the following steps were taken:

1. Start MS Office and open one document
2. Start IE and log in to an online email account
3. Open CMD and run ipcofing command and ping google.com
4. Open Windows Explorer and copy one file

After the last step, RAM imaging was performed as stated in the Table 1. The second test was performed on the same machine only with two DDR1 RAM modules.

By analyzing RAM we can considerably contribute to the progress of every forensic investigation in which it is possible to take over RAM. Forensics of RAM is also of utmost importance because it bypasses several restrictions of traditional forensic analysis such as encryption which can completely stop the investigation in case it is not possible to get the password. Since technologies continue to advance, RAM analysis will become more and more important in order to collect all information necessary for any case.

Traditional post-mortem analysis is restricted in several ways. The most important one is the encryption. Investigators cannot access encrypted data if they are not able to crack the password or recover the key used for encrypting data. Such keys are rarely stored on hard disks, however, when the user enters his password or when the data is decrypted, data is being stored in RAM. The subsequent analysis of RAM enables finding these keys. Another restriction is the inability to analyze processes which were initiated on the computer at the moment of taking over the same. The inability to review processes which were initiated on the computer makes it impossible for the investigator to get an insight in the way the computer was being used. In addition, it is also possible to hide data in RAM itself as well as to initiate processes which are endangering the functioning of the system or stealing sensible data.

Nowadays, more and more we have cases where malware was used for endangering systems. Most malware is placed exclusively in RAM and it doesn't leave any trace whatsoever on the hard disk. Consequently, by traditional forensic analysis it is challenging to detect most malware or to make it possible for the analyst to better understand the way in which the computer was used. A good example would be Stuxnet, the first computer virus which was used as a military weapon. It was placed in the computer RAM, it left very little traces on the system itself and on the hard disk, and its goal was to disable centrifuges of nuclear reactors in Iran [1].

V. INFORMATION AND DATA THAT CAN BE FOUND IN THE RANDOM ACCESS MEMORY

There is a lot of information that can be found by analyzing RAM depending on the computer and the operating system used. Most valuable being: active processes, information on open files, registry entries, information on network activities, drivers used, logged in users, passwords and cryptographic keys, hidden processes and data, malware, temporary data, portable apps (apps which are not installed on the computer itself but they are solely initiated), used dlls, open sessions and many other vital information [3][9][10]. In this section we will address only some of these information and their importance.

A. Processes

There are various types of processes which can be found in RAM. All currently active processes are stored in RAM and they can be recovered from data structure in which they are located. Additionally, hidden processes can be recovered as well. For example, malware is known for hiding its activities, therefore its activities cannot be traced by regular methods or by simply reviewing active processes through the task manager of an active computer. By analyzing processes from RAM, we can search in their Headers and Footers thus making their hiding more difficult. Additionally, advanced code or malware can hide nevertheless by changing that information as well, but with additional analysis of blocks in which these processes are located, that information can be recovered as well. The aforementioned methods are exceptionally advanced and long-lasting and they require a deep knowledge of the operational system functioning. It is important to emphasize that this data can be obtained solely by analyzing RAM and that it is not possible to get it by regular forensic procedures ("dead box analysis").

Furthermore, processes which were stopped can also be found in RAM since the computer was not restarted from the moment when a certain process was shut down, so it is possible that the process still exists in the memory because its working environment was not yet assigned to another process. This shut down processes can be singled out from RAM and analyzed just like currently active processes [2][10].

B. *Open files and "registry handles"*

Files for accessing an individual process as well as registry entries are also stored in RAM. Information about files used by the system can be of great importance. In case the process is part of a malware code, open files can point out to the investigator where the code itself is placed on the hard disk, where the data is possibly being sent to or files which were changed. By monitoring these indicators it may be possible to reach the author of the malware code and to find out what his ultimate goal is. Also, it is possible to establish which omissions the code is using and to improve the safety of the overall system by correcting the same [2][9].

C. *General files*

As already mentioned, all files used (opened, read, modified) on a PC have to pass through RAM. Usually what happens is that the program creates a temporary copy of the file, stores it in the RAM and later when modification is complete, the files is saved on the media. This temporary file can later be recovered in full or just its fragments form the RAM. This is only one of the many possible scenarios but what is important that depending on the RAM size, can in some cases hold large number of files that were opened or used on the PC. Depending on the case, information that a files was used on a certain PC can have big significance. Furthermore, one of the most basic operations done on any given PC is the copy/paste feature. Each time sometime is placed in the clipboard, it is stored in RAM and therefore it is possible to recover it [1][2].

D. *Information on network traffic*

By analyzing RAM it is possible to identify all established network connections and open ports which existed at the moment of acquiring RAM as well as local and remote information on such connections. This possibility allows the investigator to analyze the programs which were initiated on the computer itself and which can be a target of malware code that can give false information on the network status. When receiving information directly from RAM, it is much more difficult for the offender to hide his open connections or planted processes. Information on network traffic is among the most important information which can be obtained from the computer that is the subject of investigation and it is more reliable when it comes from RAM analysis or from independent scanning of IP traffic [2][10].

E. *Internet data*

In this modern times, it is very common for everybody to use the Internet in their daily work or surf the web in their spare time. All research, networking, communication and similar is usually done via Internet. Additionally, it is also very common for people to store their data in the Cloud[1] and not locally on their computers. As previously mentioned, data that is being downloaded from the Internet will always pass through RAM. Because of this fact, it is possible to retrieve Gmail or Yahoo emails, Skype conversations and many more. All this very valuable data would be impossible to recover by "dead box" analysis since it is deleted as soon the PC is turned off.

F. *Passwords and cryptographic keys*

One of the main advantages of live data forensics is the possibility to recover passwords and cryptographic keys which were potentially used for decrypting data of interest and user accounts. Passwords and cryptographic keys are generally never stored on the hard disk without additional protection. However, when they are used, they have to be stored in RAM and they remain stored there until they are overwritten with another data or until the computer is shut down. When investigator performs a RAM dump (process of obtaining data form RAM), by reviewing collected data it is also possible to find online account passwords, which can have significant impact on the case itself. As an example, we can take online data or computer storage in clouds (cloud computing) which can store up to several terabytes of data.
Many BIOS (Basic Input/Output System) will write their plain text password to the BIOS keyboard buffer during boot which automatically suggest that it is possible to retrieve from RAM [2][10].

G. *Decrypted content*

The search for passwords and cryptographic keys in RAM can often lead investigators to decrypted data which can be recovered. When the user decrypts a certain file, the same file is temporarily being stored in that unencrypted form in Random Access Memory. By analyzing RAM it is possible to find parts of documents and even whole documents that were once encrypted. Files found by post-mortem analysis of the computer, will always be encrypted.

H. *Other data that can be found in RAM*

Since it is very common to have multiple users on one PC, by analyzing RAM it is very easy to determine which user was logged on, at what time and what was he doing or what files was he using. If Microsoft Windows were used for the operating system, a complete event log can be extracted and shown through a time line view. This can create a very clear picture for the investigator what was happening on the PC at what time. This can be crucial information for multiusers workstations and while investigating malware or virus infections. Windows will by default log software, security and system related events such as plugging in a USB device. In the latest analysis of the RAM, analysts have discovered that it is possible to retrieve a screenshot of the desktop and therefore all opened windows, time and work being done at the time the RAM data was retrieved [3][9].
All before mentioned information's can be also applied to multiple operating systems such as Microsoft Windows, Apple OSX or various Linux distributions. Not all operating systems will have the same results since each operates on its own way, but the most valuable information's such as network info, processes or files can be retrieved from the RAM data. Furthermore, RAM analysis can be performed even within the virtualized environment. In this types of cases, it can be very easy to obtain virtualized RAM data since it is commonly being stored as a single file on the hard drive of the hosting machine [1][2][10].

[1] A model of data storage where the digital data is stored in logical pools, the physical storage spans multiple servers (and often locations), and the physical environment is typically owned and managed by a hosting company.

VI. RANDOM ACCESS MEMORY ACQUIRY AND ANALYSIS

Acquiring Random Access Memory

In order for an investigator to be able to analyze RAM, he must have certain tools and knowledge in order to get all information contained in RAM. There are two methods for acquiring RAM: hardware and software method.

A. Hardware method

The hardware method includes installing a special card in the computer in order to gain Direct Memory Access (DMA). Usually this process involves installing PCI or PCMCI card with FireWire connector. Once the card is installed properly, other computer is used to image the RAM of the target machine. Generally, this method is more reliable than the software method since, in case the computer is compromised, we can still obtain a copy of RAM because there is no dependence on other computer components. In addition this method is the only way to image the RAM of a live machine that is password locked. Of course, this method has its deficiencies too – the price of the hardware component is higher and more expertise is required for performing extraction and especially installing the additional card in a live machine [10][11].

B. Software method

Software extraction is a much more common way for obtaining data from RAM and it is recommended to perform it during every computer acquisition in case the computer is active. Tools used for acquiring RAM can be commercial tools such as FTK Imager, Foremost, Helix, Belkasoft Live RAM Capturer and the like and there are also many non-commercial tools such as dd, memdump or Dumpit [4][5][6][7][8][9]. The main problem of software solutions occurs in case the computer is already compromised, the offender can easily additionally hide his data from such tools or give incorrect i.e. false data. Another deficiency of software solutions is that, when they are initiated, they themselves are using RAM just like all other programs and thereby they are potentially erasing other, possibly important information. With software tools we need to pay attention that they leave the smallest possible footprint, receptively that they occupy the least possible memory space. Another important fact is that they are operating in the so called "Kernel" mode. This means that the application itself has more rights in the system and that it can reach more information in the system. The main advantage of software tools is that they are free of charge and that they simplify the takeover of RAM so that it can be performed by untrained personnel [12].

Data analysis

Among commercial tools it is possible to use tools such as AccessData FTK, Belkasoft Evidence Center, X-Ways or Digital DNA –HBGarry, however, each of these tools contains only certain options and does not provide sufficient flexibility respectively adaptability[4][5][6][7]. As already mentioned several times, in order to analyze RAM a high level of expertise and experience is required [2][3]. Volatility Framework became prominent as the main tool for "deep" respectively thorough analysis [3][4].

VIII. CONCLUSION

It can be conclude that Live data forensics and therefore RAM data analysis should become a part of regular forensic procedures. During every computer analysis (in case the situation allows it) it is necessary to acquire RAM. The acquisition itself takes much less time than the acquisition of other types of memories and the possibility that it contains information important for the case which would otherwise be unavailable, is very high. Even though the procedure itself, the methodology and analysis are neither sufficiently examined nor documented, the potential results always make up for it. As described in this paper, the data that can be found in the RAM can sometimes contain enough evidence to solve the whole case. Special attention should be given to the fact that, even though there are deficiencies in the methodology, there are ways of preventing forensics analysis (the so-called anti-forensics) and therefore, while performing RAM dump, attention should be paid to anti-forensic tools.

[1] E. M. Chan, "A framework for live forensics", https://www.ideals.illinois.edu/bitstream/handle/2142/24365/Chan_Ellick.pdf?sequence=1, 7.1.2015

[2] Techniques and Tools for Recovering and Analyzing Data from Volatile Memory, SANS Institute - InfoSec Reading Room, https://www.sans.org/reading-room/whitepapers/forensics/techniques-tools-recovering-analyzing-data-volatile-memory-33049, 7.1.2015

[3] An advanced memory forensics framework, https://code.google.com/p/volatility/, 1.1.2015

[4] AccessData ltd. FTK Imager, Relese information, http://accessdata.com/product-download/digital-forensics/ftk-imager-version-3.3.0#release-information, 10.1.2015

[5] Belkasoft Ram Capturer, release information, http://belkasoft.com/en/ram-capturer, 10.1.2015

[6] X-Ways, realese information, http://www.x-ways.net/forensics/index-m.html, 10.1.2015

[7] Digital DNA – HBGarry, http://mcsi.mantech.com/sites/default/files/MCSI%20Digital%20DNA%20Technology%20Brief.pdf, 10.1.2015

[8] Moonsols Dumpit, release information, http://www.moonsols.com/2011/07/18/moonsols-dumpit-goes-mainstream, 10.1.2015

[9] M. Hale Ligh, A. Case, J. Levy, A. Aron, "The Art of Memory Forensics: Detecting Malware and Threats in Windows, Linux, and Mac Memory", 2014

[10] M. Ligh, S. Adair, B. Hartstein, M. Richard, "Malware Analyst's Cookbook", 2010

[11] eForensics Magazine, Vol.04 no. 01, February 2015

[12] A. Borges, Memory Acquisition, https://alexandreborgesbrazil.files.wordpress.com/2015/02/memory_acquisition1.pdf

Competencies and Skills needed for Digital Forensic Trainer

D. Foit, J. Vukalović, K. Hausknecht
INsig2 Ltd., Zagreb, Croatia

davorka.foit@insig2.eu
jakob.vukalovic@insig2.eu
kresimir.hausknecht@insig2.eu

Abstract - Today we live in a time where everything is digitalized and technology moves rapidly forward. This big technology progress is closely linked to increase of cybercrime. In an effort to fight e-crime and to collect relevant digital evidence, law enforcement agencies are incorporating the collection and analysis of digital evidence, also known as computer forensics, into their infrastructure. They are challenged by the need to train officers and specialists to collect digital evidence and keep up with rapidly evolving technologies such as computer operating systems and mobile devices. With a lack of resources to send new people to the trainings, very often recently trained officers and experts become trainers to their colleagues. This role demands specific competencies and skills that can be learned on specialized courses. Based on an empirical analysis done on specimen of 20 experienced digital forensic specialists, this paper shows results of this research, offers explanation of competencies and skills needed for digital forensic trainer and helps law enforcement agencies to recognize potential trainers in their ranks, and helps them in their fight with the e-crime.

I. INTRODUCTION

Based on a long term experience in holding specialized trainings for the digital forensic specialists and law enforcement agencies, we've noticed that there is a big lack of methodology and approach in dealing with adult learning. The same thing happened when we participated some of official, vendor's trainings. The main reason to that is that people that hold trainings are not college educated people, but regular employees of companies or government agencies. Aldo this may seem strange, the explanation of such procedures lies in the fact that due to the world crises, many government institutions were forced to decrease their budgets. Because of that, companies usually send only one or two people from the department to educations and then it is on them to transfer gained knowledge to their colleagues and coworkers.

In this paper we will show what skills and competencies a digital forensic specialist needs to develop and improve to become a digital forensic trainer.

Methodology: We conducted empirical researches in January 2015 during Train the Trainers training in Pasadena. Research was conducted on a focus group of 20 experienced digital forensic specialists. Some of them work as a trainers for 5 or more years so they have enough of practical experience. Research was conducted by questionnaires, quantitative and qualitative analysis.

At the very beginning we set 3 hypothesis about digital forensics trainers that this paper should confirm or decline:

1. Digital forensic trainers understand how adults learn and acquire new skills
2. Digital forensic trainers can determine goals and purpose of a training
3. Digital forensic trainers understand and use different learning methods in order to transfer their own knowledge to participants who have different learning styles

II. ADULTS LEARNING

The main problem that appears with teaching adults lies in the fact that people who are not college educated teachers or who didn't attend some sort of train the trainers course, don't have skills and capabilities needed to perform good and successful education of adults. An important part of being an effective educator involves understanding how adults learn best. The principles and rules for adults learning are much more different than with children education with which are we all familiar. The theory that covers and handles adult learning principles is called "andragogy" (adult learning). Andragogy uses approaches to learning that are problem based and collaborative and emphasis more equality between the teacher and learner. It was defined as a study in Europe in 1950s and pioneered as a theory and model of adult learning in 1970's by Malcolm Knowles. [1]

At the beginning we asked the group what is an adult learning and what they know about adult learning principles. From 20 persons just one person knew the

term; others have never heard about the term and were not aware of adult learning principles.

III. EMPIRICAL ANALYSIS –TASK ONE: WHERE IS THE MAIN FOCUS OF THE TRAINING

The main goal of the first task is to discover where the trainer's main focus is when it comes to digital forensic trainings. It is well known that there are four main parameters on which focus can be set – trainer, participants, content, organization, – and all four of them are of equal importance to the training [2]. The purpose of a research was to determine on which of those four parameters digital forensic specialists focus the most, is that enough to improve their training performance and does this qualify them as a digital forensic trainers or they need to work more on improving their skills.

We told the members of a focus group to decide which one of four above mentioned parameters is the most important to them. When they decide, they need to go in one of four possible corners of the room and write down on flipchart (4 flipcharts on easels were already set up in four corners of the room) parameter for who they decided: TRAINER, PARTICIPANTS, CONTENT or ORGANIZATION. Focus group was composed of 20 active digital forensic specialists who have experience in holding trainings but they never got any formal education in that area. After that, members of the focus group were asked to write down list of reasons why they think that the chosen parameter is the most important for the training. The results are shown in the table below:

TABLE I. WHERE IS THE TRAINING FOCUS?

Focus on	Number of participants
Trainer	8
Participants	6
Content	3
Organization	3

Two participants remained "neutral" because they considered that all 4 parameters are of equal importance. The reason for this assignment is to discover how people think, what are their main priorities when they come to training, and what when they decide to educate other people.

Besides choosing where the focus should be, they also need to give some arguments to support their choice. Group that put focus on TRAINER listed their reasons as follows (I intentionally left original version of their handwriting):

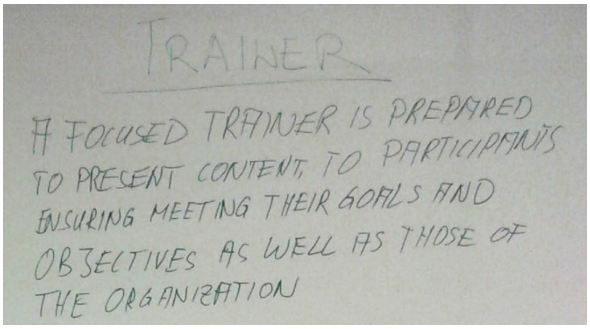


Figure 1. Arguments for "trainer" as the main focus of a training

Here is also group that decided that PARTICIPANTS are the most important for the training and list of arguments:

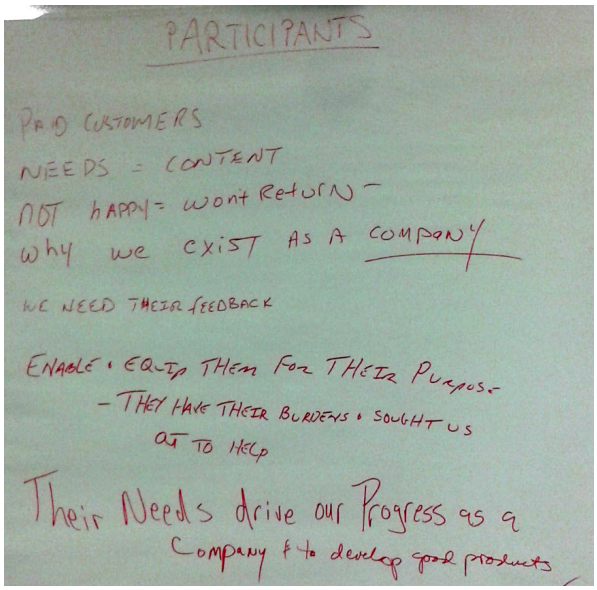


Figure 2. Arguments for "participants" as the main focus of a training

Also, in the same way are two other parameters (CONTENT and ORGANIZATION) added and described.

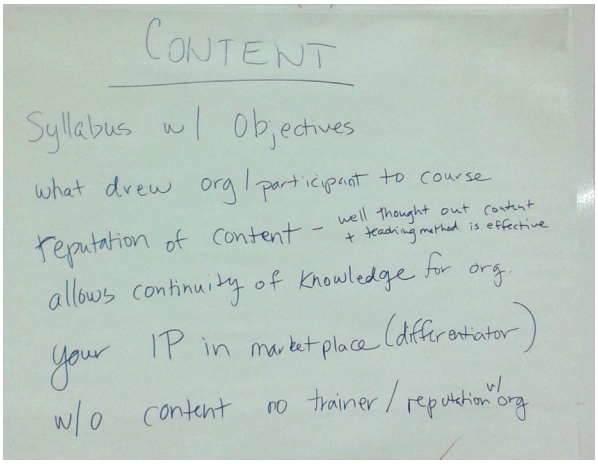


Figure 3. Arguments for "content" as the main focus of a training

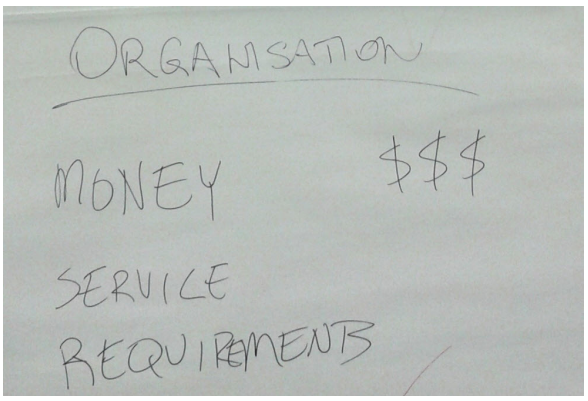


Figure 4. Arguments for "organization" as the main focus of a training

As we can see, all 4 parameters are well – reasoned.

The task was a little bit tricky, because there is no wrong answer. All four parameters are of the same importance for the training and all four need to be implemented in the adult training. Trainer needs to be focused on the organization and its need to accomplish its mission. It focuses on content that he delivers to ensure that it's accurate and practical. It also needs to be focused on himself to ensure that he is skilled and knowledgeable about the subject in hand. And finally, trainer focuses on his participants to ensure that all their expectations as a group and as individuals have been met during the training.

A. Results

The results of an empirical research show that, without proper education about adult learning principles, digital forensic specialist put their focus on trainer and participants, while content and organization are not of such importance to them. Hence, digital forensic specialists assume their own knowledge as absolute, and don't take into consideration that trainings which they perform are held in behalf of their employer or organization.

For the other part of research where thy needed to argument their decision, members of a focus group gave good explanations and arguments for each parameter.

IV. EMPIRICAL ANALYSIS – TASK TWO: LEARNING STYLES IN THE CLASSROOM

Besides determination where the training focus is, another important fact in andragogy are learning styles in the classroom. There are many modules that can be used as a reference for the learning styles, from Kolb's experiential module [3], Peter Honey and Alan Mumford's model [4], Anthony Gregorc's model [5] and many others. Module that will be used in this paper as a reference module is VAK module [6]. It is one of the most effective module of individual learning style, especially easy to applicable to adults who have already developed their own learning and working preference. This a module consists of carefully mixing all three dominant learning styles – Visual, Auditory and Kinesthetic – according to participant's preference [6].

For a good trainer, it's important to accept the fact that participants may have different learning styles, so it is important to consciously design content that is balanced between all styles. Trainer should always use a variety of approaches, techniques and activities that address the different VAK preferences such as:

Visual: creating graphic models, drawing concepts, taking notes, viewing slides and flipcharts.

Auditory: utilizing spoken words, allowing them to talk, panel discussions and music.

Kinesthetic: things to manipulate, such as modeling clay, crayons, Koosh balls and papers feels good [2].

Visual	Auditory	Kinaesthetic
Posters	YouTube clips	Note taking
Mind maps	Dictaphone	Copying out notes
Post it notes	Discussions	Games
Note taking	Question/Answer sessions	Talking-Walking Study
YouTube clips	Mnemonics	Re-enactments
Colour coding	Revision songs/rhymes	Trace words with finger
Timelines	TV Programmes	Doodling
Grids/Tables/Charts	Revision Videos	Trips and Visits
Fact Lists	Listening to music whilst revising	Construct posters and other revision resources
Facebook Profiles	Study Groups	

Figure 5. Some of techniques and activities characteristic for different learning style

It is also really important that trainer doesn't use its own preferred learning style. For instance, some people are active learners. They like to be constantly challenged, can think on their feet and enjoy the challenge of being thrown in the deep end, learning best 'on the job' through practical exposure, trial and error and direct experience. Other people are more reflective learners, they need time to plan, prepare, research and to have time to reflect on their learning before being confronted with a new challenge. They may like to be thoroughly briefed before proceeding. Some people are theoretical learners, and are stimulated by abstract ideas and concepts. They like to consider numerous viewpoints and theories and to analyze situations before selecting options and approaches to a task. They learn through observation, discussion, analysis, and enjoy logical and sophisticated reasoning. Whilst others are pragmatic learners, they enjoy learning from qualified demonstration, and need to see the practical advantage of all that they are doing. They need to know that what they are doing works and is realistic [7].

In this research we had a focus group of 20 people that work as a digital forensic specialists and, as a part of their job, hold training for their colleagues and superiors.

The group had to fill in given questionnaire (Figure 6) with a list of 25 statements that describe some behaviors during learning process. The goal was to determine learning styles of digital forensic specialists that are active and good in digital forensic investigations, and see what skills and competencies, which are related to training

performance, they need to improve or change in order to become digital forensic trainers. Assumption that was taken in this research was that the members of a focus group will expect that their learning styles are similar to those of their colleagues and supervisors (potential participants) and that, in that case, they will not have to adjust or change their skills.

The questionnaire was designed in a way that members of a focus group had to put checkmarks next to the statements that correctly and most accurately describe things they like and don't like to do while adopting new knowledge. At the end, they had to sum all the checkmarks in each column. Each column was representing one learning style [2]:

- VL (*Visual Language*) – This is the person who learns the best from reading words in books, on the chalkboard, charts or workbooks. The individual may even write down the words that are given orally in order to learn from seen them on the paper
- VN (*Visual Numerical*) – The person has to see the numbers on the board, on the paper or in the book in order to work with them. They are more likely to remember and understand number facts if they have seen them. They don't seem to need to much oral explanation.
- KL (*Kinesthetic Language*) – Learning takes place by the student actually carrying out a physical activity, rather than listening to a lecture or merely watching a demonstration. It is also referred to as a tactile learning.
- AL (*Auditory Language*) – this is the person which learns from hearing words spoken. You may hear them vocalizing or see their lips or throat moving when they read, especially when they are trying to learn some new materials. They are more capable of understanding and remembering words or facts they could only learn by hearing.
- AN (*Auditory Numerical*) – this is the person that learns from hearing numbers and oral explanations. They may remember phone and locker numbers easily and be successful with oral numbers, games and puzzles. They can probably solve problem in their head. Trainer my see participant's lips moving and hear them talking while they are trying to understand a problem.

At the end, column for which they have the highest score is their preferred learning style.

In the picture below is an example how one page of the questionnaire looked like:

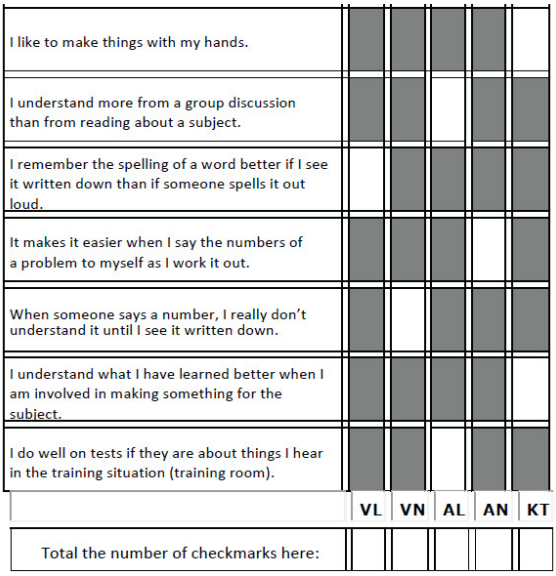


Figure 6. Part of VAK learrning style questionanire [6]

At the end we asked them to plot a spider web connecting the points with the scores on each line. We created a similar spider web based on the scores of all 20 participants.

A. Explanation of the questionnaire results

The result that focus group achieved is shown on the Figure7:

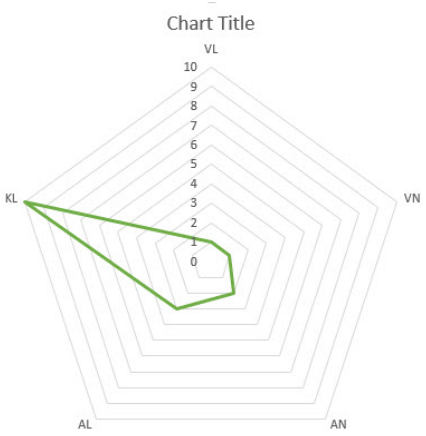


Figure 7. Spider web with connected points of scores on each line

VL stands for Visual Language and it was preference learning style for only 1 person. **VN** is Visual Numerical learning style and was represented by 1 person. **AN** is Auditory Numerical and it was preference learning style for 2 persons, and **AL** is Auditory Language and is preference learning style for 3 persons. As expected for this group of people – the most represented learning style was **KL** (Kinesthetic Language) with 10 representatives.

It is recommended for every trainer to conduct such testing at the beginning of a lecture because it gives them opportunity to combine all learning styles, but also to

encourage participants to use learning style that is not preference to them.

B. Results

With conducted research was established that the dominate style which digital forensic specialists use in their learning process is **Kinesthetic Language**.

The result is important because it shows connection between jobs that members of a focus group perform in their everyday life and learning style that they use. Digital forensic trainers have to work on understanding how to train people who have different learning styles.

Kinesthetic represents repetition of procedures and *Language* represents written part. If we look globally, that is exactly what digital forensic investigators do - they spend much of their time in laboratory settings, analyzing physical evidence from crime scenes and at the end of analysis, they write official reports on the subject.

V. Digital Forensic Specialists – General Skills and Competencies

In their everyday job, digital forensic experts should be proficient in conducting many types of tests, analyzing the results and reporting them. They are all insatiably curious about how and why computers work. They spend much of their time reading about software, hardware, registry keys and root kits. Exploration is their daily job.

Communication and learning skill which are emphasized: tenacity, persistence, dedication, patience, integrity, curiosity, investigative skills, desire to learn constantly, self-motivation, interest in researching new things.

College education is not the most important in this job – it is desirable to have a degree from a technical college to prove your competencies and expertise in the courtroom. However, the most of the knowledge that digital forensic specialists have, they obtained through professional certifications that legitimately demonstrate training, testing and practical experience and at the same time have value in helping courts, clients and potential employers assess qualifications. The ability to construct illuminating experiments and the patience to elicit the data is another hallmark of skilled computer forensic examiners. If you need to know how a file or metadata changes when a user does something or what digital evidence is left behind, you'll be best prepared to testify if you've proven your theory by competent experimentation.

Knowledge that is emphasized with digital forensic specialists: good formal education, good math aptitude, good computer background, basic computer knowledge.

Another thing that characterize digital forensic specialists is experience and sharing that experience with other colleagues. Every examiner benefits from the exchange of ideas with skilled, experienced colleagues. However, the computer forensics community has its own rules that need to be followed:

- show respect for others by first doing your homework;
- don't waste time with a question that's been asked and answered in a prior posting or one you took no steps to address on your own;
- be a learner, not a leech

And finally, all digital forensic experts know all about the tools and techniques they use in their cases. It is not enough only to know how to use some tool, but also to know how to explain and defend your report in the courtroom, in front of the people that are not familiar with digital forensic evidence and examination of the same.

The job of a computer forensics expert doesn't end when the digital evidence has been obtained and analyzed. They must document their findings in the form of one or more reports that detail not only what they found, but how they found it. It's really important that they are able to articulate the process and procedures they followed so there will be no doubt that the evidence was legally obtained, that its integrity was preserved throughout the process, and that there was no opportunity for it to be tampered with [8].

Social and cognitive skills that are emphasized with digital forensic specialists: ability to write a report on findings, ability to link one part to the whole case, ability to receive criticism.

VI. Conclusion

By empirical researches conducted on a focus group in this paper, it is determined that digital forensic specialists have some of the skills and knowledge that are necessary to become trainers in digital forensic field, but they need to change some of them in order to fulfill all prerequisites needed for digital forensic trainers.

By going back to the 3 hypothesis from the beginning of this paper, we can conclude:

1. Digital forensic specialists don't understand how adults learn and gain new knowledge. If they want to become successful digital forensic trainers, they need to learn about adults learning principles and about ways how adults adopt new skills. So, hypothesis no.1 was not confirmed by researches in this paper.
2. Digital forensic specialists can determine goals and purpose of a training. In this case they can become digital forensic trainers without any big adjustments or changes so the hypothesis no.2 is confirmed by researches in this paper.
3. Digital forensic specialists use their individual learning styles, but, at the same time, are not aware of other learning styles which their colleagues use. They need to become more aware of other's learning styles in order to be able to provide adequate knowledge and skills to their future participants. In this case, hypothesis no.3 was not confirmed by the researches in this paper.

Next research should be based on this one and it will be done in Orlando on 2 different focus groups. Groups are composed of 20 skilled and well trained digital forensic specialists who come from all over the world and are of different age.

ACKNOWLEDGMENT

I would like to thank the trainers from Guidance Software Inc. for their participation in my research and for sharing their experiences and their thoughts with me.

REFERENCES

[1] Malcolm Knowles, "The adult learner: A neglecked species", Gulf Publishing Company, April 1973.

[2] Elaine Biech, "ASTD'S Ultimate Train the Trainer: A complete guide to training success", ASTD Press, 2009, ISBN-10: 1-56286-587-0

[3] Kolb David, "Experiential learning: Experience as the source of learning and development", Englewood Cliffs, NJ: Prentice-Hall. ISBN 0-13-295261-0., 1984

[4] Honey, P & Mumford, A., "The Learning Styles Questionnaire", 80-item version. Maidenhead, UK, Peter Honey Publications, 2006

[5] Bagby, M., "Learning Style Difference vs Learning Difficulty", http://therapeuticliteracycenter.com/learning-style-difference/ , 2014

[6] Neil D. Fleming, "Teaching and Learning Styles: VARK Strategies", ISBN 0473079569, 9780473079567, 2001

[7] QOTFC The Clinical Educator's Resource Kit, „How teachers can use the Adult Learning Theory and Principles to help students", http://www.onlinehighschool.us/how-teachers-can-use-the-adult-learning-theory-and-principles-to-help-students/, August 9, 2013

[8] Craig D. Ball, „How Do I Become a Computer Forensic Specialist? Craig Ball's Seven Steps to Becoming an Expert", http://www.craigball.com/Becoming%20a%20Computer%20Forensic%20examiner.html

Challenges in Network forensics

J. Burić*, D. Delija **
INsig2 d.o.o., Zagreb, Croatia
jerko.buric@insig2.eu, damir.delija@insig2.eu

Abstract - Network forensics is a branch of digital forensics that focuses on monitoring, capturing, recording, and analysis of network traffic. More accurately, it is the use of scientifically proved techniques to collect and analyse network packets and events for investigative purposes. Network forensics is an extension of the network security model which traditionally emphasizes prevention and detection of network attacks. Current network forensics approaches are costly and time consuming. However, unlike other areas of digital forensics, network forensics deals with volatile and dynamic data. It helps organizations to investigate attacks that originated from outside and inside of the company. It's also important for law enforcement agencies when solving crimes. Paper presents different challenges that are facing investigators due to the rapid growth of network and attacker's skill, and possible framework solutions that would help to solve or minimize problems.

I. INTRODUCTION

The Internet is constantly changing, where evidences are located all over the network, there is no enough time nor staff. Expectations are unrealistic, various differences within organizations, and above all, mishandling of evidence. Network forensics investigations can be and is a very demanding process. With all challenges that traditional forensic investigators are faced, network forensic investigators need to work/interact with expert and non-expert people, learn to handle different types of equipment, both hardware and software, theirs' various versions, more than often capture evidence that exist only for a brief moment, and unfortunately, investigative teams find themselves in situations where it is not clear who's in charge [1].

An organized approach is a key to a successful investigation. With the continuous growth of the Internet, cyber-attacks and crimes are happening every day [2], with addition of increasing intruder's skills. Everything has gone that far, that it raises the fact: It's not IF you are going to be attacked, it's WHEN. Traditional tools used in investigations, such as firewalls and Intrusion Detection and Prevention Systems (IDPSs), are no longer enough since they can't provide all required evidence or data [3].

Current network forensic approaches are costly and time consuming. In addition, these approaches normally use active and reactive processes to resolve cyber-crimes, and such processes start after cyber-crime has been identified, which makes identifying useful evidence difficult [4], making network forensic investigation a post-mortem incident handling.

II. NETWORK FORENSICS

Network forensics is an act of capturing, recording, and analysing network audit trials in order to discover the source of security breaches or other information assurance problems [5]. The term "network forensics" was introduced by the computer and firewall expert Marcus J. Ranum in early 90s, but borrowed from legal and criminology fields where "*forensics"* is related to investigation of crimes.

In 2001 DFRWS (Digital Forensic Research Workshop) defined network forensics as *"The use of scientifically proven techniques to collect, fuse, identify, examine, correlate, analyse, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned intent, or measured success of unauthorized activities meant to disrupt, corrupt, and/or compromise system components as well as providing information to assist in response to/or recovery from these activities"* [6].

In essence, the concept of network forensics deals with the data found across a network connection going to and from one host to another. Network forensics analyses the traffic data logged through firewalls or intrusion detection systems or at other network devices such as routers. The goal is to trace back to the source of the attack in order to discover perpetrators [9]. One note has to be emphasized; the traffic or data which is being monitored is live so in network forensics we are no longer talking about post-mortem analysis.

The following table presents network devices and their value in network forensics.

Table 1. Network devices and their forensic value

Tool name	Value to Network forensics
Switches	Contain Content Addressable Memory (CAM) - determines which port on a switch corresponds to which MAC address.
Routers	Networking devices that forward data packets between computer networks. They store routing tables that maps ports to connected networks. They can also function as a packet filter.
IDPSs	Network security appliance that monitors network and/or system activities in real time and alerts for malicious traffic.
Firewalls	Control incoming and outgoing network traffic based on predefined rules. Analyse and log the source and destination addresses, packet payloads, and port numbers to make decision about certain traffic to help forensics in process of investigation.
Web proxies	Act as an intermediary server for requests from clients seeking resources from other servers. Can be

	configured to maintain special logs for extended period of time. Such logs store the web surfing for an entire organization.
Special type servers	The main types are DHCP, Name, Authentication, Application, and Central log servers. Data stored on each of these servers can be of the great value and can provide help in forensic investigations.

Back in 2002, Simon Garfinkel introduced two network forensic systems called "Catch it as you can" and "Stop, look and listen" [7]. "Catch it as you can" system captures and writes all the packets passing through a particular traffic point. This approach requires large amounts of data storage. On the other hand, "stop, look and listen" system, analyses each packet in a rudimentary way in memory and only certain information is saved for future analysis. This requires a faster processor to keep up with incoming traffic [2]. Network forensic systems are designed to identify unauthorized use, misuse, and attacks on information systems which were previously identified manually or in most cases, unidentified [8]. Most network forensic systems are based on audit trails. System relying on audit trails try to detect known attack patterns, deviations from normal behaviour, or security policy violations. They also try to reduce large volumes of audit trail data to small volumes of interesting data. Most of the current techniques are passive monitors, which heavily rely on network traffic, CPU usage, or I/O usage with human intervention [8]. The main focus in the area of cyber and network forensics is to automate the process of detecting the attacks and to prevent the damage caused by future security breaches.

Network forensics is not another term for network security, but it is an important part in network forensics. Data for forensic analysis can be collected from various security products which are placed onto the network to detect and prevent intrusions. Network forensics ensures that the attacker spends more time and energy to cover his tracks making the attack costly and time consuming [9].

The IT industry's ever-growing concern with security is the primary motivation for network forensics [10].

A. Network Forensic Analysis Tools (NFATs)

Although not the final-solution for security administration, NFAT is capable of saving both time and money and help administrator watch everything in the network, gather all information about suspicious traffic, and help in network forensics. Basically, NFATs capture network traffic and send data to the engine, which analyses that data and shows results. NFATs correlates with IDSs and firewalls making preservation of network traffic in long period of time possible and allows quick analysis of concerning points in the network identified by IDSs and firewalls [11, 12]. Some of the NFATs functions are:

- Intellectual property protection
- Detection of employee misuse/abuse of company networks
- Risk assessment
- Network forensics and security investigation
- Exploit attempt detection
- Data aggregation from multiple sources, including firewalls, IDSs, and sniffers
- Incident recovery
- Prediction of future attacks
- Anomaly detection
- Network traffic recording and analysis
- Network performance

Some of the commercial NFATs available in the market are: NetIntercept, NetDetector, NetFlow, SilentRunner, EnCase, and VisualRoute [11].

The open source NFATs are: TCPDump, Libpcap, WinDump, Wireshark, Snort, Nmap, P0f, Tcpstat, Tcptrace, and Tcpflow [11].

Following commands are inbuilt in many modern OSs and are very useful for Network forensics: nslookup, traceroute, netstat, nbtstat, whois, ping, wget, dig [11].

B. Network forensic Process (Generic) Model

In the paper called *"A survey about network forensic tools"* [2], the authors propose a generic framework for network forensic investigation. The proposed framework consists of many phases which are already imprinted in digital forensic model, on which they built new phases specific to network forensics. As illustrated on figure 1, generic framework consists of nine phases, which are going to be closely explained.

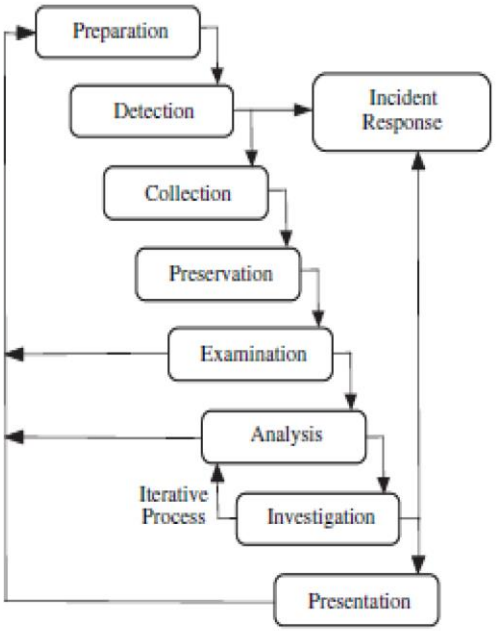


Figure 1. Network forensics Process Model [13]

- *Preparation phase* – since many tools need to be deployed (IDPSs, firewalls, packet analysers) on various points on the network, the prime duty is to obtain required authorization and legal warrants to ensure privacy.
- *Detection phase* – deployed tools generate an alert or a warning which indicates a security

breach or policy violation. A quick validation is done to assess and confirm the suspected attack.

- *Incident response phase* – the response initiated in this phase depends on the type of attack identified and also by organizational and legal policy. This phase is applicable only when investigation is initiated during the attack.
- *Collection phase* – the most difficult part because data that flows over a network changes rapidly and it is not possible to generate the same trace at later stages. It is crucial to have reliable hardware and software, along with well-defined procedures to collect maximum evidence with minimum impact on the network.
- *Preservation phase* – original data is kept safe along with computed hashes. Another copy of data will be used for analysis. This is done to ensure that there is no unauthorized use or tampering.
- *Examination phase* – this phase examines previous phase. This is done in a methodical way so no key information is lost. All hidden or altered data done by attacker needs to be uncovered. Reduction of high volume data is necessary in order to identify the least information holding the highest probable evidence.
- *Analysis phase* – collected evidence is analysed in order to find specific indicator of an intrusion. Also, statistical analysis and data mining is performed to search for data and to match it to attacking model. The attacking patterns are put together and reconstructed to understand the purpose and methodology of the attack.
- *Investigation phase* – this phase uses information gathered through analysis phase, and concentrates on identifying attacker, which is the most difficult part of analysis phase. Attacker may use many different techniques to hide their intentions or their identity, such as IP spoofing or stepping stone attack [1]. Actual approach in investigation phase depends on attack type.
- *Presentation* – is the final stage in processing model. Systematic documentation, along with observation with explanations are presented in readable and understandable format. Everything that has been performed needs to be presented in accordance to applicable legislation and security policy, along with recommendations on how to prevent future attacks.

This framework for network forensics shows how methodical and precise approach an investigator needs to take in order to have valuable, legally acceptable and forensically sound evidence to present.

III. NETWORK FORENSIC CHALLENGES

The main idea of network forensics is to identify all possible security violations and threats, and build the signatures into IDPS mechanisms to prevent further losses [8]. Many tools allow network administrators or forensic investigators to observe network traffic in real-time, but real-time monitoring at any level requires significant human and hardware resources, and doesn't scale to networks larger than a single workgroup. It is generally more practical to archive all collected traffic and then analyse it per needs [10].

When conducting network forensics, investigators frequently work with live systems that cannot be taken offline. These systems can include routers, switches, firewalls and other equipment all the way to critical servers. In network forensics, investigators also work to minimize effect on system modification due to forensic activity. However, in those cases investigators do not have luxury of an online copy. Furthermore, network-based evidence is more than often highly volatile[2] and must be collected through active means that essentially modify the system which holds necessary evidence. This impact on the system can be minimized through careful selection of appropriate tools and acquisition techniques, but it can never be eliminated entirely [1].

Network-based evidence poses special challenge in several areas [1]:

- *Acquisition* – it can be difficult to locate specific evidence in a network environment. But, even when you know where a specific piece of evidence is located, you may have difficulty gaining access to it for various reasons.
- *Content* – unlike file systems, which are designed to contain all the files and their metadata, network devices may not store evidence with that level of diversity.
- *Storage* – network devices normally do not employ secondary or persistent storage, and often have very limited storage capacity
- *Privacy* – depending on jurisdiction, there may be legal issues involving personal privacy that are unique to network-based acquisition techniques.
- *Seizure* – seizing a hard drive can inconvenience an individual or organization. Seizing a network device can be much more disruptive. In the most extreme cases, an entire network segment may be brought down indefinitely.
- *Admissibility* – file system-based evidence is now regularly admitted in both criminal and civil proceedings. In contrast, network forensics is a newer approach to digital investigations. There are sometimes conflicting or non-existing legal procedures for admission on various types of network based digital evidence.

[1] Stepping stone attack – technique used by attackers to attain anonymity and complicate their apprehension [15]

[2] Volatile – information that changes frequently and is often lost upon powering down the PC [16]

As the main idea of network forensics being to identify all possible violations and threats, a key challenge in network forensics is to ensure that network is forensically sound. That means that network needs to be properly configured, maintained, and updated by network administrator, and possibly have infrastructure that supports network forensic tools - hardware and software. When conducting network forensic investigation on a network infrastructure, following challenges should be taken in considerations:

- *Data sources* – typical network has several possible data sources, ranging from raw data to logs from network devices. The best situation would be to collect all data from all sources but this is very often not possible, especially when dealing with large networks. Therefore, selection of proper source of data is of the essence.
- *Data granularity* – as related challenge to data sources, deciding how much details we want to collect is another problem. While in smaller networks we can collect whole packets that contain packets' headers, IP addresses, port numbers, protocols, TTL, etc., collecting that much data in larger network doesn't seem to be practical or manageable.
- *Data integrity* – probably the crucial part or challenge in network forensics. The whole forensic process can be dismissed or court disproved if data collected from a network is altered or changed in any way, deliberately or accidently. Therefore, forensic investigator needs to ensure and guaranty data integrity during and after collection and analysis.
- *Data as legal evidence* – Data collected internally for different purposes within organization differs from data collected for court of law. Data/evidence as it is needs to pass strict legal procedures needs to be acceptable in the court.
- *Privacy issues* – data collected during an investigation always contains some form of personal information, whether it is an email, personal files, or metadata of a person. Collateral intrusion can also be a privacy issues, when investigators discover personal information about people who are not subject of an investigation.
- *Data analysis* – a huge challenge in an investigations due to complexity of network environment and data involved during collection process. It is quite difficult to find, analyse and present useful information or data. Therefore, investigators use variety of tools and techniques to make job easier and more accurate.

Challenges regarding network-based evidences and network infrastructure have been presented in relation to network forensic investigation. Network forensic investigators are searching for a clue or piece of evidence buried deep in a pile of data, something like searching a needle in a haystack. Keeping that in mind, following challenges occur almost every time when a network forensic analysis is taking place:

- *Time* - whether is it a deadline they must honour, or to find what actions were taken at a specific time, or just to synchronize date and time information with events being analysed, time is very important aspect in every forensic investigation.
- *Performance* – speed and accuracy are one of key components of network forensic investigations. It's all about how fast an investigator can react on alert of a security breach, and/or policy violation, and how punctual is he in his assignment. Of course, it is not all about human competence and skill; both hardware and software need to be capable of performing at top level.
- *Complexity* – bigger network, bigger problems; bigger complexity. Although not a general rule, since sometimes smaller networks can cause large problems in numerous ways, bigger networks tend to have more network devices, more end users, more IDPSs and firewalls and very large infrastructure, where investigator or team of forensic investigators can be overwhelmed by the size of data they need to process and analyse. In addition, servers can be located in different geographical areas which increases complexity and makes the job more difficult to perform.
- *Collection* – in correlation with complexity, it is true skill of an investigator to separate valuable data from terabytes of normal network traffic.
- *Law* – as being a young area of expertise, network forensics tends to have legal and court issues. They could be related to the lack of applicable legislature, a country's internal decision-making organization associated with law.
- *Hiding a breach* – many companies try to minimize exposure to the public regarding their network breaches, and 'ask for help' only when they see that there is no other option. This is another major challenge in network forensics simply because, if an attack or security breach has occurred, a skilful attacker is long gone and has successfully covered his tracks.
- *Network systems* - as already mentioned, network forensic systems are based on audit trails. One of the main setbacks with these systems is their price, which can be extremely high. In order to analyse logs, system must keep information, which results in large amount of data requiring even larger amounts of disk space with very powerful CPU resources.

Even after you have collected everything necessary for network forensic investigation, the next step is processing and converting logs in manageable format, and then comparing them with known attacks patterns and misuses,

to possibly identify security breaches can be extremely demanding. Furthermore, stored patters of attacks need to be regularly updated, which requires a lot of human expertise. A network tool that is capable of performing this automatically is the goal of a researches in cyber/network forensics.

IV. CONCLUSION

Network forensics is an art and skill. It takes a lot of knowledge and patience to learn networking, how it operates and functions, networks flaws which skilful attacker can use, as well as advantages which forensic investigator can turn into his favour. Network forensics gives an investigator an ability to trace the source of the attack to its origin and possibly to identify the perpetrator. This paper presented only some challenges concerning network forensics. Attackers, and their rivals, network investigator/administrators, constantly battle around dominance over the network. That battle is what challenges are all about: time, speed, accuracy, complexity, integrity, performance, etc. Paper also presented NFATs, collection of network tools, which investigators can use to monitor, maintain, and trace back attackers if needed. A generic framework was also presented, which provided a visual overview how network forensic procedure should look like and what are the key stones and challenges.

LITERATURE

[1] S. Davidoff, J. Ham, "Network Forensics Tracking hackers through cyberspace", Pearson Education, Inc., New Jersey, 2012.

[2] A. Lazzez, "A Survey about Network Forensics Tools", International Journal of Computer and Information Technology, vol. 2, issue 1, pp. 74-81, January 2013 .

[3] Q. Al-Mousa, Z. A. Al-Mousa, "Honeypots Aiding Network Forensics: Challenges and Notins", Journal of Communication Vol. 8, No. 11, pp. 700-707, November 2013.

[4] M. Rasmi, A. Jantan, H. Al-Mimi, "A new approach for resolving cyber crime in network forensics based on generic process model", The 6th International Conference on Information Technology, May 8th 2013.

[5] M. Rouse, http://searchsecurity.techtarget.com/definition/network-forensics

[6] Report From the First Digital Forensic Research Workshop (DFRWS), November 6th 2001.

[7] S. Garfinkel, "Network Forensics: Tapping the Internet", O'Reilly Network, April 26th 2002.

[8] S. Mukkamala, A. H. Sung, "Identifiying Significant Features for Network Forensic Analysis Using Artificial Intelligent Techniques", Internatinal Journal of Digital Evidence, vol. 1, issue 4, winter 2003.

[9] E. S. Pilli, R.C. Joshi, R. Niyogi, "A Generic Framework for Network Forensics", Internatinal Journal of Computer Applications, vol. 1, No. 11, 2010.

[10] V. Corey, C. Peterman, S. Shearin, M. S. Greenber, J. Van Bokkelen, "Network Forensics Analysis", Journal IEEE Internet Computing, vol. 6, issue 6, pp. 60-66, November 2002.

[11] B. Patel, S. M. Shah, S. S. Chauhan, "Comparative Analysis of Network Forensic Systems", International Journal of Computer Applications, Special issues on IP Multimedia Communications, vol. 1, pp. 80-83, October 2011.

[12] SANS Institute, Global Information Assurance Certification Paper, 2003.

[13] A. Chennaka, "Network Forensics: A survey", Electrical and Computer Engineering, Iowa State University, Spring 2013.

[14] A. Almulhem, "Network forensics: Notions and challenges", Signal Processing and Information Technology (ISSPIT), pp 463-466, December 2009.

[15] Y. Zhank, V. Paxson, "Detecting Stepping Stones", 9th conference on USENIX Security Symposium, vol 9, 2000.

[16] SynJunkie, "Forensics – Volatile data", http://synjunkie.blogspot.com/2007/10/forensics-volatile-data.html, date published: 6th October 2007, last accessed: 2nd February, 2015.

An Approach to Digital Evidence Collection for Successful Forensic Application: An Investigation of Blackmail Case

Kemal Hajdarevic*, Vahidin Dzaltur**,
* Faculty of Electrical Engineering, University of Sarajevo,
Sarajevo, Bosnia and Herzegovina
** American University in Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina
*khajdarevic@etf.unsa.ba, ** vdzaltur@aubih.edu.ba

Abstract – Computer forensic is the practice of collecting, analyzing and reporting evidences in a way that is legally admissible "in open court" or "public" as a part of criminal investigation process. It can be used in the detection and prevention of crime and in any dispute where evidence is stored digitally. In this paper we presented case of criminal activity in progress where forensic using hacker tools helped to proactively prevent criminal act of blackmail in progress. This approach was acceptable to company owner because public case would bring negative publicity and because incident resolution time was very important criteria for this case. Many organizations suffer great losses because of missing implementing standards for information security that includes segregation of duties, and active monitoring. In this paper we described how we used network hacking for forensic investigation for the right cause which is to prevent criminal activity in progress. This paper shows how different techniques and tools can be jointly used in step by step process to successfully perform forensic analysis and reporting.

I. INTRODUCTION

Computer forensics follows a similar process to other forensic disciplines, and faces similar issues. Purpose is to give answer to questions of a legal system related to computers. Forensics has it's role in any sort of legal issue, trial, civil court cases or any other legal processing that has computer involved.

"Computer forensics usually refers to the forensic examination of computer components and their contents..." [1]

A. Presenting the case before forensic investigation

The case presented in this paper started when forensic team was called by company's IT manager with company's owner permit with request to investigate if data base (DB) administrator was hiding important organisation's data on PC computer. There was fear that data DB administrator collects and use company information without permit. Another reason for believing that stealth blackmail is in process was the fact that DB administrator initiated negotiation for changing salary terms and conditions and refusing to provide system password to IT manager. The PC that was used by DB administrator was database administrator's device because company allows bring your own device (BYOD) and bring your own technology (BYOT) policy. The database administrator's PC was password protected at all time. Other issues included fear that database administrator will permanently destroy entire hard drive with data of the system which he manages or just quit the job and leave company without important company and user's data. Knowing that, and because there were still no elements for criminal prosecution, data or hardware was not destroyed, nor there were no formal blackmail request, it was decided to acquire access to all data without knowledge of database administrator. This case was more prevention of the possible crime. Because no one was able to tell any information regarding what operating systems, type of the software and version DB administrator used on his PC.

Decision was made with company owner that forensic team will work after office working hours. In order to keep low profile it was decide to keep the contact only with IT manager, to avoid that the other employees inform DB administrator that forensic process in progress. Because DB administrator's PC was running all the time, the idea of shutting it down and creating forensic image was out of the question. Reason for this is that DB administrator would probably be able to detect if PC was rebooted without his presence.

Network presented in this case was structured UTP wired infrastructure with central router, IP addressing of local network communication was configured by using DHCP server service located at the router.

B. Rationale for documenting the forensic case

The main reason for documenting and presenting whole process of forensic analysis was to show how hacker techniques can be used for right purpose which is stopping criminal activity in progress. Criminal are usually aware of this possibilities and they are trying to mitigate detection and prevention of their activities. While hacker tools are used to fight with criminals it is possible

to find arguments for, and against the use of "malicious" tools or code for "righteous" purposes [2].

II. PROCESS OF COLLECTING AND ANALYSING NETWORK DATA

Typically, at the crime scene first step is collecting and labelling all evidences, but in this case priority was to leave entire LAN hardware up and running so that DB administrator will not be aware that forensic process is in progress. The forensic team has arrived after working hours to use approach common to intrusion attacks shown in Figure 1. steps of intrusion attacks [3].

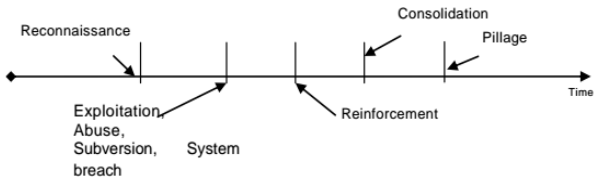


Figure 1. Steps of intrusion attacks

The first step was collection of all network which is followed by [3] of reconnaissance, exploitation, reinforcement, consolidation, and pillage to:

- Detect computers in the company network
- Detect IP addresses of devices in the network
- Detect any other type of attached network devices
- Sniff network traffic for passwords
- Gain access to PC and stored data

Because no one in the company was able to tell what software is installed on DB administrator's PC as BOYD / BYOT policy, it was decided to use available hacker tools such as Angry IP Scanner 2.16 [4] and Nmap [5] to detect DB administrator's PC. In the network diagram shown in Figure 2. presented is simplified company network diagram with several workstations, one router, and a DB administrator's PC that was a prime goal of the investigation.

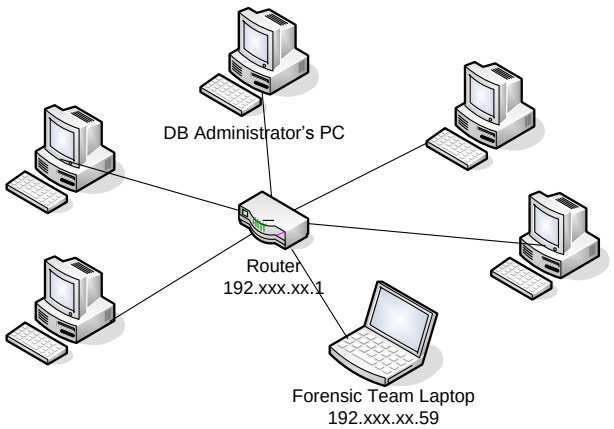


Figure 2. Simplified company network diagram segment

A. *Network scanning*

In order to scan entire Local Area Network Angry IP scanner 2.16 [4] was used and Nmap [5], with goal to find DB admin PC IP addresses. Angry IP scanner [4], and Nmap [5] are able to scan IP addresses, open ports, to detect running operating systems. They can be installed on Linux, Windows platforms

With nmap it was decide to use intense scan, and slow and comprehensive scan by using below stated nmap commands:

nmap -T4 –A –v 192.xxx.x.xx

nmap -sS -sU -T4 -A -v -PE -PP -PS80,443 -PA3389 -PU40125 -PY -g 53 --script "default or (discovery and safe)" 192.xxx.x.xx

Angry IP Scanner (or simply ipscan) is an open-source and cross-platform network scanner designed to be fast and simple to use. In Figure 3. is shown level of details which are used to gain information on available devices attached on network.

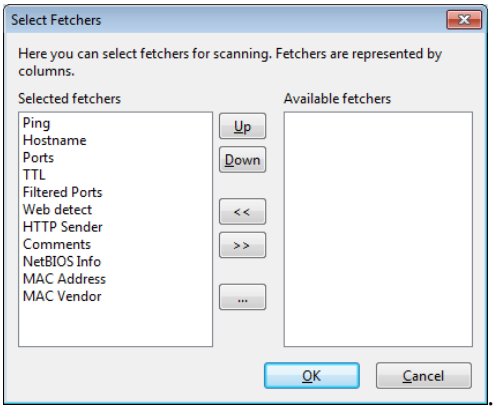


Figure 3. Result after scanning LAN with Angry IP Scanner

In Figure 4. is presented screenshot with results of IP scanning network segment with IP addresses.

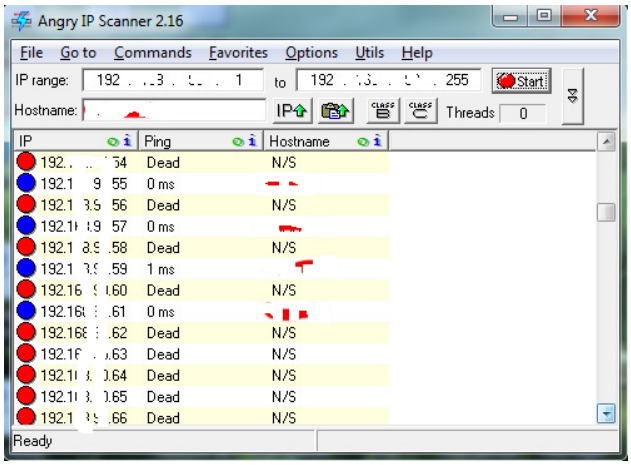


Figure 4. Result after scanning LAN with Angry IP Scanner

Result didn't show a logical host names, and it was not clear which IP addresses belongs to DB administrator's PC that forensic team was trying to investigate.

Next step was to shut down all other active computers, except a default gateway on IP 192.168.xx.1 and DB administrator PC. It was clear that only three IP addresses after all other devices were powered off were active: forensic team laptop, router and DB administrator's PC.

All scanning activities were finished in time window less than one minute.

After this step, the problem with DB administrator's PC IP address was resolved and all other network devices were again switched on. Using Nmap [5] it was detected that on this PC MS Windows operating system is installed, as shown in Figure 5. Nmap scanning results.

After reconnaissance activity which included scanning network segment where DB administrator's PC resides network topology became populated with new information.

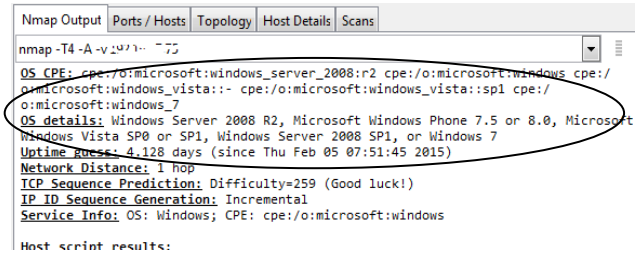


Figure 5. Nmap scanning results

After scanning network segment new information became available about network which included DB administrator's PC IP address as it is shown in Figure 6. and other network segment details with relevant IP addresses for all connected devices.

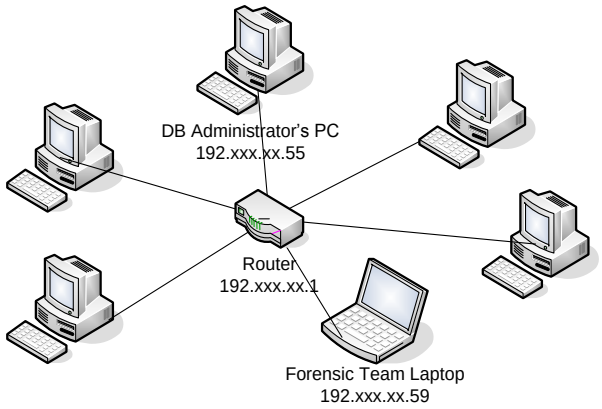


Figure 6. Network segment with relevant IP addresses

Because DB administrator worked every day, it was assumed that he will use his PC to logon with credentials at web pages, web mail, or other service. Next steps were to capture traffic and to recover passwords.

B. Network sniffing

After successful target determination another step was to capture network traffic which can be usually done by using tap device with an open source tool such as Wireshark [6] on wired networks. Tool Kismet [7] as open source can be used on wireless links, and tool called Cain (Cain and Abel) [8] which is also open source can be used on wired and even on wireless networks [9].

Next step was to choose approach to sniff LAN in order to collect all traffic between DB administrators PC and the rest of the network. For this purpose forensic team has Wireshark [6] and Cain [8]. While Whireshark [6] offers options for passive sniffing Cain [8] offers hacker tools for active attack and break security features.

This time it was agreed to use Cain [8] because it is resourceful tool for ARP poisoning [10] and is used for other purposes such as research in WIFI security [11]. Cain [8] is a well-known password cracking tool that is capable of handling different type of tasks. This tool is only available for Windows platforms. It can work as network sniffer, to crack encrypted passwords using the dictionary attack, to record VoIP conversations, to execute brute force and cryptanalysis attacks, to reveal password boxes, to uncover cached passwords, to decode scrambled passwords, and analyze routing protocols.

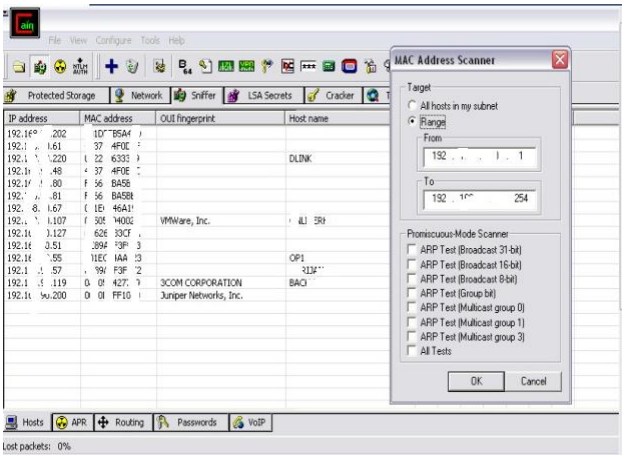


Figure 7. MAC address scanner

Cain [8] is not created to exploit any vulnerability or bug, it is created to use weakness of security I network protocols to gain passwords. After starting the program it is configured to scan MAC address of all hosts in the subnet, and then to resolve host names, just to be sure that forensic team has right information of LAN devices. Figure 7. shows active MAC address scanner in the Cain [8], same process was performed with Wireshark to confirm Cain [8] findings. Now when program collected MAC and IP addresses, host names, it was time to start with "Man in the middle attack".

III. PROCESS OF EXPLOITING PRIVILEGES

First step in process of exploiting privileges was reconnaissance and enumeration of IP and MAC address of DB administrator's PC. Second step was gaining user name and password credentials to be able to exploit privileges. One of easiest methods to gain information such as passwords is ARP poisoning [12].

A. APR poisoning

It is possible to change the Media Access Control (MAC) address of NIC card and attack an Ethernet LAN

by changing the target's computer ARP cache with a forged ARP reply packets [10]. The forensic team sends frames that are created so that they contain wrong MAC address.

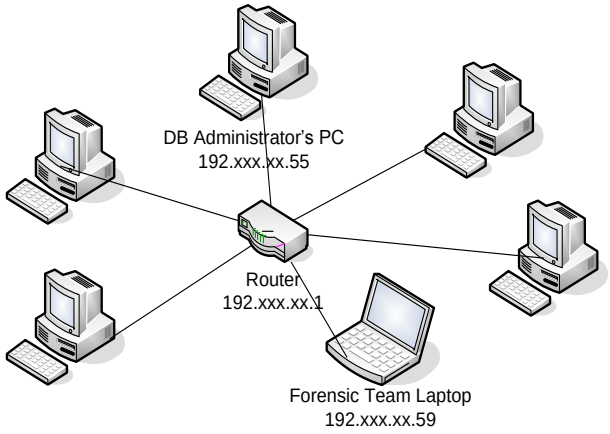


Figure 8. ARP poisoning

Because the ARP replies are forged, the target computer unintentionally sends the frames to the attacker computer instead of sending it to the genuine destination. An effective ARP poisoning attempt is undetectable to the user.

Next step was to start Cain [8], select IP address of the target devices and the gateway, and start ARP poisoning. After this step is finished system was left with running Cain [8] on the system in order to collect as more network traffic data as it is possible with goal to collect DB administrator's user name and password credentials.

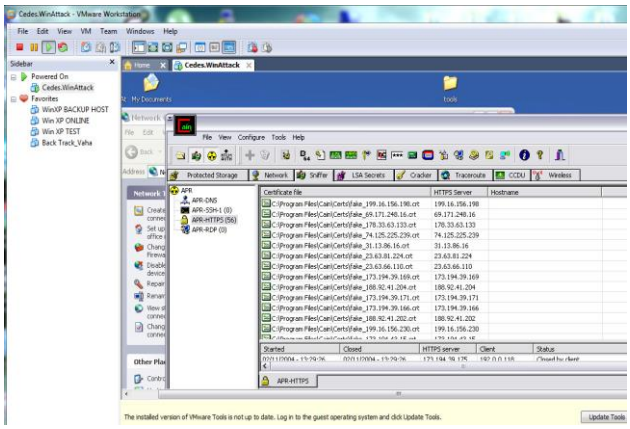


Figure 9. - Providing fake certificate by ARP poisoning [13]

B. Password collection retrieval

Four days after forensic team started with ARP poisoning using Cain [8], forensic team members came to examine what information were captured by Cain [8], it was found that program captured HTTP, POP3 and SMB encrypted passwords.

The great news for the team was that for the first two protocols, HTTP and POP3 passwords were captured in clear text format, password and username were the same for two different services.

For the SBM, sniffer captured one encrypted password used by Windows-based computers that allows systems within the same network to share files or access to other network resources. It allows computers connected to the same network or domain to access files from other local computers.

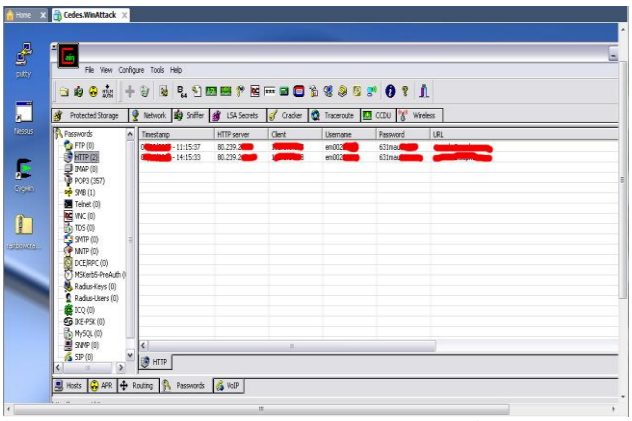


Figure 10. Captured password in clear text format

As it is shown in Figure 11. there is information available such as: timestamp, HTTP server IP address, client IP address, and the most important for the forensic analysis, username and password. The Cain [8] provides information about URL location with authentication type. Now focus of forensic team was to gain knowledge of SMB protocol and it's user-level authentication. It indicates that the client attempting to access a share on a server must provide a user name and password. When authenticated, the user can then access all shares on a server.

The only thing that's left is to try decrypt SMB hash. For this purpose Rainbow Crack [14] were used by [15, 16, 17, 18] with Rainbow table size of 64GB. This is a hacker tool that uses a large-scale time-memory trade off process for faster password cracking compared to traditional brute force tools. Time-memory trade off is a computational process in which all plain text and hash pairs are calculated by using a selected hash algorithm.

After computation, results were stored in the rainbow table. This process is very time consuming. Once the table is ready, it can crack passwords much faster than brute force tools. There is no need to generate rainbow tables every time, they can be used if someone else prepared them. Developers of Rainbow Crack have generated LM, NTLM, MD5, and SHA1 rainbow tables [19]. In the Figure 11. are shown three steps taken from article "Capturing and cracking SMB hashes with Cain [8] and Half-LM rainbow tables" [19] of loading the hashes into Cain [8], by selecting the Cracker tab step 1. then the plus symbol – step 2. to add hashes from a list – step 3.

Like Rainbow Crack [14], these tables are also freely available. It is possible to download these tables and use them for similar password cracking processes. The great thing about rainbow tables is that they can be imported in to Cain [8] in order to perform hash crack attack. This

approach was used by forensic team as an attempt to discover SMB password.

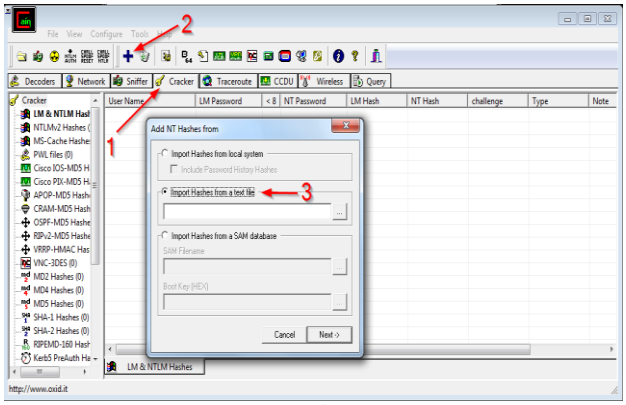
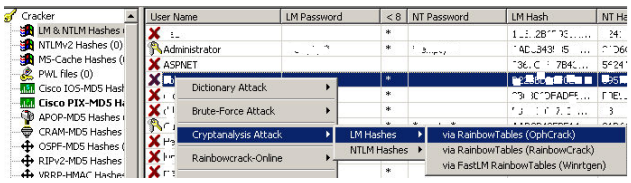


Figure 11. Adding HASH file to CAIN cracker [19]

Assuming that DB administrator was using the same password for Windows shares as he used for other two protocols, quickest solution was to convert discovered password in to hash and putting that in to rainbow table – Figure 12. After this the crypt analysis attack was started and after couple minutes program finished with analysis and it was confirmed that password was the same for all discovered network services used by DB administrator.



12. Cryptanalysis attack with rainbow tables

IV. PROCESS OF REINFORCING DATA EVIDENCES FOR FORENSIC REPORT

In order to have enough time, forensic team members came first day of the weekend to create digital copy of entire hard drive. Following procedure, DB administrator's PC was unplugged from power supply and the battery was removed. After this, hard drive was removed, labelled and placed in to antistatic bag and then carried away to forensic team office where digital image was created using EnCase [20]. Luckily, all retrieved data from digital image that were protected by password had the same password that was discovered in previous days.

This forensic analysis task was finished by returning hard drive into original place and giving system password which was object of dispute to company manager in order to set a new one.

V. CONCLUSIONS

As is the case with all evidence, it is very important to maintain a chain of custody for computer evidences. Each person who handled evidence may be required to testify that the evidence presented in court is the same as when it was processed during the investigation. Although it may not be necessary to produce at trial every individual who handled the evidence, it is best to keep the number to a minimum and maintain documentation to demonstrate that digital evidence has not been altered since it was collected. Forensic investigators must do everything possible to preserve the integrity of the digital evidences. Any mistake in the forensic process can make forensic evidence into question and it can be worthless if forensic principles were not followed. The way the forensic team handled integrity issues are numerous and include the way of seizing, labelling, transporting, coping, analysing and ate the end presenting results at court trials or for organisation needs.

It was advised by forensic team to company owner because organisational and technical vulnerabilities were detected that information security standards shall be used to prevent future problems.

VI. REFERENCES

[1] Casey Eoghan, Handbook of Digital Forensics and Investigation, Academic Press, 2005, ISBN: 978-0-12-374267-4

[2] Cobb, S.; Lee, A, Malware is called malicious for a reason: The risks of weaponizing code,. 6th International Conference On Cyber Conflict (CyCon 2014)

[3] Bejtlich R., The Tao of Network Security Monitoring, Pearson Education, Inc. Second printing 2004

[4] Angry IP Scanner, Available from: http://angryip.org/ [Accessed on 03.02.2015]

[5] Nmap, Available from: http://nmap.org/ [Accessed on 03.01.2015]

[6] Wireshark, Available from: https://www.wireshark.org/, [Accessed on 03.01.2015]

[7] Wireshark, Available from: https://www.wireshark.org/ [Accessed on 17.02.2015]

[8] Cain, Available from: http://www.oxid.it/cain.html [Accessed on 17.02.2015]

[9] Configuring Cain to Sniff Passwords on Wireless LAN Adapter, Available from: http://www.erasparsa.com/configuring-cain-to-sniff-passwords-on-wireless-lan-adapter, [Accessed on 17.02.2015]

[10] Anatomy of an ARP Poisoning Attack, Nachreiner C., Available from: http://www.watchguard.com/infocenter/editorial/135324.asp [Accessed on 17.02.2015]

[11] Wireless hacking - a WiFi hack by cracking WEP, Reddy, S.V. ; Sai Ramani, K. ; Rijutha, K. ; Ali, S.M. ; Reddy, C.P. 2nd International Conference on Education Technology and Computer (ICETC), 2010

[12] Tripathi, N. ; Mehtre, B.M. Analysis of various ARP poisoning mitigation techniques: A comparison, International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014

[13] Vahidin Dzaltur, DIGITAL FORENSIC INVESTIGATION, COLLECTION AND PRESERVATION OF DIGITAL EVIDENCE, Master Science dissertation, 2014 International Burch University

[14] Rainbow Crack, Available from: http://project-rainbowcrack.com/ [Accessed on 03.02.2015]

[15] Kumar, H. ; Kumar, S. ; Joseph, R. ; Kumar, D. ; Singh, S.K.S. ; Kumar, A. ; Kumar, P. Rainbow table to crack password using MD5 hashing algorithm IEEE Conference on Information & Communication Technologies (ICT), 2013

[16] Papantonakis, P. ; Pnevmatikatos, D. ; Papaefstathiou, I. ; Manifavas, C. Fast, FPGA-based Rainbow Table creation for attacking encrypted mobile communications. 23rd International Conference on Field Programmable Logic and Applications (FPL), 2013

[17] Kalenderi, M. ; Pnevmatikatos, D. ; Papaefstathiou, I. ; Manifavas, C. Breaking the GSM A5/1 cryptography algorithm with rainbow tables and high-end FPGAS 22nd International Conference on Field Programmable Logic and Applications (FPL), 2012

[18] Theocharoulis, K. ; Papaefstathiou, I. ; Manifavas, C. Implementing Rainbow Tables in High-End FPGAs for Super-Fast

Password Cracking International Conference on Field Programmable Logic and Applications (FPL), 2010
[19] Refresher Series - Capturing and cracking SMB hashes with Cain and Half-LM rainbow tables. Available from: http://www.infosecisland.com/blogview/22807-Refresher-Series-Capturing-and-cracking-SMB-hashes-with-Cain-and-Half-LM-rainbow-tables.html [Accessed on 18.02.2015]
[20] EnCase, Available from: https://www.guidancesoftware.com/products/Pages/encase-forensic/overview.aspx [Accessed on 06.02.2015]

miproBIS

International Conference on

BUSINESS INTELLIGENCE SYSTEMS

Steering Committee

Chairs: Boris Vrdoljak, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Cesena, Italy
Marko Banek, University of Zagreb, Croatia

Members: Alberto Abello Gamazo, Polytechnic University of Catalonia, Barcelona, Spain
Ladjel Bellatreche, University of Poitiers, France
Ljiljana Brkić, University of Zagreb, Croatia
Alfredo Cuzzocrea, University of Calabria, Rende, Italy
Jamel Feki, University of Sfax, Tunisia
Dragan Gamberger, Ruđer Bošković Institute, Zagreb, Croatia
Paolo Garza, Polytechnic University of Turin, Italy
Igor Mekterović, University of Zagreb, Croatia
Damir Pintar, University of Zagreb, Croatia
Zoran Skočir, University of Zagreb, Croatia
A Min Tjoa, Vienna University of Technology, Austria
Mladen Varga, University of Zagreb, Croatia
Robert Wrembel, Poznan University of Technology, Poland

INVITED PAPER

Spatial OLAP for agri-environmental data and analysis: Lessons learned

Sandro Bimonte
Irstea, TSCF
9 Avenue B. Pascal, 67178, Aubiere, France
Sandro.bimonte@irstea.fr

Abstract - Spatial Data warehouses and Spatial OLAP (SOLAP) systems allow analyzing huge volumes of georeferenced datasets. SOLAP has been successfully applied in several domains such as marketing, health, urbanism, etc. Few of SOLAP applications in the agricultural context have been proposed. The analysis of work reveals specific issues related to spatio-multidimensional analysis of agri-environmental data. Therefore, in this paper, we present our theoretical and implementation framework for agricultural SOLAP applications.

I. INTRODUCTION

It has been estimated that about 80% of the data stored in corporate databases integrates geographic information [12]. Spatial Data Warehouses (SDWs) and Spatial OLAP (SOLAP) systems allow analyzing huge volumes of georeferenced datasets [1][11]. SOLAP has been successfully applied in several domains such as marketing, health [2], urbanism, etc. The success of SOLAP technologies is mainly based on the capabilities of such systems to analyze large datasets in an interactive way, and provide simple and user-friendly visual analytics interfaces [2]. By a deep study of most important related work about SOLAP analysis of agri-environmental data, and our several years' experience in such kind of projects (EnergeTIC [3], EDEN [4], DispEAU [5], Fresqueau [6], VGOLAP [7], STORI [8], Myriphique [9], and FieldSOLAP [10], VagueSOLAP [17]), we describe in this paper main questions related to multidimensional analysis of agri-environmental data, highlighting theoretical and implementation issues. To remedy to these problems, we have proposed a new framework for the integration of agri-environmental data in SOLAP systems. In particular, our framework supports *conceptual modeling for vector and field spatio-multidimensional data*, and *handling quality issues related to spatial data precision and logical coherence*. Secondly, we propose a generic SOLAP architecture to deal with some implementation issues (i.e. integration, warehousing, analysis and geovisualization). Finally, to take into account skills of agri-environmental decision-makers and the complexity of their applications, we have proposed a *rapid prototyping methodology* based on our conceptual model.

The paper is organized in the following way: Section 2 presents main concepts of spatial data and SOLAP; An overview of existing work on SOLAP for agri-environmental data is detailed in Section 3; Lessons learned are detailed in Sec 4; Our framework is presented in Section 5. Section 6 concludes the paper and presents open issues.

II. MAIN CONCEPTS

In this section we provide main concepts of spatial data (Sec 2.1) and Spatial OLAP (Sec 2.2).

A. Spatial data

Geographic information is typically represented according two models: vector and field [13]. The former represents geographic data using points, lines and polygons. Vector model is well suited for representing human-built geographic objects such as parcels, buildings, etc. Vector model allows spatial analysis by means of topological, metric and ordinal predicates. Field model represents the space as a continuous field. Fields have to be discretized to be represented into computers. These representations can be grouped into two categories: incomplete (e.g. grid points) and complete (e.g. raster). Fields are very adapted for modeling spatial phenomena such as pollution, temperature, etc. They allow a point by point analysis through the Map Algebra operators [13]. Geographic Information Systems (GISs) (such as Qgis, ArcGIS, etc.) allow storing, querying, analyzing and visualizing (vector and field) geographic information [14]. Usual end-users of those technologies are first and foremost decision-makers who are rarely fully aware of the issues related to spatial data quality [15]. The (spatial) data integration of multiple heterogeneous data sources such as remote sensing systems, sensor networks, etc., may easily lead to a number of data quality problems [15] such as imprecision, incompleteness, logical inconsistencies, etc. And yet, vector data are primarily "false but useful" models about the reality of interest since they define geographic phenomena by means of crisp boundaries even if it is not always possible to define exactly where objects begin and end (due to spatial vagueness) [16]. With such a choice of representation, a clear gap is created between majority of real world phenomena and their formal representation in spatial databases [17]. Decision-makers are then exposed to erroneous analysis due to the uncertainty issues on data sources.

B. Spatial OLAP

SDW and SOLAP systems integrate GIS and OLAP functionalities in a unique analysis framework [1]. Usually, warehoused vector data are stored according to the spatio-multidimensionnal model (i.e. data cube) [11].

Data is organized in (spatial) dimensions and facts. Spatial dimensions represent the analysis axes, and are organized into hierarchies with spatial levels (for example cities, departments and regions). A spatial level is a level with geometric data. Facts represent the analysis subjects and they are described by numerical attributes, called measures (for example quantity of sold products). Measures are explored with the SOLAP operators that allow navigating into the SDW [2]. Common operators are Spatial Slice that allows select a subset of warehoused data using topological predicates, and Spatial Drill that allows navigating into hierarchies aggregating measures values with SQL aggregation operators (i.e. MIN, MAX, …) [2]. For example, in a retail application a SOLAP query can be: "*What is the total quantity of sold products by year, product and city at 50km far from Paris?*". Spatial OLAP extends OLAP technologies integrating the spatial representation of the geographic information to enhance multidimensional analysis with spatial analysis and the OLAP visualization with cartographic representations, as shown on Figure 1 using the SOLAP tool presented in [18].

Figure 1. Retail SOLAP application visualization

A typical SOLAP architecture is composed of a Spatial DataBase Management System (DBMS) to store (spatial) data (for example Oracle, PostGIS, etc.); a SOLAP server, which implements the SOLAP operators (for example GeoMondrian, etc.); and a SOLAP client (e.g. Map4Decision), which combines and synchronizes tabular, graphical, and interactive maps to visualize and trigger SOLAP queries [18]. In particular, data collected by means of several heterogeneous data sources are integrated in the SDW tier by means of Spatial Extraction, Transformation, and Loading tools (Spatial ETL) (such as Spatial Data Integrator), which provide a set of operators to work with spatial data (e.g., rename, spatial union, etc.). This kind of SOLAP architecture is called "OLAP-GIS integrated". Indeed, three kinds of SOLAP architecture have been defined [2]. "GIS dominant" SOLAP tools do not present the SOLAP server, and they define SOLAP queries using classical spatial SQL queries. Therefore, they don't take advantage from hierarchical navigation and advanced calculation features of SOLAP servers. "OLAP dominant" tools do not provide advanced GIS functionalities. They allow only simple cartographic visualization, limiting spatial analysis capabilities. Finally, it has been recognized that "OLAP-GIS" solutions are best suited for an effective SOLAP analysis [2]. Finally, SOLAP applications are based on complex multidimensional models and systems, which require particular effective design methodologies [11]. Indeed, it has been recognized that (S)OLAP projects fail usually for two reasons: data quality and correctness of the multidimensional implemented model [26]. Indeed, several DW design methodology taking into account available data sources (i.e. data-driven), users decisional needs (i.e. users driven), and both (hybrid) have been proposed [26].

III. Related Work

In this section, we detail most relevant existing work on SOLAP for the analysis of agricultural and environmental data.

In the agricultural context, the authors of [19] analyze cotton productions. They define 13 data cubes about analysis of crops according to different aspects, such as soil resources, socio-economics resources, etc., using a hybrid design methodology. Quality problems related to dimensional modeling are deeply studied. In particular authors point out the problem of multi-granular data, which consists in data sources that are not all available at the same dimensions levels. They overcome this problem by disaggregating data, when possible, using interpolation functions. The data cubes are implemented in a "OLAP dominant" solution. [20] provides a SOLAP system for the analysis of animal and crop resources. Authors conceive their multidimensional model by taking into account available data and decision-makers' needs, and validate it using a usability test (e.g. "Does the data warehouse provide sufficient information for decision-making?"). Therefore, they adopt a hybrid methodology. The authors define several data cubes providing a constellation schema based on classical dimensions (such as the product dimension) and complex dimensions such as the spatial and the temporal dimensions, which rise multi-granular problems as previously described. The multidimensional model is implemented in a "OLAP-GIS integrated" architecture. [21] studies precision dairy farming (for example the analysis of quantities of produced milk per year and farm). They define two data cubes: one for the milk analysis and another for the treatments of the animals, pointing out the versioning problem (for example cows can change groups during time). To provide decision-makers with "usable" indicators they associate some metadata (such as description, unit, etc.) to measures. Their work is based on an hybrid methodology where decision-makers exchange with DW experts using a multidimensional conceptual model (mE/RM), which is an extension of the Entity Relationship (ER) model. The implementation is based on relational data model and some examples of decisional SQL queries are described. The analysis of forestry data is briefly described in [22], where authors propose a complex multidimensional model with classical temporal and spatial dimensions to measure the area of forest. The main issue described in this work is (yet) the versioning problem related to the dependence of the spatial dimension according time.

In the environmental domain, [23] provides air and water pollution monitoring. The authors define a spatio-multidimensional model with a spatial dimension and some measures representing pollutants values. An example of SOLAP query is "What are pollutant values per year and zone?". On the top of the spatio-multidimensional model, a risk model is used to create

alerts when pollutants exceed tolerable values. [24] proposes a spatio-multidimensional model for the analysis of coastal risk erosion. They provide a complex model with several spatial, temporal and thematic dimensions, and measures. Their model allows answering queries like: "What is the total number of houses at risk, per year and municipality?". Authors define complex dimensions composed for example with alternative hierarchies and raster data (i.e. field model). Aggregation of risk measures is based on a complex risk model. A classical logical model based on the star-schema model is described. [25] also studies water quality. For that goal, authors propose a spatio-multidimensional model with seven dimensions and several measures, conceived using an ad-hoc new hybrid design methodology centered on the decision-makers. For the spatial data integration, they propose using GIS routines, and they implement the model using a ROLAP architecture coupled with a web-based GIS.

We have also conducted some projects in the agricultural and environmental domains, addressing some particular research issues. For example, in the EnergeTIC [3] and EDEN [4] projects, we investigate using SOLAP to analyze energy consummation data of agricultural farms, which comes from sensors installed on farms' agri-equipment (for example a tractor). Sensors data presents several quality problems (for example missing and incoherent data). Quality issues related to vague data (Sec. 2.1), geometries with not well defined boundaries, have been studied in a project aiming at the analysis of usage of sewage sludge produced by wastewater treatment plant used as crop fertilizer in agriculture [17]. Finally, in the FieldSOLAP project [10], we analyze urban odor propagation. Since odor is a continuous phenomenon, we represent it using grid points (i.e. field model) (Sec. 2.1).

IV. LESSONS LEARNED

In this section we discuss multidimensional modeling and architectural issues of previous described work. In particular we address issues related to dimensions (i.e. *What kind of dimensions are defined?)* to facts (i.e. *How many facts are defined?*), analysis (i.e. *What aggregation functions and derived measures are defined?*) and design (i.e. *How the spatio-multidimensional is designed?*). From an analysis of related work emerges that in the most of them dimensions are spatial and temporal ones, since agricultural and environmental data are strongly related to space and time. Dimensions and SOLAP operators are usually complex since agricultural and environmental domains need complex indicators having impacts of aggregation methods and hierarchies design. Moreover, to define a multidimensional model that answers to decision-makers needs, a hybrid design methodology is needed. We also study issues about implementation and data sources. Quality is an important issue since data source is usually "bad" quality data needing ad-hoc methodologies. About implementation, the need for a ROLAP architecture coupled with GIS also emerges, as well as adopting ad-hoc tools for the integration and loading data in the SDW. To resume, warehousing and OLAPing agri-environmental data is a complex task implying a set of design and implementation features that are at an embryonic stage yet, and consequently they are the focus of research work. In detail, warehousing and OLAPing agri-environmental data need:

- Using advanced spatio-temporal multidimensional conceptual models to allow decision-makers and DW experts to discuss and design complex dimensions, facts, and aggregation methods [11][24][27];
- Supporting vector and field data [24][10][28][29];
- Adopting a hybrid methodology to define successful multidimensional model taking into account decision-makers functional requirements and complex data sources [26] [30];
- Handling quality issues (spatial precision, logical coherence, etc.) [4] [9] [11] [15]
- Implementation in commercial OLAP-GIS systems. Integration of OLAP and GIS systems is necessary for an effective spatio-temporal analysis. However, this integration requires important implementation efforts (time, money and human resources), implying that an ad-hoc solution is not always feasible. However, actually SOLAP systems have finally reached a good maturity in terms of stability and functionalities. They move in recent years from prototypes to commercial-industrial systems [31].
- Using extendible Spatial DBMS and/or Spatial OLAP server for the integration of complex calculation methods in OLAP analysis. Indeed, some work propose to extend OLAP aggregation functions in the DBMS tier using native SQL methods [32] or in the (S)OLAP server tier using MDX user-defined functions [10], or in the SOLAP client with new advanced visual analytics features. In this way, agri-environmental decision-makers can represent their analysis indicators as aggregated measures, and transparently analyze their data by the simple exploration of warehoused spatial data.

V. SOLAP FRAMEWORK

In this section we present our framework for integration of agri-environmental data into SOLAP systems, developed in the COPAIN team of Irstea. The EDEN project aims at the multidimensional analysis of energy-use [4]. Energy data is acquired using mobile sensors, installed on the different farms' equipment (i.e. tractors), to retrieve fuel consumption. The implemented SOLAP application allows decision-makers to manage their farm activities.

A. Advanced spatio-temporal multidimensional conceptual model: ICSOLAP UML Profile

In order to support advanced conceptual modeling as described in Sec. 4 we use in our framework the ICSOLAP UML Profile presented in [27] [33]. Our UML profile allows modeling advanced multidimensional issues [27]: complex dimensions (non-covering, non-onto and multiple hierarchies) [11], spatial dimensions [11], multi-granular facts [11], complex aggregations, etc. ICSOLAP

allows representing spatial dimensions modeled using the vector [27] and field models [33].

The UML profile defines a stereotype for each multidimensional element. For example a fact is represented with a stereotyped <<Fact>> class, whose attributes are measures (<<NumericalMeasure>>). A spatial vector level is represented using the stereotype <<SpatialAggLevel>> class that designs spatial dimension levels whose geometries are represented with geometric attributes (point, line, polygon, etc.) (stereotyped <<LevelGeometry>>). An example is shown on figure 2 where (a simplified model) the SOLAP application of the EDEN project is represented. Here the spatial level "Plot" represents the plots of the agricultural farm.

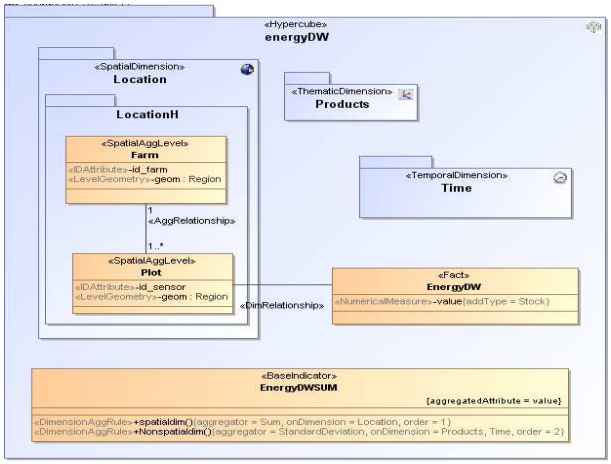


Figure 2. ICSOLAP: vector model

In order to represent field data we define <<FieldAggLevel>> as an extension of <<SpatialAggLevel>>. <<FieldAggLevel>> presents a interpolation function as a tagged value "Continuity Interpolation". It allows estimating the measure values in any point of the space. Figure 3 shows an example of a regular grid points where the *bicubic interpolation* method is used to represent the temperature phenomenon.

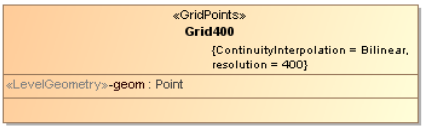


Figure 3. ICSOLAP: field model

Although, some very expressive spatio-multidimensional conceptual models based on extensions of ER have been proposed yet such as [11], we prefer to adopt the UML formalism for several reasons. Indeed, using UML eases the exchange between decision-makers and DW experts since UML model is easy to read and interpret. Moreover, UML has been adopted by the software designers' industrial community and a growing number of the UML-based tools allows for easier and quicker developments of software applications (by automating certain tasks) have been developed. Computer-aided software engineering (CASE) is the scientific application of a set of tools and methods to a software system which is meant to result in high-quality, defect-free, and maintainable software products. These tools allow automatic translation of UML models into errors free code and implementation routines (SQL, Java, etc.). In our framework, we have implemented our profile in the commercial CASE tool MagicDraw.

B. Quality: Logical coherence and spatial precision

As described in Sec 3 addressing spatial data quality issues is crucial for SOLAP projects. In this context, Integrity Constraints (ICs) play a key role in improving consistency, accuracy and completeness of the stored data [34]. ICs are assertions that can be defined in conceptual models to identify incorrect data and prevent their insertions in the database. For example, in the vector context, these constraints could be used to define the expected topological relationships between spatial objects. In order to handle with logical coherence issues in SOLAP, in [4] we propose a framework, using the ICSOLAP profile, to express IC using Spatial OCL. Spatial OCL is an extension of OCL with topological predicates [16]. The Object Constraint Language (OCL) is a declarative language for describing rules that apply to UML models. ICs are expressed at a conceptual level, and then automatically implemented in the "OLAP-GIS" architecture. An example of IC representing that the geometry of a plot must be contained in the geometry of the farm using Spatial OCL is the following:

```
context Plot inv DataIC1:
self.geo.isInside (self.Farm.geom)
```

Another important quality issue of spatial data is the geometric precision: geometries do not have exact boundaries (i.e. spatial vague data). Supporting vague data in DBMS is an open research question [37]. Since, existing SOLAP systems do not provide any tools for handling spatial vague data, in [17] we propose a new design methodology where analysis problems related spatial precision issues are replaced by risk usage policies. In particular, to each vague member we associate a risk usage in the multidimensional analysis (level risk, aggregation risk, etc.). Consequently, these risks are handled by some actions on the spatio-multidimensional model that aim at eliminating (such as delete spatial level action) or communicate the risks (for example visualize risk aggregated value). As for ICs, [17] allows to define the elements of our methodology at a conceptual level using the ICSOLAP profile. After that, the new spatio-multidimensional model is automatically implemented. An example is shown on figure 4 where the analysis problems related to plots vague geometries is handled by a visualization policy.

C. Hybrid methodology: ProtOLAP

As highlighted in the previous sections, designing SOLAP models is quite difficult because spatial data and analysis are complex. Secondly, it needs hybrid approach since data and decision-makers knowledge should be taken into account for an effective SOLAP analysis. Thirdly, agri-environmental decision-makers usually are (S)OLAP and Information Technology unskilled users. In this context in order to define a design DW methodology tackling above described issues, in [35] we propose a new methodology and a tool called ProtOLAP (Figure 5).

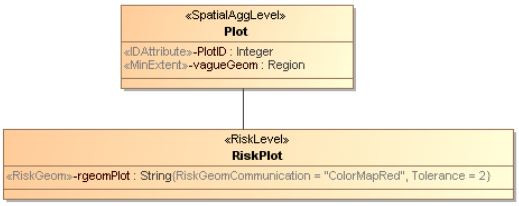


Figure 4. ICSOLAP quality: vague data

ProtOLAP supports the easy and rapid design of spatio-multidimensional models. In details, decision-makers and DW experts define their conceptual model using ICSOLAP UML Profile. After that, the schema of the SDW and the SOLAP Server tiers are automatically generated. The data warehoused is consequently fed with sample data, allowing decision-makers to play with warehoused data in an SOLAP client to validate the spatio-multidimensional model. ProtOLAP is well adapted to agri-environmental applications since: (i) automatic prototyping allows for a free-error implementation of complex spatio-multidimensional models; (ii) visualization of the resulting analysis easies unskilled decision-makers to translate (and validate) their tacit analysis needs on a spatio-multidimensional model. The integration of vector and field data as well as the support for logical coherence and spatial precision quality issues have also been integrated in ProtOLAP since these features are based on the conceptual design in ICSOLAP UML profile with its automatic implementation. Consequently, ProtOLAP can also remediate to projects failure due to spatial data quality problems. We have successfully tested ProtOLAP in the context of the EDEN project.

D. Implementation

In this section, we detail implementation issues. The architecture used in our framework is described in Figure 5. As described in section 3, agri-environmental analysis needs more complex aggregation than classical SQL ones. In a "OLAP-GIS integrated" architecture two possibilities to solve this issue can be implemented. The first one is to implement complex aggregation inside the Spatial DBMS. Indeed, when the Spatial DBMS allows for extension mechanisms, it is possible to add new functionalities directly accessible as native SQL functionalities. In our framework, we use PostGIS. PostGIS is a Spatial DBMS extending Postgresql to natively support spatial data storage and querying. PostGIS allows to defined SQL user-defined functions written in SQL, C, PL/SQL, etc., which are then usable as native SQL functions. Therefore, we implement decision-makers agri-environmental methods, such as percentile, using user-defined aggregation functions (Boulil et al., 2014). After that, these aggregation functions can transparently be used by any SOLAP server and client. The main advantage of this approach concerns performance since the Spatial DBMS natively provide optimization calculation techniques, but when analysis methods require complex navigation and calculation methods, their implementation becomes difficult. Indeed, as for the "GIS dominant" architecture, the functionalities of the SOLAP server cannot be used. The SOLAP server can be also used for the definition of such complex method. Indeed, some SOLAP servers such as GeoMondrian, which we use in our framework, allow defining user-defined functions that can be transparently used in MDX queries. GeoMondrian is a open source java-based OLAP Server that extends Mondrian with spatial operators. Indeed, GeoMondrian allows defining using Java ad-hoc user defined aggregation functions [10]. Contrary to the Spatial DBMS-based approach, this approach allows using SOLAP server functionalities for the calculation, but it presents a main limitation: calculation are performed in the main memory, which can represent an important drawback when the number of facts to aggregate is huge since response answer time becomes intolerable for on-line analysis.

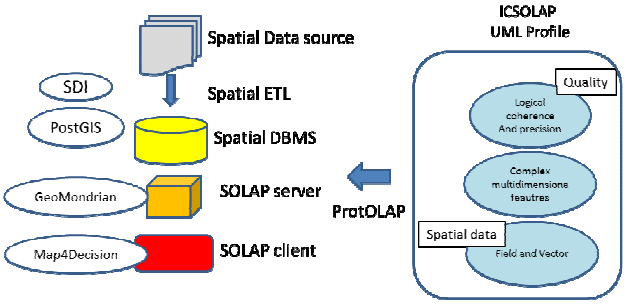


Figure 5. Architecture

As described in [31] few commercial OLAP-GIS integrated tools are available. In our framework we use the SOLAP client provided by Map4Decision (Map4Decision, 2014). According to the visual analytics mantra [38], Map4Decision integrates intelligent cartography concepts that support the automatic creation of maps while conforming to the rules of visual communication and semiology (colors, symbols, patterns, etc.). The user can easily navigate from one type of map to another (simple maps, multimaps, multivariate maps, temporal maps, etc.), explore geographic data at various levels of detail (e.g. from countries to provinces, to cities, or vice-versa), while keeping a coherent semiology. Map4Decision client configuration is very easy and it is provided by a simple "point and click" user interface. Map4Decision supports XMLA, which is the de-facto standard web-service for OLAP tools. Therefore, in our framework we connect Map4Decision on the top of GeoMondrian to provide advanced cartographic visual reports using XMLA.

VI. CONCLUSIONS AND FUTURE WORK

Spatial OLAP systems are effective GeoBusiness Intelligence technologies. Their application in agri-environmental context is few studied and is quite recent. In this paper we highlight main problems related to spatio-multidimensional analysis of agri-environmental data. We provide an overview of our new framework for this kind of SOLAP projects. (Un)lucky several issues rest unexplored. These issues represent our current work. Although, conceptual modeling is quite advanced, physical and logical modeling for SDW are not studied enough [36]. Indexes, materialized views have to be defined to take into account all particular aspects of geographic information (e.g. continuity, spatial predicates,

etc.). Finally in [18] the first generic model of geovisualization in SOLAP systems has been provided. But, an exhaustive study of well-suited geovisualization methods, and/or their adaption, should be provided to fully exploit visual analytical capabilities of SOLAP clients.

REFERENCES

[1] Y. Bédard and J. Han, "Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery," Geographic Data Mining and Knowledge Discovery, 2001.

[2] Y. Bédard, S. Rivest, and M. Proulx. 2006. "Spatial on-line analytical processing (solap): Concepts, architectures, and solutions from a geomatics engineering perspective," Data Warehouses and OLAP: Concepts, Architecture, and Solutions, vol 14, pp. 298–319, 2006.

[3] S. Bimonte, K. Boulil, J. Chanet, and M. Pradel: "Definition and Analysis of New Agricultural Farm Energetic Indicators Using Spatial OLAP, " Proc. of ICCSA 2012, pp. 373-385.

[4] K. Boulil, S. Bimonte, and F. Pinet, "Spatial OLAP integrity constraints: From UML-based specification to automatic implementation: Application to energetic data in agriculture," Journal of Decision Systems, vol 23(4), pp. 460-480, 2014.

[5] S. Bimonte, K. Boulil, F. Pinet, and M. Kang, "Design of Complex Spatio-multidimensional Models with the ICSOLAP UML Profile - An Implementation in MagicDraw," Proc. of ICEIS 2013, pp. 310-315.

[6] K. Boulil, F. Le Ber, S. Bimonte, C. Grac, and F. Cernesson, "Multidimensional modeling and analysis of large and complex watercourse data: an OLAP-based solution," Ecological Informatics, vol. 24, pp. 90-106, 2014.

[7] S. Bimonte, O. Boucelma, O. Machabert, and S. Sellami, "A new Spatial OLAP approach for the analysis of Volunteered Geographic Information," Computers, Environment and Urban Systems, vol. 48, pp. 111-123, 2014.

[8] L. Sautot, S. Bimonte, L. Journaux, and B. Faivre, "A Methodology and Tool for Rapid Prototyping of Data Warehouses Using Data Mining: Application to Birds Biodiversity," Proc. of MEDI 2014, pp. 250-257.

[9] K. Boulil, F. Pinet, S. Bimonte, N. Carluer, C. Lauvernet, B. Cheviron, A. Miralles, and J. Chanet, "Guaranteeing the quality of multidimensional analysis in data warehouses of simulation results: Application to pesticide transfer data produced by the MACRO model," Ecological Informatics, vol. 16, pp. 41-52, 2013.

[10] M. Zaamoune, S. Bimonte, F. Pinet, and P. Beaune, "A New Relational Spatial OLAP Approach for Multi-resolution and Spatio-multidimensional Analysis of Incomplete Field Data," Proc. of ICEIS 2013, pp. 145-152.

[11] E. Malinowski, and E. Zimányi, Advanced Data Warehouse Design From Conventional to Spatial and Temporal Applications, Springer, Heidelberg, 2008.

[12] C. Franklin, "An Introduction to Geographic Information Systems: Linking Maps to databases," Database, vol. 15(2), pp. 13-21, 1992.

[13] C. Tomlin, Geographic Information Systems and Cartographic Modeling, Prentice Hall, 1990.

[14] P. Longley, M. Goodchild, J. Macguire and D. Rhind, Geographic Information Systems and Science, John Wiley & Sons, 2001.

[15] M. Gervais, Y. Bedard, M. Levesque, E. Bernier and R. Devillers, "Data Quality Issues and Geographic Knowledge Discovery,". Geographic Data Mining and Knowledge Discovery, pp. 99-115, 2009

[16] L. Bejaoui, F. Pinet, M. Schneider and Y. Bédard, "OCL for formal modelling of topological constraints involving regions with broad boundaries," GeoInformatica, vol. 14(3), pp. 353-378, 2010.

[17] É. Edoh-Alove, S. Bimonte, and Y. Bédard, "A New Design Method for Managing Spatial Vagueness in Classical Relational Spatial OLAP Architectures," Proc. of ICCSA 2014, pp. 774-786.

[18] S. Bimonte, "A Generic Geovisualization Model for Spatial OLAP and its Implementation in a Standards-based Architecture," Ingénierie des Systèmes d'Informations, vol. 14(5), pp. 97-118, 2014.

[19] Chaturvedi, K., Rai, A., Dubey, V., and P. Malhotra, "On-line Analytical processing in Agriculture using Multidimensional Cubes," International Indian Society Agricultural Statistics Journal, vol. 62(2), pp. 56-64, 2008.

[20] S. Nilakanta, K. Scheibe A. and Rai, "Dimensional issues in agricultural data warehouse designs," Journal Computers and Electronics in Agriculture, vol. 60(2), pp. 263-278, 2008.

[21] C. Schulze, J. Spilke and W. Lehner, "Data modeling for Precision Dairy Farming within the competitive field of operational and analytical tasks," Computers and Electronics in Agriculture, vol. 59(1), pp. 39-55, 2007.

[22] M. Miquel, Y. Bédard, A. Brisebois, J. Pouliot, P. Marchand and J. Brodeur, "Modeling Multidimensional Spatio-temporal Data Warehouse in a Context of Evolving Specifications," Proc. of Int. Symp. International Society for Photogrammetry and Remote Sensing 2002.

[23] C. Radulescu M. and Radulescu, "A Multidimensional Data Model for Environment Protection," Proc. of 12th WSEAS International Conference on Computers 2008, pp. 1101–1106.

[24] A. Mardkheh, M. Mostafavi, Y. Bédard, B. Long, and E. Grenier, "Using geospatial business intelligence paradigm to design a multidimensional conceptual model for efficient coastal erosion risk assessment," Journal of Coastal Conservation: Planning and Management, May 2013.

[25] M. McGuire, A. Gangopadhyay, A. Komlodi, and C. Swan, "A user-centered design for a spatial data warehouse for data exploration in environmental research,". Ecol. Inform. vol. 3, pp. 273–285, 2008.

[26] O. Romero, and A. Abelló, "A Survey of Multidimensional Modeling Methodologies,". IJDWM, vol. 5(2), pp. 1-23, 2009.

[27] K. Boulil, S. Bimonte and F. Pinet, "Conceptual Model for Spatial Data Cubes: A UML Profile and its Automatic Implementation," journal of Computer Standards & Interfaces, vol 38, pp. 113-132 2015.

[28] A. Vaisman, and E. Zimányi, "A multidimensional model representing continuous fields in spatial data warehouses," Proc. of GIS 2009, pp. 168-177, 2009

[29] Leticia I. Gómez, Alejandro A. Vaisman and E.Zimányi, "Physical Design and Implementation of Spatial Data Warehouses Supporting Continuous Fields," Proc. of DaWak 2010, pp. 25-39.

[30] P. Giorgini, S. Rizzi and M. Garzetti, "GRAnD: A goal-oriented approach to requirement analysis in data warehouses," Decision Support Systems, vol. 45(1), pp. 4-21, 2008.

[31] M. Golfarelli, M. Mantovani, F. Ravaldi, and S. Rizzi, "From Business Intelligence to Location Intelligence with the Lily Library," Proc. of DOLAP 2014, pp. 33-36.

[32] Z. Chen and C. Ordonez, "Efficient OLAP with UDFs," Proc. of DOLAP 2008, pp. 41-48.

[33] M. Zaamoune, S. Bimonte, F. Pinet and P. Beaune, "Intégration des données champs continus incomplets dans l'OLAP : de la modélisation conceptuelle à l'implémentation," Proc. of EDA2013, pp. 18-33.

[34] M. Duboisset, F.Pinet, M. Kang, and M. Schneider, "Precise Modeling and Verification of Topological Integrity Constraints in Spatial Databases: From an Expressive Power Study to Code Generation Principles," Proc of. ER 2005, pp. 465-482.

[35] S. Bimonte, É. Edoh-Alove, H. Nazih, M. Kang, and S. Rizzi, "ProtOLAP: rapid OLAP prototyping with on-demand data supply," Proc of. DOLAP 2013, pp. 61-66.

[36] T.Siqueira, C. Ciferri, V. Times, and R. Ciferri, "The SB-index and the HSB-index: efficient indices for spatial data warehouses," GeoInformatica, vol. 16(1), pp. 165-205, 2012.

[37] A. Pauly, and M. Schneider, "VASA: An algebra for vague spatial data in databases," Inf. Syst., vol. 35(1), pp. 111-138, 2010.

[38] D. Keim, F. Mansmann, J. Schneidewind and H. Ziegler, "Challenges in visual data analysis," Proc. of IV 2006.

PAPERS

Accelerating Time Series Subsequence Matching on the Intel Xeon Phi Many-core Coprocessor

Ruslan Miniakhmetov, Aleksander Movchan and Mikhail Zymbler
South Ural State University, Faculty of Computational Mathematics and Informatics
Chelyabinsk, Russia
miniakhmetovrm@susu.ac.ru, movchan174@gmail.com, mzym@susu.ru

***Abstract*—The problem of time series subsequence matching occurs in a wide spectrum of subject areas. Currently Dynamic Time Warping (DTW) is the best similarity measure but despite various existing speedup techniques it is still computationally expensive. Due to this reason science community is trying to accelerate DTW calculation by means of parallel hardware. There are implementations of DTW-based subsequence matching on GPU and FPGA but there none for accelerators based on the Intel Many Integrated Core architecture. This paper presents a parallel algorithm for time series subsequence matching based on DTW distance adapted to the Intel Xeon Phi coprocessor. The experimental results on synthetic and real data sets are presented and confirm the efficiency of the algorithm.**

***Index Terms*—data mining, time series, subsequence matching, dynamic time warping, parallel computing, OpenMP, Intel Many Integrated Core architecture, Intel Xeon Phi coprocessor**

I. Introduction

A time series represents a collection of data points, measured chronologically. Currently, time series data mining touches upon a wide spectrum of scientific and applied tasks in various subject areas, e.g. medical monitoring [1], economic forecasting [2], climate modeling [3], etc. Being one of the most important problems of time series data mining, subsequence matching assumes that a query sequence and a longer time series are given, and the task is to find subsequences in the longer time series, which match the query sequence.

To measure the similarity between two time series, a number of Euclidean-based distance measures have been suggested but currently the Dynamic Time Warping (DTW) [4] is the most popular similarity measure in many applications [5]. DTW is computationally expensive and there are many approaches that have been proposed to solve this problem, e.g. pruning the subsequences using lower bound of distance [5], computation reusing [6], data indexing [7], early abandoning [8], etc. However, DTW still takes a large part of the total application runtime. That is why there are researches devoted to accelerating subsequence matching by means of parallel hardware, e.g. computer-cluster [9], multicore [10], FPGA and GPU [6].

This work was financially supported by the Ministry of education and science of Russia ("Research and development on priority directions of scientific-technological complex of Russia for 2014-2020" Federal Program, contract No. 14.574.21.0035).

In this regard, in our opinion not enough attention is paid to the Intel Xeon Phi coprocessor based on Intel Many Integrated Core (MIC) [11] architecture. This paper presents a parallel algorithm for time series subsequence matching based on DTW distance adapted for use on a central processor unit (CPU) accompanied with the Intel Xeon Phi coprocessor.

The rest of the paper is organized as follows. Section II gives the formal definition of the problem and briefly considers the Intel Xeon Phi architecture and programming model. In section III a description of the proposed algorithm is given. The results of the experiments evaluating the algorithm are presented in section IV. Section V discusses related work. Section VI contains concluding remarks and directions for future research.

II. Research Background

A. Problem Definition

A *time series* T is an ordered sequence $t_1, t_2, ..., t_N$ of real data points, measured chronologically, where N is a length of the sequence.

A *subsequence* $T_{i,k}$ of time series T is its continuous subset of length k starting at i position.

A *query* Q is a certain subsequence to be found in T.

Subsequence matching problem aims to finding all the subsequences $T_{i,j}$ such that distance $D(T_{i,j}, Q)$ is minimal.

Dynamic Time Warping (DTW) is a similarity measure between two time series which may vary in time or speed. DTW distance between two time series X and Y, where $X = x_1, x_2, ..., x_N$ and $Y = y_1, y_2, ..., y_N$, is denoted as $D(X, Y)$ and defined as follows.

$$D(X,Y) = d(N,N),$$

$$d(i,j) = |x_i - y_j| + min \begin{cases} d(i-1,j) \\ d(i,j-1) \\ d(i-1,j-1), \end{cases}$$

$$d(0,0) = 0; d(i,0) = d(0,j) = \infty;$$
$$i = 1, 2, \ldots, N; j = 1, 2, \ldots, N.$$

B. The Intel Xeon Phi Architecture and Programming Model

The Intel Xeon Phi coprocessor is an x86 many-core coprocessor of 61 cores, connected by a high-performance on-die bidirectional interconnect where each core supports 4×

hyperthreading and contains 512-bit wide vector processor unit (VPU). Each core has two levels of cache memory: a 32 Kb L1 data cache, a 32 Kb L1 instruction cache, and a core-private 512 Kb unified L2 cache. The Intel Xeon Phi coprocessor is to be connected to a host computer via a PCI Express system interface. Being based on Intel x86 architecture, the Intel Xeon Phi coprocessor supports the same programming tools and models as a regular Intel Xeon processor.

There are three programming modes to deal with the Intel Xeon Phi coprocessor: native, offload and symmetric. In native mode the application runs independently, on the coprocessor only. In offload mode the application is running on the host and offloads computationally intensive part of work to the coprocessor. The symmetric mode allows the coprocessor to communicate with other devices by means of Message Passing Interface (MPI).

III. Subsequence Matching on the Intel Xeon Phi Coprocessor

A. Research Roadmap

Development of the parallel algorithm of time series subsequence matching adapted for the Intel Xeon Phi coprocessor consists of the following steps, which will be discussed in detail further.

Firstly, we had developed a parallel version of the UCR-DTW serial algorithm [8]. Using OpenMP technology, we have obtained a parallel algorithm for CPU, keeping in mind a possibility of running this algorithm on the Intel Xeon Phi coprocessor in *native* mode. However, experiments have shown that, despite the one-order speedup of parallel algorithm, it works *slower* on the Intel Xeon Phi coprocessor in native mode than on CPU. This results from low operational intensity of our algorithm, i.e. insufficient FLOPs (floating point operations) per byte of data to be effectively processed on the Intel Xeon Phi coprocessor.

Next, we modified our algorithm combining CPU and coprocessor to process time series. In this version of the algorithm both CPU and the Intel Xeon Phi run the parallel version of the UCR-DTW algorithm developed at the first step. Here we used *offload* mode to transmit code and data to the Intel Xeon Phi. Experiments, where we varied the portion size of data to be transmitted to the coprocessor, show results similar to those obtained at the first step due to the same reason.

Finally, we developed an advanced version of the algorithm. The key point of this version is to support a queue of subsequences on CPU that are offloaded to the coprocessor to compute DTW. This significantly increased the operational intensity of the calculations on the coprocessor and experiments show acceptable performance of the algorithm.

B. Parallel Algorithm for CPU

The UCR-DTW algorithm proposed in [8] is one of the fastest existing subsequence matching algorithms. This algorithm (see Fig. 1; hereinafter, we use UML activity diagrams) uses a cascade estimation of the lower bound of DTW distance. If the lower bound has exceeded some threshold, the DTW distance also exceeds the threshold, so the subsequence can be pruned off.

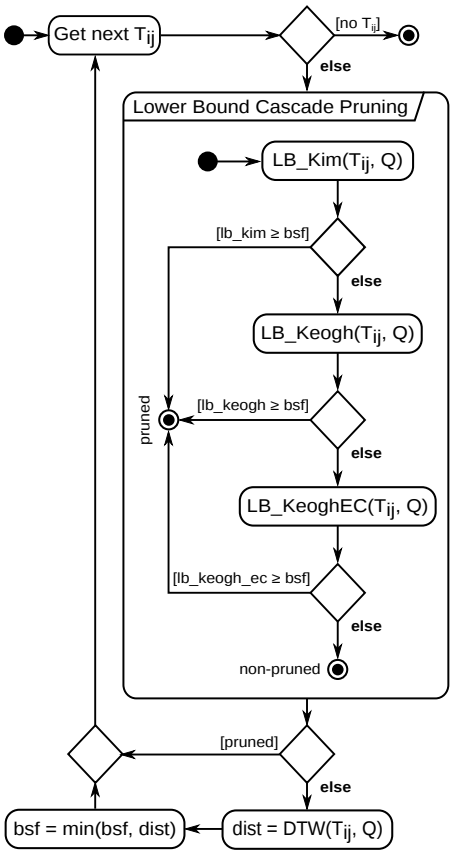


Fig. 1. Serial algorithm

A parallel version of the UCR-DTW algorithm is depicted in Fig. 2. We parallelize the original algorithm using the OpenMP technology. The time series is splitted into equal portions and each portion is processed by a separate OpenMP-thread. Here `UCR-DTW` is a subroutine that implements the original serial algorithm. It uses the `bsf` (best-so-far) shared variable that stores the distance to the nearest subsequence. This allows each thread to prune off unpromising subsequence using lower bounding. Master thread reads new data from a file simultaneously with processing of data that have been read.

We evaluated the obtained algorithm on synthetic time series both on CPU and on the Intel Xeon Phi coprocessor in native mode (see Fig. 6). Although parallel algorithm expectedly surpasses the original algorithm it works slower on the coprocessor in native mode.

This was a result of low operational intensity of our algorithm, i.e. insufficient FLOPs per byte of data to be effectively processed on the Intel Xeon Phi coprocessor.

C. Naïve Parallel Algorithm for CPU and the Intel Xeon Phi

Fig. 3 depicts the modified version of the algorithm. This version is called "naïve", because in comparison with the previous version it only distributes work among CPU and

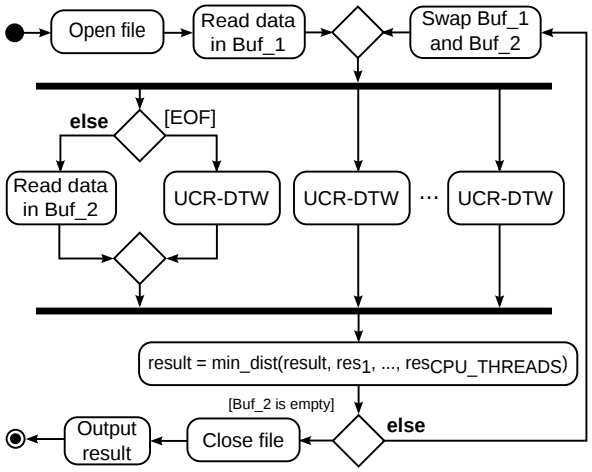


Fig. 2. Parallel algorithm for CPU

coprocessor. Here `ALPHA` is a parameter that determines a portion of data to be transmitted to the coprocessor. We use offload mode to organize data exchange between CPU and the coprocessor. The `min_dist` subroutine chooses subsequences with minimal value of DTW.

As well as in the previous step we evaluated the obtained algorithm on synthetic time series (see Fig. 6). Regardless of the `ALPHA` value this algorithm has worse performance in comparison with the parallel algorithm for CPU.

This is because we still have not increased operational intensity of calculations on the coprocessor. Additionally, `bsf` shared variable can not be synchronized between the CPU and the coprocessor while offloading is performed (the synchronization is possible at the beginning and at the end of offload section). That is why we have more non-pruned subsequences to calculate DTW.

D. Advanced Parallel Algorithm for CPU and the Intel Xeon Phi

The advanced version of the algorithm obtained at the previous step is depicted in Fig. 4.

The idea of the advanced algorithm is to support a queue of subsequences on CPU. Subsequences of the queue are candidates to be offloaded to the coprocessor to calculate DTW. One of the CPU threads is declared as a master and the rest as workers. When the queue is full master offloads it to the coprocessor.

Worker's activity looks like the following. A worker calculates cascade estimates for the subsequence. If it is dissimilar to the query then the worker prunes it off otherwise worker pushes this subsequence to the queue. If the queue is full (and data previously transmitted to the coprocessor have not been processed yet), the worker calculates DTW by itself.

The `UCR-DTW*` (see Fig. 5) subroutine implements worker's behavior as described above.

At the end of offload section the information about most similar subsequences found on the coprocessor is transmitted

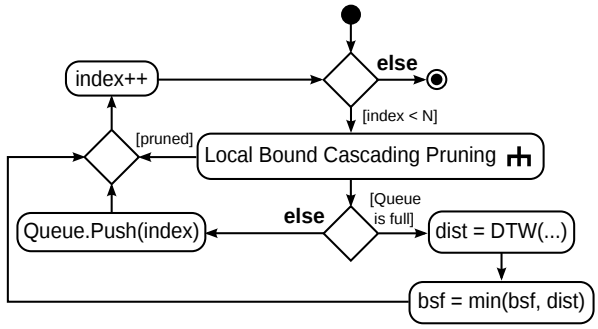


Fig. 5. UCR-DTW* subroutine

to the CPU. The final result is calculated among the most similar subsequences found on the CPU and on the coprocessor.

Additionally, we slightly modified the original source code of DTW calculation to provide vectorization of operations inside the `for` loop.

IV. Experimental Evaluation

To evaluate the algorithm developed we performed experiments on a Tornado SUSU [12] supercomputer's node (see Tab. I for its specifications). In these experiments subsequence matching runtime was measured while varying query length. Experiments have been performed on synthetic and real time series. We also investigated usage time of the coprocessor, impact of queue size on the speedup, and compared our algorithm with analogues for GPU and FPGA.

TABLE I
Specifications of Tornado SUSU supercomputer node

Specifications	Processor	Coprocessor
Model	Xeon X5680	Xeon Phi SE10X
Cores	6	61
Frequency, GHz	3.33	1.1
Threads per core	2	4
Peak performance, TFLOPS	0.371	1.076

A. Performance on Synthetic and Real Data Sets

In the first set of experiments we used synthetic time series generated by one-dimensional random walk comprising of 100 million data points.

Fig. 6 depicts the experimental results on synthetic data sets. The results show that our algorithm is more effective for longer queries. In case of shorter queries the algorithm has the same performance as parallel algorithm for CPU only.

The second set of experiments investigates the algorithm's performance on real electrocardiographic (ECG) data with about 20 million data points (approximately 22 hours of ECG sampled at 250 Hz).

Results of experiments on real data are shown in Fig. 7. Our algorithm shows a three times higher performance than the parallel algorithm for CPU only.

We evaluate what was the contribution of the coprocessor in computations as well. Fig. 8 depicts that increasing the query

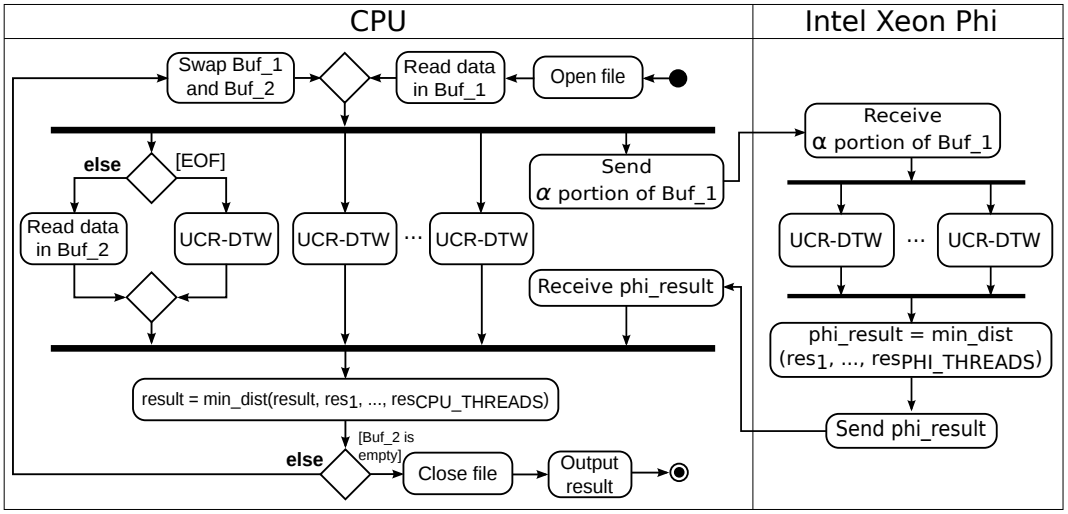


Fig. 3. Naïve parallel algorithm for CPU and the Intel Xeon Phi

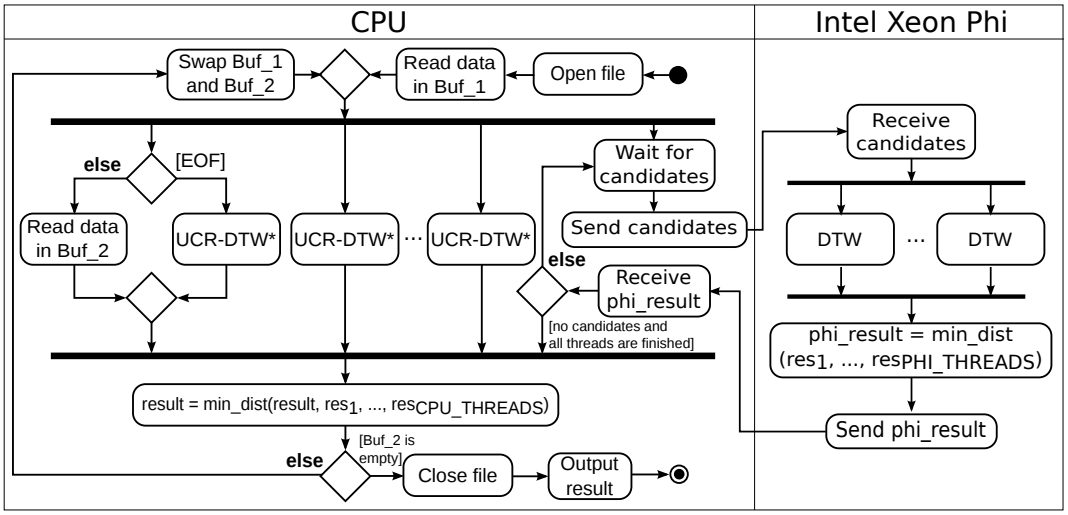


Fig. 4. Advanced parallel algorithm for CPU and the Intel Xeon Phi

length greater than 4000 leads to almost 100% usage of the Intel Xeon Phi coprocessor for both synthetic and real data sets.

B. Impact of Queue Size

The results described in subsection above have been achieved when queue size was 2400. We empirically select this value having performed some experiments beforehand (see Fig. 9).

C. Comparison with Algorithms for GPU and FPGA

We compared the performance of our algorithm with analogues for GPU and FPGA developed in [6]. We repeated the experiments presented in that paper keeping the same length of the time series and query length. The results of the experiments are depicted in Fig. 10.

We took into account that the peak performance of the hardware we used is significantly greater than its counterparts of that paper, i.e. overall peak performance of our hardware was 1.44 TFLOPS whereas GPU as NVIDIA Tesla C1060 had 77.8 GFLOPS and FPGA as Xilinx Virtex-5 LX-330 had 65 GFLOPS.

To provide a "fair" experiment we added to the chart *hypothetical* results for NVIDIA Tesla K40 (1.43 TFLOPS) [13] and Xilinx Virtex-7 980XT (0.99 TFLOPS) [14] multiplying real results of NVIDIA Tesla C1060 and Xilinx Virtex-5 LX-330 by a respective scaling factor. As we can see our algorithm provides significantly higher performance.

Finally, we compared relative costs of performance (see Fig. 11). The charts show that Intel-based solution is a bit more expensive than the latest one from NVIDIA but at the same time provides significantly higher performance.

V. Related Work

A substantial amount of work has been done in the area of time series data mining, including subsequence matching

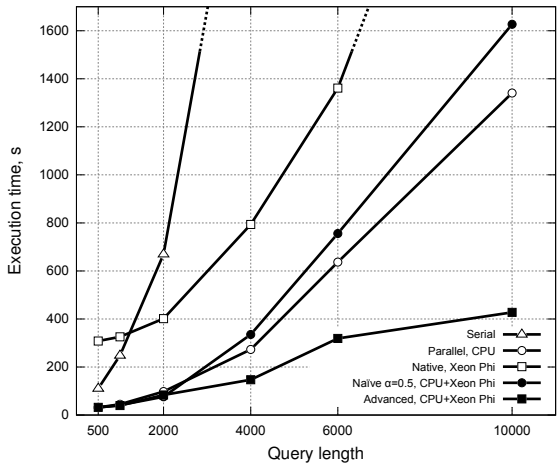


Fig. 6. Algorithm's performance on synthetic data set

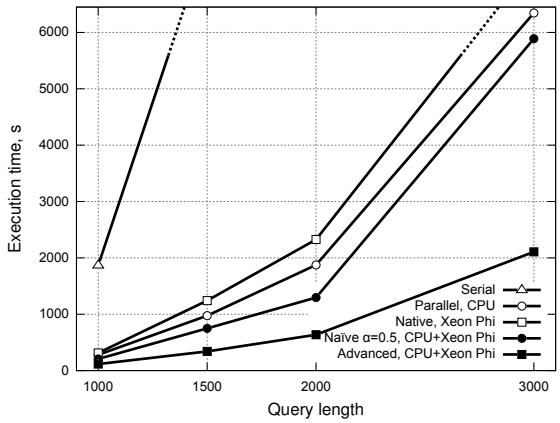


Fig. 7. Algorithm's performance on real data set

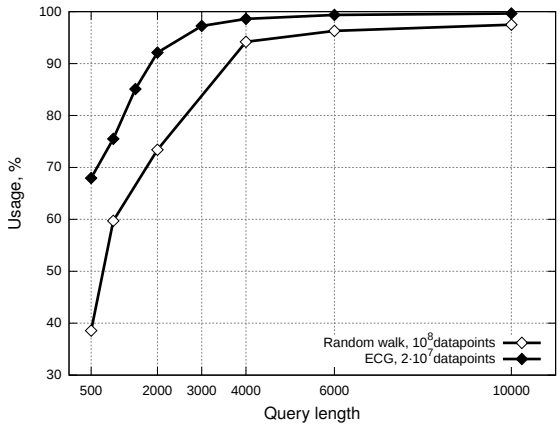


Fig. 8. Usage of coprocessor in experiments

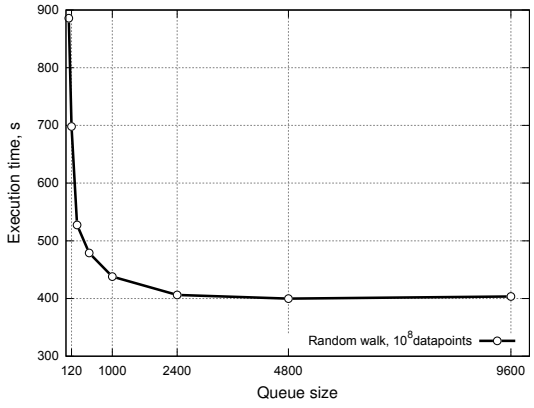


Fig. 9. Impact of queue size on the speedup

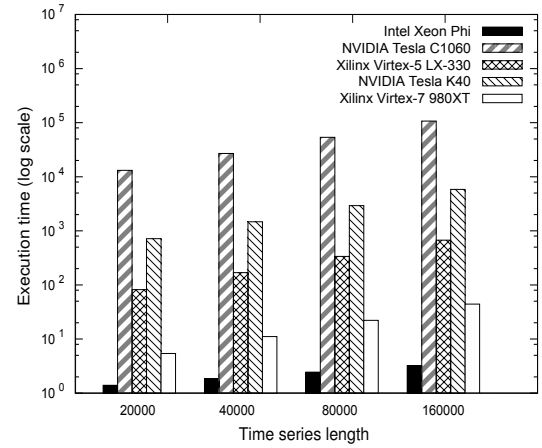


Fig. 10. Comparison of performance

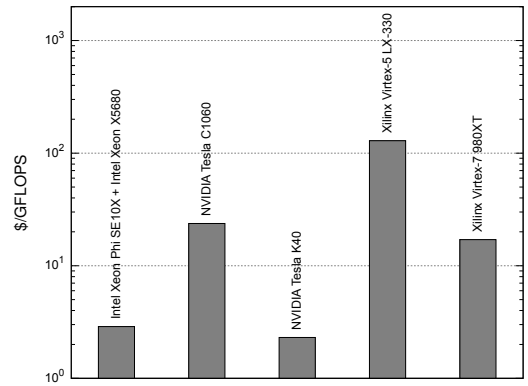


Fig. 11. Comparison of performance cost

problem [15].

One of the earliest researches [16] considers the time series as a set of multi-dimensional rectangles and constructs its spatial index to perform search of subsequences as spatial queries to the index. This work serves as a basis for many researches devoted to improving subsequence matching performance. For instance, the DualMatch algorithm [17] and its improved version, the GeneralMatch algorithm [18], split the time series into non-overlapping windows and query pattern into sliding windows. In [19] and [7] an index interpolation to overcome the storage and CPU time overhead is used. All the former techniques are based on lower bounding of the

Euclidean distance.

Nowadays the Dynamic Time Warping (DTW) [4] measure is recognized as the best similarity measure for many time series subject domains [5]. DTW-based subsequence searching techniques include a suffix tree [20], a segment-based approach based on piecewise time warping [21], an index-based approach based on prefix querying [22] and a comparison optimization approach [23]. An algorithm proposed in [8] integrates all the possible existing speedup techniques and most likely it is the fastest of the existing subsequence matching algorithms.

All the aforementioned algorithms basically focus on reducing the calling times of DTW calculation, not accelerating DTW itself. However, because of its complexity, DTW still takes a large part of the total application runtime. Due to this reason science community tries to exploit the effectiveness of parallel hardware to accelerate subsequence matching. Existing researches assume allocation of DTW calculation of different subsequences into different processing elements, e.g. computer-cluster [9], multicore [10], FPGA and GPU [6].

Despite the existing research devoted to use of the Intel Xeon Phi coprocessor in data mining [24] in our opinion not enough attention is paid to applying this coprocessor for time series subsequence matching. In this work we adapt the algorithm [8] to the Intel Xeon Phi coprocessor (what was done for the first time, to the best of our knowledge).

VI. Conclusion

In this paper we have described the parallel algorithm for time series subsequence matching adapted for CPU accompanied with the Intel Xeon Phi coprocessor. The algorithm uses data parallelism and OpenMP technology. To establish interaction between CPU and the coprocessor the *offload* execution mode is used, when the program is run by the CPU while some code and data are transmitted to the coprocessor. To provide high operational intensity of the calculations on the coprocessor the algorithm supports a queue of subsequences on CPU, where subsequences are candidates to be offloaded to the coprocessor to calculate DTW. Experiments on synthetic and real data sets have shown that our algorithm outperforms analogous algorithms for GPU and FPGA.

As future work we plan to extend our algorithm for the case of a cluster system based on nodes equipped with the Intel Xeon Phi coprocessors.

References

[1] V. Epishev, A. Isaev, R. Miniakhmetov, A. Movchan, A. Smirnov, L. Sokolinsky, M. Zymbler, and V. Ehrlich, "Physiological data mining system for elite sports," *Bull. of South Ural State University. Series: Comput. Math. and Soft. Eng.*, vol. 2, no. 1, pp. 44–54, 2013.

[2] M. Dyshaev and I. Sokolinskaya, "Representation of trading signals based on kaufman adaptive moving average as a system of linear inequalities," *Bull. of South Ural State University. Series: Comput. Math. and Soft. Eng.*, vol. 2, no. 4, pp. 103–108, 2013.

[3] S. Abdullaev, O. Lenskaya, A. Gayazova, D. Sobolev, A. Noskov, O. Ivanova, and G. Radchenko, "Short-range forecasting algorithms using radar data: Translation estimate and life-cycle composite display," *Bull. of South Ural State University. Series: Comput. Math. and Soft. Eng.*, vol. 3, no. 1, pp. 17–32, 2014.

[4] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *KDD Workshop*, U. M. Fayyad and R. Uthurusamy, Eds. AAAI Press, 1994, pp. 359–370.

[5] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. J. Keogh, "Querying and mining of time series data: experimental comparison of representations and distance measures," *PVLDB*, vol. 1, no. 2, pp. 1542–1552, 2008.

[6] D. Sart, A. Mueen, W. A. Najjar, E. J. Keogh, and V. Niennattrakul, "Accelerating dynamic time warping subsequence search with gpus and fpgas," in *ICDM*, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, Eds. IEEE Computer Society, 2010, pp. 1001–1006.

[7] S.-H. Lim, H. Park, and S.-W. Kim, "Using multiple indexes for efficient subsequence matching in time-series databases," in *DASFAA*, ser. Lecture Notes in Computer Science, M.-L. Lee, K.-L. Tan, and V. Wuwongse, Eds., vol. 3882. Springer, 2006, pp. 65–79.

[8] T. Rakthanmanon, B. J. L. Campana, A. Mueen, G. E. A. P. A. Batista, M. B. Westover, Q. Zhu, J. Zakaria, and E. J. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *KDD*, Q. Yang, D. Agarwal, and J. Pei, Eds. ACM, 2012, pp. 262–270.

[9] N. Takahashi, T. Yoshihisa, Y. Sakurai, and M. Kanazawa, "A parallelized data stream processing system using dynamic time warping distance," in *CISIS*, L. Barolli, F. Xhafa, and H.-H. Hsu, Eds. IEEE Computer Society, 2009, pp. 1100–1105.

[10] S. Sharanyan, K. Arvind, and G. Rajeev, "Implementing the dynamic time warping algorithm in multithreaded environments for real time and unsupervised pattern discovery," in *ICCCT*, D. of Computer Science and M. N. N. I. o. T. Engineering, Eds. IEEE Computer Society, 2011, pp. 394–398.

[11] A. Duran and M. Klemm, "The intel® many integrated core architecture," in *HPCS*, W. W. Smari and V. Zeljkovic, Eds. IEEE, 2012, pp. 365–366.

[12] "tornado susu" supercomputer. South Ural State University. [Online]. Available: www.supercomputer.susu.ru/en/computers/tornado/

[13] Nvidia tesla gpu accelerators. NVIDIA Corp. [Online]. Available: www.nvidia.com/content/tesla/pdf/NVIDIA-Tesla-Kepler-Family-Datasheet.pdf

[14] Virtex-7 fpgas family. Xilinx Inc. [Online]. Available: www.xilinx.com/publications/prod_mktg/Virtex7-Product-Brief.pdf

[15] T.-C. Fu, "A review on time series data mining," *Eng. Appl. of AI*, vol. 24, no. 1, pp. 164–181, 2011.

[16] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," in *SIGMOD Conference*, R. T. Snodgrass and M. Winslett, Eds. ACM Press, 1994, pp. 419–429.

[17] Y.-S. Moon, K.-Y. Whang, and W.-K. Loh, "Duality-based subsequence matching in time-series databases," in *ICDE*, D. Georgakopoulos and A. Buchmann, Eds. IEEE Computer Society, 2001, pp. 263–272.

[18] Y.-S. Moon, K.-Y. Whang, and W.-S. Han, "General match: a subsequence matching method in time-series databases based on generalized windows," in *SIGMOD Conference*, M. J. Franklin, B. Moon, and A. Ailamaki, Eds. ACM, 2002, pp. 382–393.

[19] W.-K. Loh, Y.-S. Moon, and J. Srivastava, "Distortion-free predictive streaming time-series matching," *Inf. Sci.*, vol. 180, no. 8, pp. 1458–1476, 2010.

[20] S.-W. Kim, J. Yoon, S. Park, and T.-H. Kim, "Shape-based retrieval of similar subsequences in time-series databases," in *SAC*. ACM, 2002, pp. 438–445.

[21] S. Park, S.-W. Kim, and W. W. Chu, "Segment-based approach for subsequence searches in sequence databases," in *SAC*. ACM, 2001, pp. 248–252.

[22] S. Park, S.-W. Kim, J.-S. Cho, and S. Padmanabhan, "Prefix-querying: An approach for effective subsequence matching under time warping in sequence databases," in *CIKM*. ACM, 2001, pp. 255–262.

[23] M.-S. Kim, S.-W. Kim, and M. Shin, "Optimization of subsequence matching under time warping in time-series databases," in *SAC*, H. Haddad, L. M. Liebrock, A. Omicini, and R. L. Wainwright, Eds. ACM, 2005, pp. 581–586.

[24] A. Heinecke, M. Klemm, D. Pflüger, A. Bode, and H.-J. Bungartz, "Extending a highly parallel data mining algorithm to the intel ® many integrated core architecture," in *Euro-Par Workshops (2)*, ser. Lecture Notes in Computer Science, M. Alexander, P. D'Ambra, A. Belloum, G. Bosilca, M. Cannataro, M. Danelutto, B. D. Martino, M. Gerndt, E. Jeannot, R. Namyst, J. Roman, S. L. Scott, J. L. Träff, G. Vallée, and J. Weidendorfer, Eds., vol. 7156. Springer, 2011, pp. 375–384.

Outlier Detection as the Primary Step for Promotion Planning in Retail

M. Banek*, D. Osrečki*, M. Vranić* and D. Pintar*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{ marko.banek, dinko.osrecki, mihaela.vranic, damir.pintar}@fer.hr

Abstract - Forecasting the increase of customer demand during discount promotions is a fundamental business assignment in retail, which can nowadays be performed by sophisticated data mining algorithms. The calculations are based on data gained during previous promotions. Apart from choosing the right mining algorithm, the quality of prediction models strongly depends on the quality of the training data. Outliers are points of data that do not conform to a defined notion of normal behavior and are therefore either excluded from the training set, or their impact on the model is weighted in a manner different from other data. In this paper we propose a new approach to outlier analysis, with the aim of distinguishing between the outliers associated with an outlier-generating store or product and the outliers that can be classified as noisy data. Outlier analysis is performed with a multidimensional perception of the dataset typical for data warehousing and OLAP. We also introduce measures that estimate the probability of store or product being an outlier generator and conduct experiments to determine their critical threshold values.

I. INTRODUCTION

Business information systems serve as the source of new, added-value information either by means of analytical processing (OLAP) or data mining. The work of data mining scientists is mostly aimed at constructing predictive models (by using regression, neural network, naive Bayes classifier, k-nearest neighbor classifier, support vector machines, random forests, etc., see [1]) or descriptive models (by using clustering, association rules, hidden Markov models, principal component analysis, self-organizing maps, etc., see [1]). However, modeling is just one step of the process: according to the well-known CRISP-DM process model [2], modeling is preceded by the steps of Business Understanding, Data Understanding and Data Preparation and then followed by Evaluation and Deployment. Data understanding and data preparation are frequently observed as a single process called Data Analysis or Exploratory Data Analysis [3]. The analyst first gets familiar with the data, studying attributes and their distributions as well as identifying possible quality problems, and then chooses the optimal subset for model construction, performing data transformations if necessary.

A principal task related to data quality is outlier detection. According to Anscombe and Guttman [4], "*an outlier is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed"*. Hawkins [5] defines an outlier in a similar fashion, as "*an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*". Based on the domain of application and motives for outlier detection, two general perspectives of outliers can be taken [6,7]. The first perspective is taken when the primary goal is to remove dirty data. Outliers are understood as noise that "spoils" the data and hinders further analysis. Once the outliers are detected, the easiest way is to exclude them from the dataset. However, simple exclusion may sometimes not be the best solution since the resulting dataset may become unrepresentative: in such case it is better to calculate some interpolation value for the outlier records [8,9,10]. From the second perspective outliers are understood as a useful indicator of the different ("abnormal") mechanism that creates them [6]. Thus, outliers themselves are not the purpose of the detection process, but only serve as means to detect their generator mechanism. Outliers in credit card transactions or insurance claims are strongly related to fraud attempts [6], while outliers in meteorological data may suggest a dangerous event such as storm or hurricane [6].

A part of our team has been involved with constructing predictive models of promotion sales demand for a major retail corporation in Croatia [11]. The models have been based on linear regression and support vector machines. During the exploratory analysis process it became clear that outliers occur as a result of several mechanisms and each outlier group should be processed in a different manner (i.e. not all outliers should be discarded). For instance, very low sale result for several products in a particular store during a particular promotion campaign in January, which differs from the mutually highly correlated results for the same products during other promotions and which is accompanied with the out-of-stock flag is a result of shipping problems due to a snow blizzard. A much-higher-than-usual sale result for beer in a countryside store in June is a result of a camping rally. Such cases do not express regular behavior pattern and should be discarded. On the other hand, some stores express outlying behavior (in comparison with other stores i.e. according to statistical outlier detection parameters) for many products and many promotions. Likewise, some products are outliers across the majority of stores. Stores and products can be understood as some kind of outlier-generating mechanism, obviously different from unexpected events

such as a camping rally or uncontrollable ones, such as blizzards. Removing particular stores and products from the training dataset would result in a general model that could likely be inapplicable for those particular stores and products: we would need to construct additional models (or possibly many models) or apply some heuristics. Keeping the data in the dataset and using some additional engineering (e.g. introducing of some additional parameters) may be a better solution.

In this paper we propose a new approach to outlier analysis, with the aim of distinguishing between the outliers associated with an outlier-generator store or product and the outliers that can be classified as noisy data. We adopt the multidimensional perception of the dataset typical of data warehousing and OLAP [12]. In the first step we detect outliers using the multivariate statistical technique MCD [13]. In the second step we analyze the entire dataset, understanding each record as a single OLAP cell. We introduce outlier ratio and total number of outliers as OLAP-like measures of the fact and aggregate their values across multiple dimensions.

The paper is structured as follows. Basic terms and formulae considering outliers are given in Section II. Related work is elaborated in Section III. Techniques for multivariate outlier detection are explained and applied in Section IV. Multidimensional analysis of the outlier data is performed in Section V. Conclusions are drawn in Section VI.

II. BASIC TERMS AND FORMULAE

Let the data be characterized by a set of one (*a univariate set*) or more attributes (*a multivariate set*). Each data record may be represented by a point in the *N-dimensional space* (*N* being the number of attributes). Outliers are points that are according to some distance measure "far away" from other points (Fig. 1).

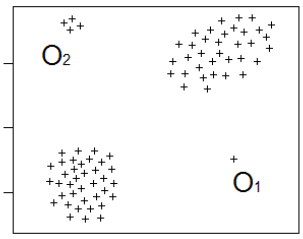


Figure 1. Outliers in two-dimensional space. A single point O_1 and a cluster of several points O_2. Adapted from [7].

In a very large number of cases attributes are continuous variables with Gaussian distribution. Considering univariate instances as the most simple case, the most common notion of outliers in statistics was long perceived in relation with the mean μ and the standard deviation σ of the population (or sample) and the confidence coefficient α (most commonly for α=0.05, 0.01, 0.005 etc.) [14]:

$$\text{out}(\alpha, \mu, \sigma^2)=\{x : |x - \mu| > z_{1-\alpha/2}\sigma\} \quad (1)$$

However, outliers themselves strongly influence the values of μ and σ. Hampel [15] suggested to use median and median absolute deviation (MAD) instead. A similar concept was used by Tukey [3], who introduced Boxplot as a convenient graphical display for outliers. Tukey's robust estimator for σ is based on the difference between the third and the first quartile also known as the interquartile range, *IQR*:

$$\hat{\sigma} = IQR = Q_3 - Q_1 \quad (2)$$

Outliers are observations smaller than $Q_1-1.5\cdot$IQR and larger than $Q_3+1.5\cdot$IQR. Fig. 2 shows the boxplot for an example dataset consisting of 12 members. The box stretches from Q_1 to Q_3. The thick line is the median. "Whiskers" denote the smallest and largest non-outlier values (50, 70). Outliers (38, 85) are represented by small circles.

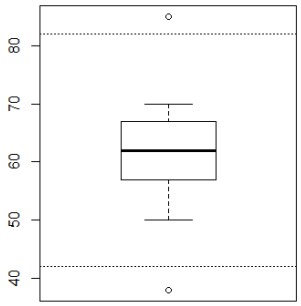


Figure 2. Boxplot for dataset {38, 50, 57, 57, 58, 61, 64, 64, 67, 67, 70, 85}. Median=62.5; Q_1=57; Q_3=67; IQR = 10; ($Q_1-1.5\cdot$IQR) = 42; ($Q_3+1.5\cdot$IQR) = 82.

III. RELATED WORK

Outlier detection became a popular topic in computer science in 1990s, with the emergence of many domain-specific classification techniques based on neural networks [16], support vector machines [17] and Bayesian models [16]. Rule-based classification techniques (i.e. classification based on pre-defined heuristic rules instead of machine learning) have also been widely adopted [18]. The main difference between "standard" classification and outlier detection is the fact that instances labeled as outliers occur very rarely in the dataset, which makes the training process particularly challenging. Certain techniques successfully solve the problem by injecting additional artificial outliers into the original dataset in order to increase their ratio [19]. Neighbor-based approaches combine the idea of calculating distance from other object in the space (which is specific for clustering), with the idea of doing the calculation locally, i.e. only for *k* nearest objects [19,20,21].

Principal component analysis (PCA) uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Since the top few principal components capture the largest portion of variability in a given dataset, the smallest principal components are expected to result in constant values. A data point that violates this expectation for the smallest components is an outlier [22].

Up to the present date outlier detection has remained in the permanent focus of scientific research in manifold and mutually diverse application domains. Price estimation in real estate market is addressed in [23]. The well-known particular case of credit card frauds can be extended to any kind of relationship between a business

provider and a consumer [24]. New visualization techniques used for clustering and outlier detection in high-dimensional spaces are proposed in [25]. Further possibilities of applying non-Gaussian models in outlier detection, with application in enterprise resource planning (in particular, executive benefits) are investigated in [26].

Particular examples of outlier detection in retail or retail promotions are [27] and [28] from the long-term econometric perspective of a marketing department, [9] from the perspective of statisticians collecting data for a government, and [29] from the perspective of an agency affiliated with the US Department of Healthcare.

IV. MULTIVARIATE OUTLIER DETECTION

A. Dataset description

The promotion planning dataset upon which the retail demand prediction models are built [11] consists of more than 50 attributes. We may adopt a multidimensional perspective of the data identifying three basic dimensions: **product**, **store** and **promotion** (Fig. 3; this is not a model of a data warehouse, since it lacks the *sine-qua-non* time dimension). Each dimension contains one or more hierarchies (for instance, there is a four-level hierarchy of products imposed by the internal organization of the retail corporation: single_product → group → subcategory → category).

A single data record i.e. a single OLAP-cell contains numerical data on the promotional performance of a single product in a single store during a single promotion campaign. There are seven numerical measures (Fig. 3).

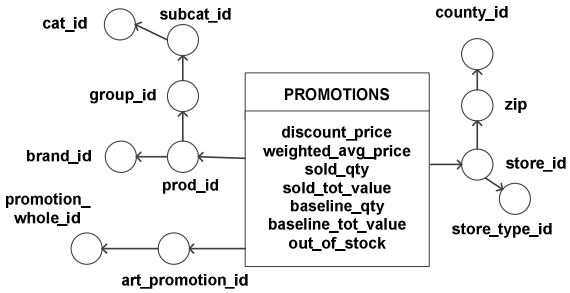


Figure 3. Dataset attributes observed from the multidimensional perspective.

Weighted average price is calculated across all stores during a longer time period. *Sold total value* is the product of *sold quantity* and *discount price*. *Baseline quantity* is the interpolated value of the sold quantity expected in case there was no promotion. *Baseline total value* is the product of baseline quantity and average price. The values of baseline quantity and baseline total value are expected to be smaller than the actual sold quantity and sold total value. *Out-of-stock* is an integer defining how many days during the promotion the product was unavailable (in case of good organization and planning it should be zero).

The purpose of the demand forecasting model described in [11] is to predict the quantity sold during promotional campaigns. Thus, our observation of outliers will be performed with respect to the values of two attributes: *sold_qty* and *baseline_qty*. We did not observe prices, since the resulting space would be very strongly influenced by the product dimension (in comparison with stores and promotion campaigns).

Observing the histograms in Fig. 4, (all the processing is performed in the R programing language), we see that the distribution of both variables is exponential, while the standard definitions of outliers (Eq. 1, 2) refer to the normal distribution. Quantile-to-quantile plots comparing the distributions of the two variables with the normal distribution display large discrepancies (points of QQ-plots for two corresponding distributions do not fit the red line; see Fig. 4). Intuitively, taking the logarithm of the variables would produce results that conform to the normal distribution. We use the Box-Cox transformation [30] which works for many distributions and turns the original variable x into normalized value:

$$x \to \begin{cases} \dfrac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases} \qquad (3)$$

R calculates the optimal value of the λ parameter by an iterative computation for values within the interval [−2, 2]. In our case the optimal values of λ are −0.02 for *sold_qty*, and −0.06 for *baseline_qty*. Histograms and QQ-plots for the transformed variables are shown in Fig. 5 (the majority of values now conform to the normal distribution). All further work is performed with the transformed values.

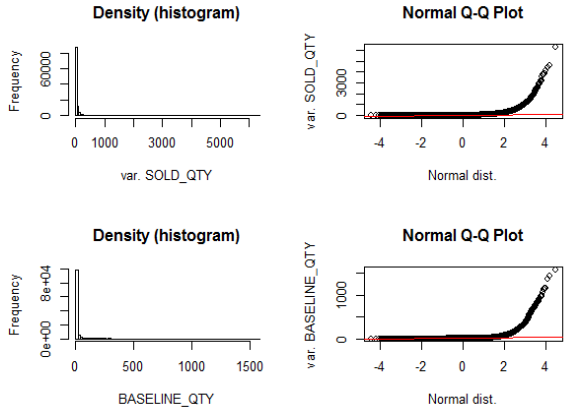


Figure 4. Histograms showing distributins of variables *sold_qty* and *baseline_qty* and quantile-to-quantile plots comparing their distribution with the normal distribution.

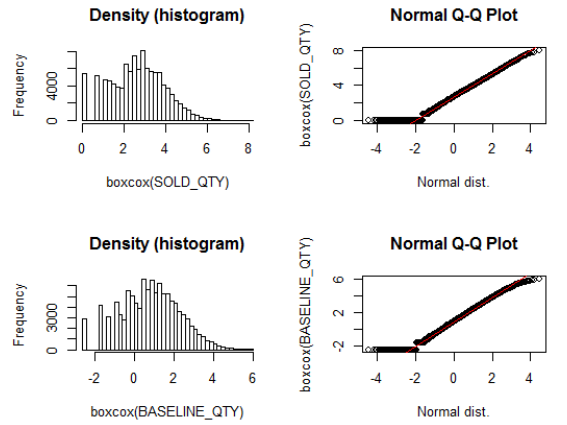


Figure 5. Histograms and quantile-to-quantile plots for variables *sold_qty* and *baseline_qty* transformed by the Box-Cox tranformation.

B. *Mahalanobis distance and MCD*

In many cases multivariable observations cannot be considered outliers when each variable is observed independently. For instance, a height of 187 cm for a grown-up male person makes him taller than average but not an outlier. Likewise, a male person with (European) shoe size 41 is not an outlier, although this is smaller than the average male size. However, a person 187 cm tall with shoe size 41 would by most people be observed as an outlier.

Given a multivariate (N-dimensional) space where all variables conform to the normal distribution, it is easy to calculate the centroid, but each of the variables has a different value of σ. Intuitively, the space is not spherical but shaped as an ellipsoid, and each dimension has to be normalized by $(x - \mu)/\sigma$ in order to achieve a spherical space using the *Mahalanobis distance*. Let $x^i = (x_1^i, x_2^i, \ldots, x_N^i)^T$ be a vector representing the *i*-th *N*-dimensional observation (let the total number of observations be *M*). Let the centroid $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)^T$ be the vector of univariate mean values for each dimension. Mahalanobis distance is defined as [13]:

$$D_{MAH}(x,\mu) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \tag{4}$$

where $\boldsymbol{S}^{-1}$ is the inverse of the covariance matrix $\boldsymbol{S}$ for observations $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}$, which is defined as [13]:

$$S = \frac{1}{M-1}\sum_{k=1}^{M}(x^k - \mu)(x^k - \mu)^T \tag{5}$$

The covariance matrix determines the ellipse which is at $D_{MAH} = 1$ from the centroid (the axes correspond to the square roots of its eigenvalues; its eigenvectors define the rotation angle). Since the distribution of the Mahalanobis distance behaves as a chi-square distribution for a large number of instances, the cutoff threshold to distinguish outliers from other data is $[D_{MAH}(\boldsymbol{x}, \boldsymbol{\mu})]^2 = \chi^2_{N,1-\alpha}$ (the number of degrees of freedom is equal to the number of dimensions in the space, *N*). The preferred value of $1-\alpha$ is 0.975 [13] which for $N=2$ gives the threshold value of $\sqrt{7.3777} = 2.716$.

Similar to the non-robust basic formula (Eq. 1) for univariate outliers that had to be replaced with a more robust estimation (Eq. 2), a group of outlying instances may skew the mean and covariance estimates towards itself. A robust estimator of the centroid and the covariance matrix is the minimum covariance determinant (MCD) [13], which instead of the original *M* observations uses its most dense subset of size $\lceil 0.5 \cdot (M + N + 1) \rceil$, which is smaller than *M* since $N < M$.

Our dataset contains 109284 records (35 promotion campaigns, 300 products, 744 stores). The scatterplot is shown in Fig. 6. Only 414 records are univariate outliers, above the horizontal line and right to the vertical line, which both denote the Q_3+1.5·IQR value. In both cases there are no outliers for Q_1−1.5·IQR. Basic multivariate method detected additional 2219 outliers (i.e. five times more than the number of univariate outliers). The threshold is denoted by the larger ellipse in Fig. 6. In comparison, the robust MCD method detected a much

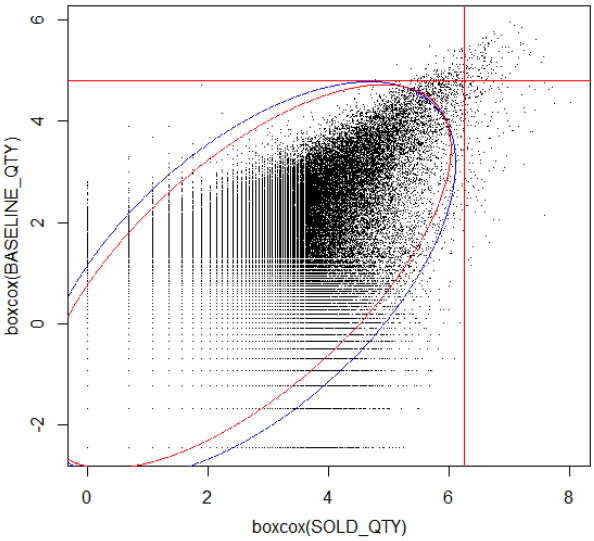


Figure 6. The scatterplot for the transformed variables *sold_qty* and *baseline_qty*. Lines denote univariate outlier thresholds. Ellipses denote multivariate threshold for basic Mahalanobis distance: the larger one for the basic calculation method, the smaller one for the MCD estimate.

larger total number of 4445 outlying points (the smaller ellipse, with a 1.15 times smaller area in comparison with the larger ellipse). When natural logarithm was used to normalize the values instead of Box-Cox transformation (Table I) the number of points identified as outliers is smaller (the normalization is of lesser quality, hiding outliers that skew the distribution parameters). It is important to note that the set of algorithms held in the literature as optimal (Box-Cox transformation and MCD) produced the largest number of outliers.

TABLE I. THE NUMBER OF OUTLYING POINTS PRODUCED BY DIFFERENT PROCESSING ALGORITHMS

	Outlier detection method		
TRANS-FORMATION TYPE	***UNIVARIATE***	***MULTIVARIATE***	
		BASIC	**MCD**
LOG	386	1996	2219
BOX-COX	414	3224	4445

V. OUTLIER ANALYSIS FROM MULTIDIMENSIONAL PERSPECTIVE

Outlier analysis from the multidimensional perspective is performed with the three dimensions already described in Section IV.A: product, store and promotion (Fig. 3). Thus, the minimum granularity corresponds to a single dataset record. Instead of the measures proper to the dataset, cells now contain two measures:

- the number of outliers, ***NO*** (in the particular cell),
- the total number of points, ***NP*** (in the cell).

Both measures are additive across all dimensions. We aggregate *NO* and *NP* and group them by each of the three dimensions: store, product and promotion, as well as by two of the three dimensions jointly (i.e. by store and product, store and promotion, product and promotion). For

each group we calculate the principal measure of *outlyingness*: **outlier ratio** (**OR**) i.e. the ratio of the number of outliers and the total number of points:

$$OR = \frac{NO}{NP} \tag{6}$$

This measure is by its definition non-additive and has to be calculated separately for each aggregated multi-cell. The outlier ratio for the entire dataset is OR_{DS} = (414+4445)/109284 = 0.0445.

Total number of outlying points is not a good measure of *outlyingness*: a large number of outliers for a particular group may correspond to an even larger total number of points i.e. a low outlier ratio. Indeed, many promotions refer to a limited number of stores, which makes *NP* for those stores larger in comparison with other stores. Table II shows that products with IDs 37 and 40, respectively, are associated with almost the same number of outliers (19 and 18, respectively). However, while the *NP* of the former is 27 points in total (making the ratio of outliers very high), the *NP* of the latter is 975, making the outlier ratio smaller than 0.02 and more twice smaller than OR_{DS}. Thus, only the former product is indeed an outlier-generator, while in the latter case the product is considered regular (the related outliers may arise due to an outlier-generator store or promotion or may even more likely be considered noise).

TABLE II. OUTLIER MEASURES AGGREGATED ACROSS THE PRODUCT DIMENSION

PROD_ID	NP	NO	OR
335	1	1	1
243	166	160	0.9639
37	27	19	0.7037
…	…	…	…
40	975	18	0.0185
…	…	…	…
72	607	1	0.0016
13	56	0	0
…	…	…	…

High outlier ratio alone can be deceiving when the total number of points is small. The highest ratio in Table II is associated with a single record for product ID 335, which happens to be an outlier. Based on the total number of detected outliers, we declare a minimum threshold of 10 outlying points (about 0.2% of the total of 4859 outliers) to consider a store or a product an outlier-generator candidate (due to a small number of promotions in comparison with stores and products, all promotions cross the threshold by far). Considering the products in Table II, only those with IDs 243 and 37 (shadowed rows in Table II) are outlier generators.

For the aggregations by two dimensions it is very useful to compare their outlier ratio with the ratios for the corresponding aggregations by a single dimension. We divide the former value with the latter two values obtaining two coefficients that illustrate whether the particular combination produces a higher *OR* than each of its parts observed separately. Table III shows several records for the joint aggregation by product and promotion. When both coefficient values (the last two columns in Table III) are significantly larger than 1 (e.g. 1.5 at least), the combination itself is a strong outlier-generator (e.g. promotion 27 and product 91). When one coefficient value is slightly larger than 1 and the other is much larger (e.g. promotion 25 and product 78), or when one coefficient is smaller than 1 (e.g. promotion 34 and product 184), the outlier-generator candidate is in the dimension related to the coefficient close to 1 or smaller than 1 (i.e. promotion 25, product 184), whereas the other one is not a candidate (product 78 and promotion 34 generate outliers only in dependence with the other dimension).

Summary characteristics of the aggregated data are shown in Table IV. We measure the size of each group, the number of groups with at least one outlier, the number of groups with at least ten outliers, the third quartile value for *OR* in each aggregation, the number of groups with *OR* higher than 0.2 and 0.3 respectively, and the maximum *OR* within the entire aggregation.

Given the minimum threshold of 10 outlying points in total (0.2% of the total number of outliers in the dataset) we decide that *OR* value of 0.2 (which is more than 4 times greater than OR_{DS}) is a good threshold to distinguish outlier-generating products and stores from the outliers that occur as result of unpredictable or uncontrollable events.

Considering higher aggregation levels, we grouped data by all cities (level *zip* in Fig. 3) with two or more stores. The highest *OR* was 0.0905, two times smaller than the critical threshold value of 0.2. The highest two *OR* values for product groups (level *group_id*) with two or more products were 0.1655 and 0.1377, respectively. The groups consisted of 2 and 14 products, respectively. The first *OR* value is smaller than the suggested threshold

TABLE III. OUTLIER MEASURES AGGREGATED ACROSS TWO DIMENSIONS: PROMOTION (PRM) AND PRODUCT (PRD)

PRM ID	PRD ID	OR	OR_{PRM}	OR_{PRD}	OR/OR_{PRM}	OR/OR_{PRD}
27	91	0.570	0.125	0.184	4.544	3.094
36	301	0.128	0.103	0.080	1.245	1.592
25	78	0.127	0.099	0.034	1.276	3.747
34	184	0.058	0.026	0.070	2.237	0.826

TABLE IV. SUMMARY PARAMETERS FOR ALL AGGREGATIONS

	AGGREGATION BY DIMENSIONS					
	PRD	STR	PRM	PRD+ STR	PRD + PRM	STR + PRM
GRP SIZE	300	744	35	89501	20412	365
# with outlr	262	671	35	4655	3590	311
# > thr10	104	132	31	0	7	114
Q_3	0.131	0.078	0.083	0	0.169	0.128
No. >0.2	17	2	1	0	1	19
No. >0.3	12	0	1	0	1	12
max OR	0.964	0.235	0.475	0	0.202	0.964

value of 0.2, although quite close to it. However, looking backward at the product level we discovered that all outliers were generated by a single product. Hence, in order to classify a group at a higher aggregation level as an outlier generator, there should be an additional threshold: at least *K* of *N* group members at the basic level should be outlier generators themselves. Focusing on this threshold is a challenging task for future work.

VI. CONCLUSION

In this paper we proposed a new approach to outlier analysis based on the multidimensional conceptual model and OLAP. We analyzed the datasets used for training promotional sale forecasting models, with the primary goal of distinguishing between outliers occurring for stores and products that behave as outlier-generator mechanisms and outliers occurring as a result of unpredictable events.

We constructed two-dimensional space and detected outliers using the MCD estimator. We counted the points (*NP*) and the outliers (*NO*) across three dimensions: product, store and promotion. Then we aggregated the data and grouped them across dimensions, associating each group with its outlier ratio (*OR*), a non-additive measure we proposed in this paper as the one that best describes outlier behavior of an aggregated group of points: a store, a product or a promotion. The analysis led to a conclusion that, regardless of its *OR* value, a group cannot be considered an outlier generator mechanism unless its absolute number of outliers (*NO*) exceeds a threshold.

Through the conducted experiments we established the threshold values to consider a group an outlier generator: the *NO* threshold is 0.2% of the total number of outliers in the dataset, whereas the *OR* threshold is 4 times greater than the outlier ratio of the entire dataset.

Our future work will be directed at studying higher aggregation levels as possible outlier generators. We also plan to analyze the outliers in the same or similar dataset from the perspective of income (i.e. sold total value measure instead of sold quantity).

REFERENCES

[1] M. Zaki and W. Meira, "Data mining and analysis. Fundamental concepts and algorithms." New York: Cambridge University Press, 2014

[2] C. Shearer, "The CRISP-DM model: the new blueprint for data mining", J Data Warehousing, vol. 5, num. 4, pp. 13-22, 2000.

[3] J. W. Tukey. "Exploratory data analysis", Addison-Wesley, 1977

[4] F. J. Anscombe and I. Guttman, "Rejection of outliers", *Technometrics*, vol. 2, no. 2, pp. 123-147, 1960.

[5] D. Hawkins, "Identification of outliers", London-New York: Chapman and Hall, 1980.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey", *ACM Computing Surveys*, vol. 41, no. 3, art. 14, 2009

[7] C. C. Aggarwal, "Outlier analysis", *Kluwer Academic Publishers*, 2013

[8] P. Rousseeuw and A. Leroy, "Robust regression and outlier detection", New York: Wiley, 2003.

[9] M. H. Mulry and R. M. Feldpausch, "Treating influential values in a monthly retail trade survey", *Proc. of the Survey Methods Section, Statistical Society of Canada Annual Meeting*, June 2007.

[10] S. Johansen and B. Nielsen, "Outlier detection in regression using an iterated one-step approximation to the Huber-Skip estimator", *Econometrics*, vol. 1, pp. 53-70, 2013

[11] M. Karan, D. Pintar, Z. Skočir, M. Vranić, A. Alajković, J. Milojević and M. Pleša, "The impact of training data tailoring on demand forecasting models in retail", *Proc. miproBIS 2014*, *in conjunction with 37th Int. Convention MIPRO*, Opatija, Croatia, May 2014, Rijeka : GRAFIK , pp. 1636-1641, 2014

[12] M Golfarelli, D Maio and S Rizzi, "The dimensional fact model: a conceptual model for data warehouses", Int. J. of Cooperative Inf. Syst., vol 7, no. 02&03, pp. 215-247, 1998

[13] P. Rousseeuw and K. van Driessen, "A fast algorithm for the minimum covariance determinant estimator", Tehnometrics, vol. 41, no 3., pp. 212-223, 1998

[14] P. L. Davies and U. Gather, "The Identification of multiple outliers", J. of the American Statistical Association, vol. 88, pp. 782-792, 1993

[15] F. R. Hampel, "The influence curve and its role in robust estimation", J. of Amer. Stat. Assoc., vol. 69, pp. 382–393, 1974.

[16] S. Singh and M. Markou, "An approach to novelty detection applied to the classification of image regions. IEEE Trans. on Knowledge and Data Engineering, vol 16, no. 4, pp. 396-407, 2004.

[17] W. Hu, Y. Liao and V. R. Vemuri, "Robust anomaly detection using support vector machines", *Proc. Int. Conf. on Machine Learning*. San Francisco : Morgan Kaufmann Publishers, pp. 282-289, 2003

[18] N. Abe, B. Zadrozny and J. Langford, "Outlier detection by active learning", Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, pp. 504-509, 2006

[19] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "Lof: identifying density-based local outliers", Proc. 2000 ACM SIGMOD Int. Conf. on Management of Data. New York: ACM Press, pp. 93-104, 2000

[20] D. Pokrajac, A. Lazarević, and L. J. Latecki, "Incremental local outlier detection for data streams". Proc. of IEEE Symposium on Computational Intelligence and Data Mining, 2007

[21] Y. Chen, D. Miao and H. Zhang, "Neighborhood outlier detection", Expert Systems with Applications, vol. 37, no. 12, pp. 8745-8749, 2010

[22] A. Lakhina, M. Crovella and C. Diot, "Mining anomalies using traffic feature distributions", Proc. 2005 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York: ACM Press, pp. 217-228, 2005

[23] P. Morano and F. Tajani, "Least median of squares regression and minimum volume ellipsoid estimator for outliers detection in housing appraisal", *Int. J. of Business Intelligence and Data Mining*, vol. 9, no. 2, pp. 91-111, 2014

[24] Q. Jiang, A. Campbell, G. Tang and J. Pei, "Multi-level relationship outlier detection", *Int. J. of Business Intelligence and Data Mining*, vol. 7, no. 4, pp. 253-273, 2012

[25] L. Boudjeloud-Assala, "Visual interactive evolutionary algorithm for high dimensional outlier detection and data clustering problems", *Int. J. of Bio-Inspired Computation*, vol. 4, no. 1, pp. 6-13, 2012

[26] P. C. Nystrom and E. S. Soofi, "Rare, outlier and extreme: beyond the Gaussian model and measures", *Int. J. of Complexity in Leadership and Management*, vol.2, no.1/2, pp. 6-38, 2012

[27] W.-S. Chan. "Outlier analysis of annual retail price inflation", *Journal of Actuarial Practice*, vol. 6, pp. 149-172, 1998

[28] P. H. Franses, T. Kloek and A. Lucas, "Outlier robust analysis of long-run marketing effects for weekly scanning data", *J. of Econometrics*, vol. 89, pp. 293-315, 1999

[29] M. Zodet, S. C. Hill and S. Zuvekas, "Evaluating an alternative data source for editing MEPS drug prices", Agency for Healthcare Research and Quality Working Paper No. 12004, July 2012, URL: http://meps.ahrq.gov/data_files/publications/workingpapers/wp_12004.pdf

[30] G. E. P. Box and D. R. Cox, "An Analysis of Transformations", J. of the Royal Statistical Society. Series B (Methodological), nol. 26, no. 2, pp. 211-252, 1964

Data Warehouse in era of Big Data platforms

Tomislav Križan, Mladen Družetić
Poslovna Inteligencija d.o.o., Zagreb, Croatia
{tomislav.krizan, mladen.druzetic}@inteligencija.com

Abstract - One of emerging questions in analytical space is difference between implementations of traditional Data Warehouse systems and those based on Big Data technologies. Also, what about Data Warehouse appliances in era of fast evolving Big Data systems? Within this paper overview will be given of differences and/or similarities between traditional and Big Data option in analytic space, with focus on relevant cases for each of them.

I. INTRODUCTION

For almost over 40 years, data management solutions typically meant focusing on relational data processing and use of relational database management systems (RDBMS's). Within last decade, industry moved to a data-driven world with explosion in the data volumes generated not just with human interactions, but even more of those originating from machines communicating with other machines (M2M). With those data volumes came need to analyse underlying data sets to find information for statistical and/or legal reporting, but even more to gain insight into previous behaviour pattern within data set to predict future outcome(s) of *our* actions.

In computing, a data warehouse (DW or DWH) is a system used for reporting and data analysis. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons [1]. DW guru's try to define meaning of DW as „A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process“ [2] or „A data warehouse is a copy of transaction data specifically structured for query and analysis.“ [3]. In essence, the DW concept was intended to provide an architectural model for the flow of data from operational systems to decision support environments.

There were many problems related to analysis of large data volumes, but the term “Big Data” which spans computer science and statistics/econometrics, probably originated in lunch-table conversations at Silicon Graphics Inc. [4]. Term meant something different in 90's than in current days and moreover, we are now talking completely different scales of used data volumes which were unthinkable in 90's. Trying to solve problems related to Big Data, Apache Hadoop came to life and became most common platform for distributed storage and distributed processing of very large data sets (e.g. Big Data). The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce, Yarn, Spark) [5]. Hadoop was created in 2005 to support distribution for the Nutch search engine project [6].

II. DATA WAREHOUSE IN PRACTICE

The enterprise data warehouse (EDW) is the backbone of analytics and business intelligence for most large organizations and many midsize firms. The tools and techniques are proven, the SQL query language is well known, and there's plenty of expertise available on the market. Over the time, many applications and concepts arose around DW with different maturity levels and use-cases.

Currently, majority of DW implementations are based on use of RDBMS's from existing mega-vendors (e.g. Oracle, IBM, Microsoft, SAP, etc.), but there is also highly rising trend of replacing general-purpose built RDBMS with appliances designed exclusively for analytical purpose (same mega-vendors compete even in this space as well).

At same time, until year 2014. use of Big Data platforms as replacements for DW was not even considered due architectural and/or conceptual differences behind major distribution of those solutions (Cloudera, Hortonworks, and MapR to name some of them). Necessary components within common Hadoop distributions, gained high level of maturity and some major obstacles to use those platforms as DW, were solved within 2014. thus allowing them to step into *ad-hoc* analytical spectrum as first-line-of-defence.

In following segments of this paper we will briefly describe use of RDBMS and analytical appliance approach and put more focus on use of Big Data platforms as sole DW provider for days to come.

A. Traditional DW

Relational databases (RDBMS) are the most common choice for data warehouse implementations. Even more, record-based structure is the most common choice of data storing within RDBMS’s. Since entire records are physically written and read as single units, this means that selecting any field requires reading the entire record, and changing a single value requires reading and rewriting the entire record unless there is other mechanisms to prevent

this behaviour, but then we can end up with result where we have same attribute spanned across multiple locations on storage level (e.g. Chained Rows problem) and additional maintenance is needed to fix this issue [7]. Over time, some vendors start introducing column-based structure along record-based structure, and some of them completely switch to column-based structure and benefited from query performance and lower storage costs. To cope with always rising data volumes, from early days, RDBMS's started to introduce indexing mechanisms which would allow quicker response to ad-hoc queries (classical use of DW systems consists of set of "static" queries from reporting tools and ad-hoc queries which can span across whole DW). Also, clever method of data pruning like partitioning, field maps, etc., were introduced to further enhance analytical needs from DW system, but at same time to speed even loading mechanisms into DW.

Constant problem in majority of DW implementations is how to effectively partition tables to support any kind of ad-hoc queries on large data volumes. This is due to different partition level limitations in traditional RDBMS's implementations (in almost all cases it stops either on partition or sub-partition level). There are methods how to overcome those limitations but it means need for specific implementation of Business Intelligence solutions on top of those DW implementations (like combining multiple attributes into single one and partition on that column/attribute, etc.). Even so, DW systems are mostly designed and implemented on top of traditional RDBMS's due to their wide use in many different applications and organizational need to have only one team managing and maintaining different instances and use-cases of same RDBMS engine.

B. *DW on appliance*

In a traditional DW implementation, the database administrator (DBA) can spend a significant amount of time tuning and putting structures around the data to get the database to perform well for large sets of users. As enhancements of existing DW implementations, new generation of specially built software and/or software/hardware combinations arise to deal with analytical problems on top of rising data volumes. A data warehouse appliance consists of a DBMS mounted on specified server hardware with an included storage subsystem specifically configured and optimized for data warehousing. They were named *appliance* because first implementations came in enclosed HW with embedded implementation of RDBMS's designed and optimized for analytical purposes. Term data warehouse appliance (DWA) defines a new category of computer architecture for data warehousing (DW) specifically targeted for Data Analytics and Discovery that is (a) simple to use (not a pre-configuration) and (b) very high performance for this workload. A DWA includes an integrated set of servers, storage, operating system(s), and DBMS [8].

Most obvious difference between traditional DW implementation and DW appliance is in scaling out whole system. When you need to scale traditional DW, you need to migrate all the data to a larger and more robust server. When using appliances it is necessary to have an additional components which connects into existing environment and auto-balance data across new components, thus scaling processing power and speed. Other differences are based on storing and retrieving data from storage components. While traditional RDBMS rely on indexes, partitions and/or statistics to speed up data retrieval, DW appliances rely on massively parallel processing (MPP) and full-table (or segment in smarter implementations) scanning complete data sets, including clever pruning mechanisms based on use of appropriate data type definition for underlying tables. Traditional RDBMS's also introduced data caching in memory to speed query responses, but only repeating queries benefit from that feature. At same time, DW appliances do not rely on statistics, in majority of cases they do not even rely on index use, thus allowing controllable and linear execution times over any data volume.

Introduction of DW appliances was a great leap in analytical capabilities of DW systems but it still needs to fight for its *place-under-the-sun* from DW's built on traditional RDBMS's and overcome common fear-of-new (which can manifest as the unwillingness to try new things or break from routine) from people who will maintain DW appliances.

C. *DW on Hadoop*

New-born. Newcomer to DW scene. Needs to prove a lot and to show its strength and benefits not just towards traditional DW, but also towards DW appliances. From technological aspects, it shares the same strategy as massively parallel processing (MPP) engines used within DW appliances, but whole philosophy around MPP implementation is different. For organisations and persons going into analytical spectrum without predetermined preferences towards traditional RDBMS's, major clash will be between DW appliance and DW on Hadoop.

Figure 1. Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics [31]

Original aspects and technology foundations for Hadoop implementation were around processing large number and volumes of textual files. Thus new purpose

built file system arose (HDFS - Hadoop Distributed File System). On top of HDFS, different processing engines were created ranging from batch processing towards real-time processing capabilities which consumed data stored in HDFS. Some of them focused on implementing DW solutions on top of Big Data platforms and this paper will focus on those components.

Gartner Inc. (world's leading information technology research and advisory company) started to introduce different Hadoop distribution Vendors to their Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics (as show on Figure 1). Entering 2015, the DW has expanded to address multiple data types, processing engines and repositories. This is creating new demands on the data warehouse market - for broader data management solutions for analytics, with features and functionality that represent a significant augmentation to existing enterprise data warehouse strategies.

III.HADOOP AND DW

Hadoop was engineered to rely on the schema on read approach, in which data is parsed, reformatted, and cleansed at runtime in a manually written program, whereas RDBMS enforces schema on write approach. In RDBMS's, a table's schema is enforced at data load time. If the data being loaded doesn't conform to the schema, then it is rejected. This design is called *schema on write*. Schema on read makes for a very fast initial load, since the data does not have to be read, parsed, and serialized to disk in the database's internal format. The load operation is just a file copy or move. Schema on write makes query time performance faster, since the database can index columns and perform compression on the data but it takes longer to load data into the database [27].

Initial processing engine created at same time as HDFS was based on paradigm of parallel processing of distributed file sets (e.g. MapReduce). MapReduce is a programming model on top of HDFS for processing and generating large data sets which was developed as an abstraction of the map and reduce primitives present in many functional languages [9]. A MapReduce program is composed of a *Map()* procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a *Reduce()* procedure that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance [10]. The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job. This is show on Figure 2.

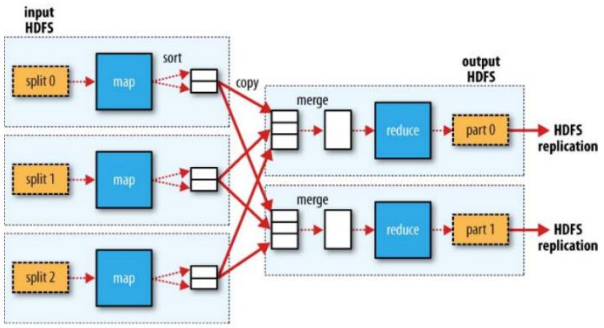


Figure 2. Map Reduce Programming Model [30]

By utilizing parallel computing and processing techniques like MapReduce, Hadoop is able to reduce long computation times to hours or even minutes. This is suitable for analytics over large volumes of historical data stored on disk, but it is not suitable for operational use of underlying data in real-time. Nevertheless, existing programming skills and infrastructure with parallel architecture of Hadoop itself can be leveraged to gain real-time data analytics on live data. Over the last few years, Hadoop has emerged as a popular way to handle both the storage and analysis of these increased volumes of data. Even so, Hadoop's parallel architecture can accelerate analytics over vast volumes of data. When it comes to real-time data, Hadoop's native batch processing and disk overheads are harder to overcome. One of the options to have real-time analytics using Hadoop is by combining an in-memory data grid (IMDG) with an integrated, stand-alone Hadoop MapReduce execution engine. With this technology fast results can be delivered for real-time data at same time as speeding-up analysis of large, historical data sets. Basically, IMDG is a data structure that resides entirely in RAM (random access memory), and is distributed among multiple servers [11]. First step from Apache community to overcome batch legacy of MapReduce were under project MapReduce 2.0 which was later renamed to YARN [12] due extensible enhancements and changes in logic than initially envisioned. On top of YARN, new extensible framework for building high performance batch and interactive data processing applications was created (Tez) [13]. This enabled new generation of Hadoop platform to appear (e.g. Hadoop 2) and at same time enabled DW solutions to be built on top of Hadoop platform.

IV. HADOOP DBMS'S FOR DW

Initial architecture over Hadoop platforms for a generic, scalable and fault-tolerant data processing architecture on distributed data processing systems was envisioned by Nathan Marz [14] and it's shown on Figure 3. This established robust architecture for distributed system with both batch processing and real-time processing capabilities, capable for long running queries over massive amount of historical data and also insight into operational real-time data events. End-user could combine those two sources without need to wait batch processing to end and get quicker results to *ad-hoc* queries.

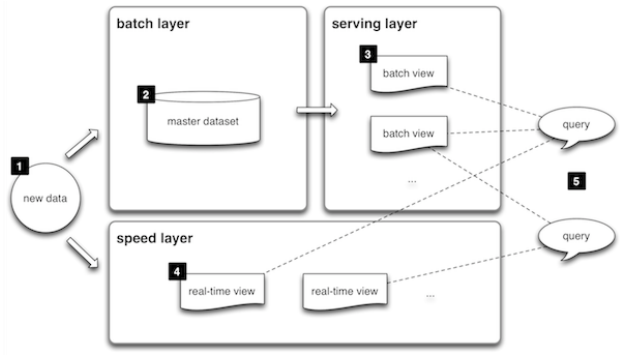


Figure 3. Lambda Architecture [14]

Better representation of Lambda Architecture is shown on Figure 4 which aligns real-time view of data with batch view. All other components prior to Serving Layer can be logically represented as Data Integration processes or ETL within traditional DW systems and will be discarded from future elaboration, thus focusing only on combined Serving Layer and different Hadoop DBMS's.

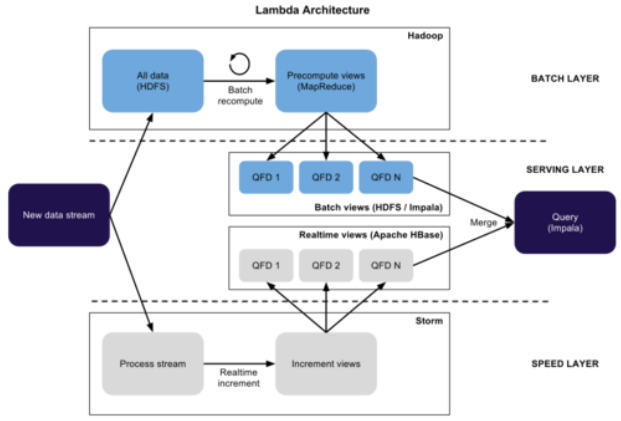


Figure 4. Consolidated Lambda Architecture [15]

Today's IT systems and networks can cause delays in capturing the data and delivering the data to a traditional DW for processing. With the explosion of data, DW's become data processing bottlenecks themselves, causing delays in processing and providing decision recommendations. As a result, the business misses an opportunity to take action (Figure 5). Major imperative is to move decision making much closer to the business event itself (e.g. to source data).

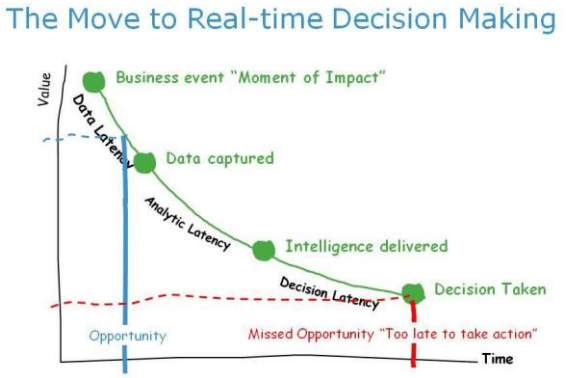


Figure 5. Correlation between Business event and Decision [28]

Some of providers within analytical appliance stack (like EMC) are starting to recognize that centralized, behemoth traditional DW are going the way of the dinosaur. For the purpose of big data analysis, they are too massive, too rigid and too far removed from the „moment of business impact"[28], so they start to evolve their business by moving analytics and decision-making to the edge of the network. They introduced solution for Real Time Intelligence (RTI) as a new kind of real time big data platform for a wide range of environments including telecom operators. The RTI system is based on the Pivotal One technologies and provides the ability to access and reason over large, diverse data sets ranging from subscriber databases to billing systems and network information. Together with in-memory real-time analytics, system is able to capture and process huge volumes of events coming from the carrier infrastructure. Those are one of many steps of existing vendors towards Hadoop platform but still with legacy within traditional spectrum which can be more of a burden than opportunity.

Newcomers without excessive technology baggage are introducing new platforms and components on daily basis with clear intent to tackle with traditional systems and take over large portion of DW business. Leading data integration vendors are already supporting Hadoop as native environment for data processing with even complete push-down capabilities to Hadoop for all functionalities. Due to large number of different solutions and implementations of DBMS engines on top of Hadoop (at the moment there is more than 80 different types of database solutions for Hadoop, e.g. Graph, Document, Object, etc.), next segments of this paper will focus on two of them that are more suitable for Data Warehouse implementations on top of Big Data platforms:

- Impala (developed by Cloudera)
- Hive (developed by Apache Foundation)

A. Impala

Cloudera Impala is a query engine that runs on Apache Hadoop. It brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache Hbase (open source, non-relational, distributed database modelled after Google's BigTable) without requiring data movement or transformation [16][17]. It features full SQL 2003-compliant analytic window functions together with external joins and aggregations using disk, sub-queries, incremental statistics and vast number of built-in functions, thus making it one of candidates for DW implementations over Hadoop. Introduction and enhancements on Impala were done by utilizing lot of underlying technologies like introduction of columnar storage format (Parquet) which can be aligned with partitioning capabilities to completely speed up *ad-hoc* query execution. Comparison of Parquet file storage format towards other file formats (with or without compression) is shown on Figure 6.

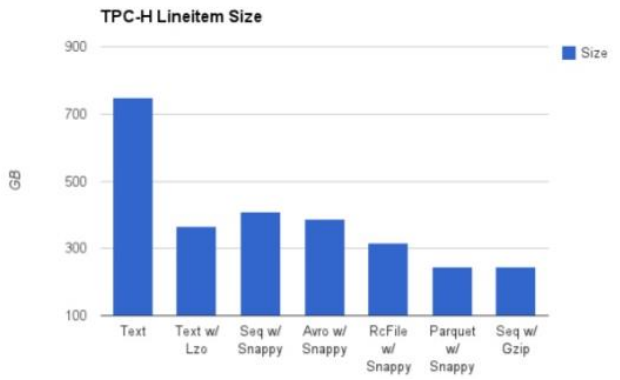


Figure 6. Parquet Storage Efficiency [29]

By utilizing Parquet storage engine with default Snappy compression, original text documents were shrink by a factor of 10, but at same time, data were redistributed based on columns within those textual documents. Performance gain based on TPC-DS like benchmark [18] is shown on Figure 7.

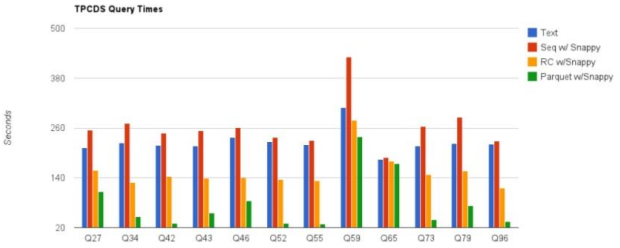


Figure 7. Parquet Scan Efficiency [29]

Currently there are obstacle related to DML SQL statements which prevents wide use of Impala as DW DBMS. It's not something that can't be overcome by modelling all phases of DW implementation (including data integration phase), but currently users of DW systems are accustomed to those features (e.g. ACID transactions). Because Impala reads data from HDFS or HBase, it can potentially have ACID support. Data in HDFS is immutable and thus ACID is unavailable unless complete files are changed, so HBase is roughly ACID for single-row transactions [19], and there are several projects to make HBase ACID for multi-row transactions. In DW implementations, fact tables are almost always immutable but *almost* part of this sentence needs to be stressed, which basically means that our fact data sometimes needs to be changed so ACID-like compliance is necessary. Impala is not a good choice for a mature enterprise DW's with strict security requirements and heavy use of exotic SQL features or proprietary extensions. In that case, Impala aims to augment the existing data warehouse by offloading high volume, low value data and workloads from the expensive enterprise data warehouse.

B. *Hive*

Apache Hive is a data warehouse infrastructure built on top of Hadoop for data summarization, query, and analysis [20]. It provides an SQL-like language called HiveQL and transparently converts queries to MapReduce, Apache Tez and, in the future, Spark frameworks. Transactions with ACID semantics were introduced to Hive in late 2014. which allowed users to easily modify data with inserts, updates and deletes. This extend Hive from the traditional write-once, and read-often system to support analytics over changing data and thus enabled reporting with occasional corrections and modifications and allowed operational reporting with periodic bulk updates from an operational database [21]. Most major players in BI segment are already implementing Native Hive connectivity and support for HiveQL together with traditional ANSI SQL support. Until recently, optimizing Hive queries focused mostly on data layout techniques such as partitioning and bucketing or using custom file formats [22]. Driven by Stinger initiative [23][24][25], Hive query time has improved dramatically, enabling Hive to support both batch and interactive workloads at speed and at scale. With Stinger.next initiative [25], Hive should support both transactions and SQL:2011 analytics as shown on Figure 8.

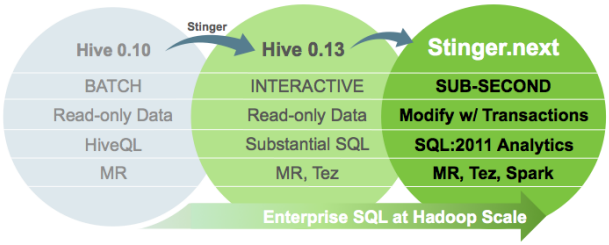


Figure 8. Stinger.next delivery [21][25]

Some of Stinger.next initiative goals are already in place, like transactions with ACID, allowing insert, update and delete, temporary tables, Cost Based Optimizer for star and bushy join queries, thus enabling Hive to become standard for DW implementations on top of Hadoop. On top of all technological advances, there is even methodological approach of optimizing queries and analytics on top of Hive. They focus on choosing appropriate execution framework (e.g. Tez in this v1.0.0, Tez or Spark in v1.1.0), selecting storage format like ORCFile [26] that sports fantastic speed improvements through techniques like predicate push-down, compression, using vectorised query execution (e.g. perform demanding operations in batches instead of single row each time), using cost-based query optimizer and most of all, write efficient SQL (same result can be achieved with different SQL statement and some of them may not be as efficient as they should be).

V.CONCLUSION

At the moment, different Hadoop distributions are implementing or have already implemented the same architecture supported by the traditional DW systems (including appliances). In next years, they will have competitive performance with much lower price tag. Analytic projects on top of Hadoop platforms are new, but the results of early projects have been impressive. The early projects have been completely business driven, and have focused on new ways of increasing revenue and reducing costs. Accelerating time-to-value has been a business imperative.

Stable and mature enterprise DW's will continue to exist due the fact that it can be cumbersome and time consuming (money as well) to migrate something which is

not broken. They will use Hadoop platforms for offloading data processing, minimize storage costs (major RDBMS vendors started to support HDFS for data storage), develop and test on Hadoop prior migrating result into traditional DW, etc.

From author's point of view, currently the best solution for DW on Hadoop and across different distribution is Apache Hive. This is due to comprehensive roadmap of new functionalities and enhancements and ANSI SQL standard implementation including analytical window functions. Other solutions are add-on on top of native Hadoop components thus adding some uncertainty to adoption of latest Hadoop features and/or visionary aspects of companies who are developing those solutions.

Nevertheless, DW on Hadoop is still not imaginable by majority of users and they will continue to use Hadoop for offloading tasks and processing for traditional DW until Hadoop platform gains some maturity.

VI. REFERENCES

[1] Data warehouse, Wikipedia, http://en.wikipedia.org/wiki/Data_warehouse

[2] W. H. Inmon, What is a Data Warehouse?

[3] Ralph Kimball, The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses

[4] Francis X. Diebold, A Personal Perspective on the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline∗

[5] Apache Hadoop, Wikipedia, http://en.wikipedia.org/wiki/Apache_Hadoop

[6] Hadoop contains the distributed computing platform that was formerly a part of Nutch. This includes the Hadoop Distributed File System (HDFS) and an implementation of MapReduce

[7] Don Burleson, Identifying Oracle Tables with Migrated/Chained Rows

[8] Data warehouse appliance, Wikipedia, http://en.wikipedia.org/wiki/Data_warehouse_appliance

[9] J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters

[10] MapReduce, Wikipedia, http://en.wikipedia.org/wiki/MapReduce

[11] Ki Sun Song, "Introduction to In-Memory Data Grid: Main Features", 04/2013, http://www.cubrid.org/blog/dev-platform/introduction-to-in-memory-data-grid-main-features/

[12] "Apache Hadoop NextGen MapReduce (YARN)", 11/2014, http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

[13] "Apache Tez: A Framework for YARN-based, Data Processing Applications In Hadoop", http://hortonworks.com/hadoop/tez/

[14] Michael Hausenblas, Nathan Bijnens, "Lambda Architecture", http://lambda-architecture.net/

[15] James Kinley, "The Lambda architecture: principles for architecting realtime Big Data systems", 04/2013, http://jameskinley.tumblr.com/post/37398560534/the-lambda-architecture-principles-for

[16] Cloudera Impala, Wikipedia, http://en.wikipedia.org/wiki/Cloudera_Impala

[17] "Cloudera Impala: The Leading Open Source Analytic Database for Apache Hadoop", http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html

[18] A TPC-DS like benchmark for Cloudera Impala, https://github.com/cloudera/impala-tpcds-kit

[19] Apache Hbase (official), http://hbase.apache.org/acid-semantics.html

[20] Apache Hive, Wikipedia, http://en.wikipedia.org/wiki/Apache_Hive

[21] Raj Bains, "Apache Hive ACID Transactions in HDP 2.2", 12/2014, http://hortonworks.com/blog/apache-hive-acid-transactions-hdp-2-2/

[22] "5 Tips for efficient Hive queries", 10/2013, http://www.qubole.com/blog/big-data/5-tips-for-efficient-hive-queries/

[23] Alan Gates, "The Stinger Initiative: Making Apache Hive 100 Times Faster", 02/2013, http://hortonworks.com/blog/100x-faster-hive/

[24] Harish Butani, "Announcing Apache Hive 0.13 and Completion of the Stinger Initiative!", 04/2014, http://hortonworks.com/blog/announcing-apache-hive-0-13-completion-stinger-initiative/

[25] Apache Hive: The standard for SQL queries in Hadoop, http://hortonworks.com/labs/stinger/

[26] Carter Shanklin, "ORCFile in HDP 2: Better Compression, Better Performance", 09/2013, http://hortonworks.com/blog/orcfile-in-hdp-2-better-compression-better-performance/

[27] Joe Pasqua, "Schema-on-Read vs Schema-on-Write", 02/2014, http://www.marklogic.com/blog/schema-on-read-vs-schema-on-write/

[28] Laddie Suk, "It's not just Big Data … It's Gigantic Data! Are you ready?", 04/2013, https://infocus.emc.com/laddie_suk/its-not-just-big-data-its-gigantic-data-are-you-ready/

[29] Graham Gear, "Building a Hadoop Warehouse with Impala", 10/2014, http://www.itproportal.com/wp-content/uploads/presentations/pdf/Building%20a%20Hadoop%20Warehouse%20with%20Impala.pdf

[30] Apache Hadoop , 12/2014, https://tedwon.atlassian.net/wiki/display/SE/Apache+Hadoop

[31] Mark A. Beyer, Roxane Edjlali, "Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics", 02/2015, http://www.gartner.com/technology/reprints.do?id=1-29X7GVV&ct=150212&st=sb

Delta View Generation for Incremental Loading of Large Dimensions in a Data Warehouse

Igor Mekterović*, Ljiljana Brkić *
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
igor.mekterovic@fer.hr
ljiljana.brkic@fer.hr

Abstract - Incremental load is the preferred approach in efficient ETL processes. Fact tables are the ones who benefit the most from this approach, since they are large in terms of row count. For the sake of simplicity, dimension tables are often ignored and populated in a full reload manner. However, big dimensions (e.g. Client) can also have a significant impact on the ETL process and should also be considered for incremental load. Although they have much smaller cardinality than a typical fact table, it usually takes much more resources to calculate one dimension table row than to calculate one fact table row. Large dimension tables are based on multiple source tables, and it is not trivial to determine the changed records that should be considered for the incremental load because changes in any and all of underlying source tables must be considered. In this paper, we present an algorithm for the dimension's delta view generation. Delta view for a dimension encompasses all its source tables and produces a set of keys (e.g. ClientIds) that should be incrementally processed. We have employed this approach in a real world project and have noticed a significant reduction in the loading time of big dimensions.

I. INTRODUCTION

While OLTP systems are focused on supporting operational business processes requiring short response time, the intention of data warehouses (DW) is strategic planning and decision-making based on data integrated from heterogeneous systems. Multidimensional database technology is the key factor in interactive analysis of large amounts of data for decision-making. A multidimensional data model is commonly used in data warehouses that integrate enterprise's data from several sources into large repositories. The data acquisition process represents one of the most technically demanding and one of the most expensive parts of building a data warehouse. The data needed for the data warehouse are being extracted from their sources during an extract, transfer and load (ETL) process. According to [1], ETL design and development work consumes 60 to 80 percent of an entire BI project, therefore it is essential to optimize this layer of BI environment. Typical DW architecture includes a data staging area where the extracted data are stored temporarily, and subsequently, upon cleaning and transformation phase, transferred into the data warehouse. Commonly, there is no real-time connection between a DW and its data sources, since the massive read in decision support systems would conflict with the continuous update of operational systems and result in poor response times.

After the first or *initial loading*, DW must be periodically refreshed to keep up with source data updates. DW loading is usually done periodically, by a background process. The update patterns (daily, weekly, etc.) for traditional data warehouses and data integration process result in outdated data to a greater or lesser extent [3]. The naive approach to refreshing a data warehouse is referred to as *full reloading.* Possible refreshment scenario is to repeat the initial loading process using modified source data, compare the results with the current DW in order to determine changes that need to be done and finally perform the changes. This strategy is known as full DW reload. With the increasing size and complexity of DW, full reloading becomes inadequate, and in some cases inapplicable. More appropriate approach is a gradual change of DW in accordance with the changes that have occurred in the data sources since the last synchronization. Only the data that has changed since the previous reload needs to be transformed and processed. This approach is known as *incremental reloading*. The volume of modified data is always small compared to the entire dataset, thus, incremental reloads can be assumed to work quicker and more efficiently than full reloads. In most cases, the content of a DW is updated once a day. However, there are businesses such as health systems, e-business, stock brokering and others where such level of freshness of data is not sufficient, since they operate in a business time schedule of 24x7 and need relevant information as fast as possible in order to react in a near real-time manner. Real-Time Data Warehousing refers to an approach where DWs are updated as frequently as possible, greatly reducing data accuracy and availability gap.

The basic assumption of incremental and real-time loading is that the changes of source data can be captured and later on propagated to the data warehouse. There is a number of known techniques for changed data capture. The most commonly used ones are discussed in chapter III.

II. RELATED WORK

Incremental warehouse reloading is evidently related to incremental maintenance of materialized views [2] [3] as both areas cope with same issue – how to update physically integrated data under a given time constraint [4].

The core idea of materialized view incremental update stated in [4] assumes computing new content using materialized parent view (the outdated view needed to be refreshed) and its incremental update, often referred to as

delta view. Finding tuples constituting the delta view is addressed in a number of papers [5] [6] dealing with so called CDC techniques. This concept of capturing information about changed tuples is valuable later in the extract phase of incremental ETL process [7] [8]. Work described in [6] is to some extent similar to ours, but we go further with optimization of extract phase. The authors presented a tool for automatic creation of all necessary structures for altered data capture and subsequent extraction. They track changes by storing changed tuples identifiers (primary keys) in a special CDC tables and later on use stored identifiers for fast retrieval of tuples relevant for the next ETL cycle. Complementary to their work we addressed the issue, and proposed the solution, for finding identifiers of tuples responsible for refreshment of a particular dimension table. Since the contents of denormalized dimensions are refreshed based on the contents of numerous source tables, finding such identifiers can be a challenging task.

III. CHANGE DATA CAPTURE TECHNIQUES IN DBMS

Change Data Capture (CDC) is a generic term for techniques that monitor data sources in order to detect and capture necessary data changes [9] [10]. Different CDC techniques are used in practice. All CDC approaches assume that a complete changed record will be available to the requesting target.

The following is an overview of three most popular CDC techniques applicable in DBMS environment.

A. Timestamp based Capture

Timestamp based data capture is performed by adding a timestamp column (specifying the time and date when observed row was last modified) to each table of interest. These timestamps provide the selection criteria for the capture of changed records. The need to change the structure of each table of interest, by adding a single timestamp column, is the first disadvantage of this approach.

Second, it is not possible to capture intermediate states of data. If a record changed state more than once since the last ETL run, only the current (precisely the last) state of the respective record would be at disposal. Accordingly, when using this method, the inserted records cannot be distinguished from updated ones.

An additional problem with this technique is a lack of possibility to detect deleted records. Possible solution to detecting deletions is to label these records as inactive until they have been processed [10]. This concept leads to other problems in operational systems. For example, all applications working with operational data have to treat inactive records as deleted ones.

This method causes frequent writes of timestamp attributes on the operational systems and frequent reads whenever ETL process runs. Indexing timestamps can speed up reading but on the other hand can slow down INSERT and UPDATE operations [11].

B. Trigger-Based Capture

This approach involves the usage of triggers to track changes in the source database. Whenever a modification takes place in the source database corresponding log record is written into a dedicated log table. These dedicated log tables serve to determine rows changed since the last ETL processing.

The log tables usually contain a timestamp column that provides the exact time and date when a given row was last modified. The statements that are executed when trigger is fired, maintain the timestamp column accurate i.e. update the timestamp column with the current time.

This technique assumes that the DBMS supports active behavior i.e. triggers. The set and granularity of the events and conditions that cause a trigger to fire are database specific.

This approach enables capturing all types of events (INSERT, UPDATE, DELETE, MERGE). For UPDATEed rows both before and after image can be available.

One shortcoming of this approach is that it might significantly affect performance on the source systems. For example, if all INSERTs, UPDATEs and DELETEs are captured, this method creates twice the workload for the database [10]. This technique imposes a considerable operational overhead at the source system. This impact should be carefully considered prior to implementation on a production source system.

C. Transaction Log Capture

This technique employs the logging and recovery capabilities of a DBMS [9]. Since transaction logs are utilized by < DBMS to store transaction information for logging and recovery, these logs contain all information required to capture changed data. A specialized vendor-specific application must be written to monitor the log files and capture the data of interest.

One of the shortcomings of this approach is that it assumes that the transaction logs remain available until the changes of interest are captured. In order to prevent transaction loss, the database administrator must constantly monitor the transaction log area. The problem can occur when the recycling of the transaction log takes place, that is, when a part of the transaction log, containing transactions already committed, is reused by the DBMS [12].

The log capture method is probably the most efficient approach to incremental capture due to the fact that log writing is already well optimized and available on most DBMS platforms today [10].

IV. DELTA VIEWS

Once we know what records have changed in the source system, we can use that information to incrementally populate fact and dimension tables in the data warehouse. Incremental loading of fact tables is rather trivial and will not be further discussed. Fact tables are also the ones that benefit most from the incremental loading style since they typically contain a huge number of records. On the other

hand, dimension tables are usually relatively small and can be populated in a full reload manner. However, there are cases when dimension tables are also fairly large, both in terms of number of records and number of attributes. Such dimensions are sometimes referred to as "monster dimensions", "whose size exceed 100 million rows" [13]. Typically, those are customer or client dimensions in large enterprises. In our case, out client dimension is of 10^5 order with more than a hundred attributes. Such dimensions may take a long time to load, since they are both large in record count and have a considerable number of attributes that need to be calculated along the way (e.g. student's Grade Point Average, rank, etc.). We believe that, for such large dimensions, it is beneficial to construct the "delta view" – a view that will contain keys of records that need to be addressed in the ETL procedure. Delta view contains the keys of deleted, updated and inserted records of the relational source table, and that information is then used in accordance with the dimension type to update the dimension. For instance, for type 1 SCD (Slowly Changing Dimension), one would simply delete all dimension records that are members of delta view, and then reload those updated delta view records as well as the new ones. Type 2 SCD would handle updates and inserts differently, etc.

Unlike fact tables, delta view records are deduced not only from the base source table (for the dimensions), but also from many other tables that are being used in the dimension's ETL process. In our, real-world case, our dimension table is based on 23 relational tables. A change in any of those 23 tables can impact the content of the dimension. In the following chapter, we elaborate on the prerequisites and algorithm for the delta view construction.

A. Delta view construction

Algorithm for the delta view construction will be explained with the help of the running example. Figure 1 shows a simplified version of the relational database subset used to populate dStudent dimension table:

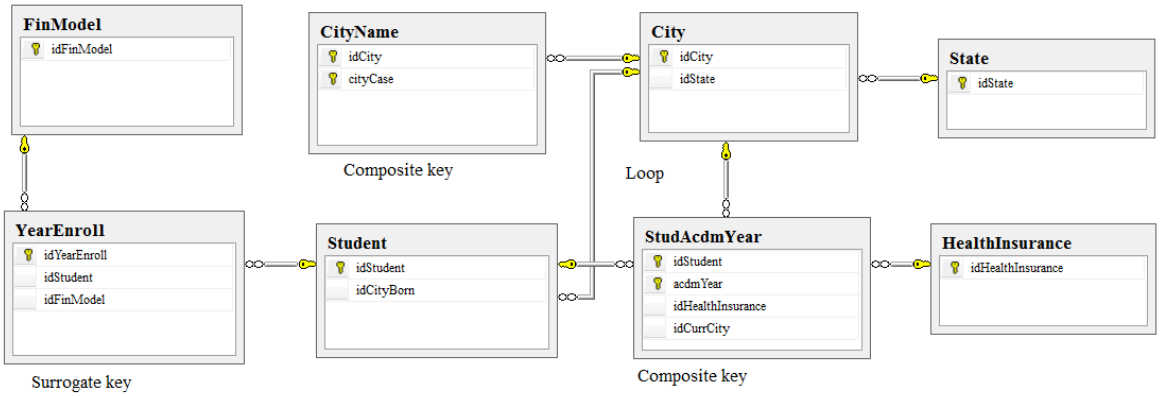


Figure 1 Relational model used to populate dStudent dimension table

For the sake of simplicity, all attributes except primary and foreign keys have been removed. Student enrolls in 0..N years using one of the supported financial models (scholarship, self-funded, etc.). Also, for each academic year we keep track of current city (and state) of residence and student's health insurance. CityName is specified in different word grammatical cases (in Croatian language there are seven different grammatical cases) because different cases are needed for different contexts (in various reports). Similar constructs appear in multilingual databases where a name for an entity is specified in various languages.

From this model, various dimension attributes are calculated. Some are trivial (e.g. dStudent.currCityName), while others have to be calculated with stored procedures (e.g. a flag indicating whether student repeated/failed any academic year, whether the first/last year enrollment was funded by the Ministry of Science, etc.). Granulation of dStudent dimensional table is the same as Student – one record per one student. So, the goal is to determine all primary keys (idStudent) that have changed in between two loads. We assume that all tables have accompanying CDC tables, having the same name but with the underscore prefix, i.e. _Student, _City, etc.

Any change in tables referenced by student, branching out, will affect the denormalized dStudent table. For instance, if the name of the city or state changes, this will affect dStudent dimension content both directly (via CityBorn relationship) and indirectly (via StudAcdmYear table, and current City/State attributes). At first, it might appear as if it is enough to start from the Student table and follow the referenced tables, however that is only part of the solution: referencing tables might also have an impact on the base table, e.g. YearEnroll is such a table. The same is true for all other tables that are iterated with the algorithm, e.g. CityName references and affects City table (in this context). These directions or ways that data from a table affects another table and consequently the base table used to populate the dimension table, cannot be determined just by inspecting the foreign key relationships. They have to be manually assigned by the domain expert. A tool can be used to assist in this task: using the foreign key relationships a graph can be constructed with default directions assigned according to the foreign key cardinalities. After that, a domain expert only has to correct directions of certain, if any, "arrows". Figure 2 shows such a graph for the example used. Arrow directions can be interpreted as "data flows" in the context of

dimension table being considered. That is, data from CityName "flows", via City, to the dimension table dStudent.

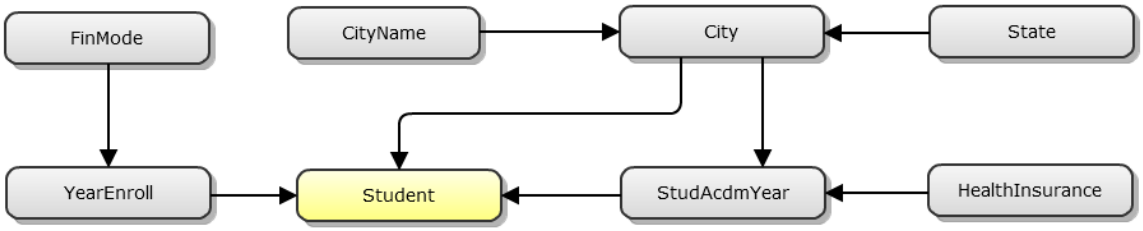


Figure 2 A data flow graph for the relational model in Figure 1

In terms of metadata, to support such setup, one only has to add a single flag to the metadata table describing the foreign keys: a flag that indicates which way the data is "flowing". In the conclusion, the standard metadata information (tables, primary key, foreign keys) enriched with a single flag in foreign key metadata table is enough information to construct the delta view for the dimension table.

Note that though we refer to relationships between tables as "foreign keys", in general, they don't have to be. The graph in fact represents ways in which tables have been joined throughout the ETL process to produce the dimension table. These can be arbitrary joins. Furthermore, the relational database might not even employ foreign keys for some reason. This only makes metadata definition process a bit less trivial, since design expert has to manually connect the tables, rather than leveraging the foreign key relationship information read from the RDBMS's system catalog.

The delta view construction idea is to, starting from the base table (e.g. Student), recursively traverse the graph and gather primary keys of the related tables. The related tables are divided in two sets:

- **REFD** – a set of tables **referenced** by the current table (e.g.
  REFD (Student) = {City},
  REFD (City) = {State})
- **REFING** – a set of tables **referencing** the current table (e.g.
  REFING (Student) = {YearEnroll, StudAcdmYear},
  REFING (City) = {CityName, StudAcdmYear, Student})

Also, we assume that every table t has a corresponding t' CDC table that contains at least primary key attributes (denoted with PK(t)) and operation attribute (update, insert or delete), e.g. for t = Student, t' = _Student and PK(t) = {idStudent}. For the sake of simplicity, in the definition of delta view we omit the operation attribute, though it should appear on every projection operation where PK(t) is used. Finally, the delta view $\boldsymbol{delta(t)}$ for the dimension table **t** is given with:

$$delta(t) = cdc(t) \bigcup_{r \in REFD(t)} refdcdc(t,r) \bigcup_{r \in REFING(t)} refingcdc(t,r) \quad (1)$$

where:

$$cdc(t) = \pi_{PK(t)}\, t' \quad (2)$$

$$refdcdc(t,r) = \pi_{PK(t)}\, \mathrm{t} \rhd\lhd delta(r) \quad (3)$$

$$refingcdc(t,r) = \begin{cases} \pi_{PK(t)}\, delta(r), & if\ PK(t) \subseteq PK(r) \\ \pi_{PK(t)}\, delta(r) \rhd\lhd (\mathrm{r} \cup \sigma_{op='del'}\mathrm{r}'), & \text{otherwise} \end{cases} \quad (4)$$

Obviously, calling $\boldsymbol{refdcdc}$ and $\boldsymbol{refingcdc}$ results in recursive calls that traverse the entire graph. We have divided the related tables in two sets because they produce different SQL constructs to attain the primary keys: **Referenced** tables must be joined to the base table to get the base table's primary keys as stated in (3). For **referencing** tables, as stated in (4), there are two possible scenarios:

- Referencing table has a composite primary key whose subset is a foreign key referencing the current table (e.g. StudAcdmYear has {idStudent, acdmYear} primary key, where {idStudent} is used to reference the Student table).
  In this case, it is superfluous to join the referencing table to the current table since current table's CDC primary keys can be attained using referencing table only, e.g.
  $\pi_{idStudent}\ (delta(StudAcmdYear))$
  We denote this case in the final delta view example with the comment /* -ABBREV- */, as this is the abbreviated SQL construct, without the current table join.
- Referencing table has a surrogate key and references the current table via the foreign key attribute that is not a part of the primary key. In this case, the CDC set for the referencing table has to be calculated and then

joined with the same table to extract the foreign key. For instance, for the YearEnroll table:

$$\pi_{idStudent}(delta(\text{YearEnroll}) \rhd\lhd (\text{YearEnroll} \cup \sigma_{op='del'}(\text{YearEnroll})))$$

We denote (alias) the table with "exp" as in "expanded", since we are using the self-join to "expand the primary key" to get the foreign key. However, since we are using the *current* data to get the foreign key, we must also include (UNION) the deleted records. We denote this case in the final delta view example with the comment **the delete exception**.

Note that, when traversing the chart, certain nodes are reached more than once (e.g. City). *cdc* function for that node returns the same keys (SQL) so there is no need to execute (traverse the subgraph) more than once. More precisely, our primary concern is not with the algorithm execution time or complexity in that respect, since these charts are small – it is hard to imagine more than few hundred nodes (tables) which presents no problem.

```
WITH WCity (idCity) AS (
      SELECT _City.idCity
        FROM _City
      UNION                          /* - REFING ABBREV- */
      SELECT CityName111.idCity              /* -ABBREV- */
        FROM  (
         SELECT _CityName.idCity, _CityName.cityCase
           FROM _CityName) AS CityName111
      UNION    /* REFD UNION */
      SELECT City.idCity
        FROM City
        JOIN  (
         SELECT _State.idState
           FROM _State) AS State112
          ON City.idState = State112.idState)
SELECT _Student.idStudent
  FROM _Student
UNION                  /* - REFING ABBREV- */
SELECT StudAcdmYear1.idStudent               /* -ABBREV- */
  FROM  (
   SELECT _StudAcdmYear.idStudent, _StudAcdmYear.acdmYear
     FROM _StudAcdmYear
   UNION    /* REFD UNION */
   SELECT StudAcdmYear.idStudent, StudAcdmYear.acdmYear
     FROM StudAcdmYear
     JOIN  WCity AS City11
       ON StudAcdmYear.idCurrCity = City11.idCity
   UNION    /* REFD UNION */
   SELECT StudAcdmYear.idStudent, StudAcdmYear.acdmYear
     FROM StudAcdmYear
     JOIN  (
      SELECT _HealthInsurance.idHealthInsurance
        FROM _HealthInsurance) AS HealthInsurance12
       ON StudAcdmYear.idHealthInsurance =
HealthInsurance12.idHealthInsurance) AS StudAcdmYear1
UNION                  /* -REFING JOIN- */
SELECT Student_Exp_Student.idStudent
  FROM  (
   SELECT _YearEnroll.idYearEnroll
     FROM _YearEnroll
   UNION    /* REFD UNION */
   SELECT YearEnroll.idYearEnroll
     FROM YearEnroll
     JOIN  (
      SELECT _FinModel.idFinModel
        FROM _FinModel) AS FinModel21
       ON YearEnroll.idFinModel = FinModel21.idFinModel) AS
YearEnroll2
  JOIN  YearEnroll AS Student_Exp_Student
    ON Student_Exp_Student.idYearEnroll = YearEnroll2.idYearEnroll
 UNION (SELECT idStudent FROM _YearEnroll WHERE operation = 'd')
/* DEL EXCPTN */
UNION    /* REFD UNION */
SELECT Student.idStudent
  FROM Student
  JOIN  WCity AS City3
    ON Student.idCityBorn = City3.idCity
```

Figure 3 Final delta view for Student dimension table

We are concerned with creating more compact and readable SQL which is, hopefully, more beneficial to the relational engine in terms of query optimization. To that end, when implementing the above algorithm, we use the ANSI SQL-99 WITH CLAUSE to define reusable SQL statements or "local views". Every node that is reached more than once will create a WITH CLAUSE that will be used more than once in the main SELECT statement. With clauses may be referencing each other, so it is important to order them properly. Nodes further from the base table should be ordered first. For the node's distance from the base table we take the maximal distance, since a node can be reached via different routes, e.g. City has distances 1 and 2 (via *StudAcmdYear*), therefore its distance is 2

Finally, **Figure 3** shows the final delta view for the Student dimension table.

We have applied this incremental loading technique to the real world project dimension table consisting of 62 attributes and two accompanying minidimensions (31 and 19 attributes), that is, 112 attributes in total. Dimension itself is based on 23 relational tables and trigger based change data capture is used to detect changes in the data.

Table I shows the average dimension load times accumulated over the years (N is the number of load times, where one load corresponds to one day). When we switched from full reload to the proposed delta load, the average duration was about 89 minutes. Currently, using delta load the average duration is only 4 minutes. We've continued to perform full reloads occasionally (on weekends), for testing purposes, and the full reload duration with the current data volume averages roughly 127 minutes. Obviously, by employing incremental loading approach we have significantly reduced the time spent to populate the dimension.

Table I Average load times in a real world case

Load type	N (loads)	Average (min)
Full	707	88,67
Delta	1276	4,03
Full (test)	170	126,97

Initially, we have constructed the delta view by hand and spent several days writing and testing the query. Using the proposed algorithm, we were able to describe and generate the same delta view in less than an hour. The view spans 227 lines of SQL using the same formatting as shown in Figure **3**.

V. CONCLUSION

In this paper we have presented an algorithm for delta view construction that enables incremental loading of dimension tables. This method is suitable for large, complicated (both in terms of attribute and row count) dimensions, , because they can have significant impact on the ETL process. We have shown that delta view, yielding the keys of records that need to be incrementally

(re)processed, can be constructed from the graph of data flow dependencies that can be fairly easy assembled.
Obviously, incremental loading of large dimensions can significantly shorten the ETL process, but writing (and maintaining!) large delta views manually is complex and error prone. On the other hand, data flow graph is simple to define, especially with the help of a tool that, for instance, uses RDBMS's metadata. We therefore find this algorithmic approach to delta view generation to be the apt solution for the given problem. Finally, we have verified it by applying it to the real world case and achieved significant reduction in loading times.

REFERENCES

[1] W. Eckerson and C. White, "Evaluating ETL and Data Integration Platforms," 2003.

[2] T. Palpanas, R. Sidle, R. Cochrane and H. Pirahesh, "Incremental Maintenance for Non-Distributive Aggregate Functions," in *Proceedings of the 28th VLDB Conference*, Hong Kong, 2002.

[3] H. Gupta and I. S. Mumick, "Incremental Maintenance of Aggregate and Outerjoin Expressions," *Information Systems,* pp. 435-464, 2006.

[4] W. Liang, H. Wang and M. Orlowska, "Materialized view selection under the maintenance time constraint," *Data & Knowledge Engineering,* vol. 37, pp. 203-216, 2001.

[5] R. Prabhu and D. Lyman, "Extracting delta for incremental data warehouse maintenance," in *IEEE 16th International Conference on Data Engineering (ICDE'00)*, San Diego, CA, USA, 2000.

[6] C. R. Valêncio, M. H. Marioto, G. F. Donega Zafalon, J. M. Machado and J. C. Momente, "Real Time Delta Extraction Based on Triggers to Support Data Warehousing," in *International Conference on Parallel and Distributed Computing, Applications and Technologies*, 2013.

[7] X. Zhang, W. Sun, W. Wang, Y. Feng and B. Shi , "Generating Incremental ETL Processes Automatically," in *Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06)*, 2006.

[8] T. Jörg and S. Deßloch, "Towards Generating ETL Processes for Incremental Loading," in *IDEAS08*, Coimbra, Portugal, 2008.

[9] S. Kozielski and R. Wrembel, New Trends in Data Warehousing and Data Analysis: Annals of Information Systems, vol. 3, Springer, 2008, pp. 19-49.

[10] M. Bokun and C. Taglienti, "Incremental Data Warehouse Updates," *Information Management Magazine,* 1998.

[11] R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit, Wiley Publishing, Inc., 2004.

[12] R. Rocha, F. Cardoso and M. Souza, "Performance tests in data warehousing ETLM process for detection of changes in data origin," in *Proceedings of Data Warehousing and Knowledge Discovery, DaWaK 2003, LNCS 2737*, 2003.

[13] R. Kimball and M. Ross, The Data Warehouse Toolkit, 2nd ed., Wiley Computer Publishing, 2002.

[14] M. Baranović, I. Mekterović and M. Madunić, "Data warehouse as a part of the higher education information system in Croatia," in *Proceedings of the 25th International Conference on Information Technology Interfaces*, 2003.

SimBusPro: A Simulation-Based Decision Support Tool used for the Optimization of Business Processes running on the Cloud

Ahmet Özkök, Ali Ercingöz, Halit A. Dönmez,Tansel Dökeroğlu, Veysi İşler
Simsoft Computer Technologies Ankara/TURKEY
{ahmet.ozkok, ali.ercingoz, halit.donmez, tansel.dokeroglu, veysi}@simsoft.com.tr

Abstract - In this study, SimBusPro, a novel business process modelling and optimisation tool that runs on Cloud, is introduced. The tool is developed during a project called SCI-BUS (Scientific Gateway-Based User Support) under the seventh Framework of the European Commission. SimBusPro has a visual editor that uses BPMN 2.0 notation. A wide range of commercial and academic workflow processes can be modelled, simulated and optimised using SimBusPro. The distributed computational architecture of the tool is capable of completing complex simulations in reasonable times. In addition to these features, with the alternative resource planning module, SimBusPro allows optimal resource assignments to be made for the processes that have been modelled.

I. INTRODUCTION

In the contemporary business world, it is necessary for firms to operate in a manner that is both cost-effective and responsive to changing conditions so as to ensure their existence. Thus, firms must continually analyse their business processes in their pursuit of continuous improvement, thereby bringing about continual improvements in their resource utilisation efficiency. One of the most prominent ways in which this continuous improvement can be achieved is via the use of business process modelling, and simulation and optimisation software [1].

SimBusPro [2], which was developed during a project called SCI-BUS (Scientific Gateway-Based User Support) under the 7th Framework of the European Commission, is an example of such a software program. In essence, SimBusPro is a simulation and optimisation-based decision support tool that can be run on cloud [3]. By using SimBusPro's modelling editor, it is possible to design and develop workflow diagrams that are in line with BPMN 2.0 [4, 5] standards. Once such a model has been established, its simulation, using parameters defined and entered by the user, can be carried out, which, in turn, will yield highly accurate and precise simulation reports containing various graphs representing miscellaneous information about the process that has been modelled. Via the results that are displayed in these reports, it is possible to make analyses that will lead to the achievement of increased efficiency in the utilisation of the resources of the processes that have been modelled.

Business process modelling software programs, despite having the ability to model business processes with consummate ease, are typically limited by their ability to attain simulation results in reasonable times. SimBusPro is an exception to this and can generate simulation reports of complex models rapidly as it is compatible with Cloud technology. SimBusPro uses a scalable, distributed architecture in its calculating infrastructure. As a result, besides possessing the capability to perform simulations rapidly, SimBusPro eliminates any need for setup and updating, meaning that it is always readily available for use. SimBusPro makes use of a pay-per-use model, thereby allowing multiple users to benefit simultaneously in a secure environment.

In July 2014, from the results that were obtained from a modelling and simulation competition in which SimBusPro was used as the modelling and simulating tool, users provided feedback about the software, thereby enabling its weaknesses to be identified. Since then, the problems identified have been addressed and SimBusPro has become more capable of successfully responding to the needs of its prospective users. SimBusPro is now extremely accomplished in the modelling and simulation of a very broad spectrum of processes [2]. Thus far in the paper, SimBusPro's various proficiencies have been discussed. In the second part of the paper, SimBusPro's programming design and architecture will be discussed. The third section of the paper will discuss the simulation-based optimisation of a pizzeria, whilst the fourth section will discuss results, and planned actions of the future.

II. SIMBUSPRO'S PROGRAMMING ARCHITECTURE

Business process modelling, simulation and optimisation software programs are designed to model the service and manufacturing processes of business and firms, and to subsequently optimise their resource utilisations. Using SimBusPro, it is possible to obtain simulation data about processes in a virtual environment [6, 9]. This will enable those who opt to use SimBusPro to test certain hypotheses about their business processes without actually having to carry them out in real-life, which, effectively, will enable them save a lot of time and money.

A. *BPMN 2.0 Based Visual Editor*

BPMN is a universal modelling notation through which businesses are presented with a means with which they can display their internal procedures [4]. BPMN is a model established by an organization called OMG (Object Management Group Inc.) in which businesses and other such entities are presented with a means, or a platform, through which their internal business procedures in a graphical notation. This model was established in the hope of giving organization a means to communicate these procedures in a standard manner, thus facilitating the business transactions and performance collaborations between different organizations. This model will ensure that businesses understand themselves and the participants in their business more clearly and will enable them to adjust to new business circumstances quickly and with ease.

The SimBusPro editor was developed using Eclipse's BPMN 2.0 library. Using this editor, it is possible to model a broad spectrum of business processes using all elements that are present in BPMN 2.0. In Figure 1, the fundamental elements of BPMN 2.0 are shown:

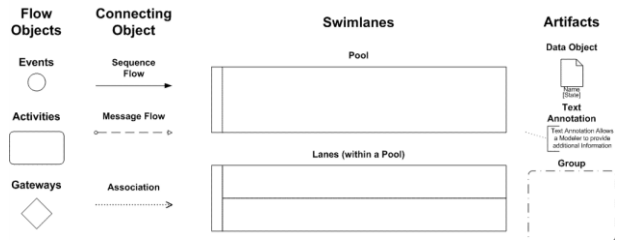


Figure 1. Fundamental elements of BPMN 2.0

Three basic elements are events, activities and gateways. Events are occurrences cannot be controlled internally - such as the arrival of a message. Events trigger the start or end of activities within a business process. Tasks are planned activities that are carried out by using the resources of business process. They may be manual based on service tool or based on the sending or receiving of messages. Gateways indicate decision points within processes where the work flow is effectively split.

B. *Infrastructure of SimBusPro*

The SimBusPro Gateway is used for optimizing business processes through simulation by utilizing the SCI-BUS framework. To simulate multiple models simultaneously, the application sends models to the portal, which are then transferred grid machines via the gUSE system. The SimBusPro gateway has used the BOINC Desktop Grid as its Distributed Computing Infrastructure [7, 10, 11, 12, 13]. SimBusPro's architecture can be seen in Figure 2:

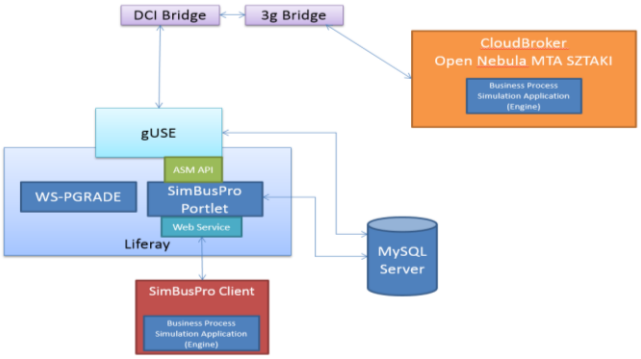


Figure 2. *SimBusPro*'s architecture

C. *Resource Utilisation Optimisation*

One prominent feature of SimBusPro is its ability to generate alternative solutions of models. This allows users to evaluate the performance of a given system under certain induced and controlled changes. SimBusPro enables users to toggle with the parameters of an existing model and display the results of the simulations of different versions of this model simultaneously on the Cloud. This endows users with the ability to easily identify superior models, thereby providing them with a means to improve their processes or systems. These capabilities will provide a multitude of benefits to firms who are seeking to improve their existing business processes or systems, and also to those who are planning the development of some system but have several alternative versions to choose from.

Thus, SimBusPro can be used to optimise the number of resources that are used in a modelled system. To illustrate the way in which SimBusPro finds an optimal solution among various alternative models, let us imagine a fictional pizzeria. Let us assume further that the number of employees is not fixed and that 1 to 10 waiters, 1 to 8 delivery boys and 2 to 4 chefs can be employed. In such a case, SimBusPro would calculate all of the possible resource combinations and identify the best model with respect to some key performance indicator – such as cost or customers served. In this scenario, SimBusPro must evaluate 240 separate models (number of waiters * number of delivery boys * number of chefs) and identify the best one among the lot. Due to the fact that SimBusPro makes use of a multitude of processors upon the Cloud in carrying out its calculations, it can generate the simulation reports corresponding to alternative models in reasonable times.

D. *Multi-Objective Optimisation Capability*

SimBusPro has the capability of carrying out both single-objective and multi-objective optimisations [8]. Multi-objective optimisations may be carried out with respect to any parameter including cost and process times (both of which it would typically seek to minimise). In the formula for multi-objective optimisation, each measure that is aimed to be optimised is multiplied by some weighted coefficient, the size of which depends on its importance. The formula is as follows:

$$\min\left(\sum_{i=1}^{n} w_i r_i\right)$$

where n is the number of measures that are to be optimised, w_i is the weighted coefficient of measure i and r_i is the value that is assumed by measure i.

III. SOLUTION OF AN OPTIMISATION PROBLEM

In this section of the paper, the optimisation of the business processes of a pizzeria will be discussed. The business processes of this pizzeria were modelled and simulated using SimBusPro. Alternative models were then generated by running simulations on the Cloud, after which an analysis of the alternatives was carried out in order to identify the optimal solution.

A. Optimisation of the Business Processes of a Pizzeria

The problem is as follows: A pizzeria is open for 10 hours (600 minutes) a day. When a customer arrives (once every 2 minutes) he/she chooses whether to eat in (80%) or take away (20%). The restaurant also accepts orders via phone or internet (which arrive once every 5 minutes). The details of the different processes that take place within the restaurant are given in Tables 1-3.

TABLE I. EAT IN

Activity	Duration	Resources used
Take a seat	1 minute	
Order	5 minutes	1 waiter
Pizza made	15 minutes	1 chef
Dine	15 minutes	
Pay and leave	5 minutes	1 waiter

TABLE II. TAKE AWAY

Activity	Duration	Resources used
Go to counter	0.5 minutes	
Order	2 minutes	1 clerk
Payment	1.5 minutes	1 clerk
Pizza made and exit	10 minutes	1 chef

TABLE III. ONLINE/ PHONE ORDERS

Activity	Duration	Resources used
Order	2 minutes	1 telephone operator
Pizza made	10 minutes	1 chef
Delivery and payment	15 minutes	1 delivery boy
Return to restaurant	13 minutes	1 delivery boy

The restaurant currently employs 5 waiters, 10 chefs, 1 clerk, 1 telephone operator and 5 delivery boys, all of whom are paid $6 per hour. Assuming that the restaurant can spend up to $1,400 daily on wages, find the optimal combination of employees that is required to maximise the number of customers served during the course of a day.

From the problem description, the following BPMN 2.0 diagram may be constructed as in Figure 3.

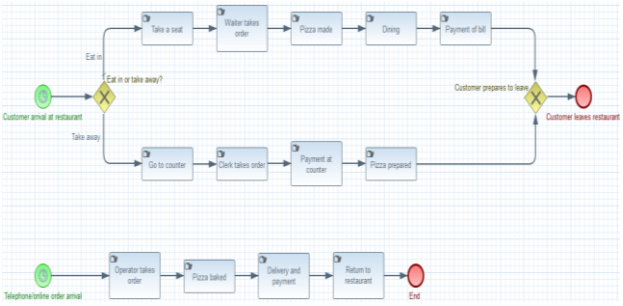


Figure 3. BPMN Diagram of the Pizzeria

Once this model was constructed on SimBusPro and the simulation data corresponding to the problem description were filled in, the model was simulated, yielding simulation reports containing miscellaneous information about the process.

B. Analysis Results

Before alternative models were developed and simulated in the search for the optimal solution, the simulation results of the existing system were analysed. The fundamental information regarding the simulation of the existing system is displayed in Table 4.

TABLE IV. SIMULATION RESULTS OF THE EXISTING SYSTEM

Simulation period	600 minutes
Tokens entering the system	420
Tokens existing the system	376
Total cost	$1,320

Of the 420 customers that entered the system, whether it be in person either to eat in or take away, or via orders placed by telephone or internet, 376 managed to get served within the simulation period, which corresponds to 89.5%. Thus, in the existing system, 89.5% of customers who enter the system get served during the course of a working day, whereas the remaining 10.5% do not. This is due to the fact that they must await the resources (the employees) that have been occupied by prior customers. The expenditure on daily wages can be seen from the table to be $1,320. The objective of this optimisation problem is to ultimately serve the maximum amount of customers without exceeding the daily wage budget of $1,400. In order to achieve this goal, simulation results from the existing system must be analysed.

The resource utilization rates must be analysed to determine which resources are being used efficiently, and which are not. The resource utilisation rates can be seen in Figure 4.

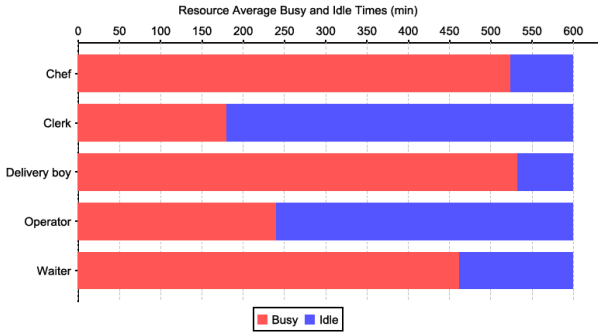


Figure 4. Utilisation Rates of the Pizzeria's Resources

From this graph, it can be seen that the waiters, chefs and delivery boys are extremely busy whereas the clerks and telephone operators are not. This shows that in the prevention of the formation of queues and as per the objective of serving the maximum amount of customers, ideally, more waiters, chefs and delivery boys should be hired. However, the management must take heed not to violate the budget constraint. Thus, the resource for which most customers are waiting for must be determined, and its numbers must be increased.

Figure 5 shows the total time waited for each resource by the customers throughout the simulation time span.

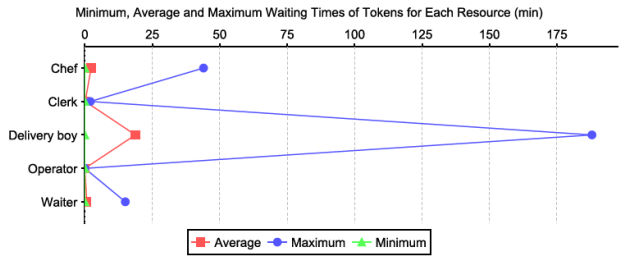


Figure 5. Minimum, Maximum and Average Times Waited by Tokens for each resource

One can see from this graph that it is the waiters, chefs and delivery boys for whom the customers wait the longest periods of time. This reflects the information shown previously in the utilisation rates graph. The total time waited for the delivery boy is particularly high (assuming an average value of approximately 20 minutes and a maximum value of over 180 minutes), thereby once again hinting at the fact that the hiring of new (a) delivery boy(s) ought to be priorities in order to improve the existing system.

In Figure 6, the total number of customers that wait for each resource throughout the simulation duration is displayed.

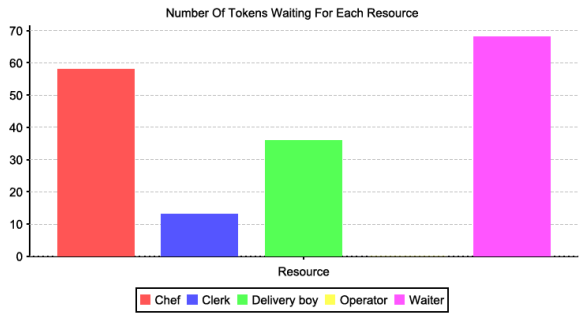


Figure 6. The Number of Customers that wait for each Resource

The graph shows that most customers are waiting for waiters, chefs and delivery boys. There are also some customers who wait for the clerk. However, the number waiting for this resource is far less when compared to the 3 aforementioned resources. This graph also shows that none of the customers wait for the operator, who is always readily available.

To summarise, the resource utilisation graphs show that the chefs, delivery boys and waiters are extremely busy, whereas the clerk and operator are not. The waiting times of tokens for each resource graph shows that tokens do not wait for clerks or operators for prolonged periods of time, whereas queues may sometimes be formed where they sometimes have to wait for the other 3 resources. This graph shows that it is the chefs, delivery boys and waiters for which most tokens have to wait. This information shows that, in the establishment of alternative models, the numbers of the chefs, waiters and delivery boys have to be modified, whereas the numbers of the operators and clerks should be kept constant.

Our analysis shows that in our search for alternative solutions, the number of chefs will either be 9, 10 or 11, whereas the numbers of the waiters and delivery boys will take the values of 4, 5 and 6. Thus, this corresponds to 27 alternative solutions. SimBusPro will generate simulation reports of the different resource combinations. Among these alternative solutions, the one that serves the most customers whilst meeting budget constraints will be chosen as the optimal.

It takes approximately 7 minutes to run each simulation. 27 separate processors were used in order to carry out the simulation of the alternative solutions, corresponding to a total simulation time of 7-8 minutes. Among the 27 alternatives, the optimal solution was identified as the one having served the most customers whilst not breaching the budget constraint. In this new system, the restaurant is to employ 5 waiters, 10 chefs, 1 clerk, 1 operator and 6 delivery boys. In this solution, the only difference from the original system is that 1 more delivery boy has been hired. The basic information corresponding to the simulation of the new system is displayed in Table 5.

TABLE V. SIMULATION RESULTS OF THE OPTIMAL SOLUTION

Simulation period	**600 minutes**
Tokens entering the system	420
Tokens exiting the system	395
Total cost	$1,380

When compared to the previous system, it can be seen that the number of customers served during a typical working day has increased to 395, which corresponds to a 5% increase and that the daily wage budget of $1,400 has not been exceeded.

Figure 7 shows the number of tokens waiting for each resource in the optimal system.

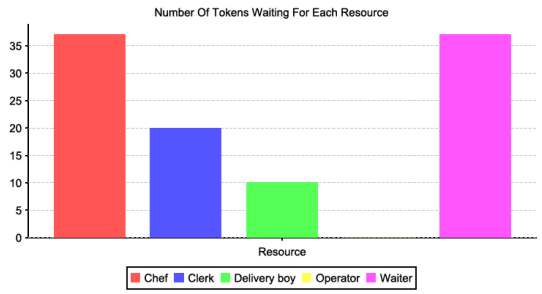


Figure 7. The Number of tokens that wait for each Resource in the Optimal System

When compared to the original system, the numbers of the customers waiting for the chefs, delivery boys and waiters have decreased by 38%, 71% and 47% respectively, whereas the number waiting for the clerk has increased by 66% (due to the fact that he/she is now being kept busier by serving an increased number of customers). The mean queue length in the system has therefore decreased by 41%.

Figure 8 shows the maximum, minimum and average time waited by tokens for each resource in the optimised system.

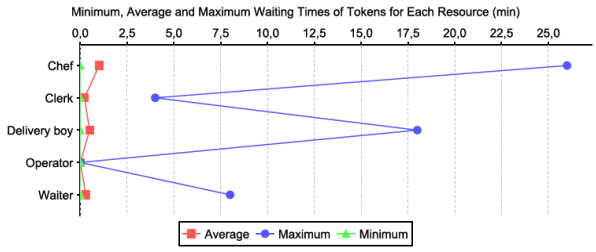


Figure 8. Time Waited by Tokens for each Resource in the Optimal System

When compared to the original data, the average waiting times in the queues of the system can be seen to have decreased by 90%, whereas the decrease in the maximum waiting times is 78%. Thus, in the new system, customer satisfaction can be expected to rise significantly as a result of no longer having to wait as much to be served.

IV. CONLUSION

In this paper, the description of SimBusPro, a simulation-based decision support tool that makes use of BPMN 2.0, has been discussed, and an optimisation problem has been solved. SimBusPro is proficient in the modelling and simulation of a broad spectrum of business processes. Being compatible with Cloud technologies, SimBusPro offers end users a plethora of advantages including the ability to perform rapid simulations. In the future, a new optimiser module that can solve integer programming and linear programming problems is planned to be added to the software. This module will allow for rapid optimisation, and will pose an alternative to the already-present simulation module.

REFERENCES

[1] Aysolmaz, B., Coşkunçay, A., Demirörs, O., & Yıldız, A. (2011). Kamuda iş süreçleri modelleme: Gereği ve yararları. *5. Ulusal Yazılım Mühendisliği Sempozyumu*, 26-28.

[2] http://simbuspro.com:8080/liferay-portal-6.1.0/ (Last visited 1 January 2015)

[3] http://www.sci-bus.eu/ (Last visited 1 January 2015)

[4] Model, B. P. (2011). Notation (BPMN). OMG Specification. Object Management Group.

[5] Chinosi, M., & Trombetta, A. (2012). BPMN: An introduction to the standard. *Computer Standards & Interfaces*, 34(1), 124-134.

[6] Baykasoğlu, A., ve Dereli, T., (2003) Proseslerin bilgisayar ortamlarında modellenmesi, analizi ve seçimi, *Endüstri Mühendisliği Dergisi*, Cilt 14, Sayı 1, 5-17.

[7] http://guse.hu/ (Last visited 1 January 2015)

[8] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. Evolutionary Computation, IEEE Transactions on, 6(2), 182-197.

[9] Baykasoğlu, A. (2013). İş Süreçleri Modelleme/Benzetim Yazılımı Seçimi İçin "Çizge Teorisi" ve "Matris Yöntemi" Temelli Bir Yaklaşım. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 28(3).

[10] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica,I, and Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58.

[11] Kacsuk, P., Terstyanszky, G., Balasko, A., Karoczkai, K., & Farkas, Z. (2013). Executing Multi-workflow simulations on a mixed grid/cloud infrastructure using the SHIWA and SCI-BUS Technology. Cloud Computing and Big Data, 23, 141.

[12] Ould, M. A., & Ould, M. A. (1995). Business Processes: Modelling and analysis for re-engineering and improvement. Chichester: Wiley.

[13] Dokeroglu T., Ozal S., Bayir M.A., Cinar M.S., Coşar A., (2014) Improving the performance of Hadoop Hive by sharing scan and computation tasks, Journal of Cloud Computing: Advances, Systems and Applications.

System for monitoring and advanced analysis of handball matches

T. Hlupić, F. Jandrijević, J. Kovačev, L. Petricioli, T. Gracin, M. Baranović
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
tomislav.hlupic@gmail.com, tomislav.hlupic@fer.hr

This paper is the result of a research project carried out by three teams of bachelor and master programme students. The research analyzes the possibility and usability of connecting different technologies. The project resulted in a system prototype based on a web application with a relational database, mobile application using a CouchDB database and data warehouse for business intelligence purposes. The three parts of the system communicate through a web service based on a node.js environment and MongoDB database. The developed system provides real-time reporting that helps decision making during handball matches and, by using online analytical processing and data mining, detailed post-match data analysis. The advanced handball statistics system offers a complete perspective of match dynamics and cumulative player and team efficiency by game segments in the time domain. This way it provides coaches with highly usable data, allowing the analysis and adaptation of training methods, coupled with efficient game dynamics monitoring and the detection of players' behavioral patterns during matches. Finally, the paper presents the challenges concerning distributed software development in the team with various levels of member's knowledge, teamwork and organization.

I. INTRODUCTION

The Advanced Handball Statistics system was envisioned as a group of several projects that would, as a whole, provide an efficient way of tracking and processing all the actions that occur during handball matches, as well as managing administrative data connected to handball.

In section II, the system architecture is presented. Section III gives the overview of web application for data administration structure and development. Section IV illustrates design and development of the mobile application. Section V presents the data warehouse design with accompanying web application and ETL processes. Section VI describes the web service components and finally, section VII depicts challenges in project management.

II. THE SYSTEM AS A WHOLE

The system consists of three parts: a web application with a transactional database, a mobile application and a web application connected to a data warehouse as shown in Fig 1. The web application with the transactional database is used for data administration, while the mobile application is used for real-time action logging, generating data that is then transferred into the data warehouse. The application connected to the data warehouse is used for calculating the statistics of different partitions (team, player, match) and generating statistical reports.

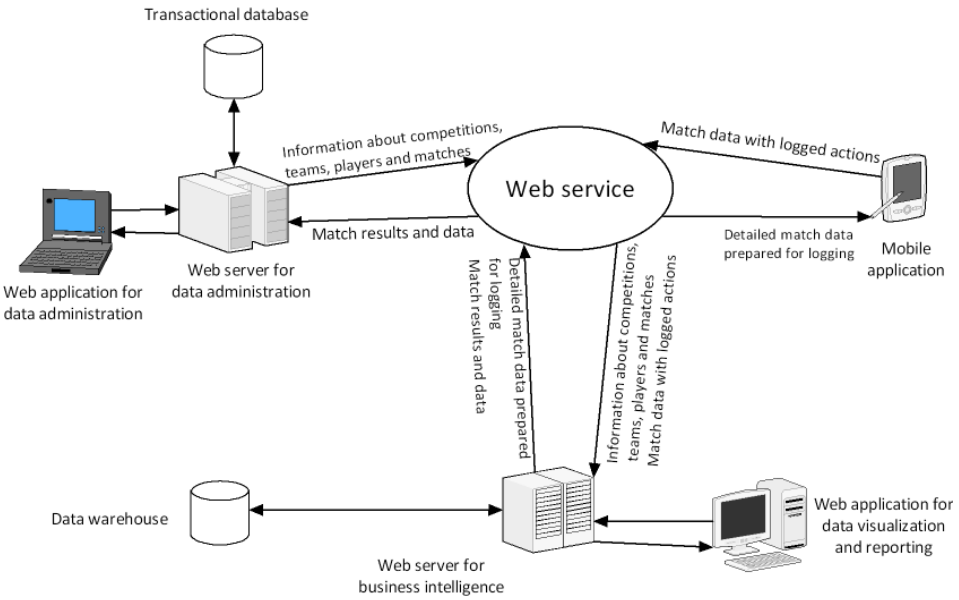


Figure 1. System diagram

None of the applications can function without the other; the one with the transactional database is used to generate a record about a match, containing basic information about it (the teams playing, the season, the competition…), which is then transferred into the data warehouse for further processing. Afterwards, the record is sent to the mobile application in order to provide the data necessary for action logging during the match. The data recorded through the mobile application is afterwards deployed into the data warehouse and it is used for statistical views provided by the application connected directly to the data warehouse. All of the communication between the applications goes through a web service.

III. WEB APPLICATION FOR DATA ADMINISTRATION

The web application for data administration is used to collect and manage administration data related to handball. Therefore, the application was connected to a corresponding transactional database.

The web application allows the user to enter data about handball clubs, competition organization and the information about players and officials. This data is then transmitted to other modules (data warehouse and mobile application). The main parts of the application are the SQL database and MVC web application, described in the following sections.

A. Used tools and technologies

The application was implemented using the object oriented programming language C# along with the development framework ASP.NET MVC [1] and view engine "Razor".

The relational database management system used was Microsoft SQL Server [2]. The database was built using the user interface provided by the program SQL Server Management Studio. For object-relational mapping the open source framework Entity Framework was used.

Also, the presentation language for web page creation called HyperText Markup Language [3] (abbrev. HTML) and the style sheet language for HTML documents - Cascading Style Sheets [3] (abbrev. CSS) were used along with the above listed main technologies.

B. Transactional database

The transactional database was organized through the relational model of data and was fully normalized in order to protect the data and to optimize the database by eliminating redundancy and inconsistency. The focus was also on choosing the appropriate type and size of data attributes within relations. Generally, smaller data types were used where possible. For example, the attributes height and weight from the table "Player" are of type tinyint which uses 1 byte instead of 4 bytes for integer values.

When deciding whether to use a variable or fixed length of character data types, the approximate average, the minimum, and maximum data length were observed.

A fixed length was used for strings whose length does not vary. Variable length was used where the average length was greater than four bytes and where the length of data often varies.

An auto-increment field was created in all tables and was declared to be the primary key. To ensure no duplicate values are entered in specific columns that do not participate in a primary key, unique indexes were created. Relations between database tables were created by specifying corresponding foreign keys.

In case of inheritance (e.g. an official person can be of the supervisor or judge subtype) one table per subclass strategy was used, meaning the common data was stored in a parent table and subclass specific fields were stored in subclass tables.

C. MVC web application

As mentioned earlier, the web application was built using the ASP.NET MVC application framework [1]. MVC is an abbreviated name for the Model-View-Controller architectural pattern which separates applications into three main components: the model, the view, and the controller.

Model objects help implement the logic needed for the applications' data domain. In this project, model objects retrieve and store data in the earlier described relational database. Entity Framework is used for mapping database entities with corresponding model classes. In this case a database first approach strategy was used, i.e. Entity Framework generated data access classes for the existing database. Entity Framework also provides an automated mechanism for accessing and storing data in the database.

Views handle and display the applications' user interface. Almost all the views in this application are created from the model data, i.e. the views are based on them. Razor syntax [4] was used for their creation. For almost each model class there are five views implemented: Index, Create, Edit, Delete, and Details. The Index view displays the main data (the data that should interest most users) from a certain table in the database and the rest of them are used according to their name.

Controllers manage user interaction. Based on user input, they can process and store entered data or retrieve data from the model, select a view to render and provide that view with the data that needs to be displayed. There are over twenty controllers in this application, each for managing a particular model entity and the views based on it.

Validation rules were created upon the database and transferred to model objects. Data types are automatically checked in views and the rest of the validation is done in the controllers.

When using the web application the user can see a sidebar menu containing links to all the relevant entities like "Player", "Club", "Club competition" etc. as shown in Fig. 2.

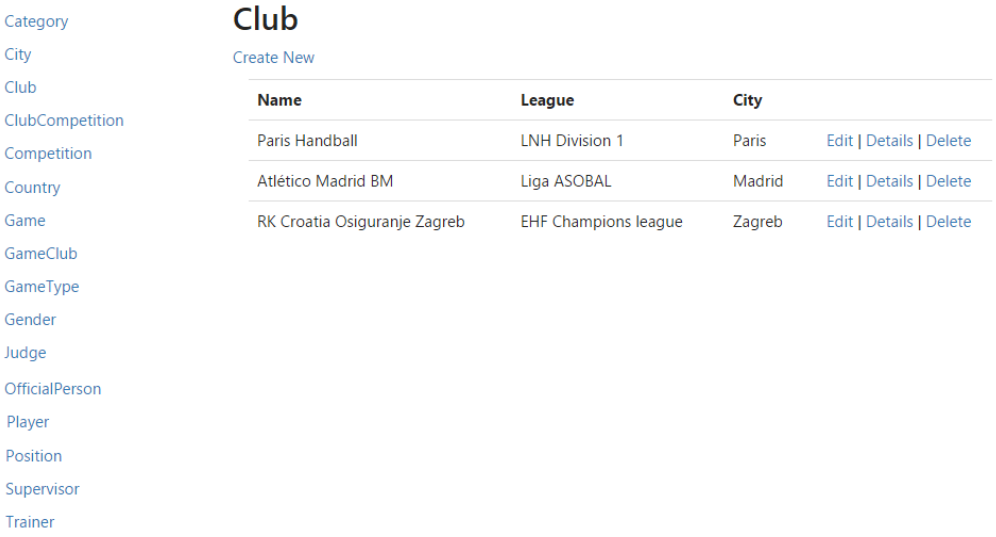


Figure 2. Screenshot of the MVC web application for data administration

Each of the links navigates to the Index view of the corresponding controller. By clicking on the chosen link, the user will be presented with all the data from the database related to that entity. In addition to that, links leading to the Edit, Create, Details and Delete view of a particular entity will also be displayed and placed in an intuitive place. When entering data, the user will be provided with user-friendly input fields like dropdown lists (where appropriate), fields for integer numbers, etc. In case of entering invalid values or not filling the required fields, the user will get corresponding validation messages.

IV. MOBILE APPLICATION

The Android operating system was used as a base for the development of the client side of this application. Through the flexibility and capabilities of Android OS the user is offered a host of options not found in comparable mobile operating systems regarding handball game analysis and statistics. The simplicity and the ease of use of the application are secured through a drag and drop graphical user interface.

The application is run on a tablet and its initial execution request asks for an internet connection. Through a web service, an updated version is received which contains the list of games whose data the user would like to analyze, monitor and process.

Despite the fact that the application communicates with other parts of the system through a web service, after loading the data from the database on the server side and a synchronization with it (if needed), further work is possible without an internet connection.

After the match has been selected, at each initial launch of the application the first step the user has to take is to select and set the player to a certain predefined position in order to help the user react faster to the action which happened on the field.

After choosing the position, a list of players appears, from which the user can select the player playing at that position.

The logging of the events during a game can begin as soon as the user clicks on the Start button, which in turn starts the timer that counts down time. The basic operation of the application is based on two cases, the first case being when the selected team attacks the goal and the second when the relevant team defends or focuses on the defense of their own goal. Depending on which selected functionality is in progress, the actions based on the user's activities in the application are generated. An example of the most often used action is a shot to the goal by a player at a specific time of the match, additionally described by shot side and height, position of the player from which he took the shot and in the case the player scored the goal from the shot, a flag marking the positive actions is set.

Every action or event is closely connected to a particular player, and all the changes are stored in a local database.

The right choice of an adequate database was essential in the design of the system. Since the data collected through the application is structured and distributed, a non-relational database called Couchbase [4] was chosen, which allows easy data loading for web services.

Data is stored in a local database on the device in JSON format, which can later be synchronized with the main database on the server. Considering that the data is presented in a very detailed way, it tends to be of very low granularity. The expandability of the system is secured through the class with defined attributes which is generated by the user in order to store data and describe the events.

V. WEB APPLICATION FOR GENERATING REPORTS

The statistical analysis of match actions is based on business intelligence tools. The basis for it is the data

warehouse, built on Kimball's bottom-up approach with a star scheme [5], part of it shown on Fig. 3. In the initial stage of system development, the data warehouse was used for generating simple reports from available data, leaving the selection of query conditions to the user's previously entered data.

The idea behind handball statistics powered by business intelligence is to enable simple access to data during and after the match through an interface which allows a fast and precise choice of data. The choices of reports differ by their depth and data scope, varying from a single match report to player and team statistics.

The match report is a list of match events (actions), entered through the mobile application and modeled through ETL (Extract-Transform-Load) [5] processes for storing in the data warehouse without any further processing. The other two reports (player statistics and team statistics) process the stored match action data to calculate the players' efficiency by game segments. Player statistics was designed to show all the data related to a specific player, whereas team statistics used only aggregated player data.

As a part of the distributed system, this application uses data from both the mobile application and the data administration application, thus completing the communication in the system. The details about matches, players, teams and competitions are gathered from the web service, which serves as an interface, and stored before they are ready to be used in the mobile application. Such a procedure not only ensures the integrity of data throughout the whole system, preventing double data entering, but also allows the restriction of data access if privacy is needed.

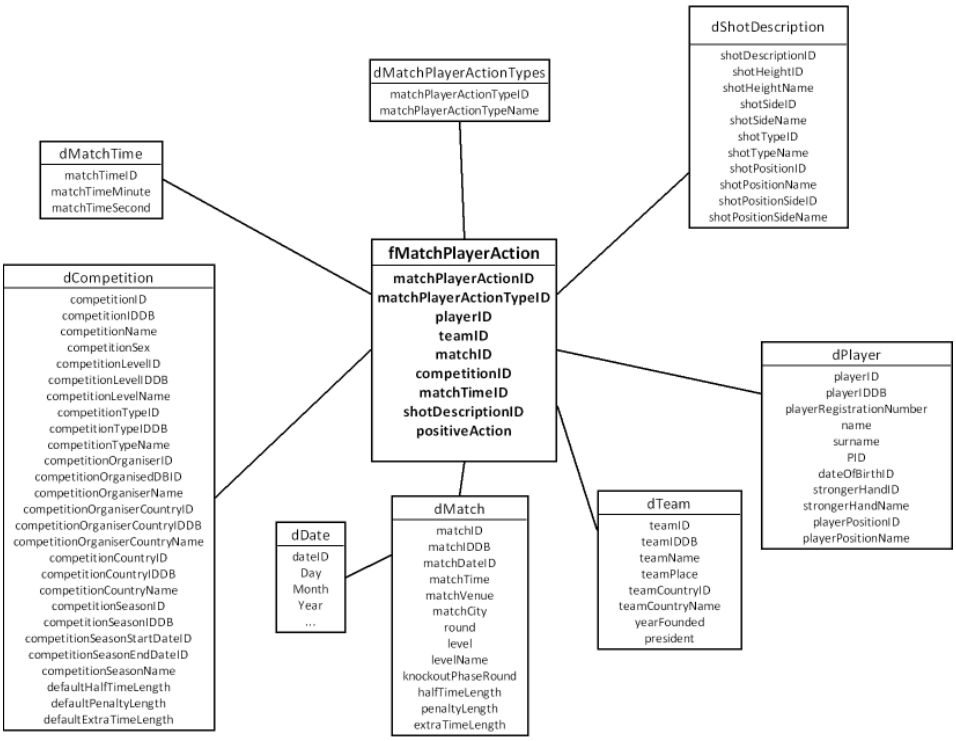


Figure 3. Diagram of the fact table fMatchPlayerAction with relevant dimension tables, in which the data with match actions is stored

Data processing was initially implemented in the web application with the processing functions developed solely as functions in the code. In the second stage of the development, detailed data analysis is enabled by introducing OLAP (Online Analytical Processing) [4], allowing slicing, dicing and drill-down operations to be executed over the data, making the initial code-based data processing implementation obsolete. OLAP, in terms of the match data analysis, give more flexibility by allowing users to focus on one or more dimensions of the game (e. g. the directions of player's shot through one or more matches and following the shots by game segments, how many goals team or player scores after the passive attack signal or who takes most responsibility for shots after the passive attack signal), thus giving more precise information suitable for adapting the game tactics.

The web application that provides the user with data access was developed in ASP.NET MVC, connected to the data warehouse through an interface on the middleware's business logic layer, which provides access to data through predefined methods, accessible in the controller part of the web application.

As the system consists of the three modules, ETL processes [5] are needed in order to store consistent data into the data warehouse. They are executed as the new data appears on a web service interface, where the extraction matches the process of collecting the data from the web service, where it's stored in JSON format.

In the transformation process, a function takes the data from JSON structure and adjusts it to suitable classes in the application's core layer, without the need for integrity checking which is ensured as previously described.

Data collected from the web application for data administration loads into all dimension tables and fact tables which follow the processes of team personnel changes (player transfers, coach assignments etc.) and assigning the match official personnel (referees, supervisors etc.). As for the data originating from the mobile application, it is loaded only into fact tables related to processes of match actions (as described before) and team match events, such as choosing team officials, following timeouts during the match and progressive penalties given to the team officials.

VI. WEB SERVICE

The web service is the main method of communication between the mobile application and the database. It is used to provide the option of update to the mobile application or synchronize data with it, and provides a way for the mobile application to enter gathered data into the database.

A. Used tools and technologies

The web service is established as a combination of a NodeJS[8] platform and a MongoDB[9] database implemented on MongoLab[10]. The RESTful[11] style was used and implemented via a node-restful library[12]. The data is sent as a JSON[13] encoded object which provides great flexibility of data type transferred, and is the favored way of transferring data while providing cross-domain flexibility by JSONP if necessary. The web service is located on the Azure server[14].

B. NodeJS and RESTful

NodeJS is a platform built on Chrome's JavaScript runtime[13] for building fast, scalable network applications easily. Its event-driven, non-blocking I/O model makes it lightweight, efficient and highly responsive. It does not matter how "complicated" the request is, the response will be instant, with a possible follow-up data stream.

RESTful is a software architecture style consisting of guidelines for creating scalable web services. Its main properties are performance, scalability, simplicity of interfaces, modifiability of components (even when the application is running), visibility of communication between components, portability and reliability. The communication with the client should be stateless – no client context should be stored on the server between requests. The communication is established via the HTTP protocol which is an application layer protocol based on the TCP transport protocol.

C. MongoDB on MongoLab

MongoDB is an open-source document database implementation of the NoSQL[5] database model made in a combination of C++, C# and Go. Its implementation of JSON-style documents is a perfect fit for JavaScript and the communication with the mobile application. It's hosted on MongoLab that manages the MongoDB database as a service.

A NoSQL database provides a mechanism for the storage and retrieval of data that is modeled in means other than the tabular relations used in traditional relational databases. That results in a simpler design, horizontal scaling and finer control of availability. Data structures in NoSQL databases (key-value, graph, document to name a few) differ from the ones used in relational databases, making some operations faster in NoSQL, and making some of them slower. Depending on the particular problem to be solved, it's crucial to select a suitable data structure. Data structures chosen in web service correspond with data structures in the system's application, making the transforming process easier without big refactoring of the previous code.

VII. PROJECT MANAGEMENT DURING DEVELOPING

The development of the advanced handball statistics project was a challenge on several levels. First of all, almost all of the members had their first contact with developing a sport statistics application, without any prior knowledge of the sport itself. By that, the students got in touch with typical development outside of the academic world, where the developed software usually is not in a domain which their knowledge covers.

Other important segments of the project were teamwork and workload distribution among the teams. Since the system was divided into three applications that could have been developed independently, only with prior given specifications, not only did the team leaders have freedom to focus just on their team, but it also allowed the teams to give more attention to the development itself, rather than data modeling for interoperability purposes.

Of the twelve students involved in the project, three are master program students and the rest are undergraduate (bachelor) program students. Each master program student became a team leader, while the undergraduate students had the opportunity to choose the part of the project which they found most interesting, thus creating four-member teams.

One of the project goals was to test whether the students that never came across technologies used in system development, but have basic programming skills, could adapt to the new development environment under the mentorship of more experienced students. The process started with gathering online materials, mostly tutorials, after which the team leaders helped the students utilize their newly gained knowledge to solve specific tasks, starting from smaller ones. By the end of the project, the undergraduate students were able to cope with more demanding tasks, which consequently led to a bigger mentors' involvement in debugging. The outcome

was satisfying, resulting in the undergraduate students being able to develop their own applications in the future.

Regarding project management, several tools were used to track the project's status and completeness of tasks. From the very beginning, the main communication method was a mailing list where all team members got the general information on the project's status and future tasks, complemented with Dropbox [16] as a main document repository, through which the members got access to system specifications, diagrams and tutorials. As for task management, Trello [17], an online application, provided enough possibilities that covered all the needs, where different project areas got their independent cards, with the possibility of restraining access to a specific card only to the team to which it belonged. It also allowed the tasks to get their deadlines assigned, according to a previously established schedule, and milestones to be set, which, in combination with the later mentioned GitHub [18], helped decision making regarding the reallocation of development resources.

Another important aspect of the development process was the usage of a distributed revision control system. Owing to the latest academic initiative, the choice got down to GitHub. From the very beginning, all team members were encouraged to track their development through their own branches, while checking for updates on the project's master branch. In order to prevent possible crashes of the master branch, a workflow of development was established, where only team leaders pushed code they previously approved through testing to the master branch, while other team members pushed their code only to their own branches, which later got merged to the leader's branch. This multiple-level hierarchy of code approving proved to be successful, since the team leaders had the opportunity to fix issues in the code or do small changes that did not require major refactoring. Also, a subsequent outcome was the development of a habit of tracking even the smallest changes during the development process, which gives opportunities not only to rollback possible errors, but also to track the status and progress of the development and, if possible, relocate development resources to project parts which require more attention or are left behind schedule.

VIII. Conclusion

The project encompassed the development of a system that provides an efficient way of tracking all handball matches, teams, and players, as well as all the player actions during a match. The system generates relevant statistics on demand.

The bachelor students learned how web applications and distributed systems are created and what the process of software development looks like. They developed their programming skills and discovered what it is like to work in teams.

The master program students honed their leadership and programming skills. They integrated several applications into one system via a web service, thus perfecting skills needed when dealing with distributed systems.

This project proved to be a challenge due to the vastness of the work involved and the number of participants, but it has also been an invaluable experience.

Further research in the project will lead to a self-learning system based on machine learning techniques which would detect and show various anomalies, usable in handball team optimization.

Acknowledgment

The authors would like to thank Microsoft Croatia for their help in acquiring access with academic benefits to the Azure platform, which ensured the system's complete functionality. Also, they would like to thank Marina Brebrić, Leon Fadljević, Ivan Golik, Velimir Lisec, Martina Martinović, Ana Matić, and Ana Šarčević for all the hard work they put into realizing this project.

References

[1] ASP.NET MVC Overview. [Online]. https://msdn.microsoft.com/en-us/library/dd381412(v=vs.108).aspx . Date of access: January 28, 2015.

[2] Microsoft SQL server Overview[Online]. http://www.microsoft.com/en-us/server-cloud/products/sql-server/. Date of access: April 2, 2015.

[3] HTML & CSS [Online]. http://www.w3.org/standards/webdesign/htmlcss. Date of access: April 2, 2015.

[4] ASP.NET Razor – Markup. [Online]. http://www.w3schools.com/aspnet/razor_intro.asp. Date of access: January 29, 2015.

[5] NoSQL. [Online]. http://en.wikipedia.org/wiki/NoSQL. Date of access: February 1, 2015.

[6] R. Kimball, M. Ross, The Data Warehouse Toolkit. Wiley Computer Publishing, 2002.

[7] What is OLAP? [Online]. http://olap.com/olap-definition/. Date of access: January 28, 2015.

[8] Node.js. [Online]. http://nodejs.org/. Date of access: February 10, 2015.

[9] Agile and Scalable. [Online]. http://www.mongodb.org/. Date of access: February 9, 2015.

[10] Fully managed MongoDB-as-a-Service. [Online]. https://mongolab.com/. Date of access: February 10, 2015.

[11] Representational state transfer. [Online]. http://en.wikipedia.org/wiki/Representational_state_transfer. Date of access: January 30, 2015.

[12] node-restful. [Online]. https://github.com/baugarten/node-restful. Date of access: January 30, 2015.

[13] Introducing JSON. [Online]. http://json.org/. Date of access: January 30, 2015.

[14] Microsoft Azure. [Online]. http://azure.microsoft.com/. Date of access: February 5, 2015.

[15] V8 JavaScript Engine. [Online]. https://code.google.com/p/v8/. Date of access: February 5, 2015.

[16] Dropbox. [Online]. www.dropbox.com. Date of access: February 7, 2015.

[17] Trello. [Online]. www.trello.com. Date of access: February 7, 2015.

[18] Github. [Online]. www.github.com. Date of access: February 7, 2015.

Planiranje i nabava energije za pokriće gubitaka u distribucijskoj mreži

Igor Žarkić*, Zdravko Lipošćak* i Matjaž Malovrh**
*HEP-Operator distribucijskog sustava d.o.o., Zagreb, Hrvatska
** Landis+Gyr, Šenčur, Slovenija
igor.zarkic@hep.hr, zdravko.liposcak@hep.hr, matjaz.malovrh@landisgyr.com

Gubici električne energije u sustavu distribucije električne energije jednaki su razlici energije koja je ušla u distribucijsku mrežu i energije predane kupcima. HEP–Operator distribucijskog sustava d.o.o. dužan je osigurati električnu energiju za pokriće gubitaka u distribucijskoj mreži, nabavom na otvorenom tržištu električne energije.

Analizom dosad primjenjivanog postupka utvrđena je potreba izrade složenijih algoritama predviđanja potrošnje te osiguranja kvalitetnije informatičke potpore, primjerene uvjetima poslovanja na tržištu električne energije. Provedeno je nekoliko studija načina potrošnje energije krajnjih korisnika distribucijske mreže, utjecaj distribuiranih izvora električne energije, predviđanja ukupne satne potrošnje s ciljem izgradnje sustava što točnijeg predviđanje iznosa gubitaka električne energije u distribucijskoj mreži. Ovim radom dan je pregled izgradnje sustava predviđanja iznosa energije gubitaka u distribucijskoj mreži, izrada plana nabave potrebne energije na tržištu te izračun ostvarenja energije gubitaka.

I. UVOD

HEP–Operator distribucijskog sustava d.o.o. dužan je osigurati električnu energiju za pokriće gubitaka u distribucijskoj mreži. Gubici električne energije u distribucijskom sustavu jednaki su razlici energije koja je ušla u distribucijsku mrežu i energije predane kupcima. Pritom se za ulaz u distribucijsku mrežu uzimaju u obzir izmjerene vrijednosti energije na ulazu u distribucijski sustav (iz prijenosne mreže, susjednih distribucijskih sustava i elektrana priključenih na distribucijsku mrežu) i izmjerene vrijednosti potrošnje energije na mjestu priključka kupca. Tako definirani gubici električne energije dijele se na dvije grupe: tehnički gubici i netehnički gubici.

Tehnički gubici (gubici u užem smislu) nastaju isključivo u elementima mreže svih naponskih razina (vodovima, transformatorima, potrošnja komponenti sustava), a netehnički gubici nastaju zbog pogrešaka uzrokovanih složenim poslovnim procesom očitanja i obračuna potrošnje električne energije te neovlaštenim korištenjem električne energije (krađa energije). Pitanje točnog izračuna razine tehničkih gubitaka složeno je zbog velikog broja elemenata mreže i vremenske promjenjivosti opterećenja koja ovisi o velikom broju potrošača energije i njihovim trenutnim potrebama te problema nemogućnosti mjerenja velikog broja izlaza energije u isto vrijeme. Parametri elemenata mreže isto su tako ovisni o radnim uvjetima i promjenjivi su s vremenom, ali se utjecaj ovih promjena zanemaruje u kratkoročnim (satnim, dnevnim, mjesečnim) analizama potrošnje i gubitaka energije u distribucijskoj mreži.

Rad ukratko opisuje osnovne korake i rezultate složenog, višegodišnjeg procesa definiranja modela planiranja ukupnog opterećenja i gubitaka energije distribucijskog sustava. Autori rada sudjelovali su u kreiranju i povezivanju pojedinih faza projekta od definiranja istraživačkog uzorka, izgradnje sustava daljinskog očitavanja brojila te u konačnici u definiranju modela i pravila izračuna gubitaka energije iz predviđenog opterećenja sustava. Tijekom projekta provedeno je nekoliko studija s ciljem analize načina potrošnje energije koje su ukratko opisane u radovima [2] i [3]. Iz ovih studija kreirana su pravila primjene nadomjesnih krivulja opterećenja [4] za potrebe rada tržišta električne energije koja omogućuju i izradu plana satne potrošnje kupaca električne energije. U konačnici provedena je i studija za definiranje matematičkog modela kratkoročnog predviđanja satne potrošnje električne energije [5].

II. ANALIZA NAČINA POTROŠNJE KUPACA ELEKTRIČNE ENERGIJE

Ukoliko je moguće predvidjeti potrošnju energije korisnika distribucijske mreže u određenim vremenskim periodima, bilo bi moguće i na velikom skupu elemenata mreže s dovoljnom točnošću predvidjeti iznose tehničkih gubitaka. Mjerni intervali u kojima se najčešće mjeri potrošnja kupaca je 15 ili 60 minuta, a nabava energije na tržištu je u 60 minutnim intervalima. Da bi se točno predvidjela energija koju je potrebno nabaviti za kupce ili za gubitke na tržištu potrebno je imati povijesne mjerne podatke u odgovarajućim intervalima. Budući se radi o velikom broju mjernih mjesta troškovi opremanja brojilima koja imaju mogućnost mjerenja potrošnje u intervalima od 60 minuta su značajni, a za provedbu cijelog projekta potrebno je i značajno vrijeme, koje je u pojedinim zemljama procijenjeno na 5 do 15 godina. Trenutno se potrošnja energije u 15-minutnim intervalima mjeri na 19.500 obračunskih mjernih mjesta na srednjem naponu i niskom naponu priključne snage iznad 30 kW. Na ovim mjernim mjestima registrira se oko 40% ukupne potrošnje svih 2,3 milijuna kupaca [1].

Dosadašnja slika pasivne distribucijske mreže postaje složenija priključivanjem sve većeg broja distribuiranih obnovljivih izvora energije. Broj ovih malih elektrana koji je danas oko 1.500, nema značajniji utjecaj u odnosu na potrošnju energije 2,3 milijuna korisnika distribucijske

mreže. Kako bi se ograničila količina podataka potrebna za planiranje ukupne potrošnje kupaca na osnovu koje se u drugom koraku izračunavaju gubici energije u distribucijskoj mreži, bilo je potrebno grupirati kupce prema sličnosti u načinu potrošnje energije. Za potrebe definiranja karakterističnih skupina kupaca kreiran je istraživački uzorak od 1.500 obračunskih mjernih mjesta koja su opremljena brojilima s mjerenjem potrošnje u 15-minutnim intervalima.

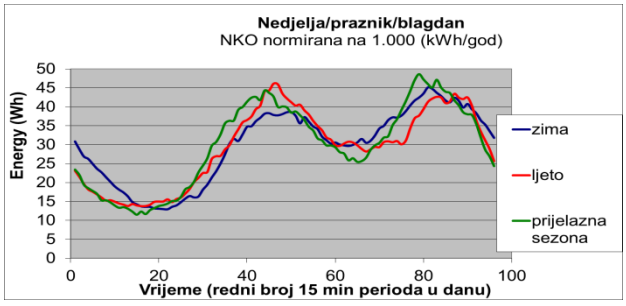
Grupe se oblikuju tako da je sličnost objekata unutar grupe veća od sličnosti između grupa. Kao mjera sličnosti udaljenosti dvaju objekata (obračunskih mjernih mjesta, OMM) koristila se Euklidova udaljenost [2].

Neka su $P_i(t)$ i $P_k(t)$ cjelogodišnje krivulje opterećenja dvaju objekata (OMM) „i“ i „k“. „t“ je brojač po svim 15 min razdobljima u promatranoj godini. Euklidova udaljenost između dviju krivulja opterećenja definirana je sljedećim izrazom (1):

$$d(i,k) = \sqrt{\sum_{t=1}^{35040} (P_i(t) - P_k(t))^2} \quad (1)$$

pri čemu je 35.040=96*365 broj 15 minutnih intervala u jednoj kalendarskoj godini (ukoliko je godina prestupna 35.136=96*366), a 96=24*4 je broj 15 minutnih intervala u jednom danu. Prepoznate su sljedeće karakterističnih skupina kupaca: K0–Kućanstva, P1–Poduzetništvo priključne snage manje ili jednake 13 kW, P2–Poduzetništvo koje ne pripada u P1, JR0–Poduzetništvo–javna rasvjeta.

Promatrajući tijek opterećenja kroz jednu godinu, svakoj karakterističnoj skupini kupaca pripada 9 dnevnih nadomjesnih krivulja opterećenja (NKO) koje su osnova za utvrđivanje godišnje krivulje opterećenja OMM: zima – radni dan, zima – subota, zima – nedjelja (slika 1), ljeto – radni dan, ljeto – subota, ljeto – nedjelja, prijelazno razdoblje – radni dan, prijelazno razdoblje – subota, prijelazno razdoblje – nedjelja.



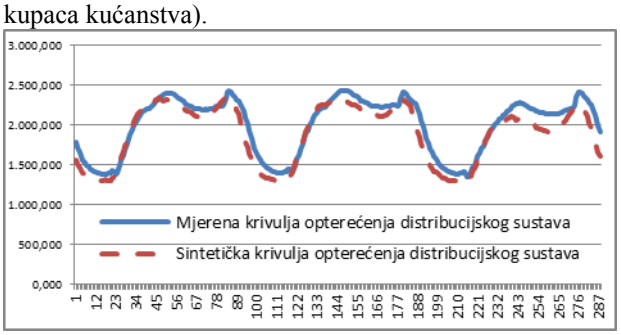
Slika 1. Nadomjesna krivulja karakteristične skupine kupaca – K0 Kućanstvo normirana na 1.000 kWh/godišnje

Analizom obilježja vremenske slijednosti opterećenja kupca kategorije kućanstva, zaključeno je kako se potonji, za razliku od ostalih karakterističnih skupina kupaca, ne mogu vjerodostojno opisati samo uporabom 9 dnevnih NKO. U tu svrhu istraživane su vremenske slijednosti dnevnih potrošnji OMM kako bi se odredila takozvana funkcija (polinom) dinamiziranja (2):

$$F_t = -2{,}454 * 10^{-10} * t^4 + 2{,}586 * 10^{-7} * t^3 - 5{,}135 * 10^{-5} * t^2 - 3{,}261 * 10^{-3} * t + 1{,}602 \quad (2)$$

gdje je **t** redni broj dana u kalendarskoj godini (t=1 za 01. siječnja) [3].

Izravnim zbrajanjem nadomjesnih krivulja opterećenja karakterističnih skupina kupaca pomnoženih s prosječnom godišnjom potrošnjom dobiva se sintetička krivulja ukupne potrošnje kupaca koji nemaju mjerenje 15-minutnih potrošnji (intervalno mjerenje). Dobivenoj krivulji dodaje se zbirna krivulja kupaca s ugrađenim intervalnim mjerenjem te dobivamo sintetičku krivulju ukupne potrošnje svih kupaca (slikom 2 prikazana je usporedba ove krivulje u odnosu na mjerenu krivulju opterećenja). Razlika dobivene krivulje potrošnje kupaca i izmjerene krivulje opterećenja cijelog sustava predstavlja krivulju gubitaka energije. Ukoliko bi se za nabavu energije koristila krivulja gubitaka dobivena na opisani način rezultat bi u sebi sadržavao netočnosti predviđanja potrošnje energije na nivou karakterističnih skupina kupaca i netočnosti određivanja nadomjesnih krivulja opterećenja (jedna krivulja K0 za 2 milijuna kupaca kućanstva).

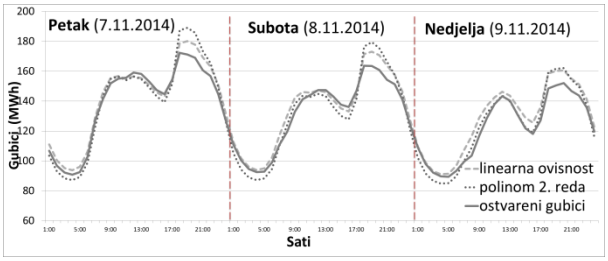


Slika 2. Usporedba mjerene i sintetičke krivulje opterećenja distribucijskog sustava (MW) za tri dana (96*3=288 15-minutnih intervala) - četvrtak, petak, subota

III. IZRADA KRIVULJE PLANIRANIH SATNIH GUBITAKA

Prema postojećim Pravilima primjene nadomjesnih krivulja opterećenja [4], gubici se računaju u linearnoj ovisnosti o opterećenju distribucijskog sustava. Dakle, za predviđanje satne krivulje gubitaka potrebno je predviđati satnu krivulju opterećenja distribucijskog sustava. Za potrebe analize gubici su se računali i u kvadratnoj ovisnosti o opterećenju distribucijskog sustava korištenjem polinoma drugog reda. Koeficijenti polinoma drugog reda dobiveni su analizom tehničkih karakteristika cjelokupne distribucijske mreže, analizom rezultata kontrole neovlaštene potrošnje električne energije i usporedbom s višegodišnjim rezultatima izračuna gubitaka.

Slika 3 prikazuje sličnost krivulja planiranih satnih gubitaka za tri karakteristična dana (radni dan, subotu i nedjelju) koje su dobivene upotrebom dva modela ovisnosti gubitaka o opterećenju (model linearne ovisnosti i model ovisnosti preko polinoma drugog reda), te stvarno ostvarenih gubitaka.



Slika 3. Ovisnost korištenog modela za planiranje gubitaka o opterećenju distribucijskog sustava uz prikaz ostvarenih gubitaka

Kao posljedica liberalizacije tržišta električne energije, predviđanje potrošnje električne energije dobiva sve veći značaj u sektoru distribucije električne energije. Točno predviđanje potrošnje električne energije omogućava značajne uštede za elektroenergetski sektor ako se vođenje elektroenergetskog sektora zasniva na (točnom) predviđanju potrošnje.

Osnovni je problem kod predviđanja potrošnje električne energije njena stohastička i nestacionarna priroda, kao i činjenica da je uvjetovana velikim brojem čimbenika poput: ekonomskih čimbenika, vremena, dana, godišnjeg doba, te hidrometeoroloških uvjeta. Postoji više različitih pristupa problemu predviđanja potrošnje električne energije, a eksperimentalni rezultati koji su dobiveni na modelima neuronskih mreža su se pokazali zadovoljavajućima i primjenjivima [5]. Privlačnost neuronskih mreža leži u pretpostavci da mogu naučiti model sustava na temelju povijesnih podataka pa je nužno poznavati potrošnju i druge čimbenike koji utječu na opterećenje do nekoliko godina unazad.

Analizom dosad primjenjivanog postupka koji se temelji na procjeni potrošnje kupaca korištenjem nadomjesnih krivulja opterećenja, utvrđena je potreba izrade složenijih algoritama predviđanja potrošnje te osiguranja kvalitetnije informatičke potpore, primjerene uvjetima poslovanja na tržištu električne energije. Cilj je bio izraditi aplikaciju za točnije predviđanje satne potrošnje i gubitaka električne energije u distribucijskoj mreži kako bi se smanjili mogući troškovi odstupanja planiranih od stvarno potrebnih količina energije. Predviđanja ukupne potrošnje i gubitaka električne energije obavljala bi se na razini pojedinog distribucijskog područja i na razini cijelog distribucijskog sustava.

IV. APLIKACIJA ZA PREDVIĐANJE SATNE POTROŠNJE ELEKTRIČNE ENERGIJE I PRORAČUN GUBITAKA

Programsko sučelje izvodi se u programskom okruženju Matlab tvrtke MathWorks sa sljedećim modulima:

1. Matlab – osnovna programska podrška.
2. NeuralNetworkToolbox – modul za modeliranje, simulaciju i učenje neuronskih modela sustava.

A. Opis osnovnih funkcionalnosti modula

Osnovne funkcionalnosti modula su:

- predviđanje potrošnje električne energije za sljedeći dan i tjedan, za pojedina distribucijska područja i za cjelokupni distribucijski sustav, na satnoj razini, na temelju povijesnih podataka, kalendarskih podataka i na temelju meteoroloških prognoza,
- predviđanje gubitaka električne energije po distribucijskom području i za cjelokupni distribucijski sustav na temelju predviđene potrošnje, gdje HEP ODS osigurava matematičku formulu za računanje koja povezuje potrošnju električne energije i gubitke.

Potrebne spomenute povijesne, meteorološke i kalendarske podatke modulu se osiguravaju iz baze podataka u informatičkom sustavu HEP ODS-a.

Na zahtjev korisnika, modul za zahtijevano distribucijsko područje može prilagoditi postupak predikcije koristeći odabrane arhivirane podatke (adaptacija parametara predikcijskog modela), tj. njihovo odabrano povijesno razdoblje. Kvaliteta predviđanja modela može se kvantificirati na povijesnim podacima prema dogovorenim kriterijima. Korisnik ima mogućnost pohranjivanja modela u svakom trenutku, te je pritom predviđena mogućnost paralelnog rada više predikcijskih modela i promjena aktivnog modela u bilo kojem trenutku. Predikcija se obavlja centralno na jednom računalu za sva distribucijska područja.

B. Opis kompletnog ulaznog prostora za svaki sat u danu

Za rad modula, čiji je mogući model kratkoročnog predviđanja satne potrošnje električne energije prikazan u sljedećem poglavlju (vidi sliku 4), potrebno je osiguravati sljedeće ulazne podatke:

1. temperatura zraka $T(i)$ za $i = 0\ldots-14$ (Pritom oznaka $T(0)$ označava temperaturu zraka u danu za koji se provodi predikcija opterećenja, $T(-1)$ označava temperaturu zraka u danu ranije, $T(-2)$ dva dana ranije itd.).
2. naoblaka $N(i)$ za $i = 0\ldots-14$.
3. potrošnja električne energije $P(i)$, za $i = -2\ldots-14$ (potrošnja $P(0)$ se ne koristit u predikciji jer je to upravo vrijednost čija se predikcija traži, dok se pretpostavlja da podatak o potrošnji od dana ranije $P(-1)$ neće biti dostupan u trenutku kad će se provoditi predikcija.
4. vremenski indeks τ_2;
5. tip dana τ_3;
6. sat za koji se radi predikcija τ_4.

Vremenski indeks τ_2 označava koje je doba godine i računa se preko formule (3):

$$\tau_2 = \sin(2\pi t/365) \quad (3)$$

gdje je t redni broj dana u godini. Pošto je sporo promjenjiv kroz godinu, njegova vrijednost uzima se samo za dan za koji se radi predikcija. Tip dana τ_3 označava je li riječ o radnom danu, suboti, nedjelji ili blagdanu.

Predikcijski model predviđa ukupnu potrošnju za odgovarajuće distribucijsko područje te će se ispadi velikih potrošača iz mreže, kao i prilagodba pripadajuće predikcije, uzimati u obzir izvan aplikacije.

C. *Vremenski horizonti*

Prema Pravilima djelovanja tržišta električne energije (NN 90/12, VII. Planiranje rada tržišta za dan unaprijed) HEP ODS obvezan je:

- najkasnije do 14:00 sati u danu trgovanja dostaviti Operatoru tržišta svoje ugovorne rasporede za dan isporuke.
- Operator tržišta do 14:30 sati u danu trgovanja izrađuje tržišni plan te u slučaju potrebnih ispravaka ugovornog rasporeda traži ispravak od tržišnog sudionika na kojeg se ispravak odnosi.
- do 14:45 sati u danu trgovanja tržišni sudionik dostavlja Operatoru tržišta ispravljeni ugovorni raspored.

Prema Pravilima djelovanja tržišta električne energije (NN 90/12, VIII. Izmjene ugovornog rasporeda tijekom dana isporuke):

tržišni sudionik može tražiti od Operatora prijenosnog sustava izmjenu ugovornog rasporeda tijekom dana isporuke najkasnije dva sata prije početka razdoblja na koje se izmjena odnosi [6].

Modul osigurava predviđanje potrošnje na satnoj razini za svako distribucijsko područje počevši od ponoći nakon dana trgovanja za sljedeća 24 sata i sljedećih 7 dana. HEP ODS će za svako distribucijsko područje iz niza modela koje će biti moguće naučiti za odabrana razdoblja povijesnih podataka odabrati jedan predikcijski model koji će time postati aktivnim. Odabir aktivnog modela omogućuje se putem korisničkog sučelja u Matlabu.

D. *Detaljne specifikacije korisničkog sučelja programskog modula*

Korisničko sučelje sadržava sljedeće funkcije:

- pokretanje i grafički prikaz predviđanja potrošnje i gubitaka za aktivni model
- pokretanje i grafički prikaz predviđanja potrošnje i gubitaka za po volji odabrani model
- spremanje predviđanja potrošnje i gubitaka u Excel (*xls*/*xlsx* format) datoteku
- kvantifikacija kriterija za ocjenjivanje predviđanja modela za odabrano razdoblje
- pokretanje prilagodbe modela za odabrano povijesno razdoblje
- odabir aktivnog modela za određeno distribucijsko područje

Sustav omogućuje memoriranje i kasniji dohvat ulaznih podataka, parametara predikcijskog modela, dobivenih rezultata te izlaznih datoteka za period od najmanje zadnja 24 mjeseca. Korištenjem povijesnih podataka moguće je ponovo generirati predviđanja za određeni dan ili tjedan u zadnjih godinu dana.

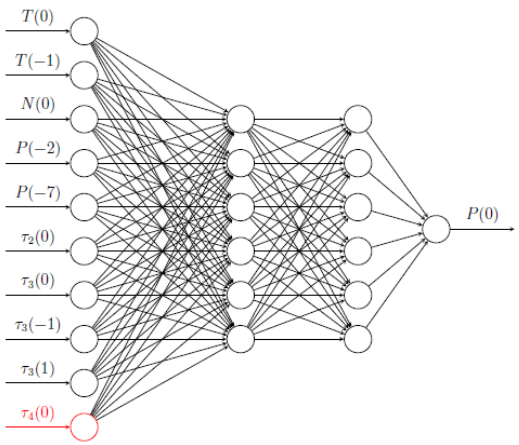
E. *Razmjena podataka*

Komunikacija modula s drugim programskim sustavima HEP ODS-a (odgovornim za podatke o potrošnji i meteorološkoj prognozi) provodi se korištenjem Excel datoteka (*xls*/*xlsx* format) formatiranih prema unaprijed utvrđenim pravilima. Pritom aplikacija provodi kontrolu potpunosti ulaznih podataka te pretpostavlja ispravnost ulaznih podataka.

V. NEURONSKA MREŽA

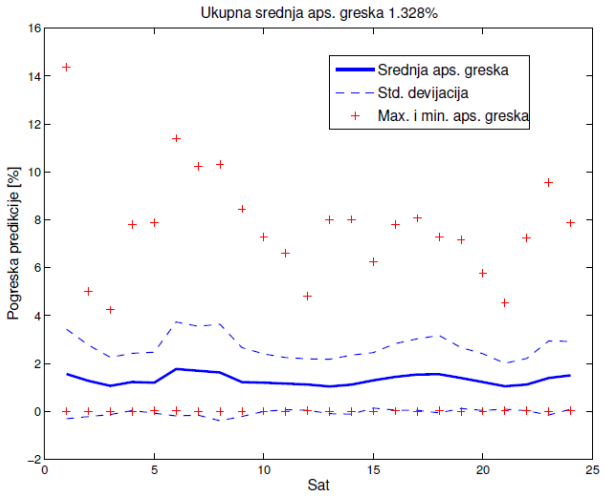
A. *Prijedlog modela predviđanja potrošnje električne energije*

Modeli predviđanja satne potrošnje za pojedina distribucijska područja dobiveni su uspoređujući rezultate predviđanja za pet različitih neuronskih mreža. Uzimajući u obzir složenost strukture, vrijeme učenja te količinu primjera za učenje na raspolaganju, zaključilo se kako je najbolja struktura dvostruka MLP (engl. *MultiLayer Perceptron network*) mreža koja koristi dva tipa predviđanja (24 i 48 sati unaprijed). Ideja ovog pristupa jest da se ne koristi jedna mreža po satu, već da jedna mreža radi predviđanje od 24 sata unaprijed, a druga 48 sati unaprijed. Potom se iz te dvije mreže dobiva predviđena dnevna krivulja. Koriste se dvije mreže zato što je pretpostavka da prilikom predviđanja za prvi sat postoji informacija o potrošnji prije 24 sata, dok za ostale sate u danu mora se koristiti mreža koja predviđa s horizontom od 48 sati, iako je najsvježiji podatak od prije 25 sati, ali sve do 48 sati unazad podaci nisu frekvencijski relevantni. Prednost ove mreže jest u tome da na raspolaganju ima 24 više primjera za učenje što bi trebalo rezultirati boljim učenjem oblika dnevne krivulje te robusnijim ponašanjem za blagdane. Nedostatak je znatno duže vrijeme treniranja te zahtjevi za memorijom. Struktura mreže prikazana je na slici 4 iz koje se može vidjeti da je dodan jedan parametar τ_4 u regresor koji označava za koji se sat u danu računa predviđanje.



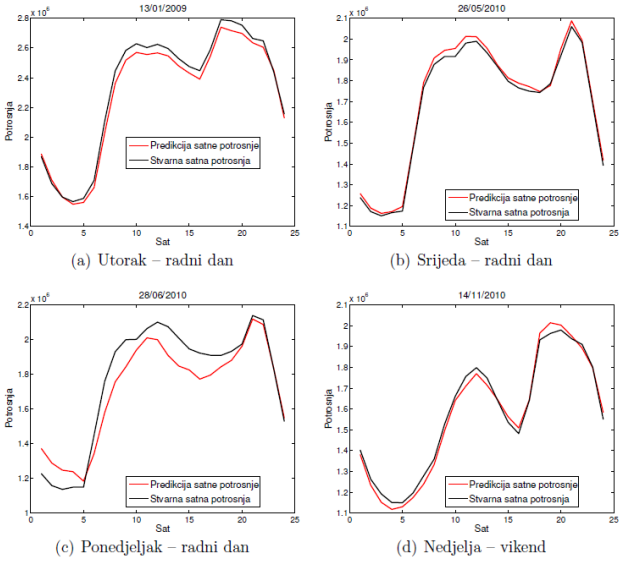
Slika 4. Struktura neuronskih mreža s horizontima od 24 i 48 sati.

Dobivena srednja apsolutna pogreška za cjelokupan sustav iznosi 1,328% (slika 5), a rezultati simulacije za nekoliko odabranih dana prikazani su slikom 6. Razlog dobrome rezultatu za cjelokupan sustav vjerojatno leži u nešto većoj robusnosti cijelog sustava na nenadane događaje te u činjenici da se predviđanje računalo za cijeli sustav zbrajanjem predviđanja za pojedine grupe distribucijskih područja od kojih su neki mogli procijeniti preveliku, a neki premalu potrošnju što u konačnici zbrajanjem može rezultirati manjom pogreškom.



Slika 5. Srednja apsolutna greška za sve sate u danu za cjelokupan sustav

Iako je odabir strukture mreža napravljen na temelju rezultata za distribucijsko područje Elektre Zagreb, pretpostavka je da se struktura modela uči radi opisivanja nelinearnog odnosa između ulaznih parametara (kalendarski, hidrometeorološki i podaci o potrošnji) te izlazne veličine (predviđena potrošnja) neovisno o strukturi distribucijskog područja. Parametri modela će se svakako mijenjati ovisno o distribucijskom području, jer različiti čimbenici utječu na potrošnju. Međutim, to ne utječe na samu strukturu modela te je korišteno jedno distribucijsko područje za usporedbu.



(a) Utorak – radni dan (b) Srijeda – radni dan

(c) Ponedjeljak – radni dan (d) Nedjelja – vikend

Slika 6. Primjer cjelodnevnog predviđanja potrošnje (u Wh) električne energije po satima za cjelokupan sustav

Također, veće netočnosti kod manjih distribucijskih područja mogu se očekivati zbog veće osjetljivosti na promjene u sustavu, ali sam izbor modela ne bi trebao imati utjecaja. Sukladno tome, za svako distribucijsko područje provodilo bi se učenje za taj slučaj te bi dobivena mreža opisivala najbolje što je moguće naučen odnos između odabranih ulaznih veličina i predviđene potrošnje [5].

VI. ZAKLJUČAK

Predviđanje potrošnje električne energije dobiva sve veći značaj u sektoru distribucije električne energije. To je ponajprije posljedica sve razvidnijeg tržišta električne energije u Republici Hrvatskoj, ali i činjenice da točno predviđanje potrošnje električne energije omogućava značajne uštede za elektroenergetski sektor. Čak i relativno mala poboljšanja u točnosti predviđanja potrošnje električne energije mogu rezultirati uštedama od nekoliko stotina tisuća do više milijuna kuna.

Da bi se postiglo što bolje predviđanje potrošnje električne energije važno je imati što točnije ulazne varijable (hidrometeorološki podaci, potrošnja električne energije prethodnih dana te tip dana i doba godine), jer neispravno odabrani skup ulaznih varijabli može znatno narušiti kvalitetu predviđanja potrošnje električne energije. Ako ulaznim skupom nisu obuhvaćene neke značajne varijable, bit će nemoguće u potpunosti opisati funkciju potrošnje električne energije. S druge strane ako ulazni skup sadrži preveliki broj (redundantnih) ulaznih varijabli, postupak učenja modela i predviđanja postaje pretjerano složen te može dovesti do toga da funkcija potrošnje električne energije ne bude ispravno naučena.

Gubici u distribucijskoj mreži predstavljaju jedan od ključnih čimbenika koji ukazuju na stupanj ekonomičnosti poslovanja i kvalitetu obavljanja djelatnosti distribucije električne energije. Ovim radom dan je pregled izgradnje sustava predviđanja iznosa energije gubitaka u distribucijskoj mreži, izrada plana nabave potrebne energije na tržištu te izračun ostvarenja energije gubitaka.

LITERATURA

[1] Internet: http://www.hep.hr/ods/publikacije/godisnje2013.pdf, pristupljeno 4. veljače 2015.

[2] D. Mileta, H. Keko, I. G. Kuliš, M. Skok, T. Baričević, M. Tot, M. Maričević, "Istraživanje i primjena krivulja opterećenja skupina karakterisičnih kupaca na tržištu električne energije", studija EIHP, svibanj 2009

[3] M. Skok, H. Keko, I. G. Kuliš, T. Baričević, D. Karavidović, T. Marijanić, Z. Lipošćak, Deriving LV customer load profiles for croatian electricity market purposes, CIRED, Prag, 2009. godina.

[4] HEP–Operator distribucijskog sustava d.o.o., *Pravila primjene nadomjesnih krivulja opterećenja*, 2011.

[5] I. Petrović, J. Matuško, S. Jurić-Kavelj, I. Marković, V. Petrović, „Matematički model kratkoročnog predviđanja satne potrošnje električne energije za distribucijska područja", studija FER, prosinac 2011.

[6] *Pravila djelovanja tržišta električne energije*, Narodne novine br. 135/06, 146/10, 90/12.

Automatska monotona kategorizacija numeričkih varijabli

Valeria Horvat i Krešimir Bokulić
Multicom d.o.o., Zagreb, Hrvatska
valeria.horvat@multicom.hr , kresimir.bokulic@multicom.hr

Sažetak - Rad se bavi područjem kategorizacije (engl. *Binning*) varijabli za dubinsku analizu podataka (engl. *Data mining*). U radu je opisano automatsko rješenje pronalaženja granice kategorizacije kontinuiranih numeričkih varijabli s obzirom na ciljnu varijablu. Rješenje se koristi u slučaju kada ciljna varijabla poprima dvije vrijednosti. Svakom razredu pridružena je WOE (engl. *Weight of Evidence*) težina, te je raspodjela težina monotona. Algoritam ima dva moguća rezultata: monotonu kategorizaciju s najvećim mogućim brojem razreda i kategorizaciju s maksimiziranom IV (engl. *Information Value*) vrijednošću. Rješenje je razvijeno u programskom jeziku R koristeći Rstudio. Sastoji se od R paketa i R skripti, koje je Oracle PL/SQL upitom moguće pozvati na Oracle Data Miner alatu.

I. UVOD

Diskretizacija, tj. kategorizacija varijabli može pružiti robusno, interpretativno rješenje u prediktivnim modelima. U području financija su to često varijable poput dobi, prihoda, nenaplaćenih zajmova. Dob je tipična karakteristika u bankarstvu, pogotovo u modelima kreditnog bodovanja(engl. *Scoring*). Takva karakteristika je podijeljena u različite razrede i svakom razredu je dodijeljena ili količina bodova ili težina, a može biti i WOE vrijednost. Na primjer, dvadesetsedmogodišnji podnositelj zahtjeva za kredit bi mogao pasti u dobnu skupinu od 24-30 godina. Na taj je način varijablu „dob“ lakše interpretirati, čak i ako osoba koja koristi model ne razumije modeliranje kreditnih rizika. Diskretizacijom kontinuirani atributi, tj. varijable postaju diskretne. Postoje razne metode pronalaženja granica diskretnih razreda: razredi iste širine, razredi po kvantilima, itd. Originalne vrijednosti atributa mogu se zamijeniti s raznim vrijednostima, npr. oznaka intervala, redni broj intervala, srednja vrijednost ciljne varijable unutar intervala, s WOE vrijednošću, itd.

Ovaj rad u drugom poglavlju daje definiciju WOE i IV vrijednosti. Zatim u trećem poglavlju opisuje općeniti problem smislenog grupiranja jedne varijable i uvjete, tj. ograničenja koja se najčešće pojavljuju prilikom traženja optimalne kategorizacije. U četvrtom upoznaje čitatelja s nekim alatima za dubinsku analizu podataka. Te u zadnjem pruža uvid u rješenje mBinning za automatsko pronalaženje monotone kategorizacije, tj. pronalaženje razreda, granica za grupiranje varijabli uz koje se dobivaju optimalne monotone WOE vrijednosti te se maksimizira IV vrijednost. U tom poglavlju detaljno je opisana implementacija rješenja. Tako je pokazano da su riješeni problemi i poteškoće koje se mogu javiti prilikom automatske kategorizacije numeričke varijable u odabranim alatima.

II. WOE I IV

Prije spomenuta metoda zamijene s WOE (engl. *Weight of Evidence*) vrijednošću uzima u obzir i ciljnu varijablu. U slučaju diskretne ciljne varijable koja može poprimiti dvije vrijednosti, npr. iz skupa (0, 1), („dobar“ , „loš“), („događaj se nije desio“, „događaj se desio“) i slično. WOE se za svaku grupu *i* dobivenu diskretizacijom računa sljedećom formulom (vidi u [2]):

$$WOE_i = \ln\left(\frac{\frac{dobri_i}{Ukupno\ dobrih}}{\frac{lošI_i}{Ukupno\ loših}}\right). \quad (1)$$

Nakon kategorizacije može se izračunati IV (engl. *Information Value*) vrijednost i druge metrike za izračun prediktivnosti varijable. IV vrijednost se za neku varijablu s *k* razreda može dobiti sljedećom formulom:

$$IV = \sum_{i=1}^{k}\left(\frac{dobri_i}{Ukupno\ dobrih} - \frac{lošI_i}{Ukupno\ loših}\right) * WOE_i\ . \quad (2)$$

III. PROBLEMI GRUPIRANJA

Proces kategorizacije je često ograničen nekim pravilima, stoga se na traženje optimalnih granica koje zadovoljavaju zadana pravila može izgubiti podosta vremena. Na primjer, kako bi se osigurao statistički značajan i primjenjiv model u nekim slučajevima, može se dati zahtjev da svaki razred sadrži obje vrijednosti ciljne varijable i „dobre“ i „loše“. Treba pružiti alternativan izračun u slučajevima kada to nije moguće. Drugi zahtjev može biti da svaki razred smije sadržavati samo određen broj opažanja, a isto tako i minimalan broj zapažanja.

Mogu se zadati i zahtjevnij uvjeti na kategorizaciju. Na primjer, u zakonu Sjedinjenih Američkih Država, točnije u Equal Credit Opportunity Act (USDOJ 2006) (vidi u [3]) zakonu, nameće se ograničenje koje kaže da se starijim podnositeljima zahtijeva za kredite ne smije dodijeliti manja količina bodova, tj. manja težina od onih dodijeljenih mlađim podnositeljima. To ograničenje zahtijeva monotonost WOE vrijednosti. Primjer, pretpostavimo da je varijabla *godine* podijeljena na tri razreda: $a = \langle 18, 25]$, $b = \langle 25, 40]$ i $c = \langle 40, 60]$, razredu a pridružena je vrijednost WOE_a , razredu b WOE_b , razredu c WOE_c . Pretpostavimo da WOE vrijednosti nisu monotone, ali da je maksimalna WOE

vrijednost u razredu s najvećim godinama, tj. WOE_c je maksimum. Kako je WOE_c maksimum, znači da su WOE_a i WOE_b manji od WOE_c, a kako WOE vrijednosti ne rastu monotono, znači da je WOE_a veće od WOE_b. No, tako je razredu s manjim godinama pridružena veća težina, odnosno Equal Credit Opportunity Act nije zadovoljen. Sličan zaključak dobiva se i kada WOE nisu monotone ali imaju minimum. Odavdje slijedi da WOE vrijednosti trebaju biti monotone. Ovakvo ograničenje je posebno zahtjevno zbog nelinearne veze između varijable koju je potrebno grupirati i ciljne varijable.

Ukratko, tražimo rješenje koje zadovoljava sljedeća ograničenja:

- maksimalan broj razreda
- minimalan broj razreda
- maksimalna veličina razreda
- minimalna veličina razreda
- monotonost WOE vrijednosti po razredima
- alternativno rješenje za razrede koje nemaju zastupljene obje vrijednosti ciljne varijable.

Kada je inicijalna podjela varijabli velika, proces nalaženja optimalnih granica može potrajati zbog mogućih kombinacija i velikog broja iteracija. Moguće je da je skup zahtijeva toliko restriktivan da ne postoji rješenje koje može zadovoljiti sve tražene uvjete. Sam postupak kategorizacije ima negativne strane, te nije primjenjiv u svim situacijama. S kategorizacijom mogu se izgubiti korisne informacije, pa tako i sam model može izgubiti na prediktivnosti.

A. *Spajanje susjednih razreda*

Jedan od pristupa pronalaženju optimalne kategorizacije je da svaku vrijednost proglasimo jednim razredom, zatim ih spajamo sa susjednim razredima dok svi zadani zahtjevi ne budu ispunjeni.

Primjerice, u tijeku kategorizacija neke karakteristike nije ispunjen uvjet minimalne veličine razreda za neki razred *k*. Rješenje bi bilo spojiti taj razred s *k+1* ili *k-1* razredom, no pritom treba paziti da se ne prijeđe uvjet maksimalne veličine razreda ili minimalnog broja razreda, te treba paziti da WOE vrijednosti rastu ili padaju monotono. Takvo spajanje razreda ima utjecaja i na druge razrede. U slučaju kada *k* i *k+2* razred ne zadovoljavaju zadane uvjete, rješenje je spajanje sa susjedima, no pitanje je s kojim razredom krenuti. Razred *k+1* je susjedni od oba prije spomenuta razreda. Ako se krene s jednim razredom, može se dogoditi da se drugi neće moći spojiti tako da su ograničenja kategorizacije zadovoljena.

Broj mogućih kombinacija raste s brojem početnih razreda inicijalne podjele. Kod neograničenih kontinuiranih varijabli taj broj može biti velik. Iz tih razloga bi trebalo bolje promisliti o inicijalnoj podjeli varijable.

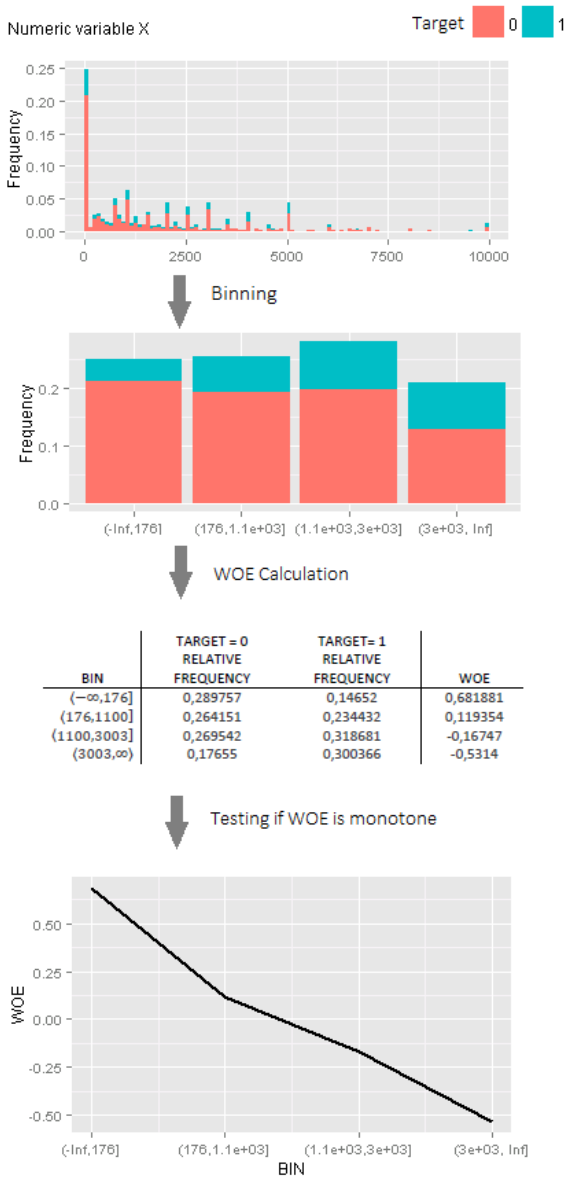
B. *Pronalaženje granica koristeći kvantile*

Umjesto prethodno spomenute inicijalne podijele, raspon varijable moguće je dijeliti na jednake razrede, odnosno razrede iste širine. No ovdje se javlja problem odlučivanja o širini razreda. Ako se uzmu preširoki razredi postoji mogućnost da se neće moći naći optimalno monotono grupiranje koje zadovoljava zadane uvijete. Ukoliko se uzmu preuski razredi, pronalaženje optimalnog grupiranja bi moglo potrajati sukladno tome dulje. Jedna od mogućih metoda je podjela na razrede s jednakom količinom podataka. Takva metoda koristi kvantile (*n* kvantila dijeli varijablu na *n* razreda pri čemu svaki sadrži najviše (100/*n*)% opažanja neke varijable) .

Nakon što se podaci podijele (Slika 1) računa se WOE vrijednost za svaki pojedini razred, provjerava se da li su početni uvjeti zadovoljeni. Ako nisu, podaci se dijele po novim kvantilima, manjem broju.

IV. R, ORE I ODM

R je programski jezik i softverska okolina za statističke obrade podataka te njihov grafički prikaz. R je GNU projekt, samim time je open source i dostupan pod



Slika 1: Prikaz procesa traženja optimalne kategorizacije za numeričku varijablu X koja poprima vrijednosti iz intervala od 0 do 10000.

GNU General Public License licencom. Izvorni kod za R okolinu je napisano prvenstveno u C, Fortranu, ali i u R-u. R koristi sučelje naredbenog retka, postoji i nekoliko alata kao Rstudio, kojim je programiranje u R-u olakšano. Prednost R-a je lako proširivanje kroz pakete, te jednostavna definicija vlastitih paketa. Njegov nedostatak je upotreba memorije. Jednostavne naredbe koje se u SQL brzo izvršavaju, ovisno o dostupnoj količini radne memorije, u R-u mogu razmjerno duže potrajati.

OAA (engl. *Oracle Advanced Analytics*) je proširenje na Oracle bazu podataka. OAA se sastoji od dva dijela: ORE (engl. *Oracle R Enterprise*) i ODM (engl. *Oracle Data Mining*).

ORE povezuje R sa Oracle 11g bazom podataka. Točnije, R je direktno instaliran na server na kojem se nalazi i baza podataka. Instalacijom se omogućuje tzv. transparentni sloj (engl. *transparency layer*) što znači da se naredbe iz klijentskog R engina na bazi izvršavaju kao SQL naredbe ukoliko postoji njihov ekvivalentni transparency oblik. Transparentni sloj omogućuje jednostavan prijenos podataka, uz to i korištenje tablica i pogleda (engl. *View*) definiranih u R-u sa ore.frame tipom koji je ekvivalent R-ovom osnovnom data.frame tipu. Zbog toga je moguće i definirati vlastite ORE skripte, koje se tada čuvaju na bazi.

ODM je skup algoritama potrebnih za dubinsku analizu podataka, klasifikaciju, predikciju, regresiju i sl. ODM se može koristiti kroz grafičko korisničko sučelje Oracle Data Miner. U tom sučelju se grade procesni sljedovi (engl. *workflow*) s raznim čvorovima ponuđenim u alatu.Oracle Data Miner se može povezati i sa ORE-om u SQL čvoru definirajući ulaznu tablicu, parametre i izlaznu tablicu ORE funkcije, odnosno skripte.

Primjer u (Slika 2) daje procesni slijed izgrađen nad INSUR_CUST_LTV_SAMPLE tablicom koja dolazi s instalacijom Oracle Data Miner repozitorija. Tablica sadrži umjetno stvorene podatke vezane uz klijente nekog fiktivnog osiguravajućeg društva. Ciljna varijabla daje informaciju o tome da li je pojedini klijent kupio osiguranje ili nije. Osim ciljne varijable tablica sadrži i identifikacijsku kolonu, broj godina klijenta, broj djece, iznos plaće, stanje na računu i slično.

V. MBINNING RJEŠENJE

U ovom poglavlju je dan opis mBinning rješenja. Opisane su pojedine komponente, važniji dijelovi njihovih algoritama, ulazni parametri i izlazni rezultati. Sve komponentne su razvijene od strane autora.

Rješenje mBinning koje služi za grupiranje kontinuirane numeričke varijable moguće je zbog svoje konstrukcije koristiti kroz R, ORE te kroz ODM pozivom skripti u SQL čvoru (Slika 2). Sastoji se od sljedećih komponenti:

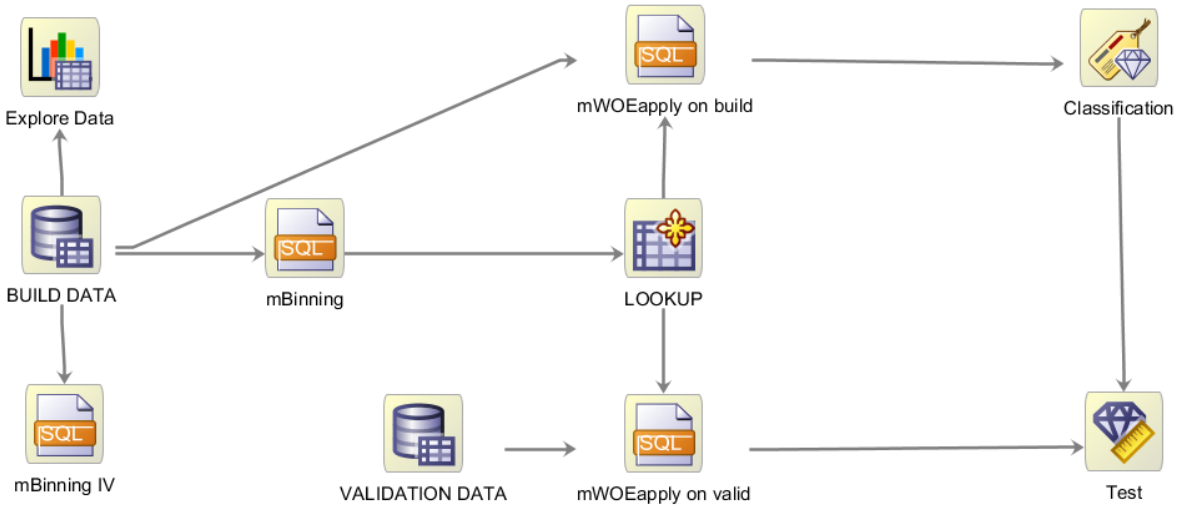
- mBinning R paketa
- m.Binning ORE skripte
- m.BinningIV ORE skripte
- m.WOEapply ORE skripte

Sve ORE skripte se čuvaju u bazi podataka. Transparency sloj omogućuje njihovo izvršavanje, koje se odvija u R-u na serveru. R mBinning paket se nalazi u R-u na serveru i tamo se izvršava (Slika 3). Moguće je paket imati na lokalnom stroju, tada transparency sloj služi za prebacivanje podataka s baze na lokalno klijentsko računalo.

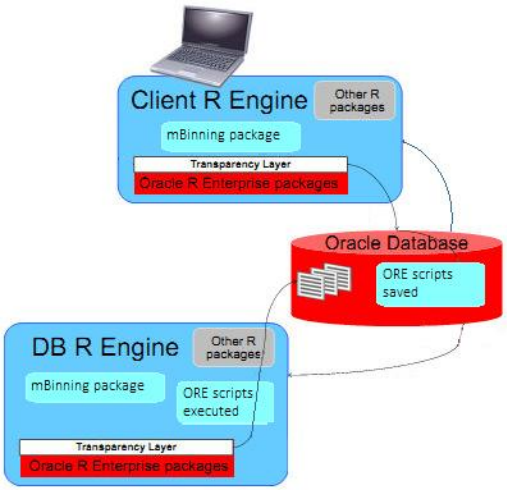
A. *mBinning R paket*

mBinning R paket je skup R funkcija. Funkcija koja se poziva iz gore navedenih ORE skripti je m.Binning funkcija. Ta funkcija za numeričku varijablu nalazi kategorizaciju koja zadovoljava zadane uvjete. Algoritam je napisan tako da iskoristi sve prednosti koje pruža R , te izbjegava moguće poteškoće koje bi se mogle javiti u ORE i ODM okruženju. U daljnjem tekstu slijedi detaljan opis rada funkcije.

Funkcija traži optimalnu kategorizaciju na način kako je opisano u poglavlju III.B (Slika 1). Slijedi detaljniji



Slika 2: Primjer procesnog slijeda iz Oracle Data Miner alata i načina poziva i redoslijeda ORE skripti iz mBinning rješenja u SQL čvorovima.



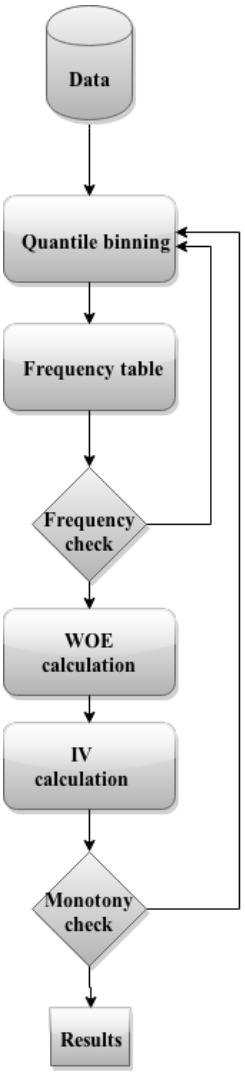
Slika 3: Prikaz transparentnog sloja i arhitektura mBinning rješenja

opis postupka (Slika 4) :

1. Da bi se odredila varijabla nad kojom se želi primijeniti kategorizacija treba navesti tablicu u kojoj se nalazi, naziv kolone, tj. varijable, naziv ciljne varijable. S dodatnim parametrima osigurava se da su prije spomenuti zahtjevi zadovoljeni. Pomoću njih se može podesiti maksimalan broj razreda, maksimalna i minimalna veličina pojedinih razreda te je moguće zadati vrijednost za korekciju razreda u kojoj nisu zastupljene obje vrijednosti ciljne varijable. Dva dodatna parametra određuju algoritme kojima se traži kategorizacija ulazne kolone. To su parametri percents i max_IV , a biti će objašnjeni u daljnjem tekstu.
2. Kategorizacija kreće od maksimalnog broja razreda *n*. Traže se granice za *n* razreda po kvantilima. Za svaki zapis traži se razred kojem pripada. Prazni zapisi se posebno tretiraju. Takvi zapisi se smještaju u *n+1*. razred. Tako korisnik može sam odlučiti kako će tretirati takve zapise, odnosno kojim razredima će ih se pripojiti.
3. Sljedeći korak je izrada frekvencijske tablice. Frekvencijska tablica u sebi sadrži informaciju o raspodjeli ciljne varijable po razredima. Jednostavnije, za sve zapise koje se nalaze u *k*-tom razredu provjerava se vrijednost ciljne varijable te se izračuna frekvencija za obje vrijednosti u tom razredu. U frekvencijskoj tablici također se nalazi informacija o relativnoj veličini razreda u odnosu na sve zapise. Ta tablica je potrebna iz dva razloga: koristi se za provjeru uvjeta na kategorizaciju (sada se lako iz tablice vidi da li su uvjeti maksimalne i minimalne veličine razreda zadovoljeni) i služi za izračun WOE vrijednosti.
4. Prije nego se krene u izračun WOE vrijednosti, u frekvencijskoj tablici potrebno je zamijeniti sve vrijednosti koje su jednake nula s vrijednošću zadanom u ulaznom parametru. Tako je zadovoljen uvjet za razrede koje nemaju zastupljene obje vrijednosti ciljne varijable. Kada bi se prethodni korak preskočio, došlo bi do greške prilikom izračuna WOE vrijednosti zbog njegove definicije. Dijeljenje s nulom nije moguće. Nakon izračuna WOE vrijednosti, može se izračunati IV metrika. Ukoliko nisu zadovoljeni uvjeti vezani uz veličinu razreda, daljnji koraci se preskaču. Sada se može provjeriti monotona raspodjela WOE vrijednosti po razredima.

Ako monotonost nije postignuta, koraci 2., 3. i 4. se ponavljaju za *n-1*. Ako se ne može ni za jedan *n>=2* naći kategorizacija koja zadovoljava zadane uvjete, tada funkcija vraća rezultatnu tablicu s posebnim ćelijama. Kolona koja daje razred grupiranja će biti prazna, a ostale kolone imat će nule.

Ulazni parametar *max_IV* utječe na algoritam tako da određuje trenutak u kojem potraga za kategorizacijom završava. Postoje dvije mogućnosti: u prvoj se nalaze sve moguće kategorizacije neke varijable koje zadovoljavaju zadane početne uvjete, te se na kraju odabire ona za koju je IV vrijednost najveća, u drugoj algoritam stane kada naleti na prvu kategorizaciju koja zadovoljava uvjete. Na



Slika 4: Blok dijagram redoslijeda radnji unutar m.Binning R funkcije

taj način ulazni parametar *max_IV* ima utjecaja na brzinu izvršavanja cijelog postupka.

Parametar *percents* također utječe na brzinu izvršavanja. Ovisno o njemu uvjeti maksimalne i minimalne veličine razreda se uzimaju u obzir ili ne. Taj parametar može usporiti cijeli proces. Ako se uvjeti o veličini razreda ne uzimaju u obzir, tada će se WOE vrijednost izračunati za svaki *k*.

Zbog parametara *max_IV* i *percents* postoji četiri različita algoritma. Razlika je vidljiva u brzini izvršavanja (Slika 5) i u rezultatu.

B. *m.Binning ORE skripta*

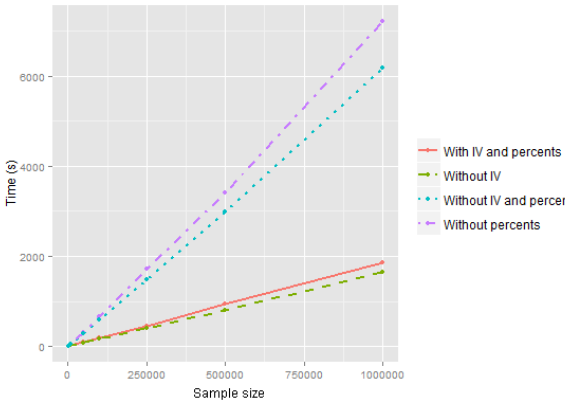
Skripta omogućuje primjenu prethodno spomenute funkcije iz mBinning R paketa na cijeloj tablici, a ne samo na pojedinoj numeričkoj koloni. Petlja prolazi kolonu po kolonu ulazne tablice te primjenjuje funkciju za nalaženje optimalnog grupiranja, s kojim se osigurava monotonost WOE vrijednosti, na svakoj koloni pojedinačno. Kolone koje nisu numeričke jednostavno se preskaču.

Rezultat skripte je tablica koja sadrži kolonu s nazivom numeričke varijable, redni broj razreda, WOE vrijednost pojedinog razreda, gornju granicu razreda, donju granicu razreda. Ako se koristi algoritam za maksimizaciju IV vrijednosti, izlazna tablica će sadržavati kolonu s IV vrijednošću grupirane varijable. Ako su zadani uvjeti na veličinu pojedinih razreda, rezultat će još sadržavati kolone s informacijama o veličini razreda te strukturu ciljne varijable u tom razredu (Tablica 1). Varijable za koje nije moguće naći grupiranje koje bi zadovoljilo zadane zahtjeve, biti će navedene u izlaznoj tablici. Sve kolone za takve varijable biti će nule, osim kolone s njihovim nazivom i razredom grupiranja. Tablica dobivena ovom ORE skriptom može služiti kao pregledna (engl. *Lookup)* tablica za primjenu u m.WOEapply ORE skripti, koja je niže objašnjena.

C. *m.BinningIV ORE skripta*

U toj ORE skripti koristi se glavna funkcija iz mBinning R paketa nad cijelom tablicom. Petlja koja to omogućuje je slična onoj iz m.Binning ORE skripte. Ona koristi spomenutu m.Binning funkciju iz mBinning R paketa na svakoj pojedinačnoj koloni. Primjenjuje funkciju nad numeričkim kolonama, kategorijske i datumske kolone se preskaču.

Razlika između ove ORE skripte i prethodne je u rezultatu. Rezultat m.BinningIV ORE skripte sadrži



Slika 5: Usporedba brzine algoritma nad tablicama s različitim brojem zapisa (*Sample size*) za sve moguće kombinacije postavki ulaznih parametara *max_IV* i *percents.*Tablice su dobivene repliciranjem kolona iz spomenute INSUR_CUST_LTV_SAMPLE tablice.

kolonu koja daje ime numeričke varijable, broj razreda za koje je nađeno monotono WOE grupiranje koje zadovoljava tražene uvjete, te rezultat sadrži i IV vrijednost takvog grupiranja. Jedna varijabla može biti navedena više puta, ukoliko za nju postoji više mogućih grupiranja koje zadovoljavaju tražene uvjete. Ako za neku numeričku varijablu nije moguće naći ni jedno grupiranje, tada je ona samo navedena, ali ostale ćelije će sadržavati nule.

D. *m.WOEapply ORE skripta*

Ova skripta služi za primjenu prethodno dobivenih preglednih tablica na tablicu koja se koristi za modeliranje.

Rezultat ove skripte je tablica s identifikacijskom kolonom i još onoliko kolona za koliko varijabli je originalna vrijednost zamijenjena s WOE vrijednošću. Dakle, kolone koje nisu numeričke i numeričke kolone za koje nije nađena kategorizacija koja zadovoljava zadane uvjete se preskaču. Njihove vrijednosti se ne zamjenjuju, te niti nisu dio rezultata skripte.

VI. ZAKLJUČAK

U području financija, pogotovo u području kreditnih rizika, kontinuirane varijable često se pretvaraju u

Tablica 1: Primjer rezultatne tablice poziva m.Binning ORE skripte.

VAR_NAME	VAR_LEVEL	WOE	LOWER_LIMIT	UPPER_LIMIT	IV	P0	P1	PTOTAL
NUM_VAR_1		0	0	0	0	0	0	0
NUM_VAR_2		0	0	0	0	0	0	0
NUM_VAR_3	1	0,6819		176	0,1752	0,2898	0,1465	0,2512
NUM_VAR_3	2	0,1194	176	1.100	0,1752	0,2642	0,2344	0,2562
NUM_VAR_3	3	-0,1675	1.100	3.000	0,1752	0,2695	0,3187	0,2828
NUM_VAR_3	4	-0,5314	3.000		0,1752	0,1765	0,3004	0,2099

kategorijske. Taj proces poznat je kao binning, kategorizacija ili grupiranje. Iako to može smanjiti moć predikcije konačnog modela, ako je učinjeno razborito gubitak je mali.

Postoje razne metode za kategorizaciju. Jedna od mogućih pristupa je dan u mBinning rješenju. U sebi sadrži automatski monotoni binning numeričkih varijabli koje se dijele po kvantilima. Omogućuje zamjenu originalnih vrijednosti varijable s pripadnim WOE vrijednostima. Korisnik ovog rješenje može birati između dva pristupa: kategorizacija s maksimalnom IV vrijednošću i kategorizacije s najvećim mogućim brojem razreda. Štoviše, ako se odluči za prvi pristup, ima uvid u preostale moguće kategorizacije za pojedinu numeričku varijablu. Ulaznim parametrima dobiveni rezultat može se podesiti tako da zadovolji neke zahtjeve na kategorizaciju poput: maksimalnog broja kategorija, veličine pojedinih kategorija itd. Rješenje mBinning može se koristiti u R-u ili ORE-u, može se pozvati kroz procesni slijed Oracle Data Miner alata.

LITERATURA

[1] G. Zeng, „A Necessary Condition for a Good Binning Algorithm in Credit Scoring“, Applied Mathematical Sciences, Vol. 8, 2014, no. 65, 3229 – 3242, HIKARI Ltd, www.m-hikari.com, http://dx.doi.org/10.12988/ams.2014.44300

[2] N. Siddiqi, “Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring”, John Wiley Sons Inc, 2005.

[3] Wikipedia Equal Credit Opportunity Act URL: http://en.wikipedia.org/wiki/Equal_Credit_Opportunity_Act

[4] J. Wu, S. Coggeshall , „ Foundations of Predictive Analytics“ , 1st ed., Chapman & Hall/ CRC Data Mining and Knowledge Discovery Series, 2012. , pp. 200-204.

[5] Wikipedia Quantile URL: http://en.wikipedia.org/wiki/Quantile , 2015.

[6] P. Burns, „The R Inferno“, URL: http://www.burns-stat.com/pages/Tutor/R_inferno.pdf, 2011.

[7] „Oracle R Enterprise 1.3 Reference Manual“, URL: http://www.oracle.com/technetwork/database/options/advanced-analytics/r-enterprise/ore-reference-manual-1882822.pdf ,2012.

[8] „Using Oracle Data Miner 11g Release 2“, URL: http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/11g/r2/prod/bidw/datamining/ODM11gR2.html

DE/GLGPS

International Conference on

DIGITAL ECONOMY AND
GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

Steering Committee

Chairs: Mladen Mauher, Polytechnic of Zagreb, Croatia
Edvard Tijan, University of Rijeka, Croatia

INVITED PAPER

Public e-Procurement as a Key-Impact Factor in the UE Welfare Context

-The Italian Public e-Markets: Research Findings and e-Solutions Analysis-

Francesco Bof*, Gianmaria Casella **
* Bocconi School of Management, Milan, Italy
** CSAmed, Cremona, Italy
Francesco.bof@sdabocconi.it, francesbof@hotmail.com

Abstract

"*For Governments thinking about the competitiveness of nations and states has focused on the overall economy, with national-level policy as the dominant influence*" (Porter, 2000).

The goal of the present paper is to help Governments Decision Makers, Public Buyers, Private Markets Representatives, and ICT Companies to understand the opportunities and the present challenges of the public processes digitalization strategy for the EU Welfare goals. Evidences on the Italian experience in public (health) e-procurement and e-markets will be brought forth.

Public (E)-Procurement of goods/services, in the EU context, is strategic because of five main reasons: a) economic relevancy (public expense can be from 10% to 20% of EU Countries GDP); b) impact on citizens' service quality c) impact on markets competitiveness and innovation capacity, d) necessity to guarantee transparency and security of transactions and e) improvement of its processes efficiency. In particular, e-procurement models and solutions play a major role in the present EU digitalization strategy. Moreover, it seems that none of them has emerged as the optimal yet.

After an analysis of the updated relevance, main factors and dynamics effecting public (e)-procurement of goods/services and public-private in the EU and in the Italian context, in particular: transactions data analysis in a four years' period on the Italian Public e-Marketplace will be presented. As the findings give evidence of criticalities connected with A. ICT infrastructures capacity of territory and, B. Public segment (specifically the Public Health), consistently, a case-study identifying potential (although partial) solutions will be presented.

I. INTRODUCTION

Nations' competitiveness depends, among the other factors "on the Government influences to act as sophisticated buyer which provide needed inputs and upgrade local demand conditions" (Porter, 1990). In particular, Public Procurement of goods and services in a public Governments context is a strategic activity aiming to provide needed inputs for improving the citizens' conditions and, so doing, to reach the welfare level goals.

In the EU, specifically, the Public Procurement strategy aims to increase EU countries' and citizens' welfare levels. Public Procurement affects the public debt, being quantitatively from 10% to 20% of a EU country public expense; consequently public debts contingencies often consider Public Procurement as a priority in the Governments Agendas. Public expense management and Public Procurement Systems should generally aim to reduce bureaucracy and administrative costs (process reengineering), to negotiate effective prices with the different market sector, increasing the quality of goods and services in a medium-long term.

Centralization of Procurements stations and tender processes, and e-procurement tools and are two strategic pillars of today's EU strategy. Centralization should reduce tender procedure and administrative costs and increase the power of demand; consequently contract prices of public-private transactions should decrease, and in the medium term, should contribute to lower the Public expense without affecting the quality of needed factors, tangibles (goods), intangibles (services).

E-procurement has a huge potential in term of administrative cost reduction, bureaucracy simplification, optimization of the employees and resources involved in the process, of time constraints, transparency level. Transparency, in particular, can help in fighting corruption and misbehaviours and in building dataset and database feeding public policies information needs and requests.

With respect to these goals, and strategy pillars, policy makers must, at first, identify the role of Central Departments in the long-term, and so doing define guidelines and mandates of the single procurement station (e.g. cities, hospitals), of possible alliances models among them (e.g. Consortium), following geographical and demographic criteria, economic parameters applied to tender models (e.g. min or max economic tender volumes per each production input), and guidelines/mandatory norms to address different public segment issues. Nevertheless the capacity of evaluating Public Procurement systems in term of socio-economic performances and, so doing, to compare different Public Procure ment models and organizations, is a complex activity, because 1. it needs a process management approach, and so the integration of different organizational professionals and tools, (e.g. in HCOs, in order to provide drugs, team of doctor, nurses, pharmacists, public procurement officers administrative employee and budget manager must work in team and often negotiate with the strategic board), 2. the expected performance can vary depending on external factors. Among these, the Nation's economic and financial "state of health" in the global context, the public services strategy and priorities, the legal set and bureaucracy impact, the politics and policy makers goals and targets, especially during electoral campaigns etc. A Public Procurement strategic management, in other words, is by definition complex and moreover it is effected by different rapidly evolving influences.

The topic of Public Procurement is worth studying for the following reasons:

a) It is economically relevant in a Public State Government balance sheet, and it contributes to its capacity to compete (see point c);

b) It has a substantial impact on Governments' public activities, and on the quality of public services (quality intended as the capacity to respond to citizens' outcome standards and specific target needs);

c) It affects competitiveness and innovation capacity of markets and companies, (as well also Nations' competitiveness and citizens' welfare).

d) Given the high management complexity, there are different institutional and management models none of which has emerged as the optimal one yet,

e) The potential improvement of its processes' efficiency, whose juridical panorama has strongly evolved in the last ten years, is extremely high, in the most of the public context of the EU;

f) research did not produce enough to substantially help top manager who are in charge of it, the academic literature on public procurement is quite poor compared with other management fields.

In particular, with respect to this last aspect (see f, above), and to the previous considerations, the debate and the research fields should aim to understand both drivers and barriers for economic and quality results of public procurement systems and tools also by improving the way to collect, and then by collecting, qualitative and quantitative data concerning the main aspects influencing the process execution and results of a Public Procurement Hub. These can be summarized as follows:

a) Public country (procurement) system,

b) Stakeholders' map,

c) Governance and organizational mission, criteria, structure, indicators (size, competences, costs etc.), swot analysis, human resources systems,

d) Financing model,

e) Legal framework,

f) Core activities processes,

g) Public clients (or users) identification and management,

h) Suppliers and bidding selection criteria,

i) (E)-business to government relationship,

l) Performance measurement and quality monitoring,

m) E-procurement solutions adoption and the evaluation of their capacity in intermediating between demand and offer, that is, between their own clients (i.e., individual public administrations) and the different market sectors.

In the present paper, after A. a review of the literature concerning the recent public procurement evolution, and B) an overview of the (E)-Public Procurement experiences in Italy and in the EU, with a specific focus on the most relevant Public Sector (Health Care), two empirical case will be presented:

The first one will present a results sum up of transactions data analysis on the Italian Public e-Marketplace in a four years' period.

II. REVIEW OF LITERATURE

In today's highly competitive business and global environment, public procurement and e-procurement need to be well equipped with all necessary knowledge, skills, and abilities, as well as innovative technologies and ICT solutions in order to survive in the regional and global markets. Researchers, practitioners and professionals continue to emphasize the need to align the curriculum, competences and ICT solutions with the needs of

national and regional economics. When the public cover a fundamental role for the territorials socio-economic growth they believe in executive education and partnership as tools for developing the capacity and flexibility to adjust highly bureaucratic systems, meet the challenges of the rapidly emerging knowledge-based global economy, and respond faster to changing needs. Therefore, rules in their different declinations (organizational regulations, regional norms, national laws, directives, International guidelines etc.) are not the only way to cope with advancing technology, transforming economy, changing demography, and more importantly dealing with the informational revolution. It is required to develop highly educated workforce and citizens who can contribute and participate in their country's economic development.

Procurement, in particular, is now seen not only as a strategic player in the value chain, but also as a major driver in the extended supply chain (Kalakota and Robinson, 2001). There is some literature expounding the benefits of innovative solutions through procurement systems, mostly connected with ICT implementation (Heywood, 2001; Minahan and Degan, 2001; Konicki, 2001;). For our field of analysis, we have outlined the benefits related to the organizational impact of efficient procurement management as follows:

- Accelerated flow of important information between buyers and suppliers, thereby shortening delivery timelines and leading to greater discounts and better services from suppliers;

- Reduced paperwork and administrative hours;

- Improved accuracy through the elimination of transaction errors and improved financial controls which match orders better;

- Improved auditing and security controls which enable staff and auditors to verify and track the movement of orders throughout the system;

- Reduced inventory levels, hence the costs associated with inventories.

Whilst these drivers provide a measure of success, it is important to consider the possible barriers that organizations may experience when adopting new procurement solutions (e.g. through ICT). A summary of these barriers (Attaran, 2001; Gebauer, 1998; Hawking et al., 2004) are the lack of specific procurement competences, the lack of relationships management with suppliers, and, related to ICT solutions, the security of transactions, the high cost of technology and the lack of technical expertise. With respect to the above, the literature (Andersen, 2004; Henriksen and Mahnke, 2004) reveals that these barriers and requirements tend to increase within the public sector, mainly because of the impact of different economic, institutional and social factors which influence the public domain with respect to the private sector (Gichoya, 2005; Maniatopoulos 2004; Tonkin 2003; Zulfiqar et al. 2001). These differences have resulted in a number of specific regulations and standards developed for public (e)-Procurement and requiring bureaucratic procedures. The bureaucracy, also related to the nature of the institutions involved, (Somasundaram and Damsgaard, 2005) embraces audit, accountability and compliance standards with national and international rules to ensure supply competition and transparency in the awarding of contracts.

Furthermore, Governments' agendas are typically more extensive and complex than the private organizations' ones, where efficiency, cost reduction and time saving are sufficient justifications for the adoption of specific procurement and e-procurement solutions (Coulthard and Castleman, 2001). The relevance of this reality means that one of the first challenges for a procurement policy and standards framework is to recognize that within a public sector context, procurement is more complicated than in the private sector and e-procurement becomes a tool for wider socio-economic goals. At the international level, public procurement rules depend on the country's legal setting and on purchase players (Central and Local Governments, public agencies, Health Care Organizations, etc.). International agreements on public tendering and procurement (WTO, OECD, EBRD) aim to regulate goods, services and work trade opportunities between public buyers and private organizations across different countries and aim to create a common framework based on social goals and key principles (e.g. Maastricht Treaty). These principles, in and of themselves, have proven to be necessary and insufficient in guaranteeing the development of a single and transparent European public market, in improving public services and, so doing, in enforcing Governments social role.

Following EU and EU countries' rules and orientations new ICT tools have been produced at a European level to facilitate the match between demand and offer (e.g., Simap website, Electronic Tendering, Directives to stimulate the use of ICT, etc.); the more the public service impact on individuals and communities is relevant (for example, health care services, environmental issues, creation of job opportunities), the more public procurement rules and ICT tools should be effectively managed. Moreover, the EU harmonization process can face difficulties depending on the different national context and the political, juridical, cultural and institutional profile of each country. The EU countries' public procurement models are basically identified by their grade of centralization/decentralization and by the extension of their public mandate (i.e., their scope and if it is mandatory or not), and on the different typologies of goods/services.

III. E-PROCUREMENT OVERVIEW (FOCUS ON HEALTH CARE)

The 2004 European directive regarding public calls for tender has been acknowledged only in 2006 in many European countries (just Denmark incorporated the European Directives in before – in 2004). However many countries have been considerably active and interested in the subject since the '90s, and have been investing resources in the adoption of e-Procurement databases/platforms for Public Administration purposes.

The Member States of the EU need organisational structures to carry out public procurement functions. These tasks range from the drafting of relevant legislation and development of public procurement policies to the training of procurement officers and publication of contract notices, ending with the structural targeted mission, and mandates delegation processe to new and already existing public procurement bodies. Particularly, many EU countries have adopted a model with a Central Procurement Department aiming to purchase goods and services for the public bodies like Cities, Regions, Provinces, Health Care Organizations (HCOs), Central Ministers etc etc. And the number is seemingly increasing due to the necessity to act more effectively on the public expense, and to optimize the public processes reducing the related costs. Still not all the European governments satisfy their needs through a central procurement agency given that there should be defined criteria, parameters and economic levels in order to identify when a procurement body is centralized and when it is not (10 institutions out of 19 were considered, Piga & Zanza, 2004, and more than 15 in year 2015 following different sources). In Italy, for example, there exists CONSIP Spa one of the biggest, most active and most representative European bodies and platforms, (around 250 employees). Moreover only Tuscany Region, out of 20 Italian Regions, has been developing a stable central procurement model since the very beginning of the territorial centralization process started in the early years of the new millennium. Although differing in terms of responsibilities, functions and tasks, as well as in implemented e-procurement projects, in Europe these bodies have several features in common. European countries tend to choose an external technology for the creation of an e-Procurement platform: The platform of a large number of projects is supplied by a provider that supplies the necessary technology whereas the purchase of a license is less frequent. Only in rarest cases European countries choose to produce internally the platform using their own resources to create the code owned by the Government. The platform organization is usually managed by agencies appointed by central entities: such agencies manage the relationships with public entities and promote the use of the platform.

The analysis of all the various instruments available on the public e-Procurement portals reveals a preference for auctions and calls for tender rather than catalogues and electronic markets. In all assessed countries there is project for the realization of an e-Sourcing platform, although such projects are all at different stages: if on one hand the initial phases of the process are all covered in most cases, (especially the phase of the call for tender publication), on the other hand the number of operating projects decreases as the further phases of the purchasing process are approached (Finland, Greece, Sweden, Ireland). However, in Europe there is a large number of public platforms that allow online negotiation and show more interest for calls managed with an open procedure. Electronic Catalogues are more and more widespread, although not as much as Calls and Auctions: these catalogues allow public entities to purchase directly from suppliers which published their proposal on the web following an authorization by the catalogue manager.

Electronic markets are becoming more relevant and strategic in EU approach to increase efficiency in public mgmt. E.g. E-market enable public entities to ask suppliers for direct orders or quotations saving time and costs.

It is interesting to assess how the different European portals offer their services to users. In some cases such as in Italy, the use of the national e-Procurement system is for free both for entities and suppliers. In other frequent cases in Europe there are several examples of platforms that charge for their services more frequently users, but sometimes also suppliers. There are different revenue policies also within each country, where different payment methods are applied to different e-Procurement instruments (in some cases suppliers pay for a catalogue update, in some others entities are requested for a fee for the use of the platform; some other time suppliers are requested for a percentage fee on the goods sold.

Italian Governments, local or not, are living a turbulent and rapidly evolving period in which the transformation of Government management (and procurement) is passing through many political, institutional and regulatory changes that began in the late ‘90s and are now still far from a stable solution. In particular, in 2006 the new legal setting concerning public procurement (“Nuovo Codice degli Appalti Pubblici”) has been approved by the Italian Parliament with the goal, among the others, to conform and align the national rules to the European ones. The Public System changes led to a model in which many central actors on a regional base are developing after the formula developed at the National level by CONSIP (Public Contracts and ICT Solutions Services Agency) for all public entities across the country (Hospitals and HCOs for which operates a specific Business Unit, Central Governments, Regions, Provinces, and other local administrations). CONSIP is a public company created by the Ministry of the Economy and Finance in 1997 in order to set up both the traditional procedures and the IT platform to carry out public acquisition processes at the central-national level in order to rationalize public expenditures, with a special focus on the Public Health Care Sector, whose national debt is extremely high, and whose public expenses are particularly critical in term of impact on welfare states. The key issues of today’s Italian National Healthcare System (SSN) are below summarized:

Constancy of changing dynamics

Institutional reintegration process

Regionalization (meaning centralization and differentiation mix)

Centrality of operations

Tension in searching Economic and financial equilibrium

In Figure 1 a representation of the health care public expense following economic classification (2012) showing how good and services (second column of Fig. 1 from left) is about a third of the Total; another third is represented by salaries, rewards and other expenses for personnel (first column from left).

After many juridical and organizational changes, today CONSIP is providing e-procurement solutions through electronic tools like the electronic catalogue, the electronic marketplace and the on-line auctions, still their mandate seem to be too challenging because of A. Italy's dimension and territory, population (60 million people), B. public administration numbers (8100 just the cities), C. welfare level and dimensions gap among the 20 Regions, especially between the Southern and Northern, D. the specificity of the Public Health Care sector in term of needs of minimum quality level requested (to the input factors). This last aspect is the most critical as Centralization imply standardization efforts on the demand needs, brought up by the different target client tipologies (Cities, HCOs, Regiones etc etc). Health care need cannot be completely standardized, especially considering the demography trends and consequent epidemiology trends (social and health care needs changes together with the population structure change).

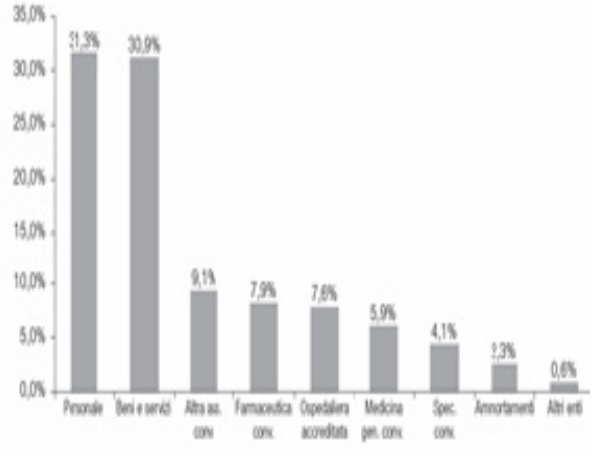


Figure 1. Public health expense following economic classification Source RGPS (2012) – Italian State Tresury

Figure 2.

Unpredictable change	Variables of a Procurement Model
1. Updatings of healthcare needs. Growth in demand of quality	a. Autonomy level (mandatory or facultative)
2 Percentage of growth in goods/services spending.	b. Mandate (which goods/services, which public segment)
3 Variable need of cost cutting and reducing expenses "without reducing the essential levels of assistance". (e.g. Spending Review)	b. budget,
4. Laws increasing pressure on responsibility and control on Public savings and transactions	c. goods and services economic volumes and innovation investments trends,
5. Market crisis, companies compete on price more than on innovation.	d. institutional model (strategy),
	e. level of aggregation of demand,
	f. organization-execution; critic point,
6. Uncertain/variable centralization process	g. ICT technologies and tools
	h. potential impact on care services;

IV. METHODOLOGY

The methodology is a case study approach. The research case study has been defined as a method for learning the "right" questions to ask (Hoaglin et al., 1982). That is, the purpose of case studies is said by some researchers to be to generate hypotheses rather than to test or confirm them. The method involves an in-depth, longitudinal examination of a single instance. In our research, we used the research case study as a method for learning about a complex instance, like public e-procurement, describing extensively the whole context in order to clarify its goals and give a comprehensive understanding.

We've adopted an "illustrative approach" to case studies, which primarily describes what is happening and why, to show what a situation is like. This can help in the interpretation of public e-procurement phenomenon, particularly because we have reason to believe most experts, professionals, business man and academics know too little about the Italian context.

We have been conducting 1. Data analysis on the Italian National Public Administration e-Marketplace transactions for the last four years, through the elaboration of transactional data and, 2. Case-study methodology (Yin, 1984; Eisenhardt, 1989) on the Calabria Italian Region Procurement Model, combining different data collection methods: archives' analysis, interviews, secondary sources, direct observation and transactions data analysis.

In the first analysis, the data compiled relate to the following dimensions:

1. value and number of transactions by six-monthly intervals and by geographic area;
2. value and number of transactions divided by modes of acquisition, specifically, Request For Quotation (RFQ) transactions and direct transactions or Direct Order of Acquisition (RFO) over half-yearly intervals and by geographic area;
3. average value of RFQ and RFO as well as the flow over six-monthly intervals;
4. number of active suppliers by geographic area and the flow over six-month intervals;
5. number of active suppliers by catalogue type (ICT, office, services, health materials, others) and the flow over six-monthly intervals;
6. number of active suppliers by type of catalogues that cover a single area per catalogue (e.g. ICT only), 2 areas per catalogue (e.g. ICT and services), etc.

In the second part, the focus will be on the single contracting office of Calabria Region, with a focus on the transactional data analysis, provided by the Calabria Contracting Office most widely used e-procurement ICT platform (Albofornitori", now Net4Market), managed by CSAmed Company.

V. THE ITALIAN E-PUBLIC MARKET STUDY (2004-2008)

The analysis here presented is divided in parts, depending on the above mentioned dimensions.

The first part of the analysis has as dimension the number of the transactions (figure 1). The first illustration shows this dimension semester by semester over the period running from January 2004 through September 2008.

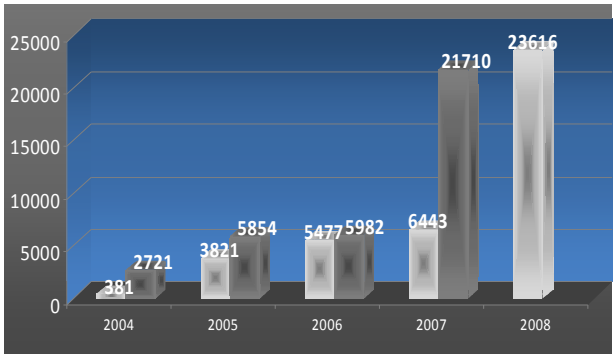


Figure 1 Number of transactions semester by semester from 2004 to 2008

Up from the first semester of 2004 - the starting point of the analysis - to the second semester of 2005, the trend in the number of transactions has always been on the rise, going from 381 to 5.854 transactions by the end of 2005. This process of steady growth has slightly decreased by the first semester of 2006 just to recover during the second semester of the same year and carry on till June 2008 - the month in which the analysis concludes. Taking into consideration the time scope that runs from the second semester of 2005 through June 2008, two different situations emerge. Up until June 2007 the process of growth in the number of transactions was slightly closer to a situation of stability, while in the second semester of 2007 witnessed acceleration tripled the number of transactions. The first half of 2008 marked another increase of the number of transactions; it may seem that, as in 2006, there was a movement towards a situation close to the stability but at a higher level, which in 2006 was around 6.000 transactions per year while in 2008 is around 22.500 transactions a year.

This reading shows a gradual learning process on the part of the supply and demand elements, leading to a stability, still, with a high growth potential of the market in late 2005, the date by which transactions had been stabilized until 2007, when a new Law (296/2006) introduced the compulsory membership to the system of conventions for Central Public Administrations (Ministers etc) and certain categories of private assets identified annually by the P.A. Electronic market. With the same budget, the P.A. Electronic Market became a binding instrument for purchases made by the central institutions. Hence the reason of the large increase in transactions since the second half of 2007 was clear.

The second illustration has still as dimensions the amount and number of transactions, divided into demand RFQ (request for quotation) and RFO (request for order), through an annual scope of time (Table 1).

RFO allows public institutions to purchase directly from the e-Catalog at pre-set prices. There is the possibility to choose a product from this catalog, by investigating the general conditions of contract, filling the order form (number, place of delivery) and signing the form with the digital signature that must then be sent to the supplier through the system. The contract is automatically and immediately a binding agreement between the parties in question; therefore the supplier is obliged to implement the agreement respecting the terms and conditions laid down.

RFQ is a more complex purchasing procedure. The RFQ is a competitive selection process through which public authorities make a request of supply to certain groups of suppliers or to all qualified suppliers so that they are able to make their bids. The suppliers in question should satisfy the money criterion and provide details related to the supply, including the technical ones from a quality point of view. The contract is awarded to those who fulfill the price-quality combination. The assignment of RFQ is carried out at the discretion of public authorities. They can, for example, pick suppliers that charge the lowest prices and promptly deliver on post-sale services.

	Numbers		Amount	
Year	RFO	RFQ	RFO	RFQ
04	2.520	426	4.391.716	4.947.443
05	7.721	1.253	12.699.547	17.144.210
06	9.862	1.597	12.920.627	24.971.498
07	23.393	4.760	30.593.134	53.021.577
08 (1°Sem)	19.199	4.507	20.861.319	42.528.345
Tot.	62.695	12.543	81.466.345	142.613.075

Table 1 Amount and number of transactions for RFO and RFQ

This table demonstrates that public institutions has been using since 2004 most of the RFO tool more than RFQ one. On the opposite the total amount of transactions through RFQ is decidedly higher than through the RFO, (just in 2004 it was slightly higher).

According to the characteristics of the two types of e-Procurement, RFO is more prompt and straightforward than RFQ, which explains why the number of transactions through RFO that supersedes the ones done through RFQ. Even if RFQ allows the making of requests on suppliers, it will require a greater commitment from both the requesting and the supplying party, which affects the element of immediacy. Public bodies seek to obtain from suppliers offers tailored to their needs and therefore buy more expensive products because of the greater focus in the purchase and willing to spend more money. RFQ transactions are of higher value than the RFO ones. Public bodies seek to obtain from suppliers offers tailored to their needs therefore linked to more expensive products. The greater focus is on the purchasing process and therefore will be at disposition to spend more money.

The third report refers to the number of suppliers in each geographic area considered for every semester from January 2004 to June 2008 (figure 2). Data were grouped by five geographic area (Central, Islands, North East, North West, and South). From the graph we see that the number of providers has been increasing in every geographic area over time. Considering the entire time horizon, the area with the largest number of providers has been the Central with a share of 37%, followed by the North West that has 24% of the number of suppliers then the areas of South and North East that each has 15% of the number of suppliers and last come the Islands with 9%. From the second half of 2007, as anticipated, there was a proliferation of suppliers and this is due to transactions occurring within that period as a result of regulatory requirements.

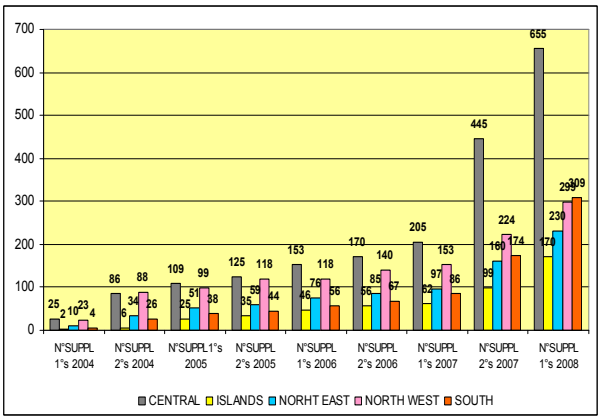


Figure 2 Number of suppliers per geographical area per semester

We can notice the growth in the number of active suppliers has been slow and steady up until June 2007 in all areas, with the Center and the North-West that stepped onto a higher level than the other areas due to the quick "sprint" made in the second half of 2005. During the second half of 2007, it is clear that it was a crucial period showing strong growth in the number of suppliers for each active area. The most significant growth occurred respectively in Central and Southern areas. The reason of the growth in the Central areas is linked to the compulsoriness of the public contracts below the EU threshold for Central Governments. The trend of growth of other areas, however, developed at a rate substantially lower (and similar) among them. Finally it is to be noted that the North West until the second half of 2005 has kept the pace of growth of the Center but paved the way after a period of stability in the first half of 2006.

The fourth report has as a dimension the number of active suppliers for each category with half-yearly flow (figure 3). Every catalog is defined according to the product it refers to. The product areas covered include ICT, Services, Health, Office supplies and "others". The total number of suppliers to be considered is 6248 that runs from January 2004 through June 2008. From the graph we see that the number of active suppliers has increased overtime for each catalog (apart from health). The product area in which the number of suppliers growth rate has increased less is Health that has been constant by that time ranging from nine to fifteen units; the second product area in which the number of suppliers growth rate has increased less is ICT (see also Figure 5). On the other hand, the office supplies area has been marked by the highest rate of growth. Other areas have instead seen both the number of suppliers and their growth rate increasing. Once again, the most growth rate was scored in 2007 (see also Figure 5). The total number of suppliers (83%) is active for Office Supplies and ICT catalogs (respectively 45% and 38%), with the remaining 17% divided among others (8%), services (7%) and Health (2%). Finally, it should be added that the services are considered from 2005 onwards, as there have been no previous data available and for the same reason Health is considered by the second half of 2004 onwards. In the following chart, catalogs have been separated from each other in order to highlight the flow in the number of suppliers during the semesters of the same period (figure 4).

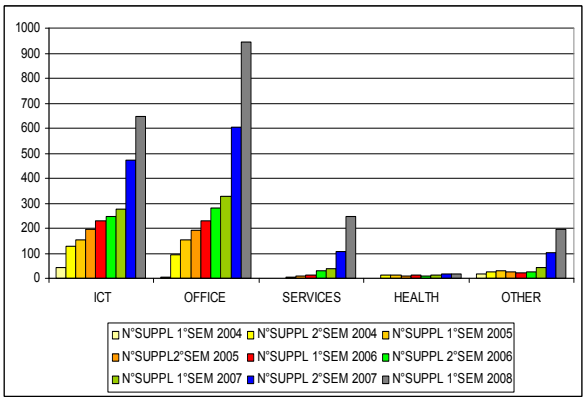


Figure 3 Number of active suppliers per category, per semester.

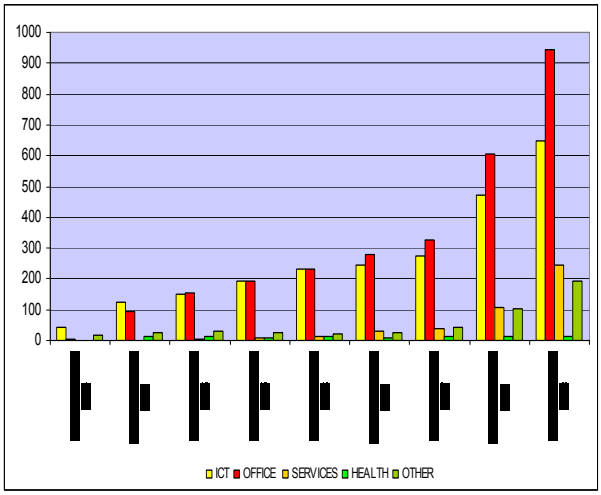


Figure 4 Evolution in the number of suppliers per semesters.

The national public bodies can be divided in Central Governments, 20 Regions, 109 Provinces, 8100 Cities, almost 276 Health Care Organizations, more then 50 Universities, Public Agencies, etc. In particular, the minimal e-transactions numbers and amounts of Health area (and related market number of suppliers) show us that the Public Health Sector and its Health Care Organisations are resistant to use the P.A. Electronic Catalogue.

The fifth illustration is the number of active suppliers catalogs by type (figure 5):

1. Mono-catalogue: suppliers that cover a single product area, are identified by the orange columns (one per catalog);
2. Bi-catalog: suppliers that cover two areas, are identified by the brown columns (one per each pair of catalogs);
3. Multi-catalog: suppliers that cover three or 4 areas, are identified by the blue columns (one for each triad of catalogs).

The graph shows that of 1.519 the considered suppliers, the major part (1.187) is operating in a single product area and thus provides a single catalog. 78% of the suppliers is mono-catalogue and focuses mainly on Office, ICT and (mostly related) Services. Bi-catalogue suppliers are the 19% that corresponds to the 288 suppliers of which 227 are involved in ICT products and Office. Only 3% of multi-catalogue suppliers (44 suppliers) are mainly engaged in ICT, Office and others.

The last illustration shows the suppliers divided in terms of regional, cross-regional and national areas (figure 6). It also shows the size of the firms which supply according to the number of employees:

⇒ Micro: less than 7 employees,

⇒ Small: from 7 to 13 employees,

⇒ Medium: from 14 to 62 employees,

⇒ Large: more than 62 employees.

You may notice that the number of suppliers on a national scale is rather small compared to those that operate on a regional or cross-regional. Providers operating at a national level represent 7% of the sample considered. The weight of the suppliers on a regional scale is 28%. 65% of the providers acted on a territorial area called multi-region, as it has embraced at least two regions. Such reasoning can be found in the size of firms that provide - especially micro and small - that are sufficiently large to provide for public entities located in at least two regions, uSUAlly neighboring, but not so much to cover the whole country. The micro-suppliers accounted for 60% of the supplying companies, small businesses 28%, those of medium size 8%, while large companies 4%.

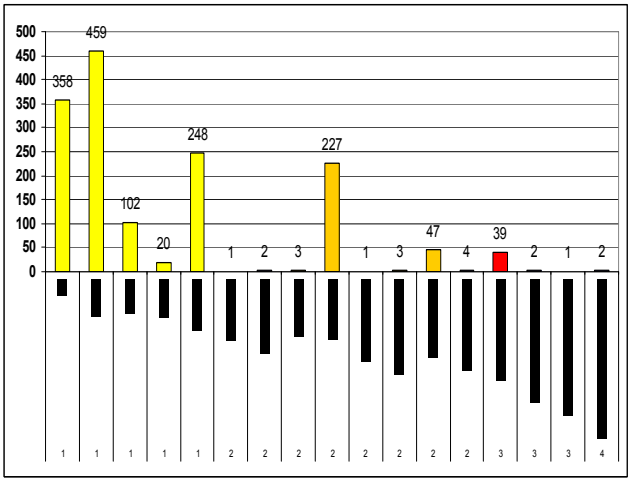


Figure 5 Number of active suppliers catalogs by type Over four and a half years of analysis, we do understand that the level its development was good passing from the initial 381 transactions of the first half of 2004 to 23.616 transactions in the first half of 2008. A major impetus to the growth was given by the Finance Act 2007 which forecast the mandatory use of the public e-market for purchases below the threshold of central institutions. The growth according to 2008 data is ongoing but at a rate far below that of 2007. In confirmation of what we just wrote is the fact that the number of active suppliers by geographic area and by type of catalog (excluding the health area) has steadily increased in these four years and a half with a decisive leap forward during the second semester of 2007. The area that has made the greatest number of transactions is the Center in that it spent the highest amount. The catalogs are the most used of ICT and Office Supplies. Health care

In addition, it is noteworthy that the Electronic Marketplace offers member administrations the possibility of concluding contracts through two different procedures for the purchase: the direct order of purchase (RFO) and the request for tender (RFQ). The data analysis shows that the number of transactions through the RFO is higher than that the ones through RFQ but that the amount of transactions is higher in the purchases made through RFQ. The total amount of purchases through RFQ in the reporting period was 1.75 times greater than the value of purchases through RFO. This is also confirmed by the fact that the average value of transactions through RFQ is significantly higher than that the ones through the RFO.

It should also be noted that 78% (1.187) of the considered suppliers (1.519) is monocatalog and focuses on Office, Information Technology and Services. Bicatalog Suppliers is 19% which corresponds to 288 suppliers of which 227 involved in ICT products and Office. Only 3% of suppliers (44 suppliers) are multicatalog and mainly engaged in ICT, Office and Others. The conclusion is that most served catalogs are those relating to office supplies and ICT and that competition among suppliers in more than one catalog is fairly limited.

In Italy during the second half of 2007 – significant year due new legislations - suppliers were distributed among 60% of micro-sized companies, 28% of small-sized companies, 8% medium and 4% larger firms. Furthermore 65% of these companies operated on a cross-regional level, 28% on a single region while the remaining 8% on a national scale. Most companies were small and operated in most regions. A result in line with the tradition of the Italian market made up of small and medium enterprises.

VI. THE CALABRIA REGION CASE- STUDY

Calabria is a region of Southern Italy, normally represented like “the tip of the boot", with a population of about 1.9 million inhabitants, (out of the 60.millions of Italy) and with its administrative capital, Catanzaro, placed between two seas (Ionian and Tyrrhenian). In Calabria, as well as in Sicily and Campania, the two closest Regions, underworld and organized crime play a relevant role, also impacting on public procurement activities.

The fight against organized crime is one of its policies key pillars. A priority is given by the attempt to exclude from such public procedures their presence. In Calabria, in particular, the fight has been conducted trying to identify ways of managing public procurement which exclude criminal interference. In 2007 the Region Institution, with special regional law, established the Station Unique Contracting Authority, a public body which, in intention, should take care of all the public contracts of its administrations (HCOs, Region, Cities, Provinces etc), including contracts for goods and services.

The purpose of this approach is, moreover, to optimize the processes, to create economy of scale by purchasing common and standardized goods/services, and so doing increasing the control of criminal infiltration.

A. The Single Station Contracting (SUA - Stazione Unica Appaltante)

SUA has been established in 2007 and has begun to take steps, through an observatory of contracts in the region, both through the direct management of the contracts: normally this is done or at the request of a specific entity for a particular contract (the SUA provides its structure to manage the contract on behalf of the Entity) or by pooling the needs of various organizations for specific product categories (the SUA manages a contract for the sum of the needs of various entities).

As often happens in the low performing Italian areas, however, the allocation of resources, has been limited and, accordingly, the results, based on professionalism and dedication of officials, are lower than those that we had in the prefixes' institutions.

Procurement of ICTs is handled partly via traditional (paper) and in part by on-line purchase, with different platforms.

"Albofornitori" (now Net4Market) provided by CSAmed Company, has been widely used by the SUA, especially in the Health Sector.

B. The results

The SUA led from 2009 to 2014, seven races of considerable economic importance, with the platform albofornitori (literally ....), For a total of about 1,440 million of Euros for a total of approximately 4,800 lots.

Among other stand out, for particular conditions, as the amount per race (between 400 and 700 million), the number of lots (well above 1,000) and by the number of participants (more than 100 per race), the regional competitions drugs for all healthcare organizations in the region.

It is currently an ongoing SDA drugs, whose values will be in addition to those already reported.

The SUA decided to use for this new purchase drugs System Dynamic Acquisition, on which the European Union is pushing a lot, in order to avoid extreme market dynamics of drugs (much needs changing, patents expire and companies that change structure very easily).

The new SDA was posted on 27/11/2014, it lasts three years, covers a demand for drugs provided about € 140 million, although it is expected that this figure will grow significantly.

The Calabria Region has instituted a Commission Station (SUA – Unique Contractor Station) whose aim is to manage from a central office the tenders of the regional authorities, in particular as far as the healthcare enterprises are concerned. The main goals of the SUA institutions are the governments' processes optimization, the economies of scale and the fight of corruption..

During the last years, the SUA has convened some tenders for the purchase of drugs of all the healthcare enterprises of the Region: in this case study we are going to analyze the very last one. Its management complexity is related to:

– the necessity of coordinating in order to focus on the needs concerning the type of drugs as well as their amount

– the total number of drugs to be purchased (2.138 different types) and the whole budget for the tender (745.377.630,69 euros)

– the number of participating enterprises (127)

A tender like this uSUAlly lasts from ten to sixteen months, adopting the traditional technique.

A special procedure was adopted in order to collect data: it allows every healthcare enterprise to insert the amount of the previous drugs and to add the new ones; this phase lasted more or less two months and included the discussion concerning the healing features and their eventual overlap. After that the call for bids was electronically published (having the documents been prepared before), the forum was activated and 127 suppliers were accredited. Afterwards the offers were electronically collected through the "electronic sealed envelope". The procedure is composed by the following 3-? 4: phases

- First , …..and the first ranking was drafted.

- Second: a second round of offers' collection: these offers were supposed to improve the previous ones.

- Third: offers and drugs were compared and the third ranking was finally published.

After that, the technical paper forms were deposited at the SUA.

This whole phase lasted more or less 160 days: the whole tender therefore lasted more or less seven months. Moreover, all the documents were automatically producted and all the contracting data were automatically sent to every commission station.

The security question is really important in tenders and was solved by using the "telematic sealed envelope©": it is based on operations temporarily articulated that make the offer of every supplier 100% secure: in fact, there is no algorithm that can fully protect a file from being read; it can only decrease this likelihood. On the contrary, this system is not based on algorithms, but on a timing, and therefore it is useless to know the contents of the various offers. This opportinuty is deeply important in such a suspicious environment.

With this tender the 84% of drugs were awarded a contract (1.801 lots awarded upon 2.138 tendering), while the average percentage is settled on the 70%. The cumulative purchase prices were decreased of 14,60%, which corresponds more or less to 100 million of euros.

VII. CONCLUSIONS

Following EU and EU countries' approach, rules and orientations, ICT tools have been empowering their role at a European level to facilitate the match the efficiency and the transparency between demand and offer (e.g., Simap website, Electronic Tendering, Directives to stimulate the use of ICT, the public process digitalization etc.); the more the public service impact on individuals and communities is relevant (for example, health care services, environmental issues, creation of job opportunities), the more public procurement rules and ICT tools should be effectively managed.

Moreover the EU harmonization process can face difficulties depending on the different national context and the political, juridical, cultural and institutional profile of each country. The EU countries public procurement models are basically identified by their grade of centralization/decentralization and by the extension of their public mandate (i.e., their scope and if it is mandatory or not), and on the different typologies of goods/services. Many EU countries adopted a model with a Central Procurement Department aiming to purchase goods and services for the public bodies. The Italian one is one of the most active and representative. Still it seems that is not sufficient as being just a part of the plans to review public spending, e-Procurement entities and ICT Software Companies have now a key role in both modernization and increasing the efficiency of both monitoring of government accounts. The importance and impact that computerized purchasing tools not only stems from the fact that public procurement accounts for about 16% of GDP (on average in EU countries), but they can also bring benefits into the management of the P.A. evolution. The P.A. Electronic Market, the virtual site for demand & supply in the form of structured and dynamic electronic catalogs, for contracts below the threshold of Community importance, is undoubtedly among the most advanced B2G tools. In Italy it has been launched in 2003 and it has now e-Catalogs with over 400,000 products with technical data sheets on which they can be searched, with automatic, direct purchases and requests for bids through the use of digital signature. Additionally, these systems also enables the management and archiving of digital streams (the contracts and bid requests that occur between P.A. and Suppliers). They are a tool that offers no pre-negotiated bids, and allows local PMI to qualify and get in touch with customers who maintain complete autonomy to choose among different suppliers and products, because the same product can be sold by several vendors, as in any traditional market.

On the base of research results, we understand how E-Procurement Central Bodies can be a guiding tool for the promotion of organizational and procedural P.A. changes which impacts and provides opportunities not only on the part of public demand but also supply, still they need to identify ICT Companies with the capacity to help them in covering strategic and potentially key role in the welfare context. That means to identify first the specific areas, (geographical, technical, merchandising, functional, service areas) on which they may need assistance in order to develop partnership with small-medium Softaware Companies that since the beginning of their life have the vocation for understanding public clients' needs and for innovation.

The picture that has been traced thus highlights what is now the potential role of E-public markets in purchases under the threshold of the whole public administration, especially in light of the requirements dictated by legislations such as EU Directives, the various EU counties Codes of Contracts and the Finance Acts.

The research also shows us that the Public Health Sector is resistant to use the P.A. Electronic Catalogue. A reason, to be anyway more investigated, can be the high relevance of quality factors for health products, and the consequent reduction of price (and administrative costs) relevance; public health purchasers consider e-procurement solutions inadequate for selecting quality. More in general, Public bodies which seek to obtain from suppliers offers tailored to their needs, are willing to select quality aspects and, therefore, either not to be so much price-oriented or to acquire more expensive products.

Finally we can state that the need to reduce public expenditure, administrative costs and timing, and to increase transparency led to new public procurement institutional models; together with the ICT evolution a relevant institutional change may be connected to e-procurement solutions and realization of ICT platforms. The Italian situation show us a high growth potential connected to these models and solutions, even if strictly related to juridical changes (e.g. compulsoriness of public contracts).

Future steps can be the analysis of the e-transactions by Government typology to check differences among non-Central Administrations (Local, and Public Health Sector Administrations, Universities etc), and understandings of the potential margins of growth in the different market sector also depending on political, cultural, economic and organizational drivers and barriers.

VIII. REFERENCES

[1] Andersen, K. V., "E-government and Public Sector Process Rebuilding (PPR): Dilettantes, Wheelbarrows, and Diamonds," Kluwer Academic Publishers, Boston, 2004, pp. 1-223.

[2] European Parliament Directive 2004/18/EC and 2004/17/EC

[3] Gichoya .D (2005) "Factors Affecting the Successful Implementation of ICT Projects in Government" The Electronic Journal of e-Government Volume 3 Issue 4, pp 175-184

[4] Henriksen, H.Z., Mahnke, V. and Hansen, J.M. (2004) "Public eProcurement adoption: Economic and political rationality", Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37), IEEE Computer Society, p50124c.

[5] Heywood J.B, Barton M. and Heywood C. (2001) E-Procurement: A Guide to Successful E-Procurement Implementation, Financial Times Prentice Hall

[6] Hoaglin, D. et al. (1982) Data for Decisions: Information Strategies for Policy Makers. Cambridge,

[7] Kalakota R. and Robinson M. (2001) E-Business 2.0: Roadmap for Success. Boston: Addison-Wesley

[8] Konicki S. (2002) E-Sourcing's Next Wave, Information Week,March 18th

[9] Maniatopoulos, G. (2004) "Enacting E-Procurement Technologies within UK Local Authorities", Proceedings of the IADIS International Conference e-Society 2004 (ES 2004), Vol 2, IADIS Press, pp850-854.

[10] Minahan T. and Degan G. (2001) Best Practices in e-Procurement, Boston: Aberdeen Group

[11] Piga G., Zanza M. (2004), An exploratory analysis of public procurement practices in Europe, Quaderni Consip, I,

[12] Somasundaram R. and Damsgaard J. (2005) Policy Recommendations for Electronic Public Procurement" The Electronic Journal of e-Government Volume 3 Issue 3, pp 147-156,

[13] Tonkin, C. (2003) "e-Procurement in the Public Sector: Story, Myth and Legend", The Policy Institute, Trinity College Dublin.

[14] Zulfiqar K. A. et al. (2001) "E-Government: An Exploratory Study of On-Iine Electronic Procurement Systems", Proceedings of the 9th European Conference on Information Systems (ECIS 2001), AIS, pp1010-1022

[15] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[16] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[17] K. Elissa, "Title of paper if known," unpublished.

[18] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[19] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[20] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

PAPERS

Microeconomic and macroeconomic overview of ICT as a primary driver behind the digital economy

Saša Aksentijević[1], Edvard Tijan[2], Gorana Stumpf[3]

[1] Aksentijević Forensics and Consulting, Ltd.
Gornji Sroki 125a, Viškovo, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr
[2] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr
[3] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: stumpf@pfri.hr

Abstract - Widespread implementation and ubiquitous beneficial impact of ICT technologies is often praised as a decisive factor demarcating the frontier between "old" (classical), and "new" models of digital economies. However, exact quantitative analysis of such impact on a micro and macro level shows some surprising results. No unique conclusions could be applied to all countries and in all timespans. ICT technologies are subject to the law of diminishing returns more than any other, their impact is notoriously difficult to measure and carries a number of hidden and sunk costs and transfer effects. Furthermore, there is a distinctive difference in yields derived from ICT technologies in national economies depending on their use or development and production. On a microeconomic level of a single company, only investments in those ICT technologies that aim to preserve and increase value of the information capital can be economically justified. In this paper, the authors will research and elaborate the impact of implementation of ICT technologies on the corporate level and aggregate measures of national economies.

I. INTRODUCTION

ICT is an important part of the fourth, quaternary sector of the national economies, along with state and society management, culture, education and science. However, none of those activities are usually attributed with such intrinsic characteristics in regard to their possibility to positively influence aggregates of the whole national economies. Macroeconomics and ICT experts are well aware of the fact that ICT possesses a number of specific characteristics that have to be taken into consideration prior to reaching final conclusions on the subject issue.

In order to analyze the impact of ICT technologies as a driver of businesses and modern (digital) economies, it is important to study and outline characteristics of the quaternary sector of the national economy where ICT dominantly operates, to describe the impact of ICT on a company level, and then translate it to aggregates of the national economies.

II. ICT AND QUARTENARY SECTOR OF THE NATIONAL ECONOMIES

The main determinant of the quaternary sector is the usage of knowledge and skills in the final product, rather than raw materials, products or simpler forms of services. It is tightly connected with intellectual services and further with use of ICT, research and development in scientific fields and consultancy. Some theorists also include pure services (those services that are the highest form of services provided to end consumers in the marketing continuum, facilitating direct contact to customers) in this sector [1].

Quaternary sector has developed inside the process of further evolution of the three-sector hypothesis of industry, developed by Alan Fisher, Colin Clark and Jean Fourastié [2], and in fact it is a derivative of the tertiary (service) sector presented in that model. These sectors are illustrated in Figure 1 where percentages of a country's economy made up by different sectors based on its level of income or development are shown. The primary sector operates primarily on activities of extraction of raw materials. The secondary sector consists of activities converting the raw materials into manufactured goods. The tertiary sector is the provision of services rather than the manufacture of goods. Countries that have achieved higher levels of socio-economic development tend to have less of their economies made up of primary and secondary sectors and relatively more employment in economic activities in tertiary sectors (and consequentially, quaternary and quinary sectors) [3]. Underdeveloped and developing countries exhibit completely inverse pattern: majority of employees work in primary and secondary sectors of national economies. Consequentially, that is where majority of economic activity happens and most of GNP is created there instead of sectors with higher percentage of value-added economic activities. These facts are shown in the following Figure 1.

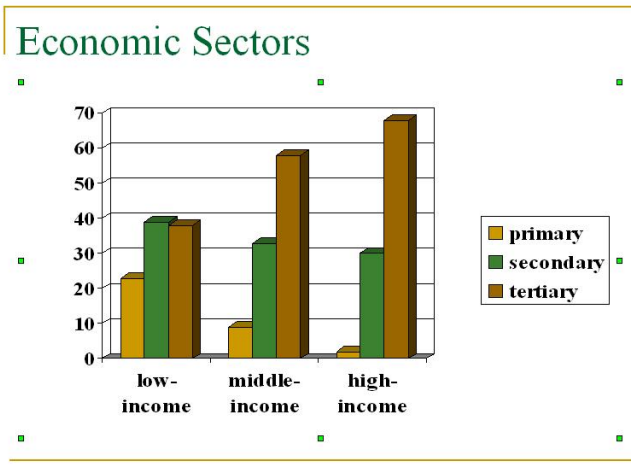


Fig 1. Economic sectors and income (adapted by the authors according to [4])

Generally, the emergence of quaternary economic sector requires significant levels of highly educated workforce. In developed countries, tertiary and quinary sectors usually make up the predominant part of the economy and employ between two thirds and three quarters of the workforce. On the microeconomic level, companies generally view investments in this sector as one of the possibilities of further expansion because returns on investments and margins are generally higher in this sector.

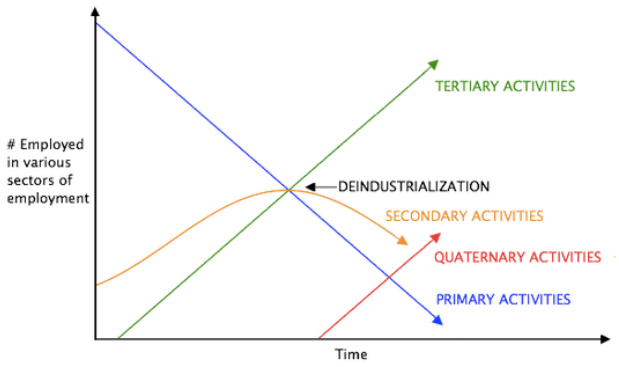


Fig 2. Visual representation of dynamics of modelled activity changes in relation to increase of tertiary and quaternary activities and decline of primary and secondary activities after peak of deindustrialization (adapted by the authors according to [5])

In Colin Clark's sector model of an economy during technological change shown in Figure 2, increase of the share of quaternary sector in national economies accelerates after the peak of the deindustrialization process and the decline of primary and secondary activities in the national economies (especially sharp decline of primary ones).

III. POSSIBILITIES OF MEASUREMENT OF INFLUENCE ON BUSINESSES AND NATIONAL ECONOMIES

Generally, the use of ICT technologies is considered to carry along significant levels of competitiveness, increased levels of employment and positive contribution to all economic measures. It is important to notice the verb "use" in previous sentence: very often, colloquially there is no distinction between "usage" and "production" or "creation" of ICT technologies; even the mere proximity of ICT technology to businesses is regarded as inherently positive. In this article, some clear delimiters will be set towards these attitudes, with the major goal of enhancing further considerations on the subject and fully understanding the fact that the impact of ICT technologies on national economies is a very complex topic, not just a sum of partial and superficial effects.

There is a total of ten considerations related to the influence of ICT technologies on the microeconomic level and the level of national economies that will be elaborated in the remainder of the paper.

1. *There are no unique conclusions about the influence of ICT technologies on national economies – conclusions can be made only based on analysis of individual national economies or groups (clusters) of similar economies.* Since the middle of the 1990s, factors of the growth productivity between Europe and USA have been divergent. The productivity rate in the USA has been growing constantly while in Europe it has been on the decline. The impact of use of ICT technologies on GDP growth and Total Factor Productivity (TFP) in the emerging and transitional economies is lower than in the developed countries [6]. Paradoxically, in some countries the increase of investments in ICT technologies and growth of GDP shows negative correlation. Furthermore, the impact of increased use of ICT technologies is different in various groups of countries, economic activities and in different timespans (decades).

2. *The impact of ICT on GDP and other aggregate indicators is notoriously difficult to measure.* On macroeconomic level, the fact that automatization of the businesses up to 1970s has caused measurable increase in productivity and direct transfer to productivity measures of the national economies is a well recognized fact. Some of the tools used to achieve this automatization were use of telephones, electronic calculators, fax machines, mainframe

computers for mass data processing and the use of vacuum transport tubes. However, this effect is not present in developed economies after 1970s. This seemingly contradictory concept is also known as Solow's paradox. Robert Solow, an American economist and Nobel prize winner, particularly known for his work on the theory of economic growth that culminated in the exogenous growth model, is also known for his phrase: "You can see the computer age everywhere but in the productivity statistics" [7]. There are several potential explanations for this paradox, one of them endorsed by Solow himself: the impact of ICT technologies on total factor productivity is difficult to measure. Other possible explanations are that ICT brings benefits on the microeconomic (company) level, but also adds cost to other stakeholders on the market, so total gains are small (or can be even negative). ICT is very complex, difficult and expensive to manage, so summary benefits are lower than initially and often subjectively expected. Furthermore, the decreased need for human labour caused by use of computers negates potential positive effects on the macroeconomic level [8].

3. *The impact of ICT technologies is visible with latency, and carries along significant sunk and hidden costs* [9]. This fact is aligned with problems of ICT technologies impact measurement: not only there is pronounced problem with multicolinearity and heteroscedasticity of variables used in quantification of impact of ICT technologies on gross national product, but the efect of their use is visible only in the long run. At the end of the investment cycle, significant new investments are required that are usually equal to the value of the initial investment, caused by technology's obsolescence. During the investment cycle, the use of ICT technologies carries significant operative costs (usually, it takes 4 to 5 years in current state of technological affairs for accumulated operative costs to reach the value of the initial capital investment).

4. *The use of ICT technologies is subject to the law of diminishing returns, both on the macro and microeconomic level.* "More" ICT does not necessarily mean "better", both in businesses and national economies. The question of optimum determination is always a very complex one. Economists typically tend to resolve it by applying an appropriate mix of labour and capital in the equation. Some go even a step further and claim that ICT is not a separate factor influencing the economic efficiency frontier, rather it can be attributed either to work or labour. There is a certain level of investments in ICT and operative costs where new additions do not produce added effects. Non- selective and literal application of the best practices and compliancy measures will certainly not create new ICT giants from micro businesses, and neither will indiscriminate investments in ICT technologies and infrastructure of national economies propel them to new levels. Every economy and business requires just exact amount of ICT investments, technologies and services and not less or more than that. Absorption capacity for ICT technologies is therefore quite rigid, and the only exception to this rule is production of ICT services and goods destined for export [10].

5. *ICT can be either used or produced: for national economies it is more beneficial if it is used.* This hard fact is not the best accepted among those who are proponents of indiscriminant increase of the native production of ICT technologies within national economies. In Figure 1., impact of the output and use effect on long term growth of GDP in a selected panel of countries is shown.

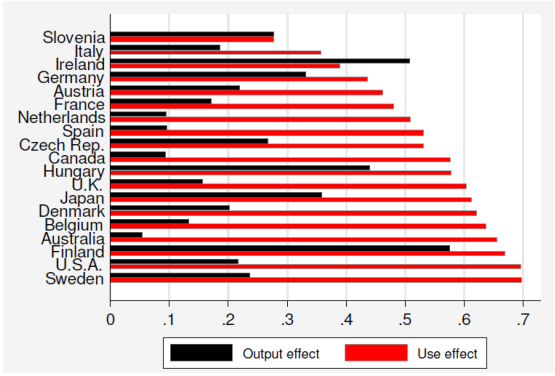


Fig 3. Effects of output and use of ICT technologies in a selected panel of countries [11]

From Figure 3 it is clearly visible that effects of production are less pronounced than use effects. Only in one country (Ireland), and caused by specific local conditions of

extremely high direct foreign investments in ICT prior to crisis eruption in 2008, the situation is inverse. In all other countries, benefits from ICT use compared to ICT production are much higher. In the case of USA or Great Britain, this ratio reaches approximately 1:3.

6. *Impact of ICT is visible in statistics of other parties (sectors and states).* Impact of ICT is not that visible in statistics of those countries that produce those technologies, but it spills over to statistics of other countries, with certain latency. Because of this reason, in macroeconomic terms ICT is considered to be so called GPT – „General Purpose Technology", a type of supporting or adjunct technology to other core processes and technologies [12]. This explains the already mentioned fact that increased levels of ICT investments in developed countries have resulted in decrease of total productivity because additional resources have to be engaged in implementation, reorganization or learning related to the adoption of new technologies.

7. *In most OECD countries, education in use of the new ICT technologies is the main cause behind their falling behind trends of business use of the same technologies.* Analysis of this apparent lag of the EU as a whole in comparison to USA shows that three main contributing factors can be identified:

   a) cost related to users' induction and implementation of systems,
   b) education, and
   c) lack of regulatory flexibility.

   These contributing factors are shown in Figure 4. for some selected developed countries.

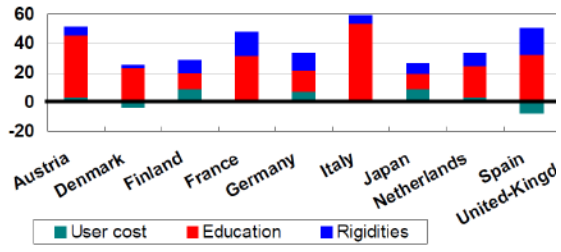


Fig 4: Factors contributing to ICT lag (adapted by the authors according to [13])

These three factors can be explained by a lower level of employed structural and organizational capital for use of ICT technologies in the EU compared to the USA. In most European countries, the main cause of the lag is education of the users, implying neccessity of additional investments in employee education and lifelong learning as support activities for successful implementation of the new ICT technologies in businesses. However, inflexibility of regulative requirements (laws and legislation related to ICT) also shows significant impact on the widening of this gap. More flexibility in law-making related to ICT, but also the labour legislation flexibility, has positive transfer effects to national level implementation of the new ICT technologies.

8. *Outsorcing enables focus to key areas.* Research that encompasses several countries and all sectors of national economies shows that those countries that use outsorcing of ICT services benefit from the increase of labour productivity in ICT intensive industries [14].

9. *Quantitative methods and econometrics can be used as a part of decision making process in ICT industries, especially related to telecommunications.* Research shows that 1% increase of broadband penetration in a general population of OECD countries between 2002 and 2007 caused direct GDP growth of 0,025% [15]. Therefore, it has been proven that on a given set of data, 40% increase of broadband penetration results in 1% increase of GDP through transfer mechanisms. In the years that have passed since this study, there has certainly been more saturation of the telecommunications market, so elasticity might have changed, but this is just an example of deployment of econometrics in decision making in ICT industries and creation of guideline for policy-makers on national levels.

10. *Conclusions related to the impact of ICT on microeconomic level cannot be directly translated to macroeconomic level.* If companies in two comparable national economies equally invest in ICT technologies, but differ in ability to introduce complementary organizational changes in order to optimally use them, increase in productivity that can be attributed to those investments will differ from one company to

another [16]. Consequentially, difference of ICT investment impact on productivity on individual companiy levels translates to macroeconomic level (national economies) in lower overall influence than the one measured or overall percieved for individual companies. Those companies that are lagging behind will either disappear from the market or catch up with respective niche leaders. However, total sector and national productivites reflect effects on both of those categories of companies: laggards and leaders.

Despite all the presented facts that might seem unexpected or contrary to the popular opinion, ICT is nevertheless a propulsive sector of national economies that proved to be very resistant to downwards trends of business cycles in the last decades. However, ICT possesses certain inherent specific characteristics and does not function well on a national level if regulatory framework is rigid, levels of employed regulatory capital are low, levels of implementation of ICT technologies are inadequate on a firm level, labor regulation is too strict and investments in employee education and training for use of new technologies is reduced.

IV. CONCLUSION

ICT technologies are typically a part of the tertiary sector of national economies and constitute most of the quaternary sector. The emergence of activities within this sector follows the peak of the process of deindustrialization and sharp decline in volume of activities in primary and secondary sectors of national economies.

However, there are some inherent specific characteristics of the ICT technologies, services and ICT based industries that separate them from other activities in the new, "digital" national economies, so it is not possible to analyze them by using the same toolbox of economic analysis that was used for previous, classical models.

The impact of ICT on national economies has to be analyzed inside the context set by overall timeframe, state of the economy and current model (transitional, mature, developed or underdeveloped countries). Usually, the countries in transition and those with underdeveloped technical penetration of ICT technologies benefit more from additional use than countries whose economies are in a more mature, developed state. However, it is important to underline that even between comparable state economies, those countries that are able to accommodate and adopt more ICT technologies because of higher deployed levels of structural capital will benefit more than countries and economies where the amount of engaged structural capital is comparably lower.

The use of ICT technologies exhibits significant spillover and transfer effects, so its impact on the levels of national economy macroeconomic aggregates is very difficult to measure, however it is indubitably tied with laws of diminishing returns, sunk and hidden costs, and especially those costs related to increased spending for education and implementation.

ICT technologies can be produced or used (implemented) to enhance business processes. National economies almost always benefit more from the use of ICT technologies than from their production. Because of aforementioned transfer effects, some costs and benefits of ICT technologies may spill across sectors or borders, so they may be present in balance sheets and accounts of other stakeholders, sectors and nations. However, it has been clearly demonstrated that econometrics and quantitative methods can be used for analysis of this phenomenon, so scholars should not be vary of deploying them in the field of analysis of the quaternary sector, while being very cautious in selection of the methods and interpretation of the output results.

REFERENCES

[1] http://www.differencebetween.net/business/difference-between-pure-and-classic-service/ (accessed 04.11.2014.)

[2] Schettkat, R, Yocarini, L: "The Shift to Services: A Review of the Literature", Discussion Paper Series, Institute for the Study of Labor. IZA DP No. 964, December 2003, p. 7

[3] Staroske, Uwe: "Die Drei-Sektoren-Hypothese. Darstellung und kritische Wuerdigung aus heutiger Sicht", S. Roderer Verlag, Regenburg, 1995, p. 443

[4] http://www.genesoc.com/introtosociology/IntroductiontoSociology.zip (accessed 04.11.2014.)

[5] Clark, C: "The Conditions of Economic Progress", Macmillan, London, 194, p. 125

[6] Havik, K. et al.: "The EU-US total factor productivity gap: An industry perspective", European Economy Economic Papers, Brussels, September 2008, p. 5

[7] Acemoglu, D: "Return of the Solow Paradox? IT, Productivity and Employment in U.S. Manufacturing", AEA Papers and Proceedings 2014, December 2013, p. 1

[8] Dreyfus, E. et al.: "The IT Productivity Paradox", http://cs.stanford.edu/people/eroberts/cs181/projects/productivity-paradox/background.html (accessed 04.11.2014.)

[9] Remenyi, D., Money, A. Bannister, F.:"The Effective Measurement and Management of ICT Costs and Benefits", CIMA Publishing, 3rd edition, 2007, p. 138

[10] Carlsson, S.A.: Enabling and enhancing potential absorptive capacity through the use of ICT",

The Fifth European Conference on Organizational Knowledge, Learning, and Capabilities, University of Innsbruck, Austria 2-3 April 2004, p. 5

[11] Oulton, N.: "Long term implications of the ICT revolution for Europe: applying the lessons of growth theory and growth accounting", ICTNET 2nd Workshop on ICT, Intangibles and Innovation, London, April 11-12.2010.

[12] Hempell, T.: "Computers and Productivity: How Firms Make a General Purpose Technology Work", Physica-Verlag, Heidelberg, 2006, p. 21

[13] Cette, G., Lopez, J: "ICT Demand Behaviour: An International Comparison", ICTNET 2nd Workshop on ICT, Intangibles and Innovation, London, April 11-12.2010.

[14] Lemmens, K.: "Recommendations for IT procurement of non-commodity outsourcing: a conceptual framework", Master Essay, University of Twente, p. 1

[15] Katz, R.: "Impact of broadband on the economy", Columbia Institute for Tele-Information, Columbia University, 2012, p. 4

[16] Aksentijević, S.: "Makroekonomija, ICT i Hrvatska", 23.06.2014., http://www.ictbusiness.info/kolumne/makroekonomija-ict-i-hrvatska (accessed 04.11.2014.)

Application of service oriented architecture of information communication system in nautical ports

Pančo Ristov[1], Ante Mrvica[2], Jelena Žanić Mikuličić[1]
[1]Pomorski fakultet Sveučilišta u Splitu,
[2]Jadrolinija d.o.o.
panco.ristov@pfst.hr, ante.mrvica@jadrolinija.hr, jzanic@pfst.hr

Abstract - Information and communication systems in support of management to manage all business processes in the ports of nautical tourism, which means unification of business processes and information and communication technologies through the use of various methods and techniques for developing, managing and analyzing business processes involving human resources, organizations, applications, documents, and various other information as an integral part of the business process in nautical ports. The introduction of Service Oriented Architecture into existing or new information systems in the ports of nautical tourism, and ways of thinking that it offers new possibilities in terms of expansion of logistics solutions, lowering costs and increasing the quality of service end users - the sailor. Service-oriented approach is no longer projecting monolithic or multi-layer applications, but services that are autonomous with the implemented security, transaction processing and interoperability with other software services in a distributed environment.

I. INTRODUCTION

In Croatia, there are 56 marinas which are member of the Association of marinas and have over 16,000 berths and over 8,500 dry berths. Nautical ports (LNT) as traditional providers of services for boaters in recent times, are forced by the increasing competition in the tourism market, as well as by increased costs caused by the introduction of new services and raising the level of existing services. They look for new solutions and services in order to increase productivity and reduce costs. One of the solutions that are proposed is introducing the concept of orientation to the service. This term, service-oriented architecture represents a model in which the logic of the system breaks up into smaller, separate pieces of logic. These smaller parts can be independently distributed across the network.

In light of the major changes and the impact of information and communication technology on the services and strategy LNT - matter is primarily related to the globalization of business, reengineering business processes, speed of response to market conditions and the connections between all LNTs and other business entities.

The application of information and communication technology enables easier management of all resources, monitoring and controlling of all processes to improve business quality and forming services according to the wish of yachtsmen.

This paper analyzes the use of information and communication system (ICS) based on architecture-oriented services in the management of LNT at all levels of management.

Service oriented solutions are not closed, nor oriented only to one vendor of software equipment. Service oriented solutions can be integrated into different environments, different platforms and save big financial resources. There is no need to change existing software, and it is possible to exchange data between the new and old software. The introduction of the service oriented solutions in the LNT is not only a temporary solution and the aim, but can be seen from the perspective of the overall strategy, especially with the marketing plan of the marina.

2. SERVICE ORIENTED ARCHITECTURE

Service Oriented Architecture (SOA) is actually a new perspective on software architecture that defines the use of services to achieve the requirements set by the user. It is important to note that SOA does not have a universal definition valid. One definition that is consistent with the work, defines SOA as "software architecture for building applications that business processes or services, by using a set of loosely coupled components box, adjusts to provide a satisfactory level of services" [2]. Also, the SOA provides that any logic or data can be used in various ways, depending on who uses them. SOA is primarily arisen from the need to improve and rationalize business systems, which will reduce costs and increase company profit.

Architecture is subjected to changes that are often required until the structure can be observed at different levels in the business system, especially in the accompanying ICS. As the process, architecture is a set of activities of designing and constructing some physical or virtual structure but includes documented occurrence. The service is an essential element of any system based on the principles of SOA. SOA system actually provides a service that comprises one or more elemental services, often called components. The components are small

logical units that perform an operation. SOA components must ensure the level of abstraction of the business logic so that: *the message* indicates the information required to do the work or any part thereof, *operation* represents the logic required to process messages in order to carry out the unit of work, *the service* is a logical set of operations able to perform related work units and *process* contains business rules that determine which services are used to perform some task. Simplified basic concept of SOA architecture is shown in Figure 1.

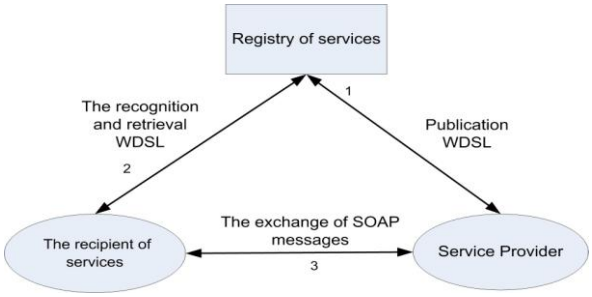


Figure 1. Main concept of SOA architecture

Figure 1. shows the way of work of Web services. The numbers on the axes indicate the order of execution of the determined steps. Step No1-service provider publishes a description of services. In step No2 a particular service on the Web is located. Step No3 is usually called a "connection with the service", and it allows SOAP (Simple Object Access Protocol).

SOA represents more than set of technologies. SOA is not directly affiliated with any technology, although it is often implemented by using Web services. Web services represent the most appropriate technology for the realization of SOA. According to available literature, the most important components of the SOA concept are: services, interfaces, messages, synchronous and asynchronous communication, registers (service provider), and the composition of services in the business process. SOA is primarily arisen from the need to improve and rationalize business systems, which will reduce costs and increase company profits. From the organization science it is well known that the business system can be divided into two levels: the level of business processes and application levels. The business level is process structured performance of business requirements along with the limitations, interconnections and external influences on these requirements. An application level solution automates business processes across different technologies. Application level contains solutions required for studying and successful flow of business processes and is usually realized through applications based on suitable technology platforms. SOA introduces the third level, the level of service interfaces. This provides the possibility of encapsulation and application and business level. Applications are separated through the platforms (.NET, J2EE, etc.) for which they are intended and reveal individual components, which now services can use through the interface. By SOA architecture, service is an essential element. SOA system actually provides a service that comprises one or more elemental services, often called the components.

To design a business process that enables the use of services by different service providers created are specialized languages that make it possible. The basis for all languages is an XML (Extensible Markup Language). Software companies that deal with this technology have become united and they created a Process Execution Language for Web Services (BPEL). BPEL is a grammar based on XML to describe control logic necessary for the coordination of services participating in a business process. Orchestration mechanism executes that grammar, coordinating and compensating the whole process in case of errors. BPEL process defines and describes the coordination of services with partners in order to achieve business objectives as well as state and the logic necessary for coordination.

Web services represent a technology that is relatively quickly and widely accepted as a way to integrate applications. Web services are based on open technologies such as:

- XML has emerged from SGML (*Standard Generalized Markup Language)* and is designed especially for Web documents. It allows users to create their own tags, allowing the transfer, validation and interpretation of data between applications and between organizations. Today is an XML language very widespread and is used for different purposes: the separation of data from presentation, data exchange, data storage, increase of data availability and creation of new specialized markup languages. XML provides an efficient, flexible and simple solution for electronic business.
- *SOAP (Simple Object Access Protocol)* is a protocol designed to transfer XML messages. SOAP is a communication mechanism for connection of Web service. SOAP consists of three parts: the first part provides a wrapper for sending Web service messages through the dormitory or an intranet, the second part has a set of rules to define special data types for the Web service, and the third part includes rules for the description of the Web service methods whereby the structure for a description of the call and corresponding Web service is included.
- *WSDL (Web Services Description Language)* is based on XML language for describing Web services. WSDL document describes how the Web service works and how to use it to communicate and where you to find it. Parts of description of a Web service can be located in multiple documents that provide more flexibility and increase the possibility of reuse.
- *UDDI (Universal Description, Discovery and Integration)* is platform-independent, based on XML, and is used to publish lists of services and mutual disclosure of the same. It also defines the way in which services or software applications collaborate over the Internet.

3. ORCHESTRATION AND CHOREOGRAPHY

In order to fully exploit the opportunities provided by the SOA architecture and provide connectivity services of different companies, it is necessary to describe in more detail the way in which services can be linked to each other. So far, there is no unified standard, but two were allocated, orchestration and choreography service (Orchestration and Choreography) [3].

The orchestration is the model for connecting various services in which all control is retained at one central process. Central process coordinates the execution of various operations of Web services according to the requirements of orchestration. Connections to Web services are being performed by the message. They contained the business logic and order of inclusion and work execution.

Choreography of service is a model based on cooperation. Choreography is not dependent on the central process. Each Web service involved in the choreography knows exactly when to execute its operations and with whom we should cooperate. It allows each Web service description part of your interaction. Choreography follows the exchange of sequences of messages between multiple parties, not just a business process that only one participant of interaction is interested in.

4. SOA APPLICATION IN BUSINESS MANAGEMENT OF NAUTICAL TOURISM PORTS

Modern and contemporary way of doing business offers to LNT huge trade and economic area, primarily related to e-business and business through the internet. Directing the business strategy of the marina to the service mode enables businesses to focus on core business and thus make improvements in competitiveness. Recognizing the importance of SOA, and the benefits that it provides (multiple service usability, scalability and manageability), in particular, a significant number of companies including also the management of the marina will have to introduce the re-engineering of existing information systems and thus fulfill expectations of sailors connected to the speed and quality of service (higher processing speed, lower cost, higher security). In this way boaters receive a completely new opportunities and the perception of service quality.

The primary objectives of the SOA in LNTs are to improve communication with subjects in order to be more efficient, more reliable, automatic exchange of information on reservations service which provides LNT and a shortened flow of business processes.

When business processes are electronically established, integrated and stored in a single database, the management of the marina is ready to orient their services to boaters through information and communication system.

A. Existing information systems in LNT

Almost all LNT management use information and communication technology in their operations. ICS are of modular construction, linked to a computer network and mostly on the Microsoft Windows platform. The main communication with guests takes place through the LNT reception. Boater can negotiate in advance, partial or total stay in the LNT, personally by direct contact with LNT (telephone, fax, e - mail or on line by filling out the form) or through an intermediary (tourist agency, agent).

Start of service execution is taken from the moment when a boater with his vessel arrives to the marina and fills in the application (information - communication center). The receptionist takes all the necessary information (personal information for boaters, vessel information, the estimated time of arrival, types of berth and determines (issues) free embroidery and turns certain supporting documentation with which boater acquires the right to use the rented berth or technical/general services. At the same time, the receptionist opens the account to pay for the agreed services. At the end of the stay, boater accesses the reception again and concludes documentation of the stay and the use of marine services, based on which he is delivered the account of the overall services for which payment is competent the reception.

Marines have their own website where they can view a detailed description of the marina (high quality images), price list, links to certain data and information (exchange rates, weather, etc.), the way of establishing the contact with the marina and others. In most marinas ICS supports i.e. available to operators are the following programming models: basic information for boaters and information on the vessel, record of berths, moorings issuance, registration fee-Ministry of Internal Affairs and tourist boards, guestbook, required reports, invoicing, cash register , subsidiary accounts, automatic accounts keeping documents, etc.). In most cases the plan of berths of marina is used i.e. visualization of state in the marina. The module is designed to control accommodation facilities and allows maximum use of capacity with a minimum number of moving vessels. In network operation, all users access the same database and can on their own computer see the data just entered or changed by the other users in the network.

B. Solution proposal

Rapid development of nautical tourism has forced the management of LNT to work continuously to improve and enrich the level of service and adapt to the modern needs and wishes of their guests. This paper proposes the introduction of SOA on existing ICS. Therefore it is necessary to create a new business process that we call ***"Reservation service",*** and it will contain all the elements necessary for the automation of marina operations. Newly created business process will support the functionality of the existing software support of ICS marina. Creating a business process consists of several models:

- Business process model, shown graphically by using BPMN diagrams.
- Ensuring communication between existing and new applications as well as adding automation of business processes by creating the following Web services:

- ✓ Web service that will allow booking and cancellation of berth in one of the offered LNT,
- ✓ Web service that will enable the booking and cancellation of technical services in one of the offered LNT,
- ✓ Web service that will enable the booking and cancellation of general services in one of the offered LNT,

- Code generating that will allow the execution of business processes, based on the created business process model.
- Creating new Web applications of portal type.
  - ✓ Web application that will provide a simple overview of the existing capacities (berths) in LNTs, organized to provide technical and general services,
  - ✓ Web application that will allow the operator: to choose the required marina, date and time of the service, enter the number of vessels that will be anchored, request for technical or general service, as well as other necessary information,
  - ✓ Web application that will allow initiating or starting a business process,
  - ✓ Web application that will be available only to authorized operators of reception and operators of agencies/agents.
- Creating BizTalk Server applications that will enable the execution and monitoring of business processes.

C. Microsoft BizTalk Server

Microsoft BizTalk Server is the leading technology platform used in manufacturing and service companies for business process automation and integration of business applications into a single business information system. In addition to Microsoft technology can be used Oracle or SAP technology. BizTalk Server enables the development, setting up, implementing, monitoring and management of business processes which are realized for the purpose of unifying the various divided applications. Such an information system is a prerequisite for the effective management of business processes, and thus to increase the competitiveness and profitability of the company, because the lack of simple ways to reach the information contained in a series of unrelated business applications is one of the main reasons for poor management of medium and large companies. An additional advantage of BizTalk is that it is, compared to similar solutions in the field of business application, significantly more favorable in price and requires less hardware resources. BizTalk communicates with the existing ICS without change of requests. The data from existing applications, BizTalk can accept in variety of formats and by using development tools, data from one format successfully convert to another, thus allowing communication between different systems. BizTalk coordinates complex business transactions and business applications by using orchestrations; it shortens the operation time, gives greater ability to update data which greatly reduces the cost of doing business.

D. Description of the execution of business process "Reservation service"

The execution of the business process "Reservation service" is done in several steps:

- The operator using a Web application, based on the demands of boaters selects information related to book berth in the marina (destination, arrival date, departure date and number of berths), reservation of technical services (servicing of the device driver, minor repairs, maintenance of vessels and cruising) and general services (parking place for car, sports and recreational services, catering and other services).
- Web application creates a corresponding XML messages and sends them to the BizTalk process machine.
- BizTalk application on the basis of the received message initiates the creation of a business process.
- Business process Web service calls to book the berth that in the existing database makes the reservation and returns information on the availability of the required berth.
- If there is a demand for technical and general services, business process calls Web service to book technical/general services that the existing database makes the reservation and returns information on the done booking.

E. Modeling of the business process "Reservation service"

Business Process Modeling (BPM) refers to the design, management and execution of the business process, and its strength lies in the unification and expansion of existing process-oriented techniques and technologies. In developing the model business processes used are standard, understandable graphic elements that significantly facilitate communication between different participants in their analysis. The model focuses attention on what is important and helps in understanding the way the business processes work. To create a graphical flowchart of the business process used are Business Process Modeling Notation (BPMN). BPMN is developed to ensure the notation for defining business processes easily understandable by all business users, from business analysts who need to outline the processes, to technical users who are responsible for implementing technologies that are executed for using the processes and ultimately to business people who will use and monitor business processes. BPMN allows you to make up the

gap between the business process design and its implementation. BPMN diagram is composed of a set of graphical elements. The elements allow easy development of simple diagrams; they are different with respect to another and look closely to the most of business analysts. To create a BPMN diagram for the implementation of the simplified process of berth booking and technical and general services in the marina which is shown in the paper, the ITP Process Modeler 5 for Microsoft Visio is used.

In a simplified model of the process "Reservation service" in the marina included are the following participants: the management of the marina/reception, boater and agency/agent. A simplified model contains all the elements necessary for the berth reservation and technical or general services at the marina (Figure 2).

The business process starts after receiving the request for insurance of services in the marina, why the start of the process is presented by Start Event element type Message. Receiving of requirements may consist of the following individual requests: berth reservation, reservation of technical or general service. Processing of the received applications is represented by parallel Gateway elements that allow passage through each branch for which the demand is fulfilled.

If the request is received to secure a berth in the marina, the task "Book berth" is performed, presented by the Task element, type Service, which calls the operation "BerthRequest" of Web service "Berth" , whose purpose is to reserve a berth in the chosen marina in the selected period of time. Web service returns a message on successful or unsuccessful completion of berth booking. In case of having to cancel already reserved berth, perform is the task "Berthcancel", presented by the Task element, type Service, which calls the operation "BerthCancel" of Web service "Berth".

If the request comes that it is needed to provide technical services, performs the task of "Technical Services", represented by the Sub - Process - the ad hoc element (a process that includes a set of activities that can be performed in any order) element, type Service, which calls operation "TechnicalServicesRequest" of Web service "TechnicalS ", whose purpose is to put aside some technical services in selected marina. Web service returns a message on successful or unsuccessful completion of booking technical services. If necessary to cancel already booked technical service, performed is the task "Cancel technical service", presented by the Task element, type Service, which calls the operation "TechnicalServicesCancel" of Web services "TechnicalS".

If there is request to provide general services, performed is the task "General Services", represented by the Sub - Process - ad hoc element, type Service, which calls the operation "GeneralServicesRequest" of Web services "GeneralS", whose purpose is to reserve a certain general service in the selected marina. Web service returns a message on successful or unsuccessful completion of booking of general service. In case of canceling already booked general service, perform is the task "Cancel general service",presented by the Task element, type Service, which calls the operation "GeneralServicesCancel" of Web service „GeneralS".

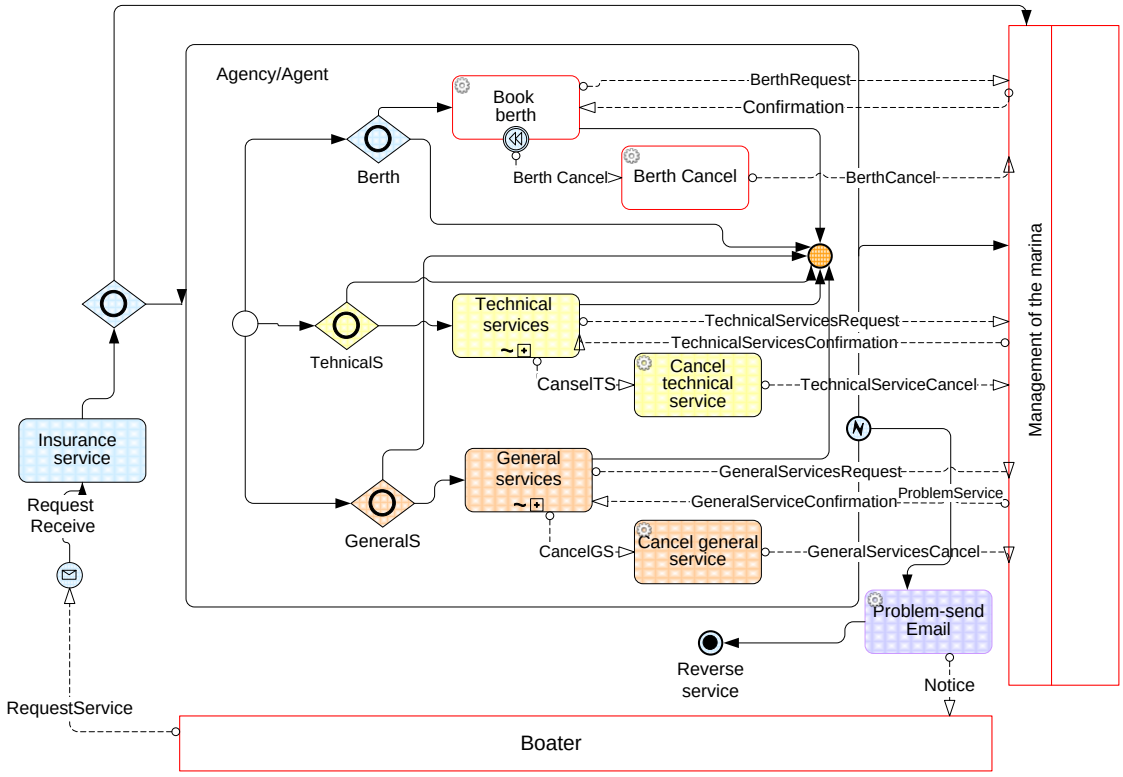


Figure 2. BPMN diagram for business process of berth reservation, technical and general service in the nautical tourism ports

Communication of Task elements, Service type, with Web Services is represented by berth objects type Message Flow, which is displayed as dotted lines with unfilled arrow. These elements are used to show the flow of messages between two separate participants in the business process. In case that happens some exception, related is the booking services, the event is raised, presented by Exception Event element. Event passes control task "Problem - Send Email to boater", who calls the operation "EmailInfo" of Web service "EmailNote", which sends an e-mail message about the existence of the problem.

E1. Web service creation

To ensure communication between existing and new applications as well as adding automation of business processes, several Web services are built. Descriptions of Web services are available through WSDL. WSDL document describes what the Web service does, how to use it and where to can find it. Parts of description of Web service can be located in multiple documents that provide more flexibility and increase the possibility of re-use. The proposed solution following describes the following Web services:

a) Berth in marina service

Web service "Berth" allows berth booking and cancellation in one of the offered LNTs. Web service "Berth" is done in several steps:

- Using a web application operator chooses: LNT, date of arrival, the size of the berth, date of departure,
- Web application creates an XML document and sends it to the destination where it will be read by BizTalk Server,
- BizTalk Server reads the received XML document and calls for the execution of Web service "Berth",
- Operation "BerthBook" of Web services "Berth" executes reservation in the required LNT ,
- Operation "BerthBook" of Web services "Berth" returns information on the availability of required berth in the LNT.

BerthBook is a synchronous request/response operation which for the required inputs (LNT, the size and number of berths, the date of the beginning of the berth's usage, date of leaving the berth) returns information of possibility of reservation of berth in Marina.

Operation Berth Cancel - using of Web services "Berth" for cancellation:

- Using a web application operator selects the number of berths in the marina,
- Web application creates an XML document and sends it to the destination where it will be read by BizTalk Server,
- BizTalk Server reads the received XML document and calls for the execution of Web service "Berth",
- Operation "BerthCancel" of Web services "Berth" cancels before created berth reservation.

Operation "BerthCancel" is the request operation that contains only the input (input) message. Operation of required inputs (LNT, the size and number of berths, the date of the beginning of berth's using, date of leaving the berth) cancels earlier created a berth reservation.

b) Technical service in marina

Web service "TechnicalS" allows booking and cancellation selected technical services in one of the offered LNT. Web service "TechnicalS" is done in several steps:

- Using a web application operator selects the offered technical service,
- Web application creates an XML document and sends it to the destination where it will be read by BizTalk Server,
- BizTalk Server reads the received XML document and calls for the execution of Web service "TechnicalS",
- Operation "TechnicalServiceRequest" of Web services "TechnicalS" does the reservation of technical services,
- Operation "TechnicalServiceRequest" of Web services "TechnicalS" returns information on the booking of required technical services.

Operation "TechnicalServiceRequest" is a synchronous request/response operation which for the required inputs (boater, number of connections, technical services) returns information regarding reservation of technical services.

Cancel technical service operation - Using Web services "TechnicalS" for cancellation of technical services:

- Using a web application operator selects the number of berths,
- Web application creates an XML document and sends it to the destination where it will be read by BizTalk Server,
- BizTalk Server reads the received XML document and calls for the execution of Web service "TechnicalS"
- Operation "Technical Service Cancel" of Web service "TechnicalS" reverses earlier created reservation of technical service.

General service will not be described in detail since it is exactly the same as the technical services.

c) Web service "EmailNote" enables sending e - mail messages to the boater. Web service has one operation EmailInfo.

Using of Web service "EmailNote" to send electronic messages:

- Using a web application operator chooses boater (number of berths),

- Web application creates an XML document and sends it to the destination where it will be read by BizTalk Server,
- BizTalk Server reads the received XML document and calls for the execution of Web service "EmailNote",
- Operation "EmailInfo" of Web services "EmailNote" sends an electronic message to the boater.

E2. Application creating

Web application "Reservation service" should allow independent berth booking and any technical or general service by the operator reception of marina or operator's agencies/agents. Functions of berth booking and technical or general services shall be executed in real-time, over a single database, which completely excludes the possibility of multiple sales of the same services. Besides the achieved consistency of database, web application development enables the following benefits:

- At any time, condition of available and rented berths and reserved services is known. That information allows the marina management to effectively manage resources, which is extremely important in the summer periods. On the other hand, agencies/agents sell only facilities that are currently available and real, which excludes cases of incorrect arrangements.
- The need for continuous oral and written communication of marina personnel/reception and agencies/agents disappears, in connection with the reservation of berths and technical and general services.
- ICS is available 7x24h week eliminating the problems of nonconformance of reception working hours, service offices and agencies/agents.

Access to the applications requires identification (username and password) users, because they are available only to certain users or registered agencies, agents and operators reception. Applications are of portal type because the main purpose of web portal is that it represents a centralized access point for finding and managing information and to execute various tasks via the Web portal Web browser is used. Also, Web portals are a very convenient way to collect information from various sources and their concentration in one place. The main window of the Web portal contains sections which allow by the operator, the selection of the required marina, date and period of the use of the berth or preferred service as well as the other necessary data.

5. CONCLUSION

By rapid development of information and communication technology, marinas have found themselves in a new environment and in order to remain competitive maritime market must adapt quickly. One of the models that will enhance and improve business is information communication systems whose architecture is based on service orientation.

This work represents the view of the model systems based on the orientation services in marina management. It describes what service oriented architecture and Web services are, that are enabling applications to communicate over the internet in an open and flexible way and the principles on which they are based. Finally, the simplified model of SOA applications in the management of the marina is given.

In developing the model it is essential to business process modeling because the goal of BPM is not just to copy the existing business processes in an information system, but also to improve them with what information technology has to offer, such as web services. Web services provide a generic, standardized client-server link to launch the application on the server. The most widespread way of using Web services is simply calling methods on remote computers.

SOA solutions can integrate in different environments, different platforms, that reduces LNT operating costs. There is no need to change existing software because it is possible to exchange data between the new and old software. In addition to financial savings, it is also a component of optimizing business processes in the proper application of SOA solutions. A key element of the SOA architecture is communication between different application services.

The application of SOA in the management of LNT obtained on flexibility, modularity, coverage of business activities, increased availability of data, easier management services and greater profits.

REFERENCES

1. Srića V., Spremić: Informacijskom tehnologijom do poslovnog uspjeha, Sinergija d.o.o., Zagreb, 2000.
2. Hurvitiz, Judith, Robin Bloor, Carol Baraudi i Maricia Kaufman, „Service Oriented Arhitecura for Dummies", Wiley Publishing, Inc., Indianapolis, Indiana, 2007.
3. C. Peltz: Web services orchestration and choreography. Computer Volume 36, Issue 10, Oct. 2003.
4. M. Jurić: Business Process Execution Language for Web Services Second Edition, Packet Publishing, 2006.
5. Web Services Description Language (WSDL) Version 2_0 Part 1 Core Language.mht.
6. Business Process Management Initiative (BPMI): Business Process Modeling Notation (BPMN), Version 1.0, BPMI.org, May 2004.
7. C. Peltz, Web services orchestration and choreography. Computer Volume 36, Issue 10, Oct. 2003.
8. T. Erl: Service Oriented Architecture: Concepts, Technology and Design, Prentice Hall, New Jersey, 2005, ISNB 0-13-1858-0.
9. Web Services Description Language (WSDL) Version 2_0 Part 1 Core Language.mht.
10. M. Havey: "Essential Business Process Modeling", O'Reilly, 2005.
11. http://www.microsoft.com/biztalk/en/us/default.aspx.
12. http://www.microsoft.com/biztalk/en/us/virtual-labs.aspx
13. http://www.infodom.hr/default.aspx?id=536
14. R. Uleman: "Service Oriented Architecture Un Veiled", Geospatial Solutions, Vol. 16, No. 6, pp. 30-33. . (2006).
15. T Erl.: "Service-Oriented Architecture: Concepts, Technology, and Design, Prentice Hall PTR" ,August 04, 2005.

Internet challenges for print media advertising practice – case of Belgrade daily newspapers

Marijana Vidas-Bubanja, PhD[1]
Belgrade Business School, Belgrade, Serbia
E-mail: marijana.bubanja@bbs.edu.rs
Iva Bubanja, Mr.sc.
ArtBoulevard, Belgrade, Osme crnogorske brigade br.8, Serbia
Heartefact, Belgrade Bulevar Kralja Aleksandra 82, Serbia
E-mail: ivabubanja@gmail.com,

Abstract – There is a growing trend of transferring commercials from print media to online format during the last decade. Rising number of users who are spending more time on Internet and social networks, are the main reason why print media are facing a problem of keeping advertisers in their media in order to provide resources for their survival. Newspapers are going through meager changes intending to adapt to business conditions in a digital environment. The conducted research showed the differences between commercials in online and offline versions of national newspapers.

I. INTRODUCTION

The media and publishing industry is undergoing a massive transformation thanks to the rise of digital technology. Over the past 30 years, changes in newspaper production have been significant: from linotype machines and hot lead type, typewriters and grease pencils, to computer terminals, satellite transmissions, digital cameras and online editions. The ability to get news online, instead of from the traditional printed newspaper, may be the most dramatic change of all.

This paper analysis how Belgrade daily newspaper industry is experiencing digital disruption process. The initial hypothesis of the performed research is that Belgrade daily print newspaper industry has to adapt to new advertising trends influenced by Internet and promotion capabilities of digital media by devoting full attention to their online versions of the daily papers in order to provide rising advertising income for their budgets.

Both retail and music are prime examples of sectors that underwent similar transitions. Many traditional retailers seamlessly translated their brick-and-mortar experience and built prosperous revenue streams through e - commerce. On the other hand, many of the traditional players in the music industry struggled to evolve and experienced a shrinking market once iTunes was born.

Digital publishing opens the door to additional income streams beyond advertising and subscriptions in content monetization. Readers will pay for content as long as it's compelling (feeds the intellect or fuels emotions) and can't be found for free anywhere else.

II. INTERNET IMPACT ON PRINT NEWSPAPER INDUSTRY

'Golden Age' for print media was a period between 1890 and 1920, when the industry grew and newspapers enjoyed the attention of being society's primary source of information. However, the Great Depression disrupted this and in the 1930s newspapers were severely hurt, hundreds of companies ended operations. This was escalated with the emerging broadcast radio, offering lower priced alternative news and as a result capturing a large portion of advertising revenues from newspapers [1].

Television was the next technology that disrupted the newspaper business. It was a more powerful medium to distribute information and it overtook the newspaper and majority of national advertising in the 1960s. In response, newspapers consolidated and existing newspapers got an opportunity to enjoy a semi-monopolistic position.

However, the early 1990s was the last time newspapers enjoyed profit margins of 20% to 25%. In 1995, the Internet was introduced to the public and by the millennium the Internet as a worldwide platform became evident. The Internet reached more people than any newspaper had and profits margins have dropped to an average of 5%. The mere ability to access information and receive instant updates threatened the purpose of a newspaper, and evidently the Internet has been a growing source of competition for the newspaper publishers [1].

The Internet opened to commercial interests in 1988, but it took a few years for the network to gain public recognition and in 1990's Internet began to take shape as the 'World Wide Web'. Against a backdrop of relentless growth, the Internet continues to change and evolve, as highlighted in Table 1. In the past ten years, the number of Internet users shot past one billion and is nearing three billion; users migrated their fixed Internet access from dial-

[1] This paper is the result of the research realized under the project 179038 „ Modelling of the development and integration of Serbia into world trends in the light of the economic, social and political commotions", Ministry of education, science and technology development of the Republic of Serbia.

up to broadband; and their usage shifted from text-based to predominantly video traffic. Globally, the number of users in developing countries now exceeds those in developed countries; there are now more mobile broadband subscribers than fixed; and mobile access has shifted to smartphones [2].

TABLE I. ICT AND INTERNET– FACTS AND FIGURES

Internet users (2014) - 2.893.587.260 Internet users worldwide. - 2/3 of Internet users are from the developing world. - Internet user penetration rate 40% globally, 78% developed countries and 32% in developing countries. - Households with Internet access: 44% globally, 31% developing countries, 78% developed countries.
Broadband Internet access (indicated years) - Global broadband subscriptions 93% of total global fixed Internet subscriptions (2012). - 673.295.648 fixed broadband subscribers worldwide and number of fixed-broadband subscriptions in developing countries overtook the number in developed countries (2013). - 2.3 bn. mobile-broadband subscription globally with penetration rate 32% (2014). - Mobile broadband connections 5.3 bn. - forcasts for 2018.
Mobile cellular subscription (2014) - The number of mobile-cellular subscriptions worldwide is approaching the number of people on the earth, almost 7 bn. - Penetration rate 96% (developing countries 90% compared with 121% in developed countries).

SOURCE:2 AND 3

The open Internet has become a medium like no other. It merges the most notable characteristics of traditional media such as broadcast and telecommunications, while also augmenting them in ways that have revolutionized aspects of civil society, business, and government [2].

As a result, the nearly 3 billion Internet users are both creators of information as well as consumers. Websites, blogs, videos, and tweets, can all be broadcast and accessed in the largest mass medium imaginable. Audio and video calls and conferences can be set up and received without regard to distance or cost.

This short analysis proves that the impact Internet has had on newspapers is two fold. Directly, the Internet has increased competition for the newspaper and revised the way news is distributed. Indirectly, the Internet has influenced advertising trends, consumer behavior and the rise of disruptive technologies. As a result, newspapers have been forced to integrate with the Web and now approximately 80% of newspaper publishers have integrated web and print operations [1].

The Internet has removed all geographical boundaries and as a worldwide platform, it has extensive global reach. Also it removed any barriers to entry in newspaper business. Initially, entry costs were very high because large investments were needed to set-up a publishing house, and so the number of players was restricted. However, the Internet gives anyone with a computer an opportunity to publish information. This has placed newspapers in an industry for news and information, and as a result they are now directly competing with all news sites online.

The Internet has revolutionized the way news content is distributed and consumed. It has the ability to deliver news immediately, thus the Internet allows readers to receive news in real time and stay constantly updated. Additionally, the Internet via e-mail and social media activity has led to news being distributed socially.

According to IAB Report for 2013 half of social network site users in US have shared news stories, images or videos, and nearly as many (46%) have discussed a news issue or event. In addition to sharing news on social media, a small number are also covering the news themselves, by posting photos or videos of news events [4].

It is evident that the existing newspaper business model necessitates change in the new digital environment whether this means: a) merely altering their existing model to include print and digital or b) a complete transformation of going purely digital what is still debatable.

III. CURRENT BUSINESS ENVIRONMENT FOR PRINT MEDIA INDUSTRY

Beside Internet, some other factors create very challenging environment for print media.

1) *Economic crises that begun in 2008/9* cause financial challenges for newspapers. On one hand, most newspapers were faced with cuts in revenues from declining subscriptions and single copy purchasers. Readers were looking for more cost effective alternatives and as a result increased consumption of online news. Since most newspapers do not charge for their online content, this shift in consumption did not drive any revenue increase for the publishers.

On the other hand, the recession resulted in deep cuts from marketing budgets in the majority of companies. Since the newspaper business model is largely dependent on advertising as a form of revenue, this advertising recession brutally hurt publishers.

In the US, the industry experienced a 23% fall in advertising revenues in 2007 and 2008 and revenue in 2012 was 45% of what it was in 2006. Print advertising losses continue to fall far exceed digital ad gains. In the US, for example, in 2012 the ratio was about $15 print lost for every digital dollar gained. In 2013 according to Newspaper association of America, print advertising revenues fell to $17.3b and they haven't been lower since 1950 [5].

In Europe the advertising share for the printed press has been decreasing continuously for more than 25 years, while the overall advertising market has increased temporarily. The financial crisis for the first time leads to a decline of advertising revenues for the press in real terms. It is expected that the level of 2007 (around $120.000 mil.), before the financial crisis will never be reached again. The revenue in 2014 was worth $88.105 mil [6].

Therefore, this overall plunge in revenue forced newspaper publishers to cut costs and scale back operations in order to survive. Internally, publishing houses began to lay-off staff and reduce the number of editors. Overall, the decade-long trend in newspaper job losses continues.

The best example of this growing phenomenon and effect of the Internet is reduction of journalists in print media. According to the annual American Society of Newspaper Editors survey, in the decade from 2003 through

2012, a total of 16,200 jobs were lost, and the similar situation is in Europe as well. [7].

2) *The Internet has played a big role in altering news consumption behavior for the average newsreader*. Today news is in a so much demand, but consumer wants to get them form different sources. In Europe 426.9 m people use the Internet (65%), and Internet was the second most popular news source in 2012. There was 25% rise of Internet users in Europe reading news online (91%). According to IAB/Mediascope Europe study each week Europeans spend 14,8h online, 16,8h watching TV, and only 4,6h reading print newspapers [8].

While newspapers with their web sites and online versions are definitely enjoying a portion of this growing online readership, the real problem is that most publishers are providing content for free. Thus, readers have become accustomed to consuming information at a zero cost, and this has resulted in subscription revenues to decline.

The rising percentage of consumer prefers news that is concise, with only the necessary information. Users have also developed a preference for a more personalized news experience as they only follow news for particular areas of interest. There are online tools that allow for such news customization.

Additionally, users have a preference for interactive tools and 36% of all Internet users take into consideration the interactive materials available on a website before selecting it as a news source [8]. Examples of interactive content are charts that users can manipulate, quizzes to test user knowledge, video news reports, and other. Internet users share the information they receive with others by forwarding links and stories. Online, sharing news is most easily facilitated via email and social media activity.

As a social media continues to be a grow space around the world, with roughly 29% of the world's population of active users (2.078 billion active social network users in the world), it is therefore important for newspaper publishers to make themselves visible to their newsreaders via social media portals as well. Research conducted by Global Web Index suggests that the average social media user spends 2,5 hours per day using social networks and microblogs. Facebook continues to dominate the global social media landscape, claiming 1.366 billion active users in January 2015. In Europe the social media penetration is 46%, what means that 387 million of Europeans are active social media users in January 2015. Among the social media users, 26% of them are accessing social media through mobile [9] [10].

There is growing trend of consumer participation in the process of online news creation and distribution. Certain publishers already adapted to this trend. For example, the New York Times launched a joint relationship with LinkedIn, where LinkedIn users are able to subscribe for news alerts on a specific industry and to share and discuss the news with users and experts via an industry specific blog.

3) N*ewspapers have a pressing need to accept the multiplatform consumption pattern that has emerged.* Digital platforms have been developed to keep up with transforming consumer demands and they include the computer, mobile and tablet. With respect to newspaper publishers, these technologies disrupt the newspaper business because they make online news more accessible for readers. Together, these devices aim to capture a larger readership for longer hours during the day. This is because now consumes are no longer limited to their work and have portable access to the news. While these devices are gradually increasing subscription and have the potential to strengthen the brand of a newspaper, they are simultaneously increasing the number of devices that users can select from. This is detracting from customer loyalty for any one device, platform or format. Newspapers are now competing in an open market for news and information sharing.

Consumers in Europe are constantly switching between devices: computer, tablets, smartphones, gaming platforms and other emerging devices are all used to access the web. This is a clear indication that European consumers are heading towards multiple platform media consumption. In fact, technology is driving the way Europeans consume media and interact with brands: 44% of all Europeans own a Smartphone and 50.9 million uses a Tablet to go online. Media multi-tasking is a large part of Europeans lives: a third (37%) of Europeans access the internet via more than one device and 48% are online while they watch TV [8].

This insight is essential to the newspaper industry because most publishers have to adapt their product to the technologies that consumers demand to ensure their presence on every available platform. In order to do so newspapers have to understand their customer segment, the news they demand and the device they prefer.

IV. PRINT AND DIGITAL ADVERTISING PRACTICE

While the Internet has severely threatened the traditional ways of news consumption, it has improved the business for an advertiser. With the Internet, advertisers are able to reach their target audience more effectively and efficiently. Marketers can accurately measure results and monitor their ad spends very closely. They can also engage with the consumer and interact with them to ensure the effectiveness of their campaignes.

With the economic recession in the world advertising also entered into the steep recession, but after 2010 media ad spending is in constant rise. According to e-Marketer media ad spending in the world rise from $475 bn. in 2010 to $575 bn. in 2014. Estimation indicates ad spending of $600 bn. in 2015 and $628 bn. in 2016 [11]. Newspapers saw a deep decline in overall advertising revenues proved statistics. On the global level there was a rising market share for Internet ad and declining for newspaper ad for the period 2005-2013. In 2005 Internet ad revenue were $13 bn. and raised to nearly $45 bn. in 2013, comparing with newspaper ad revenues declined from $47 bn. in 2003 to under $20 bn. in 2013 [11].

The same trends are present in European ad market where print ad spends per/capita fell from 60€ in 2007 to under 40€ in 2012 while online ad spends per/capita rise from 10€ to 30€ in the same period. The value of the European online ad market rise from 6,6 bn. € in 2006 to

27,3 bn. € in 2013. Online market share of ad budgets in Europe reached 25,6% in 2012 [12].

Online ad revenue in the United States grew to $42.8 bn. in 2013, according to the Interactive Advertising Bureau's Internet Advertising Revenue Report. Ad spending in US in 2013 was distributed as following: 45% TV, 22% Internet, 19% print media, 10% radio [13].

Although all the data indicate raising share of online ad in marketing budgets, newspapers can not depend on online ad as a revenue source in order to survive. In fact, online ad revenue which is approximately 10% of their total newspaper ad revenues is unexpectedly small and it therefore exposes online websites as a weak source of revenue for publishers [1].

One of the key reasons for this situation is the fact that the rate of online ads is drastically lower than the print rates. The online advertising model, unlike the newspapers flat rate, is based on the number of impressions (CPM: Cost pet 1000) and clicks (CPC: Cost per click). Since this billing model was first introduced with very low rates, the increasing competition has prohibited online rates to rise any further.

In addition, there is also evidence that news aggregators benefit disproportionately from the content they aggregate, link, index and present. This is because revenues from ad-space sold alongside re-distributed content are only paid to the news aggregators and not the news generators. As a result, newspapers have been losing a large portion of their online revenues.

V. Internet Challenges for Belgrade Daily Newspapers

The Belgrade daily newspaper industry is currently undergoing a much-needed transformation, battered and distressed by a severe recession, declining advertising revenues, major shifts in consumer behavior and emerging disruptive technologies. As a result, we have seen some major media companies face with the burden of heavy debts, while the others are still experimenting with new survival formulas to restore the equation between newspaper content and earnings. Moreover, Serbian print media are facing severe digitalization process, and every day more of them are becoming viral and providing online version of the same.

The research included media such as: Politika, Večernje novosti, Danas, Blic, Alo, Kurir, Informer, Naše novine and two specialized press Sportski žurnal and Sport. From ten analyzed national newspapers, eight of them have their online versions, and nine of them have their Facebook and Twitter accounts. On other social networks such as Pinterest, Instagram, Google+, only some of them have their profiles which are not regularly updated.

This transition of Belgrade daily newspapers to online version is closely connected with the increasing number of Internet users in Serbia, around 3.5m. They spend average 13.0 hrs online per week by IAB Europe research from 2012. The use of Internet in Serbian society keeps growing on daily basis especially among young generations, who are more mobile and informed, and who are searching for free information on the network [14].

TABLE II. INTERNET USE IN SERBIA

Facts and figures for 2014 - 3.500.047 Internet users in Serbia – penetration rate 48% - 3.800.000 Active Facebook users - penetration rate 52% - 9.138.000 Active mobile subscription - penetration rate 128% - 2.900.000 Active mobile broadband subscription, penetration 40% - 2.400.000 Active social media users accessing social media on a mobile device - penetration 33%,
The Serbian media consumption landscape 2012 - Hours per week used: TV 16,9hrs, online 13hrs, radio 13,1hrs, newspapers 4,4hrs, magazines 3,5hrs - 48% of Serbian Internet users watch TV online monthly - 61% of Serbian Internet users listen to the radio online monthly - 83 % of Serbian Internet users read news online monthly - 52% of Serbian Internet users visit news websites every day - 99% of Serbian Internet users communicate through email - 89% of Serbian Internet users have used a personal or professional Socail media website
Devices for accessing the Internet in Serbia 2012 - The computer is the most popular method, used by 3.0 million users, - 24% of all Internet users in Serbia (EU: 37%) go online via more than one device. - 2% of users go online by Tablet, - 21% of all mobile users own a smartphone.

Source: [3] for 2014 and [14] for 2012.

In line with this change in Serbian media consumption landscape are the research findings made by Gemius that the most popular websites in 2013 are online versions of national newspapers providing news from different spheres of life. The leader in 2013 in the online market was www.blic.rs, which is news portal that reached 53% of Internet population, and on the second place is www.kurir-info.rs, which is also a newspaper online portal that is providing news, after them the best ranked are www.b92.rs, www.telegraf.rs, 24sata.rs, novosti.rs etc [15].

The focus of the performed analysis devoted to commercial contents in ten Belgrade daily newspapers was to find out the relationship of their print and online versions concerning the volume and characteristics of presented commercials, as well as the possible contribution of commercial contents to print publisher's budgets and earnings. Daily newspapers are specific media with great circulation and massive and various audience what make them an efficient commercial medium.

The performed analyses included ten Belgrade newspapers of national coverage that can reflect the situation in Serbian advertising market of print media as a whole. Only small mismatches can exist. There are several reasons for this attitude. The first one is the fact that the print newspapers with national coverage are distributed all over the country and the main difference is that Belgrade news pages are not included in editions for other Serbian regions out of Belgrade. The second one is the absence of daily newspapers in local communities in Serbia, making press with national coverage the only source of daily information for the citizens out of Belgrade[2]. In this way commercials published in national daily newspapers reach the whole population of Serbia.

Final research results showed that there were only six commercials and two self-promotion contents per newspaper on a daily basis in their print version. Taken together for 10

[2] Local communities in Serbia usually have weekly newspaper editions.

newspapers there were about 62 commercials per day. Due to four time's smaller circulation compared to other newspapers, specialized press has 3 to 6 times less commercials than other daily newspapers [16].

On the same newspapers web portals situation is a bit different. There were 43 commercials and 52 self-promotion contents on daily basis. That was around 4 commercials and 5 self-promotion contents per newspaper in their on line edition. Taken together it is larger number of promotion contents in online version of the same print media. Specialized press is again in the worst position and has around 5 times fewer commercials than other daily newspapers in online form.

An increasing number of self-promotion indicates inability of a newspaper to sell its complete advertising space. Self-promotion is considered as an effective way of attracting new readership, which is necessary for the survival of newspapers. In fact, the number of self-promotion ads is rising and their contents and forms are so advanced that they are becoming equal to commercials. This opens the question regarding whether self-promotion steals readers' attention from the current commercial contents. Also large number of self-promotion in online newspaper form shows that newspaper companies are using increasing number of daily visits to increase the number of ridership by presenting a large number of self-promotion. The effect is even greater because online riders are active users and not just the passive audience.

The first conclusion from this two-month study of the selected press (online and print version) indicated small volume of the commercial content. This is not a good indicator for the Belgrade daily newspapers, having in mind that for the most print media selling advertising space is an important source of income. According to the estimation of print media publisher in researched newspapers participation level of ad resources to their budgets is around 20%. Comparing with newspapers in western developed economies where ad resources represent up to 80% of newspaper budgets, percentage in Serbia is very modest [16].

Concerning the characteristics of commercials presented in the Belgrade daily newspapers, the research results showed similarity of commercial contents in all newspaper type no matter do we talk about print or online edition. By comparing them it can be concluded the following: content types are repeated, there is a lack of variety of products and services, and narrowed choice is offered to potential buyers. Reasons for this are also connected with limited company budgets for commercials and the situation that the audiences of the different daily newspapers have the same preferences and tastes. In this way, newspapers create one public and a uniform products offer [16].

Research also showed that advertisers are choosing the same formats in online edition of different newspapers, and they place their ads in the specific page type. If advertisers' desire detailed processing of ad messages by consumers, they usually place their commercials on destination pages, even though the reach may be smaller than on navigational pages. The research identified several physical characteristics, such as visibility and time delay, as important factors in online advertising that affect perceived advertising level. Also the value of an audience to advertisers is determined by its receptiveness to advertising. Such emphasis on receptiveness towards advertising signifies the voluntary nature of media consumption and the importance of creating a receptive media environment for advertisers.

There is no regularity in repetition of commercial contents in online form. The characteristics of the texts and the photos are the same for online and print edition of newspapers commercials. The only difference is that in online world all commercials are in color no matter are they having a photo or they don't.

Besides indicated characteristics and volume relationship between commercial content in print and online version of daily newspaper, modest number of ads in ten Belgrade daily newspapers have to be considered in the respect with the current economic situation in Serbian economy.

There is significant reduction of producers' capability to allocate funds for advertising purposes due to reform and transition processes and other challenges connected with economic recession and stagnation. For this reasons, producers in Serbia choose more cautiously the media for advertising. They select media that can bring commercial message to the biggest number of potential buyers.

On the other side, complete society in Serbia has some new characteristics, as the high unemployment rate and lower monthly income force people to use the biggest part of their salary to satisfy their basic needs. Purchase of daily press became a luxury that many can not afford. Nowadays, middle and upper classes is still buying print media, but there are a growing number of readers that prefer to use online services for their news consumption what has clear impact of competitive position for commercial content between print and online version of selected newspapers [16].

General statistics of Serbian advertising market supports research results in the sample of ten Belgrade daily newspapers. Small number of ad in analyzed press is in line with the fact that Serbian advertising market is one of the poorest in Europe (average worth is about 170 million euros in the last three years). The share of print media in Serbian ad market is around 23%, what is also behind the European average [15].

Rising number of Internet users in Serbia and change in Serbian media consumption landscape opens new perspectives for digital media advertising and for online versions of print media. The Serbian online ad market is still in the phase of development, but advertisers are increasingly investing in Internet media and are becoming more interested in global advertising solutions. Rich media templates have become more popular, as well as mobile and video banners. Although the resources are limited, advertisers are trying to keeping in track with the global trends [15].

In 2013 it can be seen a positive growth of Serbian online advertising market. The total market value of digital and interactive advertising was estimated to € 16.32 million,

which includes advertising on mobile devices. The growth of the total market value is estimated at 13.41%, which means that they kept a positive trend of double-digit growth. The highest growth in 2013 had advertising on social networks, or social media marketing (27%). The market for mobile advertising in 2013 grew to as much as 150%, which means that the total value of this market segment amounted to €380,000. The total value of display advertising market segment is €8,230,000 and it remains the largest contributor of the Serbian online advertising market (50.68%) [17].

Comparing with Croatia and Slovenia as neighboring South Eastern Europe ad markets Serbia achieved in 2013 the highest growth rate in online advertising spends. The 13,4% growth rate of online advertising in Serbia was above European average growth rate of 11,9%, and higher than growth rate in Croatia of 6,2%, while Slovenian online ad spending decreased in analyzed year for 2% [17].

Regional comparisons of online ad markets are more complete by using online spend per capita across the region as an indicator that highlights the maturity of an online market irrespective of its size or absolute revenues incurred. Online ad spend per capita is a metric that helps to evaluate the maturity and scope for development of a market. By using this indicator results for 2013 are the following: 1) Slovenia with 42,6€ per capita is positioned above European average online ad spends per capita worth 37,4€; 2) Croatia spends 3,5€ for online ad per capita and is located in the lower part of the European countries list; 3) Serbia is penultimate with online ad spends of 2,3€ per capita; 4) Bulgaria with 1,5€ is the last on the list [17].

Despite positive dynamics Serbian online ad market still needs further development and maturation. Modern technology trends and new media consumption models support all traditional newspapers companies in Serbia to further invest in a digital strategy and focus their efforts on increasing their online ad revenue in future, as print ad revenues continue their decline.

VI. CONCLUSION

More than two decades after the adoption of the Internet by print media, newspaper publishers are still coming to terms with its implications. The Internet offers a novel platform for reaching audiences and has become a part of news gathering and news-processing routines. The new media ecology, with its additional agendas of interactivity, democracy, multimediality, and with a new domain of bloggers and citizen reporters, presents a set of new issues and opportunities for daily print newspapers, but also creates great challenges in the same time.

Statistics proves that the media and publishing industry is undergoing a massive transformation thanks to the rise of digital, which has made traditional revenue models based solely on subscriptions and advertising sales virtually obsolete. With an explosion of content seeking advertising support, and flat total advertising spend; publishers are scrambling to find new revenue streams and growth opportunities in digital media. The exact business model for the industry's future is still in flux, but one thing is for sure – the most successful and profitable companies will quickly find ways to capitalize on the enormous opportunities that digital offers [18].

Similar to the world trends, the competition between print and online media continues to intensify on the Serbian print media market. With decreased newspaper circulation, advertisers are turning their backs on print media in favor of the Internet. The connection between commercials and print media at the Serbian media market was researched by using a sample of ten Belgrade daily newspapers. The research results showed that the number of commercials in the daily newspapers dropped, while the numbers of ads in online version of daily Belgrade newspapers are in constant rise.

These two tendencies are connected with digitalization processes and challenges of domestic newspaper publishers, but also with general situation in the Serbian economy and society burdened with reform and transition processes, as well as economic recession and stagnation.

Research results definitely support the author's initial hypothesis that Belgrade print media need to include the Internet in their future advertising development strategy. It is also recommended that cooperation and synergy of print and digital media create a more effective way of presenting promotional contents to modern and choosy buyers, in Serbia as well.

REFERENCES

[1] Patel, A., *The Survival of the Newspapers in the Digital Age of Communication,* New York University, May, 2010.

[2] Internet Society, *Global Internet Report 2014, Open and Sustainable Acess for all*, 2014.

[3] ITU, *ICT Facts and Figures, The World in 2014*, Geneva, April, 2014.

[4] IAB and PWC, *Internet Ad Revenues Report 2013*, April, 2014.

[5] The Pew Research Center, *The State of the News Media 2013 An Annual Report on American Journalism*, 2013.

[6] ENPA, *Media on the move*, ENPA, November 2012.

[7] Jurkowitz, M., *The Losees in Legacy*, The Pew Research Center, Journalisam & Media, March 2014,

[8] IAB and Mediascope Europe, *European Media Consumption,* 2012.

[9] We are Social, Social, Digital & Mobile Worldwide Report, 2014.

[10] We are Social, Social, Digital & Mobile Worldwide Report, 2015.

[11] E-Marketer, *Media ad spending worldwide 2010-2016,* December 2012.

[12] Internet Advertising Bureau UK, *EU online market Overwie,* 25. September 2013.

[13] IAB, *Internet Ad Revenues Report 2014*, The first six months results, October, 2014.

[14] IAB Europe and Mediascope Europe, *Serbia launch Presentation,* Summary, 2012.

[15] Gemius, *Online Landscape South-East Europe,* March, 2014.

[16] Bubanja, I., *Newspapers in Modern Commercial Practice* – monography, Zadužbina Andrejevic, Belgrade, 2014.

[17] IAB Serbia, *A study about the consumption of digital and interactive advertising in Serbia in 2013,* ADEX Benchmark report, 2013.

[18] Vidas-Bubanja, M., *E-marketing,* Beogradska poslovna skola, Beograd, 2012.

The Success of Open Source Software: A Review

Domagoj Margan, Sanja Čandrlić
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {dmargan, sanjac}@uniri.hr

***Abstract*—Open source software (OSS) is software with its source code available that may be used, copied, and distributed with or without modifications, and that may be offered either with or without a fee. In this paper we present an overview of the state-of-the-art OSS-related research from different fields and disciplines of the social and information sciences. The emerging work on understanding OSS has questioned what has led to its prosperity. We present an historical review of its success, and we discuss some of the key factors that contributed to the rise of the OSS, mainly from the perspective of the software engineering and economics. We conclude with some real-life business examples of companies that achieved their profit with OSS.**

I. Introduction

The open source software (OSS) movement has over the last two decades had a significant impact on the software industry, as well as on the software organizations in both the public and private sector. Similar as closed-software projects, one can notice that some OS projects fail, while other grow, prosper and take the leading position in the market. Considering specific OSS development environment and its properties in general, we were motivated to understand and evaluate the reasons for this behaviour.

In this paper we analyze the causes of OSS success trough an historical overview of research and literature. We investigate and discuss some of the most important factors crucial for the success of the OSS.

II. Open Source Software

Open source is a term generally used in a variety of different contexts. In this paper we use the term open source software to refer to software which is distributed under terms that comply with the Open Source Definition according to the Open Source Initiative [10]. The Apache License, BSD, MIT, GPL, and Mozilla Public License are all examples of licenses that conform to the Open Source Definition. The most important distinctions refer to the extent to which they allow public property to be mixed with private property rights. An open source license is certified by the Open Source Initiative, an unincorporated nonprofit research and educational association with the mission to own and defend the open source trademark and advance the cause of OSS.

The philosophy behind open source is based on the concept of free software. In 1984, Richard Stallman founded the Free Software Foundation [54] as a charity organization for free software development. "Free software" means software that respects users' freedom and community. Basically, it means that the users have the freedom to run, copy, distribute, study, change and improve the software. From the Free Software Foundations perspective, "free software" is a matter of liberty, not price. To understand the concept, one should think of "free" as in "free speech," not as in "free beer". This belief is founded on a social ideal of collaboration through information sharing.

Open source could be seen as a movement, where communities of skilled programmers collectively develop software, usually on virtual networking on Internet and the web. They are loosely coupled communities kept together by strong common values related to hacker culture. Work in the open source often seems to be totally distributed, delegated and loosely coupled.

West and Dedrick [36] state that the success of the open source movement reflected a confluence of three factors in the mid 1990s: i) users seeking an inexpensive Unix implementation free of licensing restrictions; ii) a philosophical movement rejecting the idea of software ownership and appropriability; iii) emergence of the Internet as both an enabler and objective for collaborative software development.

But, in reality, only a small percentage of open source projects are considered to be successful [28], [29]. Healy and Schussman [35] observe that it seems clear that for every successful open source project there are thousands of unsuccessful ones. According to Fogel [30], the failure rate for open source projects is as high as 90-95%. Schweik and English [59] found that only one in six OS projects (17 %) were successful. Almost half the projects (46%) were abandoned in the initiation stage before the first software release. More than a third (37%) were abandoned after the initial release [60]. On the other hand, Miller et al. [39] found that failure rates of commercial proprietary versions of UNIX utilities ranged from 15-43% in contrast to failure rates of 9% for Linux utilities and just 6% for GNU utilities (which are also open source). When a minority of the projects is extremely successful, but most fail, it is important to ask what makes the difference.

III. OSS Development Process

Eric Raymond describes the development of OSS as a bazaar-like activity driven by volunteers [1]. Ingrained in the hacker culture, open source appears to represent the contrast of standard software engineering practices. Consequently, open source is commonly disregarded as lacking any repeatable methodology. Despite this historical stereotype, a well organised OSS project has its own rules and strategies. Crowston and Howison [6] and Scacchi [7] claim that the development

The research has been conducted under the project Extending the information system development methodology with artificial intelligence methods (reference number 13.13.1.2.01) supported by University of Rijeka (Croatia).

in OSS communities is distinctly different from traditional software development. On the other hand, the view that the development of OSS is something radically different from traditional software development is questioned by Fitzgerald [4] and Fugetta [5]. Sometimes formal methods and processes are used in OSS development as well, but there is no push towards them, and there is evidence that they are not a requirement for success, unlike generally considered true for software engineering [33].

Open source movement is using the Internet as a basis for communication, collaboration, and storage of artifacts. In the open source movement, everyone capable of writing code is welcome to join in. The Internet inherently supports globally distributed product development. Open source practices and tools have proven to be highly effective for overcoming the many problems of geographically distributed software development [11]. Typically, an open source project starts with a single programmer solving a small problem (e.g. a malfunctioning system affecting his own work) that later turns out to be significant, and then make the solution available to others.

In Fig. 1 team model for OSS development proposed by Van Vliet [64] is shown. Core team is a small team of experienced developers that plays a role of anagement and makes decisions about updates and changes in main software components. Co-developers present a larger group that analyses sourcecode and solve bugs. Active users use recent releases and report bugs, but do not program. Pasive users use stable releases. Transition of team members between layers is based stricly on credits.

According to Ljungberg [55], coordination of work in OS projects can be performed in several different ways. One of the most common models in coordinating open source projects is that several contributors work under a single benevolent dictator who owns the project. Typically this organization evolves when the founder attracts contributors. The benevolent-dictator has the right to take decisions (e.g. of what to include for redistribution) and has an obligation to credit contributors fairly. By contributing to the project you earn part of its reputation in return. When a project grows it is common that it will be split in subsystems with subsystem-owners, i.e. introducing two kinds of contributors: ordinary contributors and co-developers. An alternative to one single benevolent dictator is rotating dictatorship. Yet another alternative is to form a voting committee with the co-developers.

In the corporate model, individuals or small groups of individuals quietly and reverently develop software in isolation, without releasing a beta version before it is deemed ready. In contrast, the open source model relies on a network of volunteer programmers, with differing styles and agendas, who develop and debug the code in parallel. From the submitted modifications, the delegated leader chooses whether or not to accept one of the modifications. If the leader thinks the modification will benefit many users, he will choose the best code from all of the submittals and incorporate it into the OSS updates [1], [56].

Related research

Gurbani et al. [11] present a case study of open source software development methodology adopted by a significant

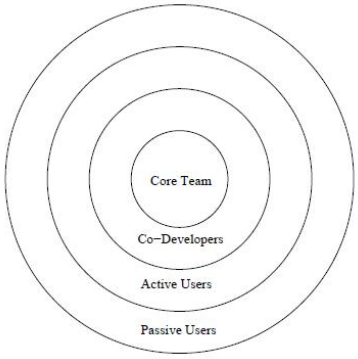


Fig. 1. Team model for OSS development

commercial software project in the telecommunications domain. Feller and Fitzgerald [50] defined a framework which is used to analyze the OS development paradigm in detail. Cowston et al. [22] made an review and empirical research of literature on development of free and open source software (FOSS) development. Weber [25] explains the open source process – how it came into being, how it works, and how it is organized. Scacchi [26] reviewed what is known about free and open source software development (FOSSD) work practices, development processes, project and community dynamics, and other socio-technical relationships. His work focuses on exploring how FOSS is developed and evolved based on an extensive review of a set of empirical studies of FOSSD projects that articulate different levels of analysis.

IV. Analysis of OSS success and quality criteria

Different measures to define the success of open source projects have been proposed in the literature. Crowston et al. [23] identified a range of measures that can be used to assess the success of open source software projects. They identified measures based on a review of the literature, a consideration of the OSS development process and an analysis of the opinions of OSS developers. They showed that the existing models of information systems success suggest a range of potential success measures for OSS projects, but a number of the measures are inapplicable, while others are difficult to apply in the OSS environment. Authors consider a software successful when its source code is of high quality, e.g. it is highly modular, correct and maintainable. In another work, Crowston et al. [52] address a key gap in Free and OS software (FOSS) research by seeking to define what ”success” means in a FOSS context. They derived a range of potential measures from the Information Systems literature, including system and information quality, user satisfaction, user experience, individual and organizational impacts, and then added a number of additional measures specific to the dynamics of the FLOSS development process.

Schweik and English [59] studied more than 174,000 open source projects shared on SourceForge, one of the largest and oldest repositories for open source code. According to them, a successful product achieves at least three software releases and has value for ”at least a few users.” As evidence of user value, they looked at project attributes such as the number of downloads and installations, the presence of continued development activity, posts to project discussion boards or

email lists and the addition of one or more developers. Midha and Palvia [51] examined two measures of project success: project popularity and developer activity, of 283 OSS projects over a span of 3 years, in order to observe changes over time. They made a comprehensive research model of OSS success which includes both extrinsic and intrinsic attributes. Maruer and Scotchmer [40] state that open source project success may depend on accident as much as quality. Comino et al. [45] provided the working definition of OS success and presented a survey of the literature describing the OS success phenomena. According to Comino et al, OS success can be classified into three categories: i) software use, ii) size of the community and/or its level of activity and iii) technical achievements of the project.

A. Software Quality Aspects

The development of open source is often perceived especially in business as being of low quality, due to its economy, team structure or philosophy. Yet, Raymond [1], [2] claims that OSS is cheaper, has fewer defects, gets improvements faster, and is generally better than other kinds of software. Kuan [44] argues that when open source programs are founded, software quality will surpass that of comparable closed source programs.

Boulanger [19] made the comparison of security and reliability between FOSS and proprietary systems, and showed FOSS systems can meet, or even exceed, the quality of their proprietary counterparts. Stamelos et al. [24] suggested that OSS code is generally of good quality, but, on the other hand, is not always as good as expected [9]. Mockus et al. [3] hypothesizes that the OSS products have lower defect density than commercial software and that OSS development rapidly responds to user requests.

Technical Aspects: Raymond [1] suggests that successful software development must be based on flexible approaches in opposition to rigid cathedral-like approaches. He states two laws for successful open source production paradigms. The first deals with the activity of bug reporting and fixing within open source projects. It states that given enough eyeballs, all bugs are shallow. The work of finding issues or solutions to them does not require much cooperation, and thus can be efficiently divided even to very large groups of people. Also finding issues doesnt require any special skills. Such practice may therefore lead to quick implementation of new features and solutions thank also to the interplay with the other open source development law: release often and release early. Early releases of the open source programs enhance the benefit of decentralized beta tests. Bug fixing and new features are quickly included in the official version of the program [14].

Johnson [38] argues that one reason the open environment can be successful is that there are many small tasks that can be completed for any particular project. With many small tasks it becomes more likely that any individual will find it worthwhile to contribute, increasing aggregate development.

Narduzzo and Rossi [43] argue that modularity, which can be regarded as an innovative manufacturing paradigm for the design and the production of complex artifacts, is a key element in explaining the development and the success of many FOSS projects, and it offers a comprehensive explanation of many key issues in FOSS development, such as how division of labor takes place within developers, how coordination is achieved and how code testing and integration is deployed, how innovation occurs, and so on. Similarly, Benkler [42] also claims that modularity is the key to making open source projects viable

Johnson [41] made a systematic investigation of the role of modularity in open source. He argues that the success rate of modular projects grows with the number of developers because programmers can self-select into working on the modules for which they are most proficient. The main drawback is that the project may founder if costs are badly distributed among modules.

Giuri et al. [49] show that modularity improves the performances of OS projects measured in terms of the number of bugs fixed, CVS commits, or files released to the community.

B. Community Aspects

According to the first two categories of OS success defined by Comino et al. [45], OS software is successful when it is widely adopted among users or when a large and active community of developers contributes to its production.

Ljungberg [55] states that the stories of the open source projects are full of descriptions of the importance of the founders as entrepreneurs, as dedicated persons, as technical innovators, and good designers. Johnson [38] claims that one major argument for why the open source system should be successful is that it engenders interactions and synergies between developers.

In Fershtman and Gandal [46] software success is measured in terms of output per contributor, i.e. the average number of source code lines written by each contributor. The analysis is based on a sample of 71 open source projects observed during a period of eighteen months, and it shows that output per contributor is significantly higher when the software is distributed under non-restrictive licensing terms. A similar result is obtained by Stewart et al. [47] in a study on 138 open source projects registered on Freshmeat. They measure project success based on the number of subscribers associated with a project, and find that this number is larger when the software is released under non-restrictive schemes and sponsored by non-market organizations.

Capiluppi et al. [8] provide evidence that the majority of OSS projects indeed struggle to attract contributors. Bonaccorsi and Rossi [14] made the comparison of motivations of individual programmers and firms to take part in the open source movement. They showed that the top reasons for firms to participate in the open source movement are the ability for small enterprises to afford innovation, contributions and feedbacks from the community, the reliability and quality of the open source software, and the independence of the price and license policies of the large software companies.

Capiluppi et al. [20] made a systematic analysis of around 400 projects from a popular OS project repository, statically and over time. They showed that only few projects were capable of attracting a meaningful community of developers, but the majority of projects is made by few (in many cases one) person with a very slow pace of evolution. Ghosh et al.

[27] showed that most open source projects are very small scale - 29% of projects have just one author, only 45% had more than two, and only one percent had more than fifty.

Fuggetta [34] argues that most of the success of products such as Linux is to be credited to the commitment and ability of the people who participate in their development.

West [37] states that one key variable for OS success may be the size of the user-programmer community. Anecdotal evidence suggests that the major reason open source projects fail is a lack of user-contributors to do the work. In this case, sponsors of open source projects (whether organizations or individuals) vendors face a particularly important adoption challenge: to attract enough of the right sort of users early enough to improve the quality and features of the software.

Kivekas suggests [33] that the successful work of a free and open source software development (FOSSD) team is largely based on creating and utilizing shared mental models. A shared mental model means that the team members have similar understanding on topics such as the project goals and limitations, the roles of different software modules, the target software architecture, the division of labor, etc. Crowston et al. [31], [32] propose that more highly developed and detailed shared mental models improve efficiency of cooperation by reducing the effort needed to collaborate.

Healy and Schussman [35] state that OSS projects are most often staffed by skilled professional software developers. They claim that the role of project leaders in mobilizing development is crucial and states that effective project leadership seems to us one of the most likely candidates for differentiating successful projects from unsuccessful ones. Authors also suggest that, despite the canonical image of OSS development as a free-for-all bazaar, hierarchical organization is central to the success of important projects. This hierarchy is not a formal organizational chart but rather a status-based pecking order which is known to project participants and serves as a way of policing members.

According to Raymond [16], the primary benefits that accrue to successful contributions of OS projects are good reputation among one's peers, attention and cooperation from others, and higher status in the exchange economy.

The so called gift economy [15] states that giving away goods and services for free allows to make and maintain social links and implies the duty to reciprocate [58]. In writings on the open source software development model, it is often argued that it is successful as a result of the gift economy that embraces activities in online communities. Raymond explicitly declares that the society of open source hackers is in fact a gift culture [16]. According to Raymond, gift giving practises rely on two factors: abundance of resources (as opposed to scarcity) and willing to signal social status. Gifting away the code allows getting new ideas and prototypes out into circulation [14], [17].

Weber [25] argues that open source software is not simply a non-rival good in the sense that it can tolerate free riding without reducing the stock of good for contributors. It is actually anti-rival in the sense that the system positively benefits from free riders. Some small percentage of free riders will provide something of value to the system, even if it is just reporting a bug out of frustration. The more free riders in this setting, the better. This argument holds only if there are a sufficient number of individuals who do not free ride - in other words, a 'core' group that contributes to the creation of the product. Open source projects have demonstrated that a large and complex system of code can be built, maintained, developed, and extended in a non-proprietary setting where many developers work in a highly parallel, relatively unstructured way and without direct monetary compensation.

C. Buisiness Aspects

Commercial businesses tied to open source projects provide additional resources for developing the free components of the software, but even more importantly it helps to promote the open source packages and drive them into the mainstream [13].

Companies make money out of open source software in at least three ways: on distributing the open source software; by adding value to the open source software by additional proprietary products; and by relating to open source software in different ways, such as bundling it with own products [12], [55].

Cook [21] argues that most of the success of open source software development is due to factors that ultimately cannot be transferred to closed-source software development in the business organization.

While traditional commercial proprietary software typically depends on monopoly support with one company providing support and "holding all the cards" (i.e., access to the code) for a piece of software, the publicly available source code for OSS enables many vendors to learn the platform and provide support [57]. Open source can create support that lasts as long as there is demand, even if one support vendor goes out of business.

An important economic incentive to take part in the open source movement deals with the emancipation from the price and license conditions imposed by large software companies. The business models of firms working with open source software are mainly based on the supply of software related services [18]. The open source framework, in fact, does not allow to profit from license fees. If these firms supplied services such as installation, integration, maintenance or support based on proprietary programs, they should pay license fees to the companies owning proprietary rights on such programs. Working with open source software these firms avoid license fees. This lowers down their production costs [14].

Foroutan [53] identified the risks of open source projects in the fields of development and licensing and offered suitable methods that can either prevent or minimize the damage inflicted by them. He states that, ultimately, it is the developers that will lead the project to success or failure. He argues that it is not that the correct license will absolutely ensure success, but rather that the wrong choice of license have a high chance of ruining a projects present and future plans for profit.

Fogel [30] states: "An open license does not guarantee that hordes of active developers will suddenly volunteer their time to your project, nor does open-sourcing a troubled project automatically cure its ills".

Fuggetta [34] states that there are particular market situations where open source is probably the only viable solution to

support successful and effective software development. Typical examples are research communities that need specific software products to support their research work. In this case, the expertise needed to develop this critical and specific kind of software, the cost of developing it, and the limited market base for the resulting products make it difficult to exploit a traditional approach to software development. Typically, in these situations software is cooperatively developed by a number of researchers worldwide. This is a situation where open source software appears to be the only viable solution.

V. Examples of Succesfull OSS

As many cases can demonstrate, open source software can also be a commercial success. Many open source software programs are widely used today. The best known is the computer operating system Linux and web server software Apache. RedHat Linux is probably the first ever corporate establishment to make it big with open source. According to Wikipedia [63], RedHat cloaked revenues of US$ 1,534.615 in 2014. Linux is the #1 server OS on the public Internet (counting by domain name), and the #2 web serving OS on the public Internet (counting by physical machine) [62]. Apache is a webserver software which can be called as one among the most important build block of Internet as we know it today. Since April 1996 Apache has been the most popular web server software in use. As of February 2015 [61] Apache served over 39.47% of all websites on the Web. Feller [50] cites O'Reilly who points out that although Linux and Apache are high profile, the more significant success of OSS can be seen behind the scenes, in the Internet "category killers" such as BIND, which runs the Domain Name Server (DNS) for the World Wide Web, and Sendmail, which is central to the Internet mail backbone [50].

VI. Conclusion

In this paper, we have conducted an analysis of various success criteria for OSS projects. We can conclude that the success of an OSS project depends on several crucial factors. Importantly, such project must attract a sufficient number of highly motivated and skilled contributors willing to invest their own time to the benefits of the project. The most successful open source projects are those not only with excellent code bases but thriving communities of users who have a need that can be met by the software. Successful OSS project must have an independent and powerful leader - a team or a individual - who will make prosperous decisions, and manage and supervise the project. A leader will have to establish well-articulated and clear goals for the project. It is crucial to have a relatively clearly defined vision and a mechanism to communicate the vision early in the projects life [59]. OS projects' collaborators can be part-time volunteers spread around the world, as well as the full-time employees (core-team) paid to develop key features and to maintain the software. For a successful OSS project, it is of vital importance to have a solid users community support and an active volunteers base. In the OS development process, volunteers play a vital role, especially in the area of software testing and bug discovery.

However, it it important to distinguish the success of the OSS project from a business accomplishments of a company which has started the OSS project.

The realization of significant business success is possible through OSS in specific business situations, where companies have developed the OS products interesting to the large number of customers, while achieving a superior quality, or have developed products of a unique value with no real business competition. For that matter, the nature of open source affects the projects' chances to thrive in the business market and achieve success. This is, to some extent, the prerequisite for the above-mentioned criteria of the OSS projects' success.

Commercially-oriented OSS can be successful and recognized as a software project, but that does not guarantee financial success for the company. Unlike proprietary software, where the success of a product usually implies the financial and business success of the company, success of a OSS do not necessary lead to the prosperity of a company or individual behind a project. If we ignore financial and time dimension of a success criteria, we can say that the OSS project is successful if it has a relatively large number of users satisfied with the quality and usability of the software. In this respect, it is of crucial importance to thoroughly understand the users' needs and demands, in a manner of improving old and implementing new product features.

References

[1] E. S. Raymond, "The cathedral and the bazaar," Knowledge, Technology & Policy, vol. 12, no. 3, pp. 23-49, 1999.

[2] E. S. Raymond, "Up from alchemy - open source development," Software, IEEE, vol. 21, no. 1, pp. 88-90, 2004.

[3] A. Mockus, R. T. Fielding, and J. D. Herbsleb, "Two case studies of open source software development: Apache and Mozilla," ACM Transactions on Software Engineering and Methodology, vol. 11, no. 3, pp. 309-346, 2002.

[4] B. Fitzgerald, "The transformation of open source software," Mis Quarterly, pp. 587-598, 2006.

[5] A. Fuggetta, "Open source software - an evaluation," Journal of Systems and Software, vol. 66, no. 1, pp. 77-90, 2003.

[6] K. Crowston, and J. Howison, "The social structure of free and open source software development," First Monday, vol. 10, no. 2, 2005.

[7] W. Scacchi, "Free and open source development practices in the game community," Software, IEEE, vol. 21, no. 1, pp. 59-66, 2004.

[8] A, Capiluppi, P. Lago, and M. Morisio, "Evidences in the evolution of OS projects through changelog analyses," Proceedings of the 3rd Workshop on Open Source Software Engineering, 2003.

[9] I. Stamelos, et al. "Code quality analysis in open source software development," Information Systems Journal, vol. 12, no. 1, pp. 43-60, 2002.

[10] Open Source Initiative: The Open Source Definition [Online]. Available: http://www.opensource.org/docs/definition.php. [Accessed: February 21, 2015].

[11] V.K. Gurbani, A. Garvert, and J. D. Herbsleb, "A case study of a corporate open source development model," in Proceedings of the 28th international conference on Software engineering, ACM, 2006.

[12] T. OReilly, "The open-source revolution, release 1.0," Ester Dysons Monthly Report, 1998.

[13] J. Ousterhout. "Free software needs profit," Communications of the ACM, vol. 42. no. 4, pp. 44-45, 1999.

[14] A. Bonaccorsi, and C. Rossi, "Comparing motivations of individual programmers and firms to take part in the open source movement: From community to business." Knowledge, Technology & Policy, vol. 18, no. 4, pp. 40-64, 2006.

[15] M. Mauss, "The gift: The form and reason for exchange in archaic societies," WW Norton & Company, 2000.

[16] E. S. Raymond, "Homesteading the noosphere," First Monday, vol. 3, no. 10, 1998.

[17] M. Bergquist, and J. Ljungberg, "The power of gifts: organizing social relationships in open source communities," Information Systems Journal, vol. 11, no. 4, pp. 305-320, 2001.

[18] T. Wichmann, "Firms open source activities: motivations and policy implications," FLOSS Final Report, International Institute of Infonomics, 2002.

[19] A. Boulanger, "Open-source versus proprietary software: Is one more reliable and secure than the other?" IBM Systems Journal, vol. 44, no. 2, pp. 239-248, 2005.

[20] A. Capiluppi, P. Lago, and M Morisio, "Characteristics of open source projects," in Proceedings of the Seventh European Conference on Software Maintenance and Reengineering, IEEE, 2003.

[21] J. E. Cook, "Open source development: An Arthurian legend," in Making sense of the bazaar: Proceedings of the 1st Workshop on Open Source Software Engineering, 2001.

[22] K. Crowston, "Free/Libre open-source software development: What we know and what we do not know," ACM Computing Surveys, vol. 44, no. 2, 2012.

[23] K. Crowston, H. Annabi, and J. Howison, "Defining open source software project success," in Proceedings of the International Conference on Information Systems, 2003.

[24] I. Stamelos, et al. "Code quality analysis in open source software development," Information Systems Journal, vol. 12, no. 1 pp. 43-60, 2002.

[25] S. Weber, "The political economy of open source softwar," Berkeley Roundtable on the International Economy, 2000.

[26] W. Scacchi. "Free/open source software development: Recent research results and methods," Advances in Computers, vol. 69, pp. 243-295, 2007.

[27] R. A. Ghosh, et al. "Free/libre and open source software: Survey and study," 2002.

[28] W. Scacchi, et al. "Understanding free/open source software development processes," Software Process: Improvement and Practice, vol. 11, no. 2, pp. 95-105, 2006.

[29] G. Madey, V. Freeh, and R. Tynan, "Modeling the free/open source software community: A quantitative investigation," Free/Open Source Software Development, pp. 203-221, 2005.

[30] K. Fogel, "Producing open source software: How to run a successful free software project," O'Reilly Media, 2005.

[31] K. Crowston, R. Heckman, H. Annabi and C. Masango, "A structurational perspective on leadership in free/libre open source software teams," in Proceedings of the First International Conference on Open Source Systems, 2005, pp. 9-15.

[32] K. Crowston, et al. "Coordination of free/libre and open source software development," 2005.

[33] O. Kiveks, "Free/open source software development: results and research methods," 2008.

[34] A. Fuggetta, "Open source software - an evaluation," Journal of Systems and Software, vol. 66, no. 1, pp. 77-90, 2003.

[35] K. Healy, and A. Schussman, "The ecology of open-source software development," Unpublished manuscript, 2003.

[36] J. West, and J. Dedrick, "Open source standardization: the rise of Linux in the network era," Knowledge, Technology & Policy, vol. 14, no. 2, pp. 88-112, 2001.

[37] J. West, "How open is open enough?: Melding proprietary and open source platform strategies," Research Policy, vol. 32, no. 7, pp. 1259-1285, 2003.

[38] J. P. Johnson, "Economics of open source software," 2001.

[39] D. Koski, et al. "Fuzz revisited: A re-examination of the reliability of UNIX utilities and services," University of Wisconsin-Madison, Computer Sciences Department, 1995.

[40] S. M. Maurer, and S. Scotchmer, "Open source software: the new intellectual property paradigm," National Bureau of Economic Research, No. w12148, 2006.

[41] J. P. Johnson, "Open source software: Private provision of a public good," Journal of Economics & Management Strategy, vol. 11, no. 4, pp. 637-662, 2002.

[42] Y. Benkler, "Coase's penguin, or, Linux and 'the nature of the firm'," Yale Law Journal pp. 369-446, 2002.

[43] A. Narduzzo, and A. Rossi, "Modularity in action: GNU/Linux and free/open source software development model unleashed," Department of Computer and Management Sciences, University of Trento, Italy, No. 020, 2008.

[44] J. Kuan, "Open source software as lead users make or buy decision: a study of open and closed source quality," Stanford Institute for Economic Policy Research, Stanford University, 2002.

[45] S. Comino, F. M. Manenti, and M. L. Parisi, "From planning to mature: On the success of open source projects," Research Policy, vol. 36, no. 10, pp. 1575-1586, 2007.

[46] C. Fershtman, and N. Gandal, "The determinants of output per contributor in open source projects: an empirical examination," Social Science Research Network, no. 515282, 2004.

[47] K. J. Stewart, A. P. Ammeter, and L. M. Maruping, "Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects," Information Systems Research, vol. 17, no. 2, pp. 126-144, 2006.

[48] R. Grewal, G. L. Lilien, and G. Mallapragada, "Location, location, location: How network embeddedness affects project success in open source systems," Management Science, vol. 52, no. 7, pp. 1043-1056, 2006.

[49] P. Giuri, et al. "Skills, division of labor and performance in collective inventions: Evidence from open source software," International Journal of Industrial Organization, vol. 28, no. 1, pp. 54-68, 2010.

[50] J. Feller, and B. Fitzgerald, "A framework analysis of the open source software development paradigm," in Proceedings of the twenty first International Conference on Information Systems, Association for Information Systems, 2000.

[51] V. Midha, and P. Palvia, "Factors affecting the success of open source software," Journal of Systems and Software, vol. 85, no. 4, pp. 895-905, 2012.

[52] K. Crowston, J. Howison, and H. Annabi, "Information systems success in free and open source software development: Theory and measures," Software Process: Improvement and Practice, vol. 11, no. 2, pp. 123-148, 2006.

[53] P. J. Foroutan, "How to successfully develop and manage open source projects," 2012.

[54] Free Software Foundation [Online]. Available: http://www.fsf.org/fsf/fsf.html [Accessed: February 22, 2015].

[55] J. Ljungberg, "Open source movements as a model for organising," European Journal of Information Systems. vol. 9, no. 4, 2000.

[56] Y. Yin, "Geographically distributed development: trends, challenges and best practices," Massachusetts Institute of Technology, 2007.

[57] C. Kenwood, "A business case study of free/open source software". in Proceedings of the Open Source Software Development Workshop, Unversity of Newcastle upon Tyne, 2002.

[58] A. Bonaccorsi, and C. Rossi, "Why open source software can succeed," Research Policy, vol. 32, no. 7, pp. 1243-1258, 2003.

[59] C. M. Schweik, and R. C. English, "Internet success: a study of open-source software commons," MIT Press, 2012.

[60] Knight Lab. "Six things to know about successful open-source software," [Online]. Available: http://knightlab.northwestern.edu/2013/07/24/six-lessons-on-success-and-failure-for-open-source-software/ [Accessed: February 20, 2015].

[61] Netcraft Web Server Survey [Online]. Available: http://news.netcraft.com/archives/category/web-server-survey/ [Accessed: February 22 2015].

[62] D. A. Wheeler. "Why open source software/free software (OSS/FS, FLOSS, or FOSS)? Look at the numbers!" 2005.

[63] Wikipedia - Red Hat. [Online]. Available: http://en.wikipedia.org/wiki/Red_Hat [Accessed: February 21, 2015].

[64] H. Van Vliet. "Software engineering," Wiley & Sons, 2008.

Analysis of e-government portal in Montenegro

PhD Vujica Lazovic
MSc Tamara Djurickovic
PhD Biljana Rondovic
PhD Dijana Kovacevic
MSc Zorana Popovic
Faculty of Economics
Jovana Tomasevica 37, Podgorica, Montenegro
e-mail: tamara.djurickovic@ef.ac.me
Tel: +382 20 241 138

Abstract - State governments, through the digitization of services that have been provided through conventional ways, have made possible to the citizens and businesses to find and use information and services electronically. Through the formation of e-government portal, as a single point of access to services, governments have greatly facilitate business and communication between relevant institutions and stakeholders. Subjects can be informed, they can download forms but also they can realize services using portal. The above is made possible with the clear objective of removing barriers that are an integral part of the administrative system.

In this paper, we will give a brief theoretical review of the concept of operations of the state administration in the provision of services through e-government portal, starting with the concept and definition of e-government, with the goals of this concept, to the benefits of its application from the perspective of the citizens and business as users, and the government itself as provider. Also, the paper aims to present and analyze the level of realization of services provided at e-government portal in Montenegro, and also to identify possible problems and give recommendations for further improvement of the same.

Key words: e-government, public services, one-stop-shop, portal

I INTRODUCTION

Significant changes that have occurred in the field of information and communication technologies development in the last two decades have influenced the change in the earlier paradigm of business. The development of the Internet and the emergence of a growing number of companies using the Internet as a means to reduce costs, but also to save time, was one of the main impulses that influenced the governments of a number of countries to launch a variety of electronic services that are needed to facilitate the operations of private business, and ultimately through the development of entrepreneurship and the private sector to become more competitive in a growing global market.

Realizing that the development of information and communication technologies opens up a space for exchange, by nature inefficient bureaucracy portals with modern electronic government, many states are deciding to withdraw the move to facilitate access to its services for a large number of citizens and companies. With digitizing a large number of processes and procedures the states have been able to influence the increase of efficiency and productivity of a large number of services, which ultimately had a positive impact on the development of the business environment in their economies as well in improving the competitiveness of the economy as a whole.

Modern public administration is significantly different from traditional public administration. Although many factors have contributed to the steady transformation of the work, it is clear that the initiator of the all changes was the use of information and communication technologies (ICT) in the daily work of the governments and its agencies. Traditional public administration is largely recognized under the concept of bureaucracy, and bureaucracy is by definition inefficient. The application of knowledge management in public administration leads to its higher productivity, and thus the competitiveness of the national economy. It is this shift that marked a new era in the work of public administration, which is now known under different terms, such as an electronic public administration, networking governance and the like. The aim of these efforts and initiatives is to move closer to services and intensifying communication of public administration with citizens and representatives of industry, bearing in mind above all, the increase of the efficiency and transparency of the process.

Governments and public sector organizations around the world are trying to combine basic, but often conflicting, demands: to provide better, more relevant public services that are focused on the needs of the citizens and businesses they serve; to reduce costs but also improve the efficiency of their operations; as well as to change and improve their supply chains in order to create processes that are quick, cheap and effective.

However, the application of technologies and advanced technological solutions is not the only solution to the implementation of electronic governance. The technology is now available especially when it comes to small countries which did not require large investments in

infrastructure and services. It is necessary to bear in mind the extent to which people own a computer, from which place they most commonly use internet services, what kind of equipment they use, which is the level of internet penetration, what kind of public key infrastructures is, whether digital certificates are available to the average user and the like.

If governments want to get liberated of the growing internal and external pressures, they need to review, redesign and optimize its services, and to put the users - citizens, businesses and NGOs - in focus, i.e. center of their operations, rather than the needs of government organizations and service providers. Also, they need to shift from outdated business models, management, culture, technology and processes that are inherited from some prior periods quickly and efficiently. This transition will involve destructive and very painful changes from closed, top-down, bureaucratic and paper-based transaction services to online, integrated digital offers that will encourage and enable a new way of interaction between citizens and the state.

II THE CONCEPT AND DEFINITION OF E-GOVERNMENT

e-Government or e-business in the public administration is the use of information and communication technologies combined with organizational changes and new skills in order to improve the delivery of public services, increase of democratic participation and involvement of the public in decision-making and policy-making processes [11]. So, e-government use modern achievements of information and communication technologies in order to provide greater productivity and heterogeneity of its activities, in order to meet the needs of citizens. Defining the concept of e-government, it is clear that it has adopted the famous marketing concept: in the focus of marketing management must always be the consumer himself, or transferred to the field of e-government: the focus of the public sector is a citizen. On that basis, it is clear that the modern public administration is highly qualified, initiative, solid, responsible and efficient network of public services. The most important consequence of the introduction of information and communication technologies in public administration is the elimination of physical contact between the users and service administration, and increased efficiency and transparency of the public sector. Performing numerous procedures online, either citizens or legal persons, represents a tremendous savings in time, and that is the most valuable resource in the possession of any individual or business entity. In addition to increasing efficiency, public administration with the introduction of ICT in the work contributes to the creation of new skills that are expected of employees in the public sector, thus changing the situation on the labor market.

There are different approaches when it comes to measuring the level of e-government development. If we focus on the concept of e-governance, it is possible to analyze the different levels of maturity of e-Services, e-Management, e-Democracy and the like. Most of the measurements tries to encompass all of these questions, but the scope of the analysis is often narrow and restricted. As long as it comes to the used methodology, the assessment is carried out by combining several techniques including web research, questionnaires, face-to-face interviews and the like. However, reliable and relevant measurements can provide key inputs to point to policy makers and practitioners to chances but also the problems they may face and to help them, in this regard, to focus in the right way.

One indicator of the level of development of e-government is e-government index, developed and administered by the United Nations, which is widely recognized as the authoritative measure of the capacity of the public sector in the provision of electronic and mobile services. According to research by the United Nations for 2014 [12], the index of development of e-government (e-Government Development Index -EGDI) for Montenegro has shown that Montenegro is in the category of high-income countries (High EGDI between 0.50 and 0.75). According to these data, it is considered that there is considerable scope and opportunity for Montenegro to continue the development of e-government. Government can achieve the development of e-government with implementation of clear strategy, continuous investments in primary, secondary and tertiary education, as well with the radical transformation in the offer of online public services available to citizens.

With the aim of easier facilitating the functioning in the world of internet, variety of platforms have been created. The purpose of their creation is to allow everyone access to services and portals. That is why it is necessery to allow everyone access to the Internet, and then make most of the work available online, for which it took a long time to finish in the real world, which will ultimately save time and make customers satisfied.

III DEVELOPMENT OF E-GOVERNMENT IN MONTENEGRO

In late 2010 the European Commission adopted the Digital Agenda for Europe as one of the seven most important initiatives of the European strategy 2020, which was established to define the key role of information and communication technologies in social and economic development of the country. In this context, the main goal of "Information Society Development Strategy 2012-2016. - Montenegro digital society", which is the most important strategic document that defines and provides a road map for Information Society Development in Montenegro, is planning activities that will maximize the social and economic potential of ICT, especially the Internet as an important means of economic and social activities [11].

As one of the priority areas of Information Society development, this document has recognized the development of e-government. Through the use of information and communication technology, substantially will be improved the delivery of public services with improvement of democratic participation and public involvement in decision-making and policy-making. The main goal of e-government is to increase the availability of public services to citizens and businesses. It should also enhance the effectiveness of state management, as well as to enable it to have better insight into the allocation of economic and social resources. Through efficient interaction of all participants - state, private and corporate customers - will stimulate political, social and economic development of the country.

Respecting the guidelines defined in the policy document for the development of information society in Montenegro [11] as well as EU standards, with a goal to implementing e-government, in Montenegro was developed web portal www.euprava.me who has aims to bring together in one place, consolidate, and enable electronic realization of services used by citizens, businesses and employees in the state administration.

IV ANALYSIS OF THE E-GOVERNMENT WEB PORTAL IN MONTENEGRO

Ministry for Information Society and Telecommunications of Montenegro has implemented the project of developement and realization of internet e-government portal - www.euprava.me. The institutions of state administration and local government units are providing via the aforementioned portal services to citizens and legal persons, as well as other institutions, by electronic means.

Web portal is a unique platform that provides (some) electronic public services offered by public administration. The basic principles on which web portal is built are: security, control, simplicity, transparency, services created by the user's needs, scalability, flexibility and adaptability looks, adapting to people with special needs, in person, and for anyone, anytime-anywhere, connect with other systems [15]. It is a unique place on the Internet where you can find all the public services offered by public administration. E-government services are available 24 hours 365 days a year. The aim of the implementation of the e-government portal is to provide an easier, cheaper, more transparent communication between government and citizens, government and companies and government agencies themselves. The government today operates through networked databases, and automated administration, which represents the essence of e-government. Therefore, the goal of e-government is to improve service delivery to citizens, and thus save time in a manner which will provide their services via the Internet.

e-government portal has started on 7th of April 2011 in cooperation with the five state institutions with the initial operation of the 12 services. Currently web portal provides 77 electronic services, which provide 25 institutions. The portal is based on a one-stop-shop system (all services in one place for the user), along with a complete or partial re-engineering processes and operations. In addition to the information and services on the portal grouped into three categories, such as services for: citizens, legal entities/businesses and government officials, information and services on the website are grouped in several ways, so that users can easily access the desired services. Information on this website is grouped according to the situations of life, in order to follow all significant situations of citizens (education, business, issuing documents, etc.) or legal entities (business and labor, document, finance, etc.).
Moreover, services within the portal euprava.me are defined through three "statuses" ie. the degree of realization, which are:

- informative services - services that provide users with information on how they can be implemented in a physical way, whether they are necessary and which documents, what it the competent institution which implements the service, etc. These services can not be realized electronically.
- electronic services that can be realized through the portal euprava.me, and can be done in two ways:
    - registering by e-mail adress
    - signing with the digital certificate

Depending on the type of the service, the portal supports the following system users: *anonymous users* - users who access the portal anonymously, without prior registration and they are allowed use of the services of the state administration on the level of information about procedures, contact details for the implementation of procedures and other necessary information about the institutions, as well as all the information of public importance which have the character of free access; *registered users by e-mail* - users who access the portal by registering the mail address, without the use of qualified and valid certificates; portal enables the use of government services, in addition to obtaining information, at the level applying for services that do not require digital signatures; *and registered users* - users who are accessing the portal with qualified and valid digital certificates. Registered users may be individuals, legal entities and civil servant who use the services of the portal as well as authorized officials of the relevant institutions (employees of public institutions) responsible for providing services and the provision of services on the e-government portal. Registered users can register via qualified digital certificates issued by any of the registered certification authorities (CA[1]) in Montenegro

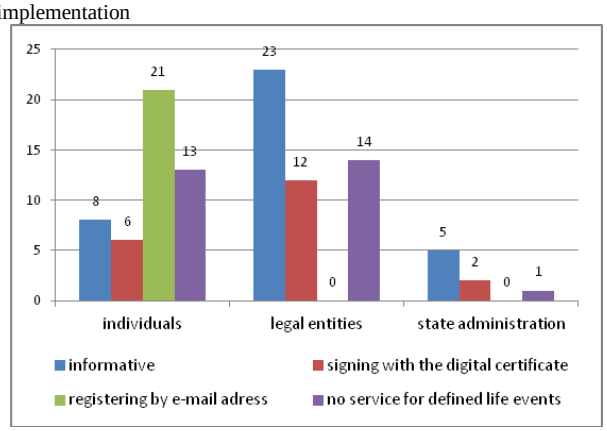
[1] Post of Montenegro is qualified Certification Body in Montenegro managing public key infrastructure (PKI-PoštaCG) for public needs of Montenegro. As part of the PoštaCG-PKI for the purpose of providing certification services is established certification body with a Single Rooted Certification Authority - PoštaCG-CA, which issues qualified

and can use the site's services at the highest level that has been implemented. For registered users, the portal identifies what type of users they belong to (citizen, business entity, employed in public administration, public administration) and assign rights that the user may get on the portal. Registered users can use the services of the Portal at the highest level of implementation, for example, citizens and businesses can, for certain services, request form necessary for receiving the services, fill it, sign electronically, electronically pay the administrative fee and electronically submit the request. Portal assigns each request to the competent institution, where it is processed classical and elecronically. Registered users can at any time check the status of their application. This procedure applies to both physical and legal persons. Registered users employed by the institutions in charge of public services have assigned corresponding rights that are recognized by the portal, and depending on the rights, they can set up the service of the state administration, create electronic forms on the site, track user requests, administer electronic public debates on the Portal and the like. Portal keeps a history of all activities and identities of all participants in the process [19].

Taking into account the level of development of e-government portal services, for this study was conducted web research of the euprava.me portal, through an analysis of the number of available services, as well as the extent to which the services are developed. Bearing in mind that the services can be informative, and that can be electronic services that can be implemented by logging with e-mail or using a digital certificate, we have, for this study, used these three categories for "evaluation" the level of implementation. In addition, we have analyzed it in terms of categories of users that use the available services.

Bearing in mind the above mentioned, although that 77 services are available on the portal euprava.me, information on the degree of realization of these services show that 36 service are developed just to the level of providing information, while the 41 service can be realized by electronic means (including 20 service using digital signatures and 21 service through registration by e-mail). When it comes to the services that are informative, 8 services are related to the services for citizens, 23 are services for business entities, while only 5 services are for employees in the state administration. For services that can be executed electronically and using a digital certificate, the six services are for citizens, 12 service are to local businesses and 2 are realted to the services of Public Administration. On the other hand, of the services that can be done electronically by registering by e-mail, 21 are for individuals while for businesses and government officials are still no services in this category. Although it is expected that the businesses are the main users of electronic services, more of such services (which may be executed electronically) are developed for individuals. On the other hand if we analyze the types of these services, it is not surprising, bearing in mind that the services on the portal and divided by the institutions that offer them, and that a significant number of these services for individuals that can be realized by registering an e-mail are just service *request for free access to information* (portal has 14 such services). The data are presented in Table 1.

Figure 1. Distribution of the services based on users and level of implementation



The fact that there are more services that can be implemented using digital certificates are developed for businesses and employees in the state administration is fully correlated with the current state of development and use of digital certificates in Montenegro, bearing in mind the high cost of certificates, as well as the level of use and the citizens need to use a certificate.

Thus, the same number of services is developed for individuals and legal entities, although the distribution by way of provision of services is different. In addition to these services, which can be implemented in any of the three modes, the portal also lists the services for which there is no information, following: 13 for individuals, 14 for businesses and one for employees in the state administration. Also noticeable is that the particular service in parts of legal entities and individuals coincide, as for example Statistical data. In the category of individuals this service is not developed while in the category of legal persons it is developed. When it comes to the category of administration that is employed in the state administration, there are many services that are supported by the portal, but the state administration only

certificates to interested individuals and corporate entities. PoštaCG-CA uses one registration authority, which operates within the Post of Montenegro Ltd. and is authorized to verify the identity of users in the Management of certificates such as first issue of the certificates, renewal of certificates, certificate revocation, to grant a request for the certificate. The registration authority forwards the approved requests to the operational. Certificates issued by PoštaCG-CA can be used for different purposes depending on the certificate policy. Certificate policy is in every issues certificate marked in the extension certificatePolicies in accordance with the specification in RFC (Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation LICT - CRL Profile). Certificates issued by PoštaCG-Ca are allowed to be used for verification of digital or electronic signature, verification of identity and ensuring confidentiality [19].

needs to work on its expansion, which would imply better coordination between the state institutions.

Figure 2. Distribution of the services based on deegree of realisation for individuals

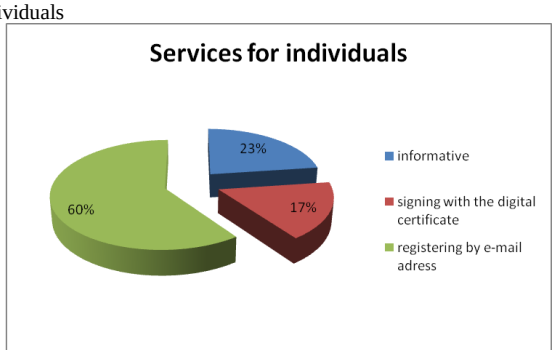


Figure 3. Distribution of the services based on deegree of realisation for legal entities

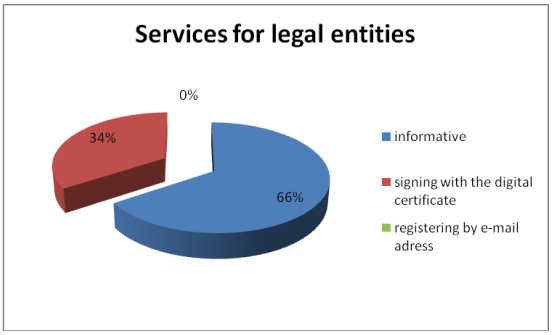
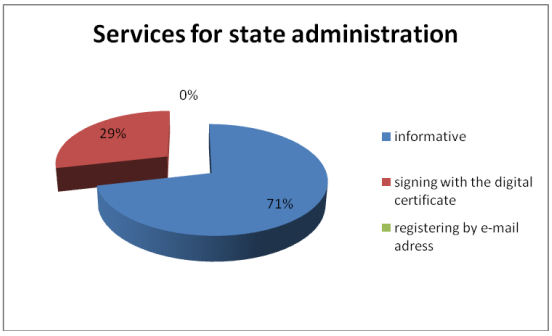


Figure 2. Distribution of the services based on deegree of realisation for state administration



Analyzing Monstat data, which state that in Montenegro Internet penetration is 60.3%, this means that around 400,000 citizens of Montenegro have access to the internet and the possibilities it offers [17]. In accordance with these data, the number of laptops and PCs in households was around 80% and the number of mobile phone users as much as 90%. All these data suggest that in Montenegro all households do not own a computer, but has at least one mobile phone. When talking about the e-government portal, it can be concluded that the approach to it is not possible for about 40% of the citizens of Montenegro due to lack of internet access. However, the fact that the penetration rate of mobile telephony is constantly growing, and that penetration in the fixed-line declines, shows that an increasing number of users opt exclusively for the use of mobile telephony. Every citizen has an average of 1.6 mobile phone which, from the aspect of the e-government portal, is a good predisposition to increase the degree of usability services. Also, while retaining all the functionality of Web versions of the portal, there is developed a mobile version of the portal which give access to the portal via WAP and XHTML web browser with mobile phones. At this time, the portal offers access to the system via the email address and password. Full realization of the mobile version of the portal will be realized after such a system (electronic certificate on the SIM card mobile phone) is developed by some of the operators in Montenegro.

Action Plan for the development of information society [14], that follows strategic document, defines the objectives for the development of e-government, which implies that by 2016, 50% of the citizens of Montenegro and 80% of companies need to have benefited from e-government services. Also, one of the tasks defined in the "strategy of Information Society 2012-2016" is that the 100 most frequently used government services will be available on the website until 2014 and the 200 most commonly used services will be available until 2016 [11]. Although the portal of public administration is faraway from the defined guidelines and objectives, it is clear that such activities are conducted both in the promotion and improvement in order to be recognized as a unique access point, as well as to make portal and services available for all categories of users to use the services at the highest level. This is supported by the fact that the Government of Montenegro, in January 2015, adopted a "Program of Economic Reforms in Montenegro for 2015-2017" which defines specific objectives and measures towards the development of e-government in Montenegro. Although the analysis of the state of e-government from 2013 showed that the total level of sophistication of e-Services in Montenegro is below the average of all EU countries, the commitment to the realization of defined measures from the Program of Economic Reforms, technical and normative preconditions for its implementation (the portal of electronic government and the Law on e-government) that are created, are the guarantor for improvements in this area. In contrast to the positive trends stand resistance of the administration to change and resistance to centralized access to services in the state administration over the e-government portal as well as the identified risks to the achievement measures. According to the Law on e-government, state authorities, state administration, local government, public services and legal persons exercising public powers are obliged to, within about 18 months from the date of entry into force, all the services they provide to citizens and businesses traditionaly to make available and electronically through the e-government portal. Although there are currently available 77 services on the e-government portal, which provide 25 institutions, the structure of the services are not satisfactory, and insufficient number of services relating to business and civil rights is available, which is the most measurable in terms of administrative barriers. In this sense, through strategic documents, action plans, reports on implementation of action plans as well as economic reform plans, the intention is that the more

services (relating to key life events, both for individuals and for businesses and employees in the state administration) is available online through the e-government portal.

V Conclusion

Technological progress has influenced to a governments facing the challenges of maintaining and directing the development of e-government until there is great uncertainty to a rapid technological change. Therefore, it is very difficult to predict the detailed impact of future policies. New technologies are very attractive because they often promise better solutions and create an environment in which certain business changes can happen. More often may, however, they promise solutions that include facilitating the implementation of information technologies without changes of their own business processes. Experience shows, however, that systems that are built on the unknown and emerging technologies are very susceptible to failure. In some cases, potential benefits may justify taking such big risks; but in most situations this is not the case.

E-government is a government that recognizes the opportunities and impact of information technology on management in terms of the digital economy and that makes maximum use of them to improve internal processes in order to provide high quality services and the establishment of management oriented towards users. As such it has to provide: effective realization of the rights and interests of citizens, the protection of human rights, the protection of public interest, improving the quality of life of every citizen, efficient and professional self-control and quality preparation of the law.

Montenegro is following current trends in the information economy which is confirmed by a strategic documents and concrete activities on implementation of the objectives defined by these documents. However, in order to let the country became part of the information age it is necessary that in addition to the institution, which prepares and delivers strategy in the field of information society, there are other institutions that will focus on particular segments of the electronic development. No matter what the e-government portal now has 77 services, it can be concluded that the administrative authorities have not recognized the importance of this portal and the benefits it provides in communication with citizens and businesses. The number of existing services in Montenegro is not entirely satisfactory and requires dynamic development of electronic services. As for the assessment of public administration services, they can be improved by adding the most current information and forms on websites. However, for significant progress in the development of e-government it is necessary to develop the back-office applications, which are reflected in the fact that it will provide services for users, realization of registers, integration of information from databases as well as a larger number of users of a service which will ultimately increase the value of the portal.

References

[1] Analyzing and Comparing European e-government Strategies, Konstantinos Parisopoulos, Efthimios Tambouris, Konstantinos Tarabanis, University of Macedonia, Thessaloniki, Greece, Informatics and Telematics Institute, Center for Research and Technology Hellas, (ITI/CERTH), Thessaloniki, Greece

[2] Benchmarking e-Government: Improving the national and international measurement, evaluation and comparison of e-Government, R. Heeks file: http://www.download-it.org/learning-resources.php?promoCode=&partnerID=&content=story&storyID=1994

[3] Benchmarking eGovernment: tools, theory, and practice, Cristiano Codagnone, Trond Arne Undheim, European Journal of ePractice · www.epracticejournal.eu 1 Nº 4 · August 2008

[4] Capgemini (2007). The User Challenge: Benchmarking The Supply Of Online Public Services, European Commission

[5] Communication From The Commission - Europe 2020: A strategy for smart, sustainable and inclusive growth, European Commission, Brussels 2010; http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF

[6] E-Government – The Digital Divide And Information Sharing: Examining The Issues, Thomas B. Riley, Commonwealth Centre for E-Governance

[7] e-Government benchmarking 2010, Malta information technology agency, 20 years of transformation, innovation, excellence, 2009

[8] E-Government in Digital Era: Concept, Practice, and Development, Zhiyuan Fang, Ph.D., International Journal of The Computer, The Internet and Management, Vol. 10, No.2, 200

[9] i2010 High Level Group - Benchmarking Digital Europe 2011-2015 - a conceptual framework, European Commission, Information Society Development, 27 October 2009

[10] Strategija razvoja Informacionog društva u Crnoj Gori od 2009. do 2013. godine, Ministarstvo za informaciono društvo, Vlada Crne Gore, 2009

[11] Strategija razvoja Informacionog društva 2012-2016 - Crna Gora digitalno društvo, Ministarstvo za informaciono društvo, Vlada Crne Gore, 2012

[12] United Nations E-Government Survey 2014: E-Government for the Future We Want , Department of Economic and Social Affairs, United Nations New York 2014

[13] Predlog mjera iz nadležnosti Ministarstva za informaciono društvo i telekomunikacije koje će direktno uticati na rast konkurentnosti crnogorske ekonomije, Ministarstvo za informaciono društvo i telekomunikacije, Vlada Crne Gore, mart 2015. godine

[14] Akcioni plan za 2012. godinu za implementaciju Strategije razvoja informacionog društva 2012-2016 godine, Ministarstvo za informaciono društvo i telekomunikacije, Vlada Crne Gore, 2012

[15] http://www.cesifo-group.de/ifoHome/facts/DICE/Public-Sector/Public-Governance-and-Law/Transparency/E-Gov-Dev-Index_DICE-Report_4_2012/fileBinary/E-Gov-Dev-Index_DICE-Report_4_2012.pdf

[16] http://unpan3.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2014

[17] www.monstat.org

[18] www.euprava.me

[19] http://www.euprava.me/WebResources/Documents/User_manual_me_lat.pdf

Determinants of electronic commerce - the factors of acceptance in Montenegro

Phd Biljana Rondović
PhD Vujica Lazović
MSc Zorana Popović
MSc Tamara Đuričković
Phd Dijana Kovačević
Faculty of Economics
Jovana Tomasevica 37, Podgorica, Montenegro
e-mail: tamara.djurickovic@ef.ac.me
Tel: +382 20 241 138

Abstract - The paper presents the results of the level of adoption of e-commerce in Montenegro and provides an overview of the factors which slows the use of its full potential. The aim of the research is to create a picture of the perception of e-commerce in terms of consumer and to identify the determinants that can be affected in order to create a better environment for electronic transactions in Montenegro.
Quantitative research (conducted through survey) and qualitative research (conducted through assessment of web sites of companies that are on the supply side), are associated with the subjective judgment of the authors, in order to better examine the key issues that affect the decision on the online store.
Research has shown that the potentials of electronic commerce in Montenegro are great, but that there is still not a critical mass of users.
On the other hand, in companies that are on the supply side, there are ambitions, but there is not enough knowledge to further examine the potential of electronic commerce and properly direct the offer to the different categories of clients.

Key words: e-commerce, determinants, perception, consumers, factors

I INTRODUCTION

Two last decades have been marked by a significant penetration and dispersion of electronic commerce. If we exclude 2008, when the beginning of the economic crisis, which was met without preparation, reflected on the negative growth rate of electronic commerce, throughout all other years the growth rate of its application on the global level has been evident. Whereas at the beginning of this century the leading positions considering its acceptance belonged to highly developed countries of the world, especially to USA, nowadays the high growth rates are recorded in the developing countries as well [1]. The reasons for such state could be found in a large number of factors:

- Accelerated technological penetrations;
- Growth of trade volume within both national and international frameworks;
- Readiness of network configuration to support commercial transactions on all levels of different strength and complexity, which is followed by change in supply and demand in electronic environment;
- General economic terms favoring electronic commerce;
- Favorable world monetary and fiscal ambience for transferring business processes into electronic platforms;
- Change of legal regulation and its adjustment to technological changes;
- Activities of governmental bodies and certain international organizations and institutions to provide support to affirmation of electronic business and within it to electronic commerce as well;
- Development of strategies aimed at making effects of globalizations move into positive direction;
- Increased number of online payment methods designed to support business on open networks, primarily on the Internet.

From experience and all conducted researches so far, it is clear that delay in accepting electronic commerce may lead to isolation of national economies. On the other side, electronic commerce offers numerous advantages such as improvement of productivity, decrease in prices, cost savings, change of size and structure of the market, elimination of mediators and direct access to end-users [2].
For success in the electronic world, the first and most important step is planning to identify barriers to adoption of e-commerce [3].
The growth of electronic commerce has led to increased number of academic works that deal with this very issue and put focus on two interest groups – customers and companies. It has been noticed that other actors, such as suppliers, investors, regulation bodies etc, are disproportionally less treated in this field of research. On this degree of EC development their influence is not that considerable, but there is evidence that they will deserve more attention in the recent future [4].
Although it is believed that the electronic commerce is a global phenomenon, in the most of the studies the focus is on developed market economies [5]. We agree that the experience of developed countries is valuable, but we

think that developing countries should not be disregarded and that the complete picture can be gained only by parallel treatment of this issue on all levels of development. EC are seen in different lights in developed versus developing economies. The pattern of influence and use of EC differs according to a country's level of socio-economic development [6].

Montenegro, as a typical developing country, has an incentive environment and great economic potential for more massive usage of electronic commerce technologies, but the low degree of usage of those technologies and a great lack of awareness of potential benefits they offer.

The aim of this work is to identify factors that influence the low degree of acceptance of the electronic commerce in Montenegro from both perspectives – of consumers and companies.

This research has a triple purpose: 1) To create the picture of customers' perceptions about electronic commerce and recognize their expectations; 2) To identify factors determining degree of electronic commerce technology usage by companies; 3) To create data base pursuant to which the electronic transaction volume can be increased and strong base of satisfied internet users can be created.

II METHODOLOGY OF RESEARCH

We collected the data for this study during the one comprehensive research which had the goal to gain insight into the state of electronic business in Montenegro. For the purpose of this paper we used a part of that survey and analyze relevant data from submitted questionnaires.

In the survey we used two types of questionnaire, while one was forwarded to consumers (citizens of Montenegro), the other collected answers from companies, or precisely from employees and managers.

Perception and experience of different profiles of consumers about electronic commerce was important issue of this research. According to that, one of questionnaire made in way to collect as much as possible information about demographic characteristics of consumers (sex, age, employment, education, income level and geographic affiliation), their experience in using of the Internet, and electronic transactions, as well as their opinion about security in transactions and offers on the sites.

The second questionnaire consisted of two parts. The first section was intended for employees, and the second for managers of the company. The aim of this questionnaire was to gather information about organizational and technical factors which determine the degree of acceptance of electronic commerce technology from the perspective of supply.

During consumer survey, we used the method of random sampling, which proportionately included population of central, southern and northern regions. About 70% of the questionnaire is collected from the central region which can be considered as a representative sample. From 500 submitted questionnaires, we analyzed 350 properly completed.

From 100 questionnaires forwarded to the addresses of companies (randomly selected from the Central Registry of Companies), we received 87 completely filled. A total of 84 surveys were received from small and medium enterprises (SMEs). This fact justifies the efforts of the professional public to explore the acceptance of e-commerce technology particularly in the SME category.

The survey was anonymous in both cases and persons who participated in the survey were familiar with the purpose of research.

After collecting the data and their analyzing by a pre-defined plan, we approached to interpretation of research results.

For the purposes of this study, we used basic theoretical methods, appropriate to the social sciences: definition, classification, induction, deduction, analysis and synthesis.

III INTERPRETATION OF RESEARCH RESULTS

Although the Strategy for Information Society Development in Montenegro [9] that is currently in force sets the goals of digital environment development for operation of innovative companies, increase of the number of customers who are familiar with and use electronic commerce technology and growth of modern public sector participation that supports all of that, Montenegro is still at the bottom of the list of the European countries in the field of electronic commerce technology usage.

From aggregated data gained from the Montenegrin banks at the end of 2014, it could be seen that around 11 million euro was spent for online transactions in that year, which was about three millions more than in 2013, but the growth rate was still far from the one existing in neighboring countries.

From four examined categories of transactions (services of electronic banking, services of electronic tourism, electronic commerce and online auctions), online commerce occupies the penultimate position (Table 1).

The research has shown that 98% of citizens do the online purchase using foreign sites. As reasons for such decisions they state: poor offer on the sites of the Montenegrin companies, relatively high prices, bad functioning of sites and lack of confidence into companies that have no tradition of electronic commerce. Without doubt this leads to conclusion that the efforts made so far for implementation of electronic commerce in the Montenegrin companies are not sufficient and they give no significant results.

Participants in electronic transactions		Percentage
Use the services of electronic banking	Yes	15.14%
	No	84.86%
Use the services of electronic tourism	Yes	17.43%
	No	82.58%
Buy products over Internet	Yes	7.70%
	No	92.3%
Participate in electronic auctions	Yes	2.00%
	No	98.00%

TABLE 1. PERCENTAGE PARTICIPATION IN ELECTRONIC TRANSACTIONS

From the total number of those who do the online purchase, 81% of citizens belong to the group of highly educated, 14 % are students and 5 % accounts for other educational structures.

When it comes to online purchase, salaries exceeding 700€ are the breaking point. It has been shown that the amount of income is in direct correlation with participation in electronic transactions and by that, for now, the amount of income is in direct correlation with the level of education. Pursuant to the economic situation in Montenegro, 91% of those doing online purchase are employed, where difference in sector employment (public and private sector) has no statistical importance. If we analyze the number of online purchases done, it can be concluded that it is unevenly distributed from the aspect of age structure since 65% of online customers are people of middle age (35-50 years old). This is the category of customers who have more steady income. Besides 65% of online customers from that category, only 14% is from the category 18-25 years of age, 19% is from 25 to 35 years of age and less than 2% accounts for customers older than 50 years.

Differences between the sexes in Montenegro have no importance in examining the influence of sex on decision to carry out electronic transaction (from 58 YES answers, 27 gave women and 31 men).

Significant digital gap on relation north-central part-south of Montenegro is visible in gained data as well. From the total number of internet users doing online shopping 63% come from central region, 31% account for coastal region and only 6% for north region (which is recognized as the region with the smallest number of employees, lowest income and smallest number of educated inhabitants).

For ranging the importance of factors influencing satisfaction, trust and loyalty we used 9 criteria:

1. prices of service;
2. the quality of offered service;
3. the quality of information on services;
4. the quality of post selling service;
5. evaluation of previous experience;
6. brand effect;
7. design of user's interface;
8. perception of security;
9. perception of privacy of data.

In survey we used Likert-type scale as a criterion. The analysis has shown that when making decision on online purchase customers pay attention to (in the following sequence): trust, price, brand effect, the quality of offered service, the quality of information on service and the quality of post selling service. The trust is mostly influenced by belief about security and data privacy as well as the evaluation of previous experience. As it has been shown in previously published researches, in situation of greater space and time separation of the act of purchasing and the act of receiving goods the perceived risks grow higher. For that reason trust has the key role in developing this way of commerce and it is even more important than the price of the product to be bought [10]. Dehkordi and the associates have reached a similar conclusion [11]. In the research conducted in developed market economies they have shown that perceptions of customers regarding price and trust remain the same regardless of the level of development of the economy from which consumers come. Also, the information that 93% of customers from Montenegro give the highest grade (on scale from 1 to 5) to the importance of previous experience confirms the fact that the psychology of users on the web is the same regardless of demographic differences. [12].

Data gained for design of user's interface have shown the modest influence of this factor on the level of carried online purchases.

From the survey it can be seen that the examinees have been familiar with the importance of the Internet, they have modest knowledge about benefits of electronic commerce, insufficient knowledge about instruments of electronic payment and deeply rooted fear of negative consequences that can come out from such way of paying.

Generally, all reasons for inertness of citizens of Montenegro regarding participation in online commerce we can group into several units:

- There is a lack of trust into online transactions by consumers. In that regard, both companies and the Government must have understanding for consumers' perception regarding this issue. The problem must be approached at least from three aspects: a) electronic commerce should be promoted through development of training programs which can raise customer's awareness of problems regarding security and safety and which can inform consumers about the means that can be used for avoiding those problems; b) tools which customers can use to assess the reliability of service which is used during online commerce should be presented; c) national legal framework protecting rights to privacy and anonymity of online customers and punishing those who violate those rights should be secured.
- There are issues that arise from social differences. In Montenegro there is a huge difference in economic level of development on relation north – south part of Montenegro.

Government's programs should be used for removing digital gap existing on this relation, as well as on relation village – town so that all citizens could have the equal access to ICT and electronic commerce technologies. As long as telecommunication environment isn't promoted along with investments into Internet infrastructure by coordinated action in all regions, electronic business will remain at low level and in domain of conventional order.

- There is insufficient information regarding potentials of ICT in commerce. Promotional and educational activities should be used to help consumers understand the Internet environment, to realize new aspects of virtual economy (benefits and possible barriers) and to meet the so called Gift Economy on the Internet.
- There is an incomplete legal framework dealing with consumers and protecting their interests. Besides amendments to currently valid Law on Electronic Commerce and the Law on Electronic Signature, based on the model used in neighboring countries, the Law on Foreign Current Operations should be changed and a new law on payment transactions should be adopted. The Law on Foreign Current Operations could secure Montenegrin companies mechanism for sale on foreign markets. To the State it would secure foreign exchange revenue. To consumers it would bring trust into online commerce using national sites. Since the laws defining electronic business in Montenegro adopted so far do not define the concept of electronic money, this could be done by the law on payment transactions, by which citizens, besides a wide branch of instruments of electronic payment, would also get legal protection.
- There is a lack of purchasing power. This issue should be treated within overall economic policy aimed at creating conditions for reducing unemployment, increasing incomes and raising standard.

It is necessary to map the basic critical factors for the growth of electronic commerce from the consumers' aspect and suggest measures for resolving these problems so that companies by their activities could increase the base of satisfied online customers and the Government by strategic plans could create a favorable environment for acting towards other institutions and establish better cooperation with companies regarding mentioned issues.

After analyzing data gained from employees and managers in the Montenegrin companies on the side of offer we have got results which are not optimistic at all.

Companies are aware that on a long-term basis the usage of electronic commerce technologies is related to profitability and long-term growth of company, but the participation of electronic transactions in overall transactions is still very low. Reasons for such state managers see in discouraging economic environment and employees in feeble effort to attract and keep online customers.

More than 75% of persons surveyed on the side of offer have their own website, mostly informative and static one. Based on answers got from employees it could be concluded that in all companies maximum efforts have been made regarding designing and functioning of website. As regards quality, there is no backwardness in comparison to other companies that have online offer.

However, when we did a quantitative evaluation applying Web Assessment Index (WAI), we discovered that there were shortcomings and still a lot had be done regarding functionality and personalization of the sites.

WAI model was developed in 2001 and so far it has been used in a great number of researches aimed at evaluating the quality of websites of companies coming from different branches of industries [13,14].

The quality of websites of the surveyed Montenegrin companies we measured using 4 criteria: access, speed of access, navigation and site content.

As sub criteria for access we used the presence on search engines and link popularity. For measuring the presence on search engines we decided to use Google and for measuring popularity we used data got from AlexaWebsite.

For measuring the speed of access we carried out independent testing of the opening speed of websites of companies by approaching from one computer at the same time. We repeated testing three times in 15 days and got results which were slightly different.

For testing navigation we used a subjective standpoint of the authors, evaluating the presence of the site map, simplicity of navigation, linkage of the basic page with other site pages, easy orientation of users on the site etc.

For evaluating the content, following the pattern of previous works, we split content into three subcategories:

1. Information content
2. Content for the purpose of communication with consumers
3. Content related to transactions themselves

From the total of 100 points for evaluation of the site quality, 55 were intended for content (by subjective assessment as the most important criterion for site quality evaluation) and 15 for each – access, speed and navigation. From 15 points for access, 5 were given for presence on search engines and 10 for the number of visits on the site.

When the testing was done, we got the following results:

- Out of total of 100 points, only two companies have between 80 and 90 points, 4 companies are in the range of 70-80, 7 companies are in the range of 60-70, 21 companies are in the range of 50-60 and even 32 companies have less than 50 points.
- When analyzing the presence on search engines, none company appears on more than 60 search results previews, 14 companies have 50-60, 27 companies have 30-50 previews and the greatest number is of those having less than 30 search results previews;

- By measuring the speed of access we established the quickest access of 3.4 seconds and slowest of 41.6 seconds. The greatest number of companies has the speed of access between 6 and 8 seconds. We have also noticed that the speed of access is in direct correlation with quantity of content on the site.
- We have seen that the companies with a greater number of achieved points regarding access have a better navigation. Out of 66 companies 27 have and 39 do not have the site map.
- Regarding the content assessment, all companies got maximum points for information content and content related to contact and communication. Very small number of points has been given for content related to transactions and information on carrying out online transactions, which leads us to conclusion that the level of online transactions is very low. Regarding the assessment of content quality, 26 companies got between 40 and 50 points, 11 between 30 and 40 points and 29 between 20 and 30 points.
- 27% of analyzed websites support bilingualism (the English and Montenegrin language).

By comparison of these data with data got by other researches (by application of the same technique) it can be concluded that managers and marketing experts of the Montenegrin companies should make more efforts in order to improve the functionality of their sites. At the first place we think on more serious approach to SEO technique so that visibility could be increased, then on improvement of the quality of content in part of online transactions that can be offered to consumers, on improvement of quality of navigation, on more frequent refreshing of the available information and on higher level of encouraging customers on two-way direction.

Among the Montenegrin companies that have been analyzed there is a strong competition, for the purpose of gaining comparative advantage, employees in marketing sectors are currently focused on finding new products and services. On the other side, development of web presence strategy, on cite, forums, social communities etc, in 53 out of 87 companies is partially left to IT sector, where logically exists a lack of interdisciplinary knowledge.

Our impression is that all the efforts of managers and employees in the area of usage of electronic commerce technologies are aimed at stimulating customers to come to a physical location, point of sale, which is still considered as the most secure place for transactions.

As it is the case with a great number of companies from other branches of industry, in Montenegro, instead of adopting an insufficiently known ground, the management its development strategies mostly directs towards reducing costs of traditional business and traditional communication.

Answers of all persons surveyed on the side of offer show that there are expectations regarding support from the State.

Analyzing the experience from the area of electronic commerce in developed market economies, the support would be significant in the following fields [15]:

- The level of information on electronic services should be increased by educational activities, as well as the awareness of their potentials;
- Digital gap on the relation north – other regions, village – town, younger – older population should be reduced;
- Implementation of measures aimed at further cutting the prices of access to new technologies should be continued.

Also, in companies there is no ambition to separate the effects of traditional from effects of electronic types of communication and carrying out transactions, which leads to conclusion that there is no serious consideration of the role of ICT.

In all analyzed companies some mutual problems have been noticed:

- It frequently happens that important decisions on transition from traditional to electronic business are made by managers without cooperation with other experts for organizational and technological changes;
- Very often, decisions are made based on the feelings, intuition, bad experience and by neglecting key facts on business;
- Decisions are made without researching the market of customers and their needs, with no knowledge of competition and its activities;
- There is no knowledge of electronic commerce technology usage phases, and the importance of transition from one phase into another. Therefore, often there is unnecessary long or permanent remaining in the existing phase;
- After decision on transferring into electronic market is made, it often happens that affairs of transition are entrusted to persons or institutions out of company, which as a consequence can have a lack of knowledge of business processes, business climate, culture, mission and vision of a specific company and the problems may arise;
- None of the filled questionnaires gave positive answer to question on possessing methodology or any internal activity by which metrics of the importance of application of ICT in business is done.

It is clear that successful usage of ICT implies the existence of suitable stimulating politics that will be based on strong scientific basis, appropriate human and financial resources and adequate technological and business infrastructure, which is not the case in Montenegro.

Positive developments in the area of electronic commerce technologies usage can happen only if:

- Traditional business procedures and programs are continuously adjusted to new instruments of digital environment;

- Competence of all staff and management is adjusted to potentials of the Internet environment;
- Criteria for successful implementation of the Internet activities are defined;
- Requests of the clients are permanently observed and, in accordance with that, intuitive analyses are replaced by analyses based on micro segmentation;
- Benefits of existing "a large number of data" are used, and which can be gained by application of up to date instruments of marketing;
- The web offer is promoted amongst the clients;
- The experience of successful companies from developed market economies are used and bottlenecks in business practice of the Montenegrin banks are removed;
- The methodology for measuring the success of the usage of the Internet services is elaborated.

IV CONCLUSION

At this moment in Montenegro there is a need for better understanding of electronic commerce concept, its potentials and factors inhibiting its faster development. Also, there is no holistic approach of the analysis of the influence of the identified factors on adoption of ICT on national level, level of industry and level of organization.
There are many factors influencing slow adoption of ICT and consequently slow adoption of electronic commerce technology: a lack of technologies, insufficient usage of existing technologies, IT illiteracy, incomplete legal infrastructure, problems arising from economic potential of the country that belongs to a group of small and insufficiently developed economies etc.
Results of the analyzed data got from the companies on the side of offer show that the affirmation of electronic commerce can be achieved only by coordinated action aimed at solving the fundamental problems that can be divided into following groups:

- financial problems;
- organizational problems;
- technical problems;
- problems whose solving requires the assistance of the State;

On the other side, results of the analyzed data got from the individuals on the side of demand show that the growth of electronic commerce volume can be secured only when the following issues are solved:

- when a perceived digital gap is reduced;
- when the legal framework protecting the rights of customers in electronic environment is completed;
- when the purchasing power of customers is increased;
- when IT and telecommunications environment is promoted and programs of supports to the Internet users securing the system of information and education of participants in online transactions are designed

It would be illusory to expect that all perceived barriers would be removed in short time, but we think that identification of the problems can accelerate their resolving. If we consider the speed of solving these problems during last two decades, it could be concluded that the process of their removal or mitigation in Montenegro might last until the beginning of the second half of this century.

REFERENCES

[1] http://www.emarketer.com/Article/Global-B2C Ecommerce-Sales-Hit-15-Trillion-is-Year-Driven-by-GrowthEmerging-Markets/1010575 (datum pristupa sajtu:11.06.2014.)
[2] Salmani, B., 2005. Proposed criteria for measuring e-commerce in Iran. In the proceedings of third National Conference on E-Commerce, Department of Planning and Economic Affairs. Iran
[3] Kamalabadi, I. Nakhai, et al. "Identifying and prioritization of challenges and barriers of e-commerce implementation in Iran." *World Applied Sciences Journal* 5.5 (2008): 590-597. Strana 464
[4] Chua, Cecil Eng Huang, et al. "The evolution of e-commerce research: A stakeholder perspective." *Journal of Electronic Commerce Research* 6.4 (2005): 262-279.
[5] Molla, A. and Licker, P.S., "Perceived E-Readiness Factors in E-commerce Adoption: An Empirical Investigation in a Developing Country". International Journal of Electronic Commerce, 10(1), 2005, pp. 83-110.
[6] Van Slyke, C., France, B. and Varadharajen, S. (2005), "A comparison of American and Indianconsumers perceptions of electronic commerce", Information Resources Management Journal, Vol. 18 No. 2, pp. 24-41.
[7] Nasco, S.A., Toledo, E.G., Mykytyn Jr., and Peter P.,
"Predicting electronic commerce adoption in Chilean SMEs". Journal of Business Research Strategic Management in Latin America, 61(6), 2008, pp. 697-705.
[8] Sparling, L., Toleman, M., and Cater-Steel A. SME
Adoption of e-Commerce in the Central Okanagan Region of Canada. *In Proceedings of the 18th Australasian Conference on Information Systems*, Toowoomba, Australia, 2007. pp. 1046-1059.
[9] www.mid.gov.me/.../FileDownload.aspx?rid=84067 (datum pristupa sajtu:11.01.2015.)
[10] Kim, H.W., Xu, Y. and Gupta, S (2012). Which is more important in Internet shopping, perceived price
or trust? *Electronic Commerce Research and Applications*, 11:241–252,
http://dx.doi.org/10.1016/j.elerap.2011.06.003
[11] Dehkordi, L. F., ShaHnazari, A. and Noroozi, A. (2011). A Study of the Factors that Influence the Acceptance of e-Commerce in Developing Countries: A Comparative Survey between Iran and United Arab Emirates, *Interdisciplinary Journal of Research Business*, 1(6):44-49
[12] Abbasi, Parvin, Bahram Sadeghi Bigham, and Saeed Sarencheh. "Good's history and trust in electronic commerce." *Procedia Computer Science* 3 (2011): 827-832.
[13] Miranda, Francisco Javier, Rosa Cortés, and Cristina Barriuso. "Quantitative evaluation of e-banking web sites: an empirical study of Spanish banks." *The Electronic Journal Information Systems Evaluation* 9.2 (2006): 73-82.
[14] Ranković, Marko, and Vojkan Vasković. "Quantitative website analysis: Case study of the Serbian banks." *Industrija* 39.2 (2011): 29-49
[15] Biljana Rondović. 2012. "The degree of the application of e-business in small And medium-sized enterprises". *Facta universitatis Series: Economics and Organization* Vol. 9, No1: 139 – 148

Laws of Network's Values and Values of Its Users

Emina Pirić*, Merima Bajrić and Himzo Bajrić*****
***BH Telecom dd Sarajevo, RD Tuzla, BiH**
****BH Telecom dd Sarajevo, ID EFP Sarajevo, BiH**
*****BH Telecom dd Sarajevo, ID TIRS Sarajevo, BiH**
emina.piric@bhtelecom.ba
merima.bajric@bhtelecom.ba
himzo.bajric@bhtelecom.ba

***Summary:** Network industries are very important in today's economy. The reason for this lies in the fact that in each network operates certain external or extra power so-called network effect. Bearing in mind that the laws are set for the value of the network, but not for the value of the users, this paper analyzes the possible functional value of the network and network users to another network.*

***Keywords:** network industries, fixed telephone network, the network effect, the value of the user.*

I. INTRODUCTION

Network industries play a vital role in the life and economy of modern society. Modern economy would be greatly impoverished without transportation, communications, information and network of roads and railways. This paper will analyze the specific features related to the networks and users. Yet the attention will be focused on the telecommunications network.

Specifically, each network has its own value (word "value" here means a functional value). Discovering the links between this value and the parameters of the network is for many years subject of scientific and expert analysis. It was found that the increase in the value of the network is a function of the number of its members or nodes. Furthermore it is logical that the increase in the value of the network is different for different networks. It is expected that one law is applied for a broadcast network, such as TV broadcasting, while for networks in which users communicate with each other it is faster than linear when number of nodes or members increases. But what is the speed of increase or what is the actual relation between the number of members and the value of the network has maybe finally been revealed. The importance of this relation is basis for making business models in development of network.

Let us remember the famous sentence: "Companies that leverage the power of networks will dominate the Internet in the future" - pronounced by Sean Parker, President of the Facebook. This brings us to the question of whether the network has something special or so-called "*power of the network*", what other business platforms don't have, and it brings us to to the next question: what is the real *power of the network* - or so-called *network effect*?

II. TYPES OF NETWORK EFFECTS

The best is to start from its definition: the network effect (also known as network externality) exists when the functional value of the product for the user increases and when the number of users of a given product grows. A classic example for this is the telephone. The more people own a telephone the functional value of telephone network for each of its owners is more valid.

When you buy a DVD player and a few DVDs, it encourages the industry to more DVDs marketed. In fact, the concept can be easily extended: perhaps the most dominant network effect exists in the language. If you are writing a website in a language of a small, fictional nation Kynuk, only few people (its members) can read it. If your targeted audience is in this community, that's fine. But if you want to increase attention, you'll probably write in English: scientific journals, for example, are now predominantly in English, especially because the Internet network has increased the network effect.

Network effects are fundamental to business strategies, intellectual property and technical applications in a wide range of industries [1], [2], [3], [4]. The main objective of any company is to immediately establish action network effects in its operations in the appearance on the market. The main issue, which is considered here, is how really is powerful network effect, namely how network effect influences quantitatively on functional network value for the user, and functional value of the user for the network owner maybe finally found [6], [7], [8], [9], [10], [11]?

Direct network effect: the simplest network effects are direct: an increase in the use of the product will lead to a direct increase in its functional value. The original example of telephone service is a good illustration of the product that shows the direct network effects.

Indirect network effects: network effects may be indirect, where an increase in the use of the product causes the production of more valuable complementary assets, resulting in an increase in the value of the original product. For example, although there are some direct network effects associated with Windows OS (arising from files' compatibility), probably more important network effects arising indirectly from an increase in the quality and availability of complementary software applications.

Two-sided network effects: network effects can be two-sided; increase in the use by one set of users, increases the value of complementary products for another different set of users, and vice versa.

Network effects occur in the platform mediated network - PMN, which includes networks of customers - often known as users - who want to communicate with each other, together with one or more intermediaries providing a platform, which includes infrastructure and rules to facilitate the mutual interactions of users. PMNs form the heart of computers, telecommunications, media and Internet sectors.

However, PMNs are not limited to the information and communication industry. They can be found in financial services (e.g., stock exchanges, credit cards, ATMs), health care, energy (e.g., connection to the electrical network), transportation (e.g. air transport, container transportation, gas stations) and retail (e.g., shopping centers, barcodes). There is a diverse range of appropriate business where network transactions mediate, including transportation, auctions and travel agencies.

Rank by market value, sixty out of the 100 world largest companies, which earn more than 50% of its revenue from online network - mediated platforms, include Apple, American Express, Cisco, Citigroup, Time Warner, UPS and Vodafone. In short, network-mediated platforms (PMN) are the essential characteristics of a modern economy.

Many beginners in the market entry aim either to build these new platforms (e.g. Facebook) or to profit by offering updated leverage of existing platforms (e.g. Zynga, which relies on Facebook as host of its game). Business models based on such a platform can be extremely complex, and the errors in their designs are often common.

III. LAWS OF FUNCTIONAL NETWORKS' VALUE

Each network has its own value. Discovering the link between that value and the network parameters has been, for many years, the subject of analysis, mainly in the American scientific community. It was found that the increase in the network value is a function of the number of its members or hosts. Furthermore, it is logical that the increase in the network value is different for different networks. It is expected that one law applies for a broadcast network such as a television, while for the networks in which users communicate with each other it is faster than linear when the number of participants increases.

But what is the speed of increase or what is the actual relation between the number of members and the network value is maybe finally established. The importance of this relation is the basis for making business models in the development and operation of the network. The fact is that the telecommunications network has never been isolated (at least not in the last 20 years), but has been connected with other networks through interconnections. It is expected that such a procedure increases the value of both networks but also is realized profit for the users of these networks [9], [12].

The network value is explained in Metcalfe's Law and in the Law presented by Bob Briscoe, Andrew Odlyzko, and Benjamin Tilly in their paper of 2006 [9], further referred to as Odlyzko's Law. Hereinafter we will try to provide answers to some questions about the value of the networks for users and users' value for the network owners.

IV. FRAMEWORKS FOR NETWORK EFFECTS AND THEIR VALUES

During the last 25 years, in the United States, an extensive academic attention has been dedicated to calculating the gains and benefits of connectivity in networks, which resulted in simple "laws" designed to capture the distributive effects of membership in the network, or "network effects" (Metcalfe's Law, Reed's Law, etc.) (Table I) [6], [7], [8], [9], [10], [11]. Currently proposed laws on network values, their inventors and chronology of proposals are listed in Table I.

All proposed "laws" show monotonous increase of network value, with the growth of the volume or the number of network members from linear (the slowest) to factorial (the fastest), but the Laws of Sarnoff, Metcalfe and Odlyzko are best-known ones. Like all such "laws", they are based on the application of simplification and assumptions. Key aspects of leading network models are identified and compared in Table I. [8].

Metcalfe's Law has been known for more than a decade, when it is was named as such by George Gilder in early 90's. According to Metcalfe's remembrance his intention was: "The original purpose of my law was to establish the existence of the price value of the critical coupling point - under which networks mutually do not pay to each other."

Metcalfe had an ideal opportunity for observation and analysis of the growth of networks and their profitability. In 1970, the first in his Harvard Ph.D. thesis, and then at the legendary Xerox Palo Alto Research Center, Metcalfe developed Ethernet protocol, which has managed to become dominant in the telecommunications networks.

The basis of Metcalfe's Law is the observation that in the communications network with n members, each member can make a connection with (*n-1*) other members. If all these connections are equally valuable - and there is a big "if", the total value of a network is proportional to $n^*(n-1)$, which is approximately n^2, for large *n*. So, for example, if the network has 5 members (Fig. 1), there are 20 different possible

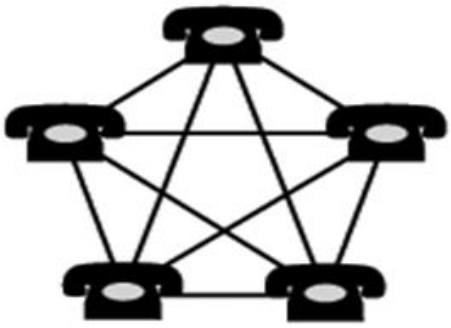


Fig.1. Illustration of a network with a network effect (the number of connections increase as the number of the network members increases).

connections that a member can make with the other one. If the network size increases four times, to 20, the number of connections increases not only four times, to 80, but increases to 380 - or in other words it is approximately eighteen times increase. This approach has helped the growth of dot-com companies, but also their collapse at the end of the last millennium in the United States, because, after all, this type of law was not appropriate, and used business models were mainly based on it.

Reed, on the other hand, introduced a law that recognized the value of the group within the network, not just couples, so certain group of four people can be connected not only to form pairs, but also in groups of three or even superset of all four persons. Adding 4 groups of 3, plus the entire group of four, all together as $2^n - n - 1$ and it is approximately proportional to a 2^n. This law is too optimistic and has no real stronghold in the telecom network, which mainly has communication between two members, but it partly has a stronghold in social networks, such as Facebook.

Briscoe, Odlyzko and Tilly [9] in their paper pointed out that presented network laws are probably too optimistic in their values, and it can be intuitively recognized that the growth rate of the network value must be reduced as future members join - because probably the most valuable connections have been formed first. This is parallel to the concept of "diminishing returns", the central position of neo-classical economics. Such a reduction in incremental values is modeled as total $n*log\ (n)$, (where the logarithm has base 2) where future membership reduces the growth rate values (but still ensures positive growth). This frame also coincided with visible power law in real-world of networks, which was highlighted by Barabasi and colleagues [5].

There is a big "but", because each member of the network establishes connection only with a relatively small number of members who are close to him and with whom he has a certain social relations, although there is the possibility of establishing a connection with each of the members of the network. Therefore it can be concluded that the value of the network is not proportional to the ability to establish a connection, but to established connections, or to be more precise, to the traffic realized when establishing connections. In fact, we all have definite directory in our phone, consisting of several hundred members, where some of them we call very rarely while some other close family members very often, and we belong to a network of over one million members.

Therefore, the paper [9] has proposed a new form of law on numerical value of network effects in telephony, in the form that the value of a network is proportional to $n*log_2(n)$ and it is much softer in the increase and it provides a new view on the issue of interconnections as analyzed in [12]. Fig.2. graphically illustrates increase in the network value in accordance with the laws listed in Table I. The paper [12] has shown that this law reflects well changes in the value of networks during their interconnection.

V. THE VALUE OF THE NETWORK FOR THE USER AND THE VALUE OF THE USER FOR THE NETWORK OWNER

There is a question: if the product or network is more valuable for the user when it is also used by others, should he pay accordingly? When it comes to material product, there is lot of such situations where one pays name/brand. When it comes to communications, then there is dealing with services. In the telecommunications network, there are different services and different payment models for using the services. Payment models can be compared with the parameter of functional network value and it can be seen whether models include the application of the network effect. As seen, in principle, there are two realistic models of the network value law: linear (Sarnoff) and logarithmic (Odlyzko) law.

Telecom operator can, in different ways and unconsciously, use the power of the network by application of some of network value laws in its business models of service provision. The most common model is the first service model, developed as broadcasting with TV and radio, and it applies linear law of network value. Therefore, the value of each member of the network is mutually equal and constant and it does not depend on the number of network members.

Total value (proportion. with)	***Hron-ology***	***Inventor***	***Model***	***Example***
n	1	Sarnoff	Broadcast	TV, Radio
$n*log(n)$	5	Odlyzko	Practical Metcalfe	Telephony
n^2	2	Metcalfe	Network	Telephony
n^c	6	Nivi	Practical Reed	Google, Facebook groups
2^n	3	Reed	Social networks	Google, Facebook groups
$n!$	4	Haque	?	?

TABLE I. THE PROPOSED NETWORK VALUE LAWS [8]

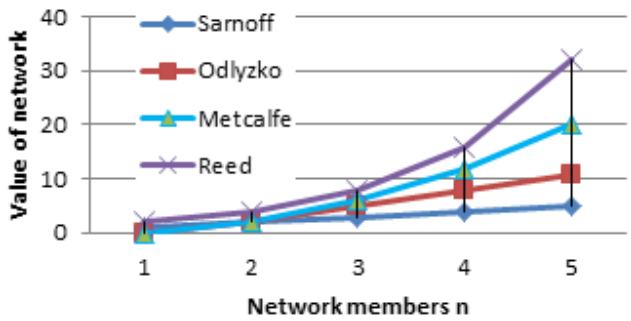


Fig.2. The increase in the value of the network

Historically, following service model that was developed and offered to the market is the voice service in fixed telephony. Such service completely follows a logarithmic (Odlyzko) model of the network value. In this case, the value of the network and each user increases as the number of connected users to the network increases.

In all business models where the principle of a flat rate charging (ADSL or other Internet access) is introduced, operates linear law of the network value. In the ADSL example, users have certain mean value of K_p and proportional to that value they pay to the owner of the network (you can put that monetary proportionality constant is 1 for easier calculation). In the network dominated by linear law of value of network connectivity, "content is king." In fact, in such networks, there are fewer sources of content that each user chooses from the available sources. There is a competition for customers based on the source of their content, i.e. based on the value of their content (published stories, published pictures, standardized consumer goods).

Where Odlyzko's Law dominates, the number of established connections, respectively transactions, becomes a central point. In cases where it is traded with transactions (either via e-mail, text mail or voice mail, money, securities, contracted services, or else) the transactions are key. It can be reasonably assumed that the average number of connections established by the user - member of the network is proportional to the logarithm of the total number of network members, proportional to a constant K_1 (Odlyzko's Law). It is evident that the value of member to the network increases as the number of members in the network increases. Direct examples include the provision of telephony and messaging services. From this it can be seen that the value of a network member increases by connecting new member to network, and it is clear that there is a network effect precisely operating in accordance with its definition.

However, in many cases with no lower limit of network user's payment to the owner, which is the case of telephony and all other services with Odlyzko's Law, network owners introduce a minimal fee for membership in the network, so - called subscription - K_{P1}. In such cases the value of the network member and the appropriate business model are determined by cumulative action of linear and logarithmic laws of the network value, with appropriate constant. From a business perspective, this is very favorable situation for the network owner. Table II. summarizes some typical cases of value related to the network and users and appropriate example of the telecommunications services.

Table II. shows:

- Column 1: the value of the network for the member;
- Column 2: the total value of the network;
- Column 3: the value of the member of the network;
- Column 4: the value of the network member in the business model of service; and
- Column 5: an appropriate example of the telecommunications services.

In principle, business models involving logarithmic law are economically better than linear, and the best business models are these where both laws operate simultaneously. It is evident that lately, operators in mobile communications use business models without subscriptions, which significantly reduces the value of a member itself, and then the overall economic effects of the network as well.

TABLE II. THE VALUES RELATED TO THE NETWORKS AND USERS

Value of the networks for members	***The total value of the network***	***The value of the networks member***	***The value of the networks member in the business model of service***	***Appropriate example***
Kp	$n*Kp$	K_p	$n*K_p$	ADSL Internet access, IPTV
$K_1*log(n)$	$n*K_1*log(n)$	$K_{p1}+K_1*log(n)$	$n*[K_{p1}+K_1*log(n)]$	Fixed telephony, VoIP, Postpaid mobile telephony
$K_2*log(n)$	$n*K_2*log(n)$	$K_2*log(n)$	$n*K_2*log(n)$	Prepaid and postpaid mobile telephony
$K_3*log(n)$	$n*K_3*log(n)$	$K_3*log(n)$	$n*K_3*log(n)$	SMS, MMS
$K_5*log(n)$	$n*K_5*log(n)$	$K_{p4}+K_5*log(n)$	$n*[K_{p4}+K_5*log(n)]$	Mobile Internet

With web access to the content on the Internet, network effects are actually closed-loops generating increasing feedback, the higher the number of users is. In the case of web, feedback results in more content and more possibilities for their users. Current Web 2.0 technology exactly allows such feedbacks, by allowing users to be editors, namely to place their own content on the Web. Such ideas has also started to follow business activities which aim to bring together its users in the business communities that share certain interest and create bigger value when having more members.

Regardless of the law used to describe the network effect, conclusion remains the same - all Web applications or services used daily by lot of people, allowing only passive usage for users, without any other value except just service or application (example of website that can only be browsed), lose a large part of the value, they could have if correctly use the network effects, for example, by allowing users' communications, or allowing users to change the very application or service. An example of such a service is Facebook, Digg or YouTube that not only attract an enormous amount of visitors every day, but also use a visitors as (the only) source of the content.

The question is how this phenomenon will affect further development of the telecommunications, the Internet and the Web in general, as well as business companies from this sector, but the fact remains that these who ignore it will have negative effects on their business.

VI. CONCLUSION

Modern economy would be very impoverished without transportation, communications, information and network of roads and railways. All these systems are based on a network-mediated platform. PMN can be found in financial services, health care, energy, transportation and retail.

Specifically, each network has its own functional value. This paper analyzed specific characteristics related to the telecommunications networks and users. In the analysis was used a new mathematical formulation of the network effect.

The development of the telecommunications operators is supported by strong externalities - network effect, which is more or less successfully implemented in their business models of individual services. In the telecommunications network, there are different services and different payment models for using the services. As seen, in principle, there are two realistic models of the network value law: linear (Sarnoff) and a new mathematical formulation of the network effect in the form of logarithmic or Odlyzko`s Law.

The business value of a network member is defined by various network value laws, only linear, only logarithmic, and in some business models both linear and logarithmic (best cases). In the case when a logarithmic network value law is applied in a network, presence of network effects is obvious.

It is expected that the content will get even more importance in the future. That is why more attention must be paid to the content and its creating, so own users should be involved in content creation, which is fully supported by existing Web 2.0 technology, and what in return, introduces activity of the network effect.

All Web applications or services used daily by lot of people, allowing only passive usage, without any other value than just service or application, lose a large part of the possible value they could have if correctly use the network effects, for example, by allowing communication among users or allowing users to change the application or service.

REFERENCES

[1] Katz, M. L./Shapiro, C.: „Network externalities, competition, and compatibility", The American Economic Review, Vol. 75, No. 3 (June 1985), pp. 424-440.

[2] Katz, M. L./Shapiro, C.: „Network externalities, competition, and compatibility", The American Economic Review, Vol. 75, No. 3 (June 1985), pp. 424-440.

[3] F. Inceoglu, M Park: „Diffusion of a new product under network effects: The U.S. DVD market", Applaid Economics, 2010.

[4] R, Tongia, E. J. Wilson III: „The flip side of Metcalfe's law: Multiple and growing costs of network exclusion", International Journal of Communication 5 (2011), pp. 665–681.

[5] Albert-László Barabási: „Linked: The new science of networks", 2002. ISBN 0-452-28439-2

[6] B. Metcalfe: „There oughta be a law", New York Times, July 15 1996.

[7] B. Metcalfe: „Metcalfe's Law: A network becomes more valuable as it reaches more users", Infoworld, Oct. 2, 1995.

[8] http://www.infoworld.com/cgi-bin/displayNew.pl?/metcalfe/bm050696.htmi.

[9] Briscoe, B., Odlyzko, B., & Tilly, B.: „Metcalfe's law is wrong", IEEE Spectrum, (2006) 43(7), pp. 34–39.

[10] B. Nivi: "Between Metcalfe and Reed Laws" (2005). http://www.nivi.com

[11] D. P. Reed: "That sneaky exponential beyond Metcalfe's law to the power of community building", http://www.reed.com/contactinfo.

[12] H. Bajrić, E. Pirić, M. Bajrić: "Network effects and interconnection", Magazine for telecommunications, Vol. 13, No. 40, BHTEL 2014, pp. 3-12.

Open issues of open data in Croatia

Dalibor Bužić, Jurica Đurić
Visoka škola za informacijske tehnologije, Klaićeva 7, Zagreb, Croatia
dalibor.buzic@vsite.hr, jurica.duric@vsite.hr

Optimism and euphoria about open government data as a silver bullet which should lead economy and society to significantly higher levels have deflated after few years in collision with a reality. Expectations from open data usage were great, but in exploitation it is more than questionable at which rate open data generate new economic values. Many issues appear on both sides – government as a supplier and on demand side which lags behind the supply. In this paper we analyze and comment open data published by some of the public authority bodies in Croatia and we present an overview of mobile application based on those data.

I. INTRODUCTION

Governments all over the world generate large amounts of data. Movement for free access to government data is based on the fact those data are collected and processed by taxpayers' money. The goal of free access to data is enabling citizens, companies, media and academic community to use those data for economic and social progress. Published data should be in genuine form (without aggregation), up-to-date, ready for easy machine processing, available to everyone and license-free.

Open data promoters claim that the useful value of open government data can be created by citizens, associations, companies, academic community, and government itself. Citizens can find their interest in information which helps while making decisions such as choosing city for living with regard to air pollution or climate elements (areas with high average humidity, for instance) or choosing schools for children (based on the number of children per one teacher or a success in the later stages of education) and so on. Non-government associations can improve status of vulnerable civil groups. Academic community can conduct new researches. Companies can create new services and products based on open data. At the end, whole state can have benefits – new employments in the private sector, tax incomes, raise of the economy, and one entirely different spectrum of benefits such as increasing transparency or involving citizens in activities of wider social interest.

Economic gain of open government data usage one European commission study estimates on about 40 billion euros per year in the EU [1]. But this business activity domain is relatively new, and there is a lack of knowledge about potential economic effects of usage and creating new products and services based on open data. Some of the greatest challenges for existing and new companies interested in creating new values from open data are which precisely new products and services should they make, and how to make it lucrative. Applications related to meteorological data (such as temperature, humidity, wind speed and direction, UV index) and to traffic data (such as railway timetables, delays, road works and traffic jams) are well accepted by a wide range of users. Statistic data processed and presented in an attractive way (ecpecially graphics) can be a good opportunity. Information can be related to pupils' success in classes and competitions, water and air quality, police crime maps or government management (expenses share by type, supplier list, miscellaneous comparisons of cities or municipalities).

II. OPEN ISSUES

Optimism and euphoria about open government data as a silver bullet which will raise the economy and society on a significantly higher level had been deflated a few years after colliding with a reality. Web site data.gov, central place for publishing open US government data, was established in May 2009. when USA (and many other countries worldwide) experienced big crisis. Expectations from open data were large, but in practice it is very questionable at which rate open data generate new economic values.

The gap between promises and barriers are formulated in five myths [2]:

- publishing data will automatically yield benefits,
- all information should be published without any restriction,
- it is enough to publish data,
- everyone can make use of open data,
- open data will result to open government.

A. Supply-side issues

Issues arise at a very first part of the value chain: government. Bad design, flawed execution and adverse consequences are three main problem areas [3]. Bad design manifests in the initial uncertainty of a concept (what do open data really mean?), unrealistic goal of absolute transparency and focus on a technology as a transparency indicator.

Flawed execution reflects in three elements: resistance in agencies, data decontextualization and expense side. Resistance appeared, among other reasons, because of unrewarding extra work in preparation of data for publishing, but also because trading data were the significant income source. The question to be answered can be, for instance, what consequences brings public access to Croatian Meteorological and hydrological service (local acronym: DHMZ) which makes earnings from selling collected data? Decreasing income can lead to cutting down investment in equipment and decreased experts' salaries. That could result in negative employee

selection and, finally, in poorer service. DHMZ's income can be compensated from government budgets, but in that case, all taxpayers suffer that load. The serious question is then: who has benefits from free data? Right now some of DHMZ's data purchasers are largest media, telecom and insurance companies. On the other side, research about opening hydrometeorologic data in Netherland revealed that at the time when data has been sold, there were ten purchasers which brought income of four million euros. After data opening, number of users has increased, and government's tax revenue was about 30 million euros, which is 750% more [4]. In any case, cost-benefit analysis should be done before opening data which generate income to government agencies.

Decontextualization is the process of converting useful complete records into useless datasets by separating data from context. Agencies in the USA, just to fulfill the 'norm', often repacked data already published somewhere else. Data were sometimes without description, and in some cases even could not be opened.

Expenses side, as the third aspect of flawed execution, is conscious or unintentional ignoring of opening data costs. Some of the costs are related to adjusting data to new standards, employee training, improving the quality of data, but also hand-written and verbal data converting into digital records. Those expenses simply were not planned in agencies' budgets.

Among adverse consequences [3] emphasizes the fact that the open data programme did not decrease the information divide between developed and developing countries. Causes are in fact that developed countries have good mechanisms for data collecting and experts who use them. Undeveloped countries haven't enough such experts, and data are often of low quality or incomplete. In such countries citizens rarely participate in efforts about using of open data. It is a consequence of lack of knowledge and motivation. Of course, some kind of solidarity toward 'weaker' societies should not be obstacle in development.

Open data empower already powerful corporations and software companies who posses funds and skills for data usage. In other words, rich corporations (such as Lloyds of London, Google, Daily Mail, etc.) are subsidized by open data. It is almost absurd that sometimes government agencies buy its own data, but also there is a danger of corporative monopolization which could shape urban infrastructure guided by interests of profit [5].

Government bodies sometimes don't focus on processes outside their boundaries. That is why the process of opening data they don't see as interaction of government and public. While politicians are focused on stories about economic growth and citizen participation, officials are concerned about issues such as privacy, confidence, quality and data completeness.

Securing personal data privacy is a very sensitive issue. A fact is that most of open data (such as geo or budget data) are unrelated to people. But, personal data from which are deleted names, surnames and other identifiers don't guarantee absolute secrecy. Using only data about sex, date of birth and city in the USA, 53% of the population can be uniquely identified [6].

B. Demand-side issues

The issues are present also on demand side which lags after supply side. Part of causes lies in published data itself. Government agencies often under pression of data opening have taken passive-aggressive attitude by publishing minimal quantity of almost useless data. At first developers chose interesting data for new applications, but interest fell off quickly. Developers from New Zealand even refused to develop applications with useless data [7]. The second part of causes lies in lack of knowledge how to use those data which are pretty useless, until someone transforms it into something valuable. Lack of technical and statistical knowledge are serious barriers for exploiting potentials. Unknowing tools for effective open data processing is one more obstacle. That is why need for education in fields of statistics has appeared. It is necessary for avoiding erroneous analysis and results misinterpretations. Great Britain's Office for National Statistics, in cooperation with Durham University, trains interested journalists and politicians for using tools which enable data synthesizing from various sources, multivariate data interpretation, and so on.

Process of visualization, as a very important aspect of open data usage, demands special applications which should be easy to use, enable saving intermediate states of processing (due to repeatability), visual display of differences.

Even finding published datasets can be an issue. In New Zealand it is common that every agency publishes datasets on its own website. Since websites differ in design, finding data can be challenging [8].

III. TOWARDS OVERCOMING OF ISSUES

With a goal of relieved use and alleviation of some topical issues, dedicated infrastructure to work with open data are being developed. One of them is ENGAGE e-infrastructure. ENGAGE is based on literature, interviews, questionnaires and workshops, and is developed by Design Science Research methodology (DSRM). ENGAGE supports supply and usage of open data in five categories [9]:

- data provision – except data, it enables metadata acception, data cleansing, validation, conversion into other formats and metadata enhancement,
- searching and using data – uses SQL and SPARQL, has module for data demands and version management,
- linking data – manual and automated,
- user ratings – assessment system and quality control,
- user collaboration – user profiles, three user roles: supplier, consumer and coworker.

It is hard to understand how open data are used without previous understanding who are the users of open data and which are their motivations. Research conducted in Great Britain identified six motivation groups [10]: one has the government into focus, the second has focus on technology and innovation, third is focused on making a

profit, fourth deems that government digitalization will have reform of government as a consequence, the fifth represents users to whom open data will be a tool for achieving predefined goals, while the last group consists of entrepreneurs who cooperate with government.

Researchers in [11] identified six types of companies who are primarily oriented to:

- collecting and aggregating data,
- analytics as a service,
- creating data and analysis,
- knowledge discovering,
- data aggregation as a service,
- mult-source mash-up and analysis.

The main goal of a company is gaining a profit, which basically means making larger incomes relative to expenses. Web and mobile applications, as well as various services generate expenses in development phase, but also later on – during operation and changes. So, it is not only a matter of thinking about resources, what to offer, how to create it and who are the users. A company must think about profit side, i.e. how to charge a service. Income model exposed in [12] differentiates five main ways to generate an income:

- subscription model – recommended for frequently used service, but suitable for specialized or (semi) monopolistic information (such as geo maps in large scale) because internet users unwillingly pay online services,
- usage model - means service charge measured in bytes, time or session; suitable for ad hoc users who need access to services (possession is not important to them),
- royalty model – customer will pay a price (in the form of percentage or royalty) only after he creates and sells his product with added value
- free model – suitable especially for government services or products
- hybrid model – such as enticement model in which part of the content is free and the other part is to be charged, or community model (like Wikipedia or OpenStreetMap) where income is generated by subsidiary products/services or by donations.

Besides above-mentioned, income model that can be applied on mobile applications is one-time charge on download. Price admissibility combined with usefulness can generate significant income if application becomes popular. Mobile applications can find one exclusive buyer who will, by further free of charge distribution, create a new marketing channel. Web sites which offer free information generated by open data and have a large number of visitors can attract sponsors or advertisers and ensure income. Start-up companies or individuals can attract partners and find their place on the market. Therefore, the application itself doesn't create benefit, but enables easier finding jobs in the future.

IV. OPEN DATA IN CROATIA

Based on surveys, the Open Data Census has created Open Data Index for more than 100 countries worldwide. Table 1. shows ranking of Croatia and its neighbours according to mentioned index [13]. As we can see, of all Croatia's neighbours, only Bosnia and Herzegovina has poorer standing (no data exist for Monte Negro).

Table 1. OPEN DATA INDEX RANKINGS

Position	Country	Score
26.	Italy	545
28.	Slovenia	540
39.	Hungary	475
49.	Serbia	420
52.	Croatia	405
94.	Bosnia and Herzegovina	205
---	Monte Negro	no data

Information Commissioner (IC), as an independent government body for protection of the right of access to information, highlights following issues of open data in Croatia [14]:

- ignoring requests for information (2/3 of the complaints were about the silence of public authority bodies),
- tardiness of public authority bodies,
- information officers' lack of professionalism,
- overcharging of providing information (when actual material expenses incurred),
- insufficient proactive data publishing,
- low citizens' awareness about the right of access to information,
- uncoordinated regulations (about data privacy, trade secret etc.).

IC has a very important role in resolving complaints related to the right of access to information reuse. It also supervises the implementation of the Law on the right to access information. IC can submit an indictment proposal and can issue a misdemeanour order. At this moment, IC office is underfunded and understaffed. That fact negatively impacts on significant progress in realization of the right of access to information and information reuse.

At the time of finishing this paper (in mid-February) we are waiting for launching data.gov.hr portal (we can expect it in a few weeks, according to announcements). The portal is being developed by use of open source code of British data.gov.uk available at GitHub. It uses CKAN platform for data publishing. At the beginning, about 70 datasets will be published. Like in the USA or Great Britain, the intention is data.gov.hr would be the central portal for publishing open data, so all government bodies and agencies, every city and municipality will be able to put data in one place. This concept has advantages for

both publishers (no need for investment in infrastructure) and users (easier finding datasets). On Croatian open data portal there will be a request form for asking new (unpublished) datasets. The intention is to listen to datasets users' needs to avoid efforts about publishing datasets which will not be used.

Before presenting our findings it is important to emphasize that we analyzed only some publishers – so the list of datasets is not comprehensive. Municipalities and counties were not in our focus at all.

In the current situation, we found that some government bodies have a clear understanding of the need for open data. One of them is Ministry of administration which on its website https://uprava.gov.hr/ clearly states the terms of use for their 12 datasets published in CSV format. The Croatian public health institute publishes lot's of reports and publications in PDF format, and eight XLS datasets at http://www.hzjz.hr. The Croatian health insurance fund published more than 40 datasets in XLS format available at http://www.hzzo.hr/. Ministry of economy has recently developed web site for fuel prices tracking (http://www.min-go.hr) which helps citizens to find a cheaper gas station. It is expected that raw data will be available in acceptable formats for interested parties soon.

Geoportal is a service which is available for citizens recently. With it, we can see in every moment and examine everything related to cadastral and land registry business. Also via additional links, users can view documents such as proprietary and title deed, and view the history of documents.

Croatian bureau of statistics at http://www.dzs.hr/ provides national statistics about a large set of various data on a national level. They offer users to download and use their program for viewing and formatting datasets and downloading datasets directly from the website in the format they want.

Another site that offers open data to the public is Croatian waters. Croatian waters on their pages http://vodostaji.voda.hr/ offer information that is updated every hour. All data are complemented with graphical icons so it is very easily to notice water levels that are higher than normal limits.

Easy access to their data, gives Meteorological and hydrological institute of Croatia (DHMZ). On their site http://meteo.hr there are current weather data for a lot of cities in Croatia and Europe. Beside all the weather data that can be retrieved, data on water levels of rivers as well as information about the index danger of forest fires is also available. What DHMZ explicitly offers is dataset in XML format containing data about temperature, humidity, pressure, wind and condition for many Croatian cities.

The Croatian environment agency offers lots of data contained in Environment Information System at http://iszo.azo.hr/.

Croatian Automobile Club at http://www.hak.hr/ provides access to information about road conditions as well as other various services related to road trip with possibility of refreshing rate so it can be said that the data is updated in real time. There is also a map that offers a view of all information that are essential together with an estimate of travel time, toll rates, the list of restaurants, resorts, hotels and cameras that are installed on highways.

The HAK mobile application offers a very transparent access to all data that are essential for the journey. In addition to what has already been stated, that is available on the website, mobile application offers finding the car if it is parked in a place which is not known to the owner. It also provides reference assistance both on the road and at sea by giving the exact location of which can help in case of a damage or accident with the vehicle.

Croatian Railways at http://www.hzpp.hr/voznired provides monitoring service trains in real time. All data can be found on their website and is very easy to find the data that are important for the train journey. The HZnet mobile application provides more readable and easier information about trains in Croatia.

Jadrolinija provides sailing schedule for Croatian ferries across the Adriatic Sea at http://www.jadrolinija.hr/. The data is in pdf document so if developers want to use that information, they must convert this data in more accessible format.

Most Croatian bus carriers provide their timetables on their websites. Some of that data is on website as image so they can't be reused. Some of data is in PDF document so it can be reused but it requires reformatting.

Mobile application provides information about domestic and international intercity bus schedule of Croatian bus carriers. We can find departures from each bus terminal in Croatia and phone numbers, locations and other contact information of bus terminals and carriers.

Croatia Airlines provides on its website (http://www.croatiaairlines.com/) all information about timetable and flight statuses of planes. There is also a mobile application that mobile users can use for getting information and checking their flights.

All post codes and post offices can be found on the Croatian Post website http://posta.hr/. Mobile application of Croatian Post also provides information about post codes and all post offices in Croatia. Users can use their GPS to locate the nearest post office, which is very usable if user doesn't know the area.

All major cities in Croatia provide their city budget open to the public. Most of that data is in PDF or Word document so it is easy to read for the user, but for other purpose, users need to reformat data. All data are in low level of transparency because they show only a rough picture of real costs for some activities and for every kind of deeper data you need to contact city office to give you that data.

City of Zagreb at http://data.zagreb.hr/ has published 12 datasets. Five of them are related to (un)employment statistics. Two datasets are published in XLS(X) formats, while others are available in preferred non-proprietary CSV format.

Part of research about open government data mobile applications was also challenging. First few applications

we found on webpages of datasets publishers. Then we concentrated on Google Play app store because it has good search engine, some data of our interest (number of downloads, average rating) and last, but not least important, according to [15], Android operating system is present in 69,5% mobile devices in Croatia. Since there is no „open data“ category of applications, we searched terms like Croatia, names of Croatian cities, weather, travel, timetable, budget, health, ecology and many other terms in Croatian language (sometimes also in English). When we found application of interest, we also searched for other applications of same developer. Sometimes we found applications in „Similar“ category. When we exhausted search in play store, we used web search engines and we searched on few Croatian web portals. We stopped when we found 50 apps, because after about 40 apps found, every new was hard and time consuming to find. So we are aware the application list is not final, but it gave us some interesting findings.

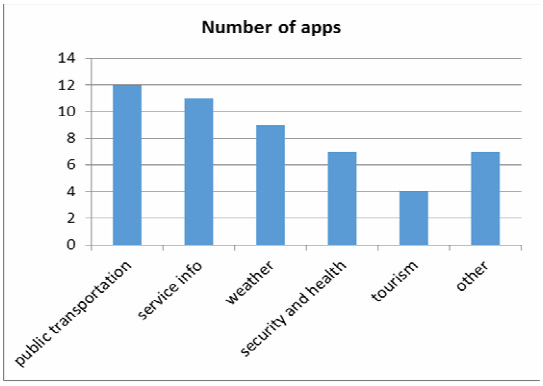


Figure 1. Number of apps per category

Figure 1 shows the number of applications within certain categories. Most of the apps are related to public transportation (bus and train timetables, delays, taxi and tram), various service information (holidays, post numbers and offices, road traffic, ATM machines etc.), weather, and security and health (emergency services, pollen forecasts, bike security, etc.). Our research didn't include classic tourist city guides, so number of tourist applications is quite low. All other apps are related to finance, ecology, city public services and social themes.

Table 2. Downloads and ratings of applications

Number of downloads range	Number of applications	Average grade
<100	4	4,0
100 – 500	7	3,9
500 – 1.000	2	4,6
1.000 – 5.000	9	4,0
5.000 – 10.000	8	4,1
10.000 – 50.000	16	4,1
>50.000	4	4,2

Table 2 shows that one third of apps are downloaded between 10.000 and 50.000, while only four are downloaded more than 50.000. Two of them are related to weather, one to road traffic and one to, interestingly, finding information about car (type, producer, insurance details etc.) based on registration plates. Only app about earthquakes didn't get any grade, but it is probably because it has been downloaded less than 50 times.

Only one app is not free of charge (it is related to Croatian marines), while all others are free to download. It is hard to say would it be different if Google allowed Croatian developers to sell their apps on Google Play. At this moment, Croatian developers can freely distribute apps, but if they want to sell, they have to open company in one of the countries which are acceptable to Google.

Fourteen apps were developed by individual users, and other by companies and institutions. Almost every app is in Croatian language, many of them are in English (and Croatian), and only a few are in Italian or German language.

V. CONCLUSION

Open data didn't empower economy as it was expected. Publishers of open data meet many issues such as unrealistic goals, expenses, decontextualization and lack of knowledge about what data is interesting to publish. Developers still have to learn how to use this data to create applications that are recognized and wanted in the market. They have to find an income model if they want to succeed. Developing mobile applications is not the only possible activity with open data; there are others – such as analysis creation, knowledge discovery, multi source mash-ups and so on.

Croatian IT specialists are still in a lack of valuable open government datasets, clear licenses and data formats. Despite that, they are quite active. They fight against hard-to-use formats, captchas, misunderstanding, and impossibility to charge app downloads on most popular app store. Portal data.gov.hr will make easier access to open data. About 70 datasets provided on portal maybe doesn't seem as a lot, but the quality (or better to say: usability) is more important than quantity. Developers should ask for opening other interesting data with understanding that publishers cannot always know needs for data in advance and also that publishers sometimes have rational constraints not to open every asked dataset.

REFERENCES

[1] C. M. Bonina, “New business models and the value of open data: definitions, challenges and opportunities”, NEMODE – 3K Small Grants Call, 2013.

[2] M. Jansen, Y. Charalabidis and A. Zuiderwijk, “Benefits, Adoption Barriers and Myths of Open Data and Open Government”, Information Systems Management, vol. 29, issue 4, pp. 258-268, 2012

[3] A. Peled, “Re-Designing Open Data 2.0.”, in Conference for E-Democracy and Open Governement, vol. 5 (2), pp. 187-199, 2013.

[4] G. Eibl and B. Lutz, “Money for nothing – data for free”, in Conference For Democracy & Open Government, 2013.

[5] J. Bates, “This is what modern deregulation looks like: co-optation Data Initiative”, The Journal of Community Informatics, vol 8 (2), 2012.

[6] D. J. Hand, “Data, Not Dogma: Big Data, Open Data, and the Opportunities Ahead”, Advances in Intelligent Data Analysis XII, 12th International Symposium, IDA 2013, London, 2013.

[7] A. Peled, “When transparency and collaboration collide: The USA Open Data program”, Journal of the American Society for Information Science and Technology, vol 62, 11, pp. 2085–2094, 2011.

[8] J. Oh, “Literature Review – Overview of Open Data in New Zealand”, University of Auckland, 2013.

[9] A. Zuiderwijk, M. Janssen and K. Jeffery, “Towards an e-infrastructure to support the provision and use of open data”, in Proceedings of the International Conference for E-Democracy and Open Government, pp. 275-291, Danube University Krems, 2013.

[10] T. Davies, “Open data, democracy and public sector reform: A look at open government data use from data.gov.uk”, Masters thesis, University of Oxford, 2010.

and contestation in the shaping of the UK’s Open Government

[11] P. M. Hartmann, M. Zaki, N. Feldmann and A. Neely, “Big Data for Big Business? A Taxonomy of Data-Driven Business Models Used by Start-Up Firms”, University of Cambridge, 2014.

[12] F. Welle Donker, “Public sector geo web services: which business model will pay for a free lunch?”, in SDI Convergence: Research, Emerging Trends, and Critical Assessment. Nederlandse Commissie voor Geodesie, pp. 35-51, 2009.

[13] Global Open Data Index: Survey, http://global.census.okfn.org/

[14] A. Musa, “Pravo na pristup informacijama u Republici Hrvatskoj: novi izazovi, nova rješenja”, http://www.pristupinfo.hr/wp-content/uploads/2015/01/01-Anamarija-Musa-Predstavljanje-rada-ureda-PI-240914.pdf

[15] Top 8 Mobile Operating Systems in Croatia from Feb 2014 to Jan 2015., http://gs.statcounter.com/#mobile_os-HR-monthly-201402-201501-bar

Cloud services and Marketing

Jelena Jeknić* and Bojan Kraut**
*Alcyone, Maribor, Slovenia
jelena.jeknic@alcyone.si
**Alcyone, Maribor, Slovenia
bojan.kraut@alcyone.si

Nowadays, most of the companies use cloud tools. Well-known services, such as Gmail, Google Docs and Dropbox are used every day. Salesforce is commonly used in companies' business routine. Through social networks (Facebook, LinkedIn, Twitter), significant part of personal and business data is collected and saved in the cloud. So, all these services and their data are easily accessible and available for different purposes.

On the other hand, users have to give up their privacy and control. Anyone with moderate analytical skills can research, profile and use all published data for their own gain and interest (advertising, sales, different analysis, etc.).

Nothing excites managers and marketers more than the massive application of cloud services and infinite flow of fresh information about consumers, which is then used to improve their products and services. In this paper we present process improvements in marketing automation, lead tracking, and inbound marketing as a result of cloud services.

No matter which cloud supported services are used, they are less time and money consuming and easier to access and manage. They are accepted by more users than traditional desktop tools. In the future, processes could be even more automated which could put businesses on autopilot. This might be the Holy Grail everyone is seeking.

I. INTRODUCTION

The theme of this paper is application of cloud services in business, especially in marketing. Our aim is to present how certain marketing concepts were not even plausible without cloud services.

Firstly, we present advantages and disadvantages of cloud services. So, we address issues of privacy and safety and then we present a couple of great implications of cloud services in marketing. We want to remind readers of all negative issues of cloud services, as we feel that users often embrace new applications of cloud services without taking these safety and privacy questions into considerations. We present marketing cloud implementations, because we want to show the vast area of cloud implication in business, and significant improvements it brought. We think that more and more services will migrate to the cloud in the years to come.

Our hypothesis is: "Marketing landscape has changed since cloud services were invented."

We use methods of description and compilation of secondary data to research the hypothesis.

Before we go any further, we would like to shortly define certain terms and terminology concerning cloud computing.

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared poll of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management efforts or service provider interaction. This cloud model is composed of five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service [4]."

In the simplest terms, *cloud computing* and *working in the cloud* refer to performing computer tasks using shared resources for services running in public, private or combined cloud.

According to the reference [5], cloud model is composed of four service models. These services are referred to as:

- *IaasS* (Infrastructure as a service – servers, networks, virtual machines),
- *SaaS* (Software as a service – virtual desktops, Email, CRM),
- *PaaS* (Platform as a service – databases, development tools) and
- *UCaaS* (Unified communication as a service – instant messaging, audio, web & video conferencing, desktop sharing).

The lines between different service models are not always clearly defined as there is no universally accepted distinction between lower-level and higher-level infrastructure.

According to reference [4] cloud model is composed of four deployment models:

- *Private cloud* is used by an organization for its completely internal use. So, it is built exclusively for an individual enterprise. Private cloud may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

- *Community cloud* is provisioned for use by a specific community. It may be owned, managed, and operated by one or more organizations, a third party, or some combination of them. It exists on the premises of the cloud provider.
- *Public cloud* is made available to the general public. Customers benefit from economies of scale because infrastructure costs are spread across all users.
- *Hybrid cloud* is composed of two or more distinct cloud infrastructures (public, community, or private).

When the cloud's service is being sold, that is *utility* or *on-demand computing*.

So, cloud computing is the sum of service model (SaaS, PaaS, IaaS, or UCaaS), deployment model (private, public, community, or hybrid) and utility computing.

II. ADVANTAGES OF CLOUD SERVICES

Broad application of cloud technology in our everyday life, especially in terms of business, is often related to instant access to the tools used by the millions for affordable service fees.

Developing great IT solutions takes extensive knowledge, financing, risk taking and time. It is not an easy task. However, everyone wants to consume IT without constructing it – and finally, with cloud services, it is possible!

Businesses that use cloud implementations, no longer need to invest large capital in hardware or human resources to build their services. They can even choose among different solutions from several certified cloud service developers. Companies of all shapes and sizes are rapidly adopting cloud technology, since it is cost efficient and the usage of cloud services presents minimal risk for business. Cloud services are usually developed in a way that they can be applicable in different companies regardless of their size, shape or industry. Therefore, development costs are distributed over many users, positively influencing the pricing policy. So, prices of cloud services are relatively low.

Another benefit of cloud services is device and location independence. Modern cloud systems are constructed of sophisticated server geo-clusters. So, a specific piece of information is stored on several cluster nodes worldwide and on-demand, it is transmitted to the user as soon as possible. This system offers greater level of service comparing to the traditional system architecture. Moreover, it is also pretty common that cloud services offer application programming interface (API), allowing users to expand the functionality of the service and tailor it to their needs. Most of solutions are Web-based and even older computers can be used to access cloud services.

Cloud computing facilitates easier collaboration, since distributed teams or combinations of mobile workers and in-office staff can all work together on shared information stored centrally in the cloud. From the employee point of view, cloud's greatest beneficiaries may be remote workers, as Web-based applications empower them to be truly mobile and still accomplish their work.

Key characteristics of cloud computing mentioned above are just a few of many advantages cloud computing offers.

III. DISADVANTAGES OF CLOUD SERVICES

Along with pretty great advantages, cloud services have disadvantages, as well.

The first shortcoming of cloud computing is the fact that an Internet connection is necessary to take full advantage of a cloud service. Therefore, when users are offline, or if there are some disruptions with the cloud service the data may not be accessible at all. So, if the service is disrupted or stopped, it can result in additional costs.

However, there are applications, such as Evernote (note-taking application), that offer a good in-between or hybrid solution, with both desktop or phone software, and an online service that synchronizes data to the cloud. So, there are applications (e.g. Gmail) that offer some basic offline functionality, as well.

Far more serious and not easily solvable issue with cloud computing is security – both from providers' and consumers' point of view.

The location data processes and storage determinates the security in terms of data safekeeping. Organizations have a full control over data only when the data is stored on their own physical computers and servers. On the other hand, when organizations have service and data in the cloud, then they are not fully aware of where their processes are running or where their data is stored, as they have no control over it. They have to trust cloud service providers to respect their privacy and follow the regulations. So, individuals and/or companies may not be comfortable storing their information, especially proprietary or sensitive data, on someone else's server on the Internet.

End users have to agree to the privacy policies before they start using cloud services – it is "take it or leave it" contract. So, if it is not encrypted, cloud services providers have access to all the data and they could accidentally or deliberately disclose it to third parties or use it for unauthorized purposes. Of course, this raises privacy and confidentiality concerns.

The social networks' privacy issue is one of the biggest, as some applications are explicitly centered on "cyberstalking".

For instance, when Facebook recently announced changes to its default privacy settings regarding user's content sharing with friends, it was also revealed that Facebook can access smartphone's microphone to capture and analyze songs, TV shows and other content users listen to. This sounds disturbing.

Google suffered a court ruling in Europe last year, forcing it to allow petitioners to be "forgotten" by its search engine.

Yahoo recently changed its policy, so that it can deny users the ability to request that their behavior is not tracked, and called it an improvement in "personalized experience".

It is crystal clear that the main goal of cloud service providers is profitability which depends on users' shared data and on users remaining unaware of the extent to which their privacy is exploited.

Most of cloud service providers' profit comes from businesses, that want to acquire as much data as possible about their customers, because it enables them to not only better recognize customers' needs, but to anticipate them, ideally before they have been consciously acknowledged by customers themselves.

These issues of trust, confidentiality and reliability are critical for cloud services to resolve before everyone truly moves to the cloud. Certain security assurance is provided through service level agreements (SLA) or other standardized agreements. However, cloud services need way more reliable assurances of encryption technologies, privacy protection and solutions for offline accessibility.

There are a couple of suggested solutions to privacy issues in cloud computing. These suggestions include policy and legislation, as well as end users' choice for how data is stored [5]. The cloud service providers need to establish clear and relevant policies that describe how the data of each cloud user will be accessed and used. Cloud service users should be able to encrypt their data when it is processed or stored within the cloud, as encryption is one of the solutions for preventing unauthorized data access [3].

IV. CLOUD APPLICATIONS

According to reference [1], public services linked with cloud computing grew from $9 billion to $40 billion over the last five years. This convergence is aggravated by the increased usage of e-Commerce, social media, smart phones and mobile commerce.

Reference [6] has divided all cloud computing business models into eight types: (1) service provider and service orientation, (2) support and services contracts, (3) in-house private clouds, (4) all-in-one enterprise cloud, (5) one-stop resources and services, (6) government funding, (7) venture capitals, and (8) entertainment and social networking.

Today, companies can find almost all desired applications in the cloud. Following services are just some of many useful tools that can speed up daily work and lower operational costs, especially in small and medium enterprises (SMEs).

Solutions like Openstack offer private cloud building tools, so companies can create their own clouds. Such services often have strong online community, so users can get help and support quite easily.

Applications such as Gmail or Hotmail refer to software solutions provided over the Internet, or SaaS. They support communication within and outside the company. Other cloud computing services might include server virtualization and data storage, or IaaS, such application is Amazon Web Services, offering wide variety of services. If a cloud service includes software and product development tools then it is PaaS, e.g. Google Apps.

Mozy, Dropbox and Goggle Drive are online backup and synchronization services that continuously back up the files on a computer or server. They give small businesses the space to back up all their computer and server files for a very reasonable price, so owners of small or medium enterprises know their files are retrievable, even during a data loss crisis.

Finance cloud applications, like Outright help small businesses with their business accounting. They allow enterprises to track income, expenses, tax obligations, and profits or losses, in real time. This type of service is ideal for small companies or entrepreneurs that are looking to get a hold of their finances, as it provides access to finance data whilst on the move.

Here are other well-known cloud services that don't need much introduction:

- Salesforce – it supports most of company's marketing and sales activities and it can be easily expanded,
- Google Docs and Google+ – they support data storing, sharing and communication,
- Facebook, Twitter, Pinterest and other social networks – they support networking, communication, sharing, content marketing, etc., and
- Skype – it supports communication through instant messaging, video conferencing, as well as data sharing.

In spite of all the benefits these services offer, the integration of all different cloud services into the organization's core business software might not be possible or it might cost a lot of money, so organizations should choose their cloud services carefully.

V. APPLICATION OF CLOUD SERVICES IN MARKETING

In today's knowledge economy, the marketing landscape for businesses has changed dramatically especially thanks to the cloud services. Companies that hope to effectively: spread the word about their brands and products, attract customers, engage with them and ultimately retain them, need to adapt to mentioned market changes as soon as possible. Those that fail to do so will fall behind.

There are some marketing concepts that are impossible without cloud computing, and before cloud services, they existed only as mere ideas.

A. Marketing automation and lead tracking

Marketing automation is a process that enables marketers to nurture and quantify leads.

It includes: lead generation, segmentation, lead nurturing and lead scoring, relationship marketing, cross-sell and up-sell, customer retention, and marketing return on investment (ROI) measurement.

First and foremost, marketing automation software helps organizations discover more about their leads and customers. Organizations can track and expand their leads' profile data, including demographic and firmographic information that can be used to better segment and target the marketplace. Profile data can even include attitudinal information obtained from different surveys and form questions that provide powerful insights into leads' intentions and authority. By tracking what was previously invisible – the interactions leads have with organization's content – companies gain a much richer overview of the audience and how content meets their needs.

Armed with a greater understanding of their leads, organizations are in a position to expand their market reach. With automated outbound messages, whether in the form of newsletters, messages triggered by site registrations or complex automated nurture campaigns, they can reach more prospects in easier and more effective manner.

More touch points with leads mean more interactions and deeper understanding of leads, so organizations can prime them for engagement. The engagement starts with understanding the buyer's needs and the whole process buyer goes from awareness through research to a purchase decision. So, with marketing automation tools each customer step is carefully recorded by the IT infrastructure, and then compared to existing data, so next step can be predicted, often more accurately than sales personnel does it. Organizations can now create content to better meet their leads' information needs at every stage of their purchase decision process, and track their level of engagement.

Marketers' ultimate goal is to convert visitors to leads, leads to sales opportunities and finally, profitable customers. Marketing automation software enables organizations to convert anonymous visitors into qualified prospects using features such as reverse IP look-up and real-time website personalization with collaboration filtering. Organizations track and score these leads and convert them into sales opportunities.

Marketing automation is all about measuring. When using marketing automation service for sending out an e-mail (e.g. newsletter), firstly, companies can send it to more e-mail addresses, and they can also see how many e-mail receivers have actually opened the e-mail in question and/or specific (offer) link within the e-mail. They can also identify those receivers and keep records of them that can be used for personalization of future marketing activities. Companies can also measure the conversion of leads into potential customers, and finally buyers, which is their ultimate goal, of course.

Furthermore, before executing an e-mail campaign, several different versions of the e-mail can be prepared and tested, varying in title, content, day and time of sending. So, different versions of e-mail are sent out as a pilot testing, feedback is measured, analyzed and then the most successful scenario is chosen and used for the final e-mail campaign.

Services like MailChimp and Mandrill offer detailed analytics and control over sent e-mails and with their new mobile friendly templates they can reach almost everyone.

Well-known marketing automation cloud services with very similar sets of solutions offer wide application possibilities and they support many marketing techniques. Here they are:

- Oracle Eloqua – it offers content creation templates for e-mails, landing pages, social media marketing, lead segmentation, lead scoring/analysis, and personalization of marketing activities,
- Marketo – it also supports generating demand, building relationships with prospects, driving sales, measuring and optimizing,
- Google Analytics – it offers all web analytics, reporting and it measures conversions and sales,
- Adobe Marketing Cloud – it offers analytics, social, media optimization, targeting, web experience management, and cross-channel campaign management.

B. Inbound marketing

Marketing automation can be both outbound (e.g. e-mail marketing) and inbound (e.g. social media marketing).

Problems with some outbound marketing techniques include: difficulty in tracking return on investment (for TV or print ads), increasing blocking techniques ("do not call" list in phone marketing, spam filters in e-mail marketing, etc.) and high cost – low yield as a result.

The greatest advantage of inbound marketing lies in its extensive possibility of tracking and measuring. As mentioned before, the tracking of what was before social networking invisible, helped companies get a much larger picture of all their prospects and potential customers, and considerably richer view of their customers' needs, wants and interests.

Social networks tend to make users more open. Spending significant time online, using social networks, users become much invested in social media and they feel

connected through them with their friends. They follow their online friends' content and advices about products and services, which then consequently raises their level of trust in those products. Social network marketing campaigns are becoming more and more efficient, because it became easy to personalize those campaigns, lower marketing costs and improve conversions, all as a result of insights from considerable user data analysis.

Inbound marketing, in the simplest words, consists of creating content and posting it on the Internet. Inbound marketing strategy focuses on attracting customers, or leads with original content, thereby having potential customers come to the company themselves, rather than marketers vying for their attention. The beauty of inbound marketing lies in this very element that content readers are interested in the content and they voluntarily look for it, so they immediately become potential customers and even future promoters.

Inbound marketing is one of the hottest marketing techniques in business today. Be that as it may, managing outstanding marketing campaigns requires some human and technical resources and knowledge.

Quality content is not easily created and it takes much time and energy. Moreover, content marketing requires contemporary information sharing, active educational approach, significant general knowledge about product and related issues, and to top all up, excellent writing skills. All these competences are equally important, if marketers want users to perceive the content as reliable and significant, and to relate it to the product or company and get involved with the whole idea, so that they will share it with their friends and hence, become promoters the brand or product.

Inbound marketing content can be in form of: blog and blogging, forum activity, comment marketing, video and/or audio materials (webinars, online video, tutorials, etc.), podcasting, document or links sharing, questions and answers sites, social media marketing, self-published books, etc. With social networks, sharing such content is easy, as social networks facilitate user interactions.

The impact of the published content is usually measured in user interactions. Social networks offer different ways of sharing content, choosing right audience, choosing language, scheduling date and time of posts, reposting and post-editing, adding links, photos and videos, and performing descriptive statistics on users, posts, interactions, etc. As end users also have considerable set of tools to make their social network experience more agreeable, they can choose how they want to be served with the content. They can choose their interests and ways to get notified, as well as frequency and importance of the certain notifications and information, so basically, content can be personalized and with appropriate analytics and marketing skills, users can be served with targeted content and advertisements.

YouTube, Twitter, Pinterest and Facebook are all examples of cloud hosted services that have enabled the rise of inbound marketing.

Successful inbound marketing softwares simplify organizations' marketing efforts by automating content creation and distribution, lead capture and management, and measuring of ROI.

C. The art of engineering which makes it all work seamlessly

In general, marketing oriented cloud services address users that are potential customers and they promote users' engagement in company's content.

As mentioned, the most commonly used cloud services are Facebook, Youtube, and Twitter. They all allow users to share, comment and like the content. Sophisticated algorithms allow indexation of graphical, audio and video content, then they process the data in specific cloud services to extract all kind of metadata, from simple keywords, to human faces (Facebook acquired DeepMind, a small company that developed a software, which can recognize human faces with higher accuracy then a human being) and according to the results, algorithms can identify and "calculate" users' interests, personality and their mood [2].

Sophisticated algorithms can also perform predictions for larger group of people, and for the society as a whole [2]. So, well calibrated sophisticated prediction algorithms can, for example, predict society mood and satisfaction, and with high accuracy they can predict major world events such as strikes and protest, crises, even wars.

To increase the volume of analytics, social networks (e.g. Facebook), create widgets such as "Like", "Share" or "Comment" buttons which can then be integrated into external sites, as well. So, beside richer user experience and smooth content sharing mechanisms, these "buttons" facilitate monitoring of the user's activities all over the Internet. A user can be identified on the Internet in several ways. Most common way is through login, during which a cookie is saved on user's device, so each and every time specific user goes to pages with Facebook wigets, no matter from which device, the information about user's activity is gathered and saved. Furthermore, along with monitoring users' activity on the site, cookies also get information about which external pages are visited during and after Facebook sessions. Based on that, sophisticated algorithms can calculate user behavioral patterns.

With all the information gathered, marketers can thoroughly profile users, and according to all relevant information their experience can be nicely personalized. Personalization leads to one-to-one marketing or targeted marketing which is an extreme form of database marketing.

Marketing used to be based on product differentiation. Namely, marketers tried to differentiate their products from competing ones. Personalization has changed these marketing principles, since it tries to make a unique product offering for each customer individually. Consequently, users are served with the content, which is exactly what they want to see, hear or buy.

As these days most of the people from developed countries have smartphones – it is even simpler to perform personalization. Namely, smartphones send their location information when sending audio or video data from the phone to the cloud services, and new marketing strategies are based on proximity services, so users receive push notifications according to their current location. Smartphones represent a huge potential and great advantage for marketers.

With software platforms like HP Autonomy's HAVEn almost all the information can be extracted and categorized, forming unimaginable rich context for a specific user. As these platforms analyze and find meaning in "big data", this means that marketers can know more about users than users are aware of, which gives them a huge advantage in sales processes.

Thanks to "big data" and algorithms, it has become possible to anticipate individual's needs and tailor online advertisements and content according to a specific user. Additionally, services that help users personalize their devices and make it easier to automate tasks across applications are adding even more layers of functionality to mobile gadgets. In the future, these services will embrace emerging technologies such as augmented reality (AR) and then, the options for further development and application will be endless. For example, with AR all the data will be streamed directly from the cloud, according to user location and presented on next generation gadgets.

VI. CONCLUSION

The evolution of cloud computing over the past few years is potentially one of the major advances in computing.

Before cloud services, marketers used concept of four Ps (product, price, place and promotion), outbound marketing (TV and outdoor ads, direct marketing, etc.), segmentation of customers and differentiation of the products (figure 1) but now, they use new cloud based marketing activities that are build on completely different marketing philosophy.

Old marketing concepts were difficult to track and evaluate. Segmentation was long-lasting and never ending process that took place hand in hand with differentiation of the product, but was never flexible enough or quick enough to follow all market changes. Furthermore, it was quite demanding to calculate marketing return on investment (ROI).

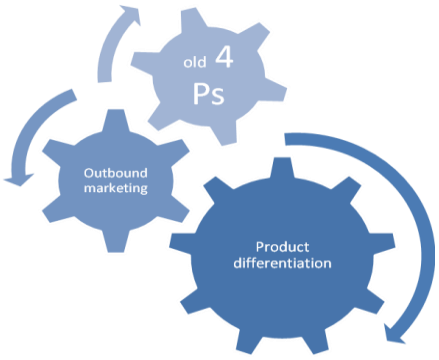


Figure 1. Pre-cloud marketing

These new concepts are less time and money consuming and more effective than older ones.

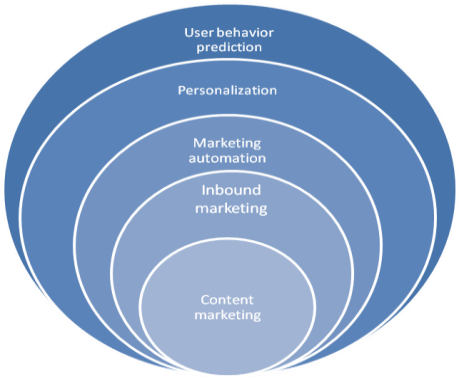


Figure 2. Post-cloud marketing

Cloud based content marketing, inbound marketing and marketing automation tools help organizations find out more about their leads and customers. They also provide great deal of functionality and measurement, so personalization and even user behavior prediction are possible (figure 2).

Cloud based marketing tools made it possible for organizations to bring customers on their sites, monitor their behavior, compare the results in real time and adjust marketing strategies accordingly. Furthermore, it is possible to create several different strategies and learn from a pilot test which one is the most effective. So, through better lead and customer analysis and personalization help organizations offer such products and services that are exactly what customers want. So, marketing principles have changed drastically. Marketers have adopted these new concepts and they exploit them in their everyday work. The philosophy has changed.

Therefore, we confirm our hypothesis: "Marketing landscape has changed since cloud services were invented."

Emerging technologies will blur the boundary between digital and real, and merge the two worlds together. Our every move, interaction and behavior will be recorded and processed in the cloud, and then used to augment our world.

For this reason, if cloud computing is to achieve its potential, everyone needs to research and address security and personal privacy issues with more care. It is much needed that various raised questions (both from the perspective of the providers and the consumers of the technology) are resolved, sooner rather than later.

Cloud services remain not just one of the most interesting contemporary computing issues, but one of the most challenging issues in general.

REFERENCES

[1] D. C. Lakshmi, "Impact study of cloud computing on business development", Operations research and applications: an international journal, 2014, 1(1), pp.1-7.

[2] J. Anderson and L. Rainie, "The future of big data", Pew research center's Internet and American Life Project, 2012, [available online: http://www.pewInternet.org/topics/Future-of-the-Internet.aspx].

[3] M. D. Ryan, "Cloud computing security: The scientific challenge, and a survey of solutions", Journal of Systems and Software, 2013, 86(9), pp.2263-2268.

[4] P. Mell and T. Grance, "Final version of NIST definition of cloud computing," Special publication 800-145, National institute of standards and technology, U.S. Department of commerce, 2011 [available online: csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf].

[5] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang and A. Ghalsasi, "Cloud computing — the business perspective", Decision support systems and electronic commerce, Elsevier, 2011, Vol. 51, Issue 1, pp.176-189.

[6] V. Chang, D. Bacigalupo, G. Wills and D. De Roure, "A categorization of cloud computing business models", CCGRID '10 Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, IEEE Computer Society, Washington DC, 2010, pp.509-512.

RFID based access control system

Pančo Ristov*, Toni Mišković* and Ante Mrvica**
*Faculty of Maritime Studies, University of Split, Split, Croatia
**Jadrolinija d.o.o. Rijeka, Croatia
panco.ristov@pfst.hr; tmiskovic@pfst.hr; antemrvica@jadrolinija.hr

Abstract – Radio frequency identification (RFID) provides the ability for unmanned automatic real-time data acquisition. The aim of this paper is to open discussion about implementation of RFID technology in Croatian ports. The authors propose RFID as one of the technologies that can accelerate and improve ferry loading process and make it more efficient and reliable. Basic RFID system was built to introduce its capabilities to perform as one of the suitable solutions for ferry tickets check-in process within certain technical constraints.

I. INTRODUCTION

Radio frequency identification – or, as it is more commonly known, RFID – is an automated data acquisition technology. Nowadays, RFID technology has become integral part of our daily lives routines and process, such as: restrict access to buildings or rooms within buildings, livestock identification, asset tracking, pet ownership identification, warehouse management and logistics, product tracking in a supply chain, product security raw material tracking/parts movement within factories, automatic vehicle identification (AVI) systems, library books check-in/check-out, railroad car tracking, luggage tracking at airport, electronic toll systems, automobile ignition systems, tracking containers in an integrated traffic.

The developments in the semiconductor industry, as well cost decrease, trigger the mass usage of RFID in the global business world. RFID can be used for redesigning business processes, cutting costs, and improving the operational performance by increasing factors such as: efficiency, accuracy, visibility, and security [1].

Figure 1 presents applications of RFID by purpose in enterprises in the EU and Croatia, first quarter 2009

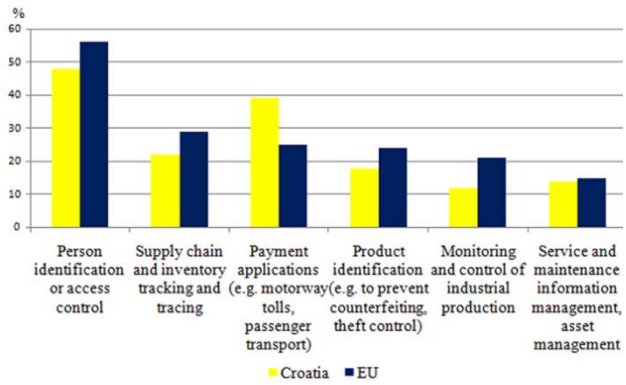


Figure 1. Applications of RFID usage in enterprises by purpose

(Eurostat- the statistical office of the European Union and Croatian Bureau of Statistics, 2010) [2][3]. RFID technology was mostly used in Croatian enterprises for person identification and/or access control and for payment applications (e.g. toll collection). Application of RFID technology is increasing in various industries while prices of RFID elements are continuously decreasing. According to data from Eurostat, the most common application of RFID technology in enterprises in the EU (in January 2009, sample size: 152 900 enterprises out of 1.6 million in EU-27 were surveyed) was in the area of person identification or access control (56%), supply chain management and inventory tracking and tracing (29%), payment (25%), product identification (24%), monitoring of industrial production (21%) and service and maintenance information management (15%). In 2011 (sample size: 146 000 enterprises out of 1.5 million in EU27 were surveyed), among all enterprises in the EU27, the use of RFID devices remained limited (4%) and at a similar level to 2009 (3%) [4].

RFID technology has been used in public transport for many years ago. RFID system, which is used in Helsinki, is the largest system and it comprises all sorts of public transport, as well ferry transport (e.g. M/S Suomentlinna II). Norwegian ship operator Fjord Line uses RFID technology to manage ticketing and cabin access on one of its ferries (MS Stavangerfjord), and plans to expand the technology's use to the other ferries and as well on other onboard services, such as at restaurants [5].

In the most applications RFID technology achieve from 99,5% to 100% accuracy of reading in first scan. Furthermore, without moving elements, maintenance of RFID resources is far simpler and more efficient [6].

Integration of RFID technology into integrated information systems of shipping companies enables: secure and real-time passengers/vehicles tracking, automated data entry, provides insight into real situation of passengers/vehicles on each ship, ticket control without presence of the ship's officers, increase the capacity of port in a way that accelerate the process of verification, monitoring and management of passengers/vehicles, reduction of operating costs, timely and accurate information's about passengers/vehicles in front of berths in order to plan the needs of port resources, increase control functions at all levels of managements and optimization of the vessel traffic in the port.

This paper presents the RFID technology, its practical application, with special emphasis on application in the ferry ticket check-in process.

A. Basic physics of RFID

Radio frequency identification is based in discoveries made by Faraday during the mid-nineteenth century, as well in discoveries made between 1900 and 1940 in radio and radar technologies. Faraday discovered the concept of mutual induction, which is the basic principle for powering short-range passive tags operating in the near field. During the first half of twentieth century, due to technological advances, the development of far-field tags occurred. Far-range passive tags must harvest energy to operate and backscatter the interrogation signal transmitted by the reader to communicate with it. The voltage induced in a coil consisting of N loops immersed in a time-variant magnetic field can be calculated with equation of Faraday's law (1):

$$V = -N\frac{d\Psi}{dt} \tag{1}$$

with Ψ being the magnetic flux.

The magnetic flux can be calculated as (2):

$$\Psi = \int B \cdot dS \tag{2}$$

where:

- B - the magnitude of the time-variant magnetic field,
- S - the surface area of the coil.

Reflection is a property of the matching of the impedance of the antenna to the impedance of the load (chip) on the RFID tag [7]. The reflection coefficient is calculated in equation (3):

$$\Gamma = (Z_L - Z_A)/(Z_L + Z_A) \tag{3}$$

where:

- Z_L – chip (load) impedance,
- Z_A – antenna impedance.

If the load impedance is shorted (0), $\Gamma_{short} = -1$. If the load impedance is open (∞), $\Gamma_{open} = 1$, and if the two impedances are matched, that is $\Gamma_{match} = 0$. Thus, all other variations of matching the antenna and load impedances result in a Γ somewhere between + 1 and – 1.

II. SYSTEM OVERVIEW

RFID technology does not require line of sight and provides fast and accurate reading from a great distance without human intervention. Besides hardware and middleware cost, service cost factors, including compliance cost, cost of training users, and transition cost make the total RFID system cost structure. Benefit factors are more complex and almost impossible to calculate than cost factors as the limited pilot studies or studies from other application areas are unable to reveal the full set of benefits. Also, intangible benefit factors such as increased customer satisfaction are not easily measured. A simple RFID system is composed from three main components (tag, reader and middleware) shown in Figure 2.

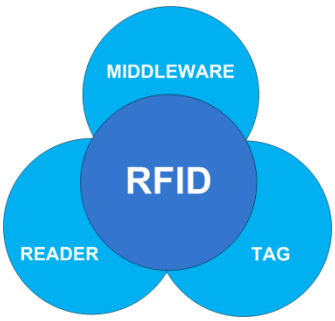


Figure 2. Three main components of RFID system

A. Tag

Tag (***trans***mitter/res***ponder-***transponder) is composed of a small microchip and an antenna and can be embedded into or attached to any kind of objects to be identified in an RFID system. Tags have different shapes and sizes and have a memory for data storage. These are specific data such as a unique identification number, ticket price, service location, date and time of transaction.

For such systems multi-application smart cards are mostly used. The card belongs to a group of passive RFID tags.

Passive tags are also called pure passive or reflective. Passive tags obtain operating power from the reader which sends electromagnetic waves that induce current in the tag's antenna and the tag reflects the RF signal transmitted and adds information by modulating the reflected signal. Chip cost and assembly cost define the tag cost and as the manufacturing volume increases chip cost decreases [8].

The advantage of RFID tag is that user can use one tag repeatedly (reusable) during many years. This advantage is significant for tourists who come every year. The data in the card can be changed only by authorized personnel. Devices for amendment are located at several locations in the port and/or at an authorized agency.

B. Reader

RFID readers communicate with tags through an RF channel to obtain identifying information. The basic components of the reader are: control unit (microchip), antenna and communication interface. In environments with many tags, a reader may have to perform an anti-collision protocol to ensure that communication conflicts to not occur. Anti-collision protocols permit readers to rapidly communicate with many tags in serial order. Readers often power what are called passive tags through their RF communication channel. These types of tags carry no on-board power and rely only on a reader to operate.

Reader cost, together with tag cost make up the hardware cost of an RFID system. Reader and related infrastructure cost are often regarded as fixed cost in contrast to tag cost which is variable. RFID readers may

also be integrated into hand-held mobile devices. Readers can have an integrated antenna, or the antenna can be separate device.

Communication between reader and computer takes place over communication interface, which can be RS-232 and/or RS 485, USB, WiFi and TCP/IP.

RFID readers are connected through the network (Ethernet protocol) with database server and application server. Furthermore, RFID readers can be connected with long range wireless connection (WiFi) or/and shorter – range wireless connection (Bluetooth). The choice of connection methods depends on depends on the configuration of berths and on the existing network in port. These solutions can be combined (wireless and wired devices at the same time in the network).

C. Middleware

Middleware software manages the readers and the data coming from the tags, and passes it to the backend database system. Figure 3 presents RFID system with middleware included. Middleware fills the gap in the middle of the data flow between the readers and the processing unit, and manages the flow of information between the readers and the processing unit. In addition to extracting data from the RFID tags and managing data flow to the processing unit, middleware performs functions such as basic filtering and reader integration and control. An RFID tag sleeps in normal mode: it does not perform any activities and does not use any power. This will change as soon as the tag moves through the reader's field. An RFID reader emits radio waves at a frequency similar to those of the tag. The tag antenna retrieves enough energy from waves so as to exit the tag from the sleep mode and to power the microchip. The tag's microchip initiates a wireless communication with the reader and receives a command. The chip will execute the command, resulting in a data transfer through the tag's antenna. In case of the passive tag, the data consists of the unique serial number of the tag in binary format emitted as radio waves. The antenna will pick up the data emission and sends it through a cable and in the form of

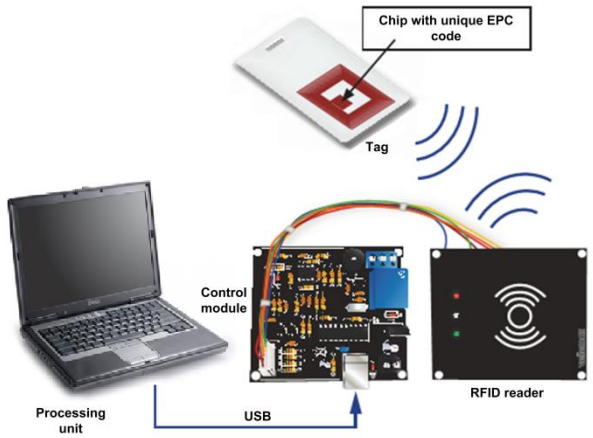


Figure 3. RFID system (including middleware)

electrical signals to the reader's control module. The module will in turn decode the radio signals and retrieve the binary data. The reader now has the tag's ID in digital format, which can be processed or printed immediately. This data will mostly be completed with additional data (e.g. time and location) and quickly forwarded to a processing unit or information system [9].

RFID middleware is software that facilitates communication between RFID readers and information system of enterprise. Middleware collects, filters, aggregates and applies business rules on data received from readers. It is also responsible for providing management and monitoring functionality - ensuring that the readers are connected, functioning properly, and are configured the correct way. Middleware may also contain a localized data store for data archives. Middleware may be implemented on a host computer, a central server, or on intelligent readers.

III. DISCUSSION

The authors propose RFID as one of the suitable ID technologies that can accelerate and improve ferry loading process and make it more efficient and reliable without burdening employees. Among automatic ID systems shown in Figure 4, RFID technology is optimal solution for check-in procedures. RFID technology provides support for e-ticketing and creating passenger and vehicle database.

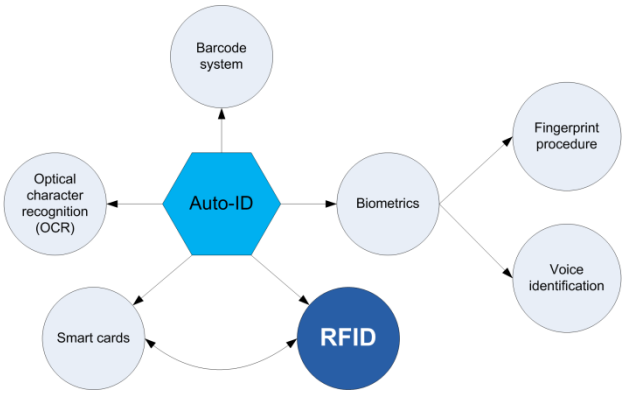


Figure 4. Overview of the automatic ID procedures

The barcode is a binary code comprising a field of bars and gaps arranged in a parallel configuration. They are arranged according to a predetermined pattern and represent data elements that refer to an associated symbol. Optical character recognition (OCR) systems have failed to become universally applicable because of their high price and the complicated readers that they require in comparison with other ID procedures. Biometrics is defined as the science of counting and (body) measurement procedures involving living beings. In the context of identification systems, biometry is the general term for all procedures that identify people by comparing unmistakable and individual physical characteristics. Voice identification systems have become available to identify individuals using speaker verification (speaker recognition). In such systems, the user talks into a microphone linked to a computer. This equipment converts the spoken words into digital signals, which are

evaluated by the identification software. Fingerprinting procedures (dactyloscopy) are based upon the comparison of papillae and dermal ridges of the fingertips, which can be obtained not only from the finger itself, but also from objects that the individual in question has touched. When fingerprinting procedures are used for personal identification, usually for entrance procedures, the fingertip is placed upon a special reader. The system calculates a data record from the pattern it has read and compares this with a stored reference pattern. Modern fingerprint ID systems require less than half a second to recognize and check a fingerprint. A smart card is an electronic data storage system, possibly with additional computing capacity (microprocessor card), which – for convenience – is incorporated into a plastic card the size of a credit card. Smart cards are placed in a reader, which makes a galvanic connection to the contact surfaces of the smart card using contact springs. The smart card is supplied with energy and a clock pulse from the reader via the contact surfaces. One disadvantage of contact-based smart cards is the vulnerability of the contacts to wear, corrosion and dirt. Readers that are used frequently are expensive to maintain due to their tendency to malfunction. RFID systems are closely related to the smart cards. Like smart card systems, data is stored on an electronic data-carrying device – the transponder. However, unlike the smart card, the power supply to the data-carrying device and the data exchange between the data-carrying device and the reader is achieved without the use of galvanic contacts, using instead magnetic or electromagnetic fields.

Contactless smart cards are designed for a lifetime of 10 years. Rain, cold, dirt and dust are not a problem for either smart card or reader. Contactless smart cards can be kept in a briefcase or handbag and are therefore extremely convenient to use. Although the purchase costs of a contactless smart card system are still higher than those of a conventional system, the investment should repay itself within a short period. Usually, in payment systems contactless smart cards apply standard ISO/IEC 14443 for data transmission [10]. The replacement of conventional paper tickets by a modern electronic fare management system based on contactless smart cards provides a multitude of benefits to all those involved.

Use of RFID technology facilitates automation of processes and improves operations management through reduction of workload and elimination of human errors. The most of Croatian enterprises still use labor-intensive methods (enter data manually and using bar-code systems) for products-related data acquisition. However, RFID technology has some advantages over the bar code:

- there is no need for line of sight (tags can be read through a different materials),
- suitable for harsh environments (e.g. dirt, moisture, dust, chemicals, high temperatures),
- the RFID reader could interact with multiple tags and automatically receive information from the tags,
- greater distances of reading with RFID readers,
- tags contain more data than bar codes,
- tags can have read/write memory capability, while barcodes do not (bar codes are not able to be modified once they are printed),
- form and dimensions of tags could be various, depending on the application.

Authors propose following main attributes as possible data composition for RFID ticket/tag:

- passenger ID,
- vehicle ID (e.g. license plate),
- maritime company ID,
- code of navigable area,
- credit balance,
- ticket ID,
- departure time of ferry,
- date and time of transaction.

Benefits of using RFID ticketing are presented in Table 1.

TABLE I. BENEFITS OF RFID TICKETING

Benefits for passengers
• cash is no longer necessary, RFID tag can be loaded with large amounts of money, passengers no longer need to carry the correct change • prepaid RFID tag remain valid even if fares are changed • the passenger no longer needs to know the precise fare; the system automatically deducts the correct fare from the tag • monthly tickets can begin on any day of the month, the period of validity begins after the first deduction from the tag
Benefits for the driver
• passes are no longer sold, resulting in less distraction of driving staff • no cash in vehicle
Benefits for the maritime company
• very secure against vandalism • it is easy to change fares; no new tickets need to be printed • ticket holders are known so passenger/vehicle reports can be created • expressive statistical data is obtained • waiting time reduction • transparency

Figure 5 presents overview of RFID ticket check-in procedure which is created by authors. RFID check-in procedure can be performed without impact of weather conditions, thus whole process become more robust and reliable. Authors suggest separate new user registration and check-in procedures to make concept smoothly and transparent. Without RFID ticketing roughly more than half ticket holders are unknown so passenger and vehicle reports created from known RFID ticket holders are of great importance for maritime company. Tags can be interrogated by RFID hand (mobile) readers or, more complex, by stationary fixed readers for realizing unmanned solution without employee burdening (human errors are reduced to minimum).

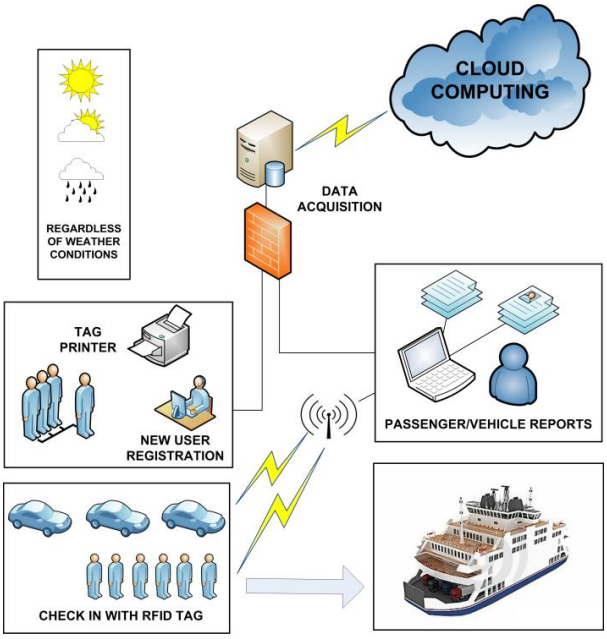


Figure 5. Overview of RFID check-in procedure

The collected data can be stored on the local database server (agency), on the maritime company's database server, or nowadays, more reliable way is to store data in the cloud. This paper gives detailed description of the data storage in the cloud.

Cloud computing storage is provided for backup of passenger and vehicle reports. Cloud computing offers maritime company back office online (real-time) data acquisition and inspection.

Cloud computing represents a combination of many different existing technologies comprising the virtualization, data storage, connectivity and processing power to create a new environment. Storing data in the cloud is an easy and automated backup mechanism for real-time storage of data to a remote location. This mechanism provides a reliable solution with a centrally managed backup that facilitates easy storage of data and enables company managers and owners to protect their data as the most important corporate assets. Furthermore, this mechanism provides a very efficient compression, global data deduplication, significantly reducing the time frame for storage, energy storage capacity and loading on the network infrastructure. Data deduplication eliminates duplicate data at the entire port, detached offices and mobile users. In addition, the mechanism verifies the integrity (Data Integrity Verification) with automatic recovery and the possibility of advanced rescue (Advance Recovery) that provides immediate availability and eliminates the need for manual data recovery activities.

Cloud system is an ideal way to effectively protect important data on a remote location against unexpected events such as theft, fires, earthquakes, floods, etc. Cloud systems are software solutions that perform storage of data at several remote locations at certain time intervals. In a secure manner the software collects, compresses, encrypts and archives data on the servers of the service provider, which are located at several different remote locations.

Before any piece of data is transferred, the data storage system in the clouds encrypts it with the most secure encryption mechanism. The data remain encrypted during storage and during all transfers. The system has a special module enabling that all data storage and recovery processes are registered, therefore allowing a clear overview of data operations. In addition, such a system features a multiple factor authentication of identity. In other words, the data storage system in the cloud is a set of regulations for digital storage and access to confidential data.

Data storage systems in the clouds are professionally maintained servers and disk systems, providing support 24 hours a day, seven days a week. Data are stored in three or more geographically remote data centers, thus providing unmatched availability and reliability. If there is a failure in the system, the data can be accessed in a secure manner until the re-establishment of the system's full availability.

Systems of data storage in the clouds will be increasingly present within the integrated information and communication systems because they are cost effective in financial and material terms. There is no need for a database administrator and there are no new or additional costs regarding the acquisition and maintenance of the software and hardware... The only cost is related to paying for the service which eliminates the need for any subsequent significant investment.

The advantages of storing data in the cloud include: fixed costs of space rental, expandability (scalability), saving to another location, speed of deployment, ease of administration, low energy consumption, reduction of electronic waste and decreased environmental impact. The disadvantages of storing data in the cloud include data security, control, openness, scalability, backup duration, making contract with the service provider, etc.

The implementation of storing data in the cloud requires lower initial costs and shorter time to achieve productivity compared to traditional ways of storing data. Storing data in the cloud will reduce costs, increase availability and reliability of computer resources and increase profits of maritime companies. The core concept of data storage in the cloud is that the customer pays only for the service of data storage in the cloud. Regardless of the advantages, users are still concerned about the security of their own data stored in the cloud. In the near future the security problems as well as ship-shore communication difficulties will be solved and the servers of cloud computing will enable the use of almost all IT services in the cloud [11].

Figure 6 presents economic impact, developments and application areas of RFID technology. Numerous factors make RFID technology the best choice for implementation in maritime company's business processes. RFID technology is suitable for monitoring and supervision of spare parts and technical equipment, access control, tracking, supervision and monitoring of container handling.

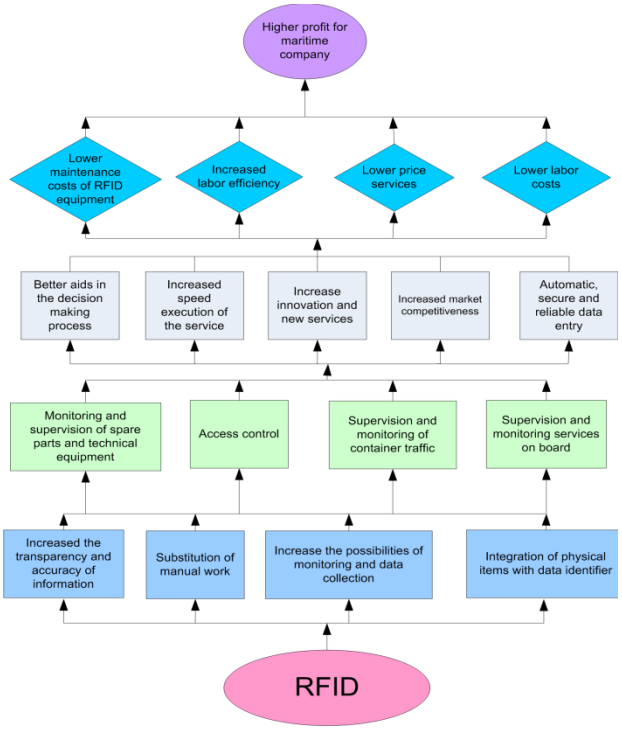


Figure 6. Economic impact and developments due to use of RFID

Monitoring and supervision of onboard services (such as cabin access control or restaurant services) could be also improved by implementing RFID technology. Individual and customized services (breakfast, Internet access, cocktails, etc.) can be offered via RFID prepaid solutions so there will be no need to use vouchers (additional discounts for frequent travelers are becoming available). All these factors contribute to reduction of human resources, paperwork, manual labor, accelerate the execution of business processes and contribute to increase of financial profits of maritime company.

IV. CONCLUSION

Although the market for RFID is already well established, because of the developments in the semiconductor industry, as well cost decrease, the mass usage of RFID will redesign existing processes in the global business world.

The advantage of using RFID technology in comparison to other technologies are: adaptability to working conditions (moisture, dirt, high temperature, mechanical resistance), long lifetime (standards prescribe a 10-year duration of RFID media with minimum 100,000 readings), protection of data on the card, all favorable price of RFID components, etc.

The aim of this paper was to start discussion about implementing RFID technology in Croatian ports by representing its capabilities to accelerate and improve ferry loading process and make it more efficient and reliable. RFID technology combined with advantages of cloud computing offers higher profit to maritime company. Additional promising features of RFID technology and its security aspects will be the subject of future work.

The use of cloud computing services (Database as a service -DBaaS) provides for cost savings in the costs of developing and maintaining an internal infrastructure, operating expenses such as those allocated for the maintenance of IT resources, administration costs and costs of scaling information infrastructure. In addition, it gives the shipping company a possibility of faster and greater development, a possibility of upgrading existing services or investments into new and thus further development of its IT sector.

REFERENCES

[1] A. Ustundag, "The Value of RFID," Springer-Verlag London, 2013, pp. 3–5.

[2] Croatian Bureau of Statistics, "Usage of Information and Communication Technologies (ICT) in enterprises," 2009, First Results, First Release. http://www.dzs.hr/Eng/Publication/2009/2-1-8_1e2009.htm (05.01.2015.)

[3] G. Matičević, M. Dabić, and T.U. Daim, "Using RFID for real-time production management in the digital enterprise", 2010, p.214.

[4] http://ec.europa.eu/eurostat/documents/3433488/5579940/KS-SF-11-065-EN.PDF/b359b480-9454-4790-b169-b304eb76dcfe?version=1.0 (05.01.2015.)

[5] C. Swedberg, "Fjord Line Puts RFID Aboard Ferry," RFID Jurnal, 2013, http://www.rfidjournal.com/articles/view?11087/

[6] P. Ristov, A. Mrvica, "Primjena RFID tehnologije u pomorstvu," Proceeding of 3th International Maritime Science Conference (ISSN 1847 1498), Split, 2011.

[7] M. H. Mickle, L. Mats, and P. J. Hawrylak, "Physics and Geometry of RFID," in RFID handbook: applications, technology, security, and privacy / edited by S. Ahson and M. Ilyas, 2008, pp. 3-17.

[8] E. Ilie-Zudor, Z. Kemény, P. Egri, L. Monostori, "The RFID technology and its current applications," in Proceedings of The Modern Information Technology in the Innovation Processes of the Industrial Enterprises-MITIP, 2006, pp. 29-36.

[9] https://www.velleman.eu/downloads/0/infosheets/datasheet_k8019_uk.pdf (05.01.2015.)

[10] K. Finkenzeller, "RFID handbook," 2010, pp. 2-6.

[11] P. Ristov, A. Mrvica and T. Mišković, "Secure data storage," Proc. of 37th International Convention, Digital Economy and Government, Local Government, Public Services, MIPRO 2014, Opatija, Croatia, May 26-30, 2014, pp. 1803-1808.

Interoperablity of bank statements: A case study

Rosica Chejkova-Nikolov*, Marjan Gusev**, Magdalena Kostoska**, Sasko Ristov**
*National Bank of R. of Macedonia, Skopje, Macedonia
** University Ss Cyril and Methodius, Faculty of Computer Science and Engineering, Skopje, Macedonia
rosica_ii@gmail.com; {marjan.gushev, magdalena.kostoska, sashko.ristov}@finki.ukim.mk

***Abstract* - There are different types of bank statements offered by various banks. So far there are no standards, and each bank builds its own type of statements. Most of these statements given to the citizens are in pdf format, and lately several banks offer statements in excel formats. Very rarely one can find XML types of statements, which can be further processed.**

To use data from statements in other applications, one has to build their own interpreter of bank statements. To solve the existing problem with interoperability, as a main motivation for this paper, we introduce an adapter, which can read various types of statements and return an XML file as output. Our realization of the adapter works with statements offered by most of the banks in our country as a proof of concept. The adapter was efficiently used in a special program for analysis of consumer costs, where one can access accounts in several banks and calculate taxes, income statements and analyze the costs.

I. INTRODUCTION

The development of IT technology and its fast update and improvement enable an increased use of software solutions in businesses. All business transactions require extensive communication between the company and the banks. These financial transactions are usually recorded on the company's account. Similarly, the citizens too use financial transactions with banks and relevant data is stored on their accounts.

Internet allowed the banks to offer almost all the information and services online 24/7, including desk operations to their customers. This step contributed to the rapid development of modules that enable integration of e-business or ERP software with the banking systems. The final goal is recording and tracking of all realized transactions.

This paper analyzes some of the more familiar systems that enable management of expenses and bank transactions in a company. These systems also allow the users (companies or citizens) a complete insight into their banking transactions. Users can analyze, compare and make conclusions about the realized costs or give them an opportunity to perform management of the costs, which would lead to more rational spending of the funds.

Various banks deliver different types of bank statements. Since there is no recognized standard, each bank builds its own type of statements. Most statements are pdf documents, spreadsheet (excel) or word documents. These documents do not give a structural organization that can be used for automatic processing in computer programs, such as the xml format, used only by a small number of banks. This problem is mostly interpreted as interoperability problem, when two different systems exchange data.

When the problem is exposed only to two entities, then the solution can be realized as a simple interpreter between them, which processes the data coming from one system prior to use of the other. However, when one wants to build a more universal solution in cases when there is no official standard, one has to use other approaches.

To solve the analyzed problem with interoperability of bank statements we introduce a software adapter, which can read various types of statements and give XML file as output. The proof of concept is realized as a case study on examples of several banks in Macedonia. The realized solution uses an adapter to access various accounts in different banks and analyze the consumer costs.

The rest of the paper uses the following structure. Related work is presented in Section II and the analysis of existing solutions in Section III. Section IV presents the new interoperability framework and finally, the conclusion is given in Section V.

II. RELATED WORK

This section presents similar projects and approaches that tackle the analyzed problem. Later on we discuss interoperability solutions.

A. Similar ongoing projects

During work on this paper, we have tried to find some similar projects that are currently developed or operate in this area.

The first analyzed project is about integration of credit card transactions into a financial management system used by a company to track and control budgets [1]. The system realizes an accounting service for credit card transactions within the financial management system. The project uses various card transactions limits, regularly not available to a credit card issuer, and ensures that the transactions correspond with the financial system controls. The transactions can be obligated prior, to or during the actual transaction with the bank and thereby subjected to the controls of the financial management system. Obligated transactions can be authorized for immediate payment. The invention provides for the complete reconciliation of the credit card transactions with bank records after the transactions occur using the obligation function to capture

the transaction before it occurs, even the transactions that are immediately paid. The system reconciles the transactions recorded by the bank with those recorded in the financial system and updates budget, plan, project, and ledger entries accordingly. The system also allows cardholders to identify disputes and track the correspondence with the card issuer over the dispute.

The second analyzed project is about a system and method that automatically tracks player gambling transactions in a casino [2]. The system includes a casino database that stores betting summary records for multiple players, where each betting summary record is associated with a player identification code, and includes the player's betting rating. One or more gaming tables have multiple player positions and multiple code readers. A code reader initiates a betting session in response to reading a player identification card encoded with a player identification code. A communications network couples the casino database and gaming table to an automatic tracking and management unit (ATMU). The ATMU has a display unit, and is coupled to the code readers at the gaming table to receive a player identification code for a player at the initiation of a betting session. The ATMU retrieves from the casino database the betting summary record of the player, and displays it to pit personnel at the gaming table on the display unit. The ATMU also collects real time data of the player's betting transactions, including the player's identification code, and an average bet by the player during the betting session. The ATMU updates the betting summary record with the collected real time data for the player, and provides the updated betting summary record to the casino database via the communications network.

The next project is about control system and method for interfacing a control device to a vending machine having input/output control lines along which are transmitted signals to control the operation of the vending machine [3]. The control device includes means for connecting to the control lines of the vending machine; means for reading the control signal; means for determining operational responses/outputs of the vending machine, which correspond to the control signals; means for storing the control signals and information concerning the operational responses; and means for reading the stored information to control the vending machine. Control and monitoring of vending transactions comprises means for reading stored information from a magnetic card; means for obtaining an approved code for a specified authorization amount for a transaction; means for storing the approval code; means for operating the vending machine without exceeding the authorization amount; and means for storing transaction information.

The last project that was analyzed, and aligns with our goal of interest, is about a system and process for determining the terms of the transaction and establishing a function having a proper function identifier [4]. The function is adapted to be processed during a transaction such as a point-of-sale transaction that may involve a credit card. The system and process involve identifying an account having an account identifier, establishing a function having a function identifier, and receiving a transaction request including the account identifier, a transaction amount, and the function identifier. The system also processes the transaction in accordance with the function to affect the transaction and, possibly, the transaction amount. Additionally, the system and process are adapted to allow message to be presented by customers and credit card holders during a transaction for later appearance on a credit card statement or other notice.

All these projects work on a specific problem, but no one analyzes bank transactions per user from different bank documents.

B. Interoperability

Communication in this modern age is something that we take for granted. Interoperability is the ability of a system or a product to work with other systems or products without special effort on the part of the customer. Interoperability describes the extent to which systems and devices can exchange data, and interpret that shared data. In order two systems to be interoperable, they have to be able to exchange data and subsequently present that data in a way to be understood by a user [5][6].

In IEEE [5] dictionary interoperability is defined as the ability of two or more systems or components to exchange information and to use the exchanged information.

O'Brien and Marakas [6] define interoperability as the ability to achieve end user application, using different types of computer systems, operating systems, and application software, interconnected with different types of local and wide spacious networks.

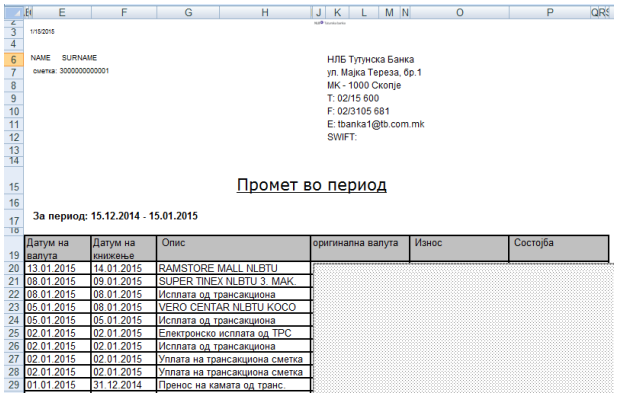
The purpose of interoperability is calculated data exchange between computer systems that can be owned by different organizations that cooperate with each other in order to meet specific business process / workflow.

Interoperability can be distinguished on four different levels: technical, syntactic, semantic, and organizational [7]. The technical interoperability enables technical connection between entities needed to exchange information [8]. Syntactic interoperability enables appropriate format for data exchange [9]. Semantic interoperability ensures that meaningful exchange of data, i.e. the exchanged data are correctly understood by the entities included in the process [8]. Finally, the organizational interoperability deals with exchange information between entities with different internal structures and processes [8].

Interoperability commonly can be achieved in three ways: establishing a standard in the given area, accepting often used method/structure as a de-facto standard or building adapting software between different interfaces.

III. ANALYSIS

We have realized a detailed analysis of the files received via the system of e-banking of several banks in Macedonia (NLB Tutunska Banka, Ohridska Banka, Komercijalna Banka and Stopanska Banka). Further, these files are processed and inserted into database, from where then user draws the necessary data reports. The analysis will be categorized by various data formats, including spreadsheet and word documents, and XML formats.



1/15/2015

NAME SURNAME
сметка: 3000000000001

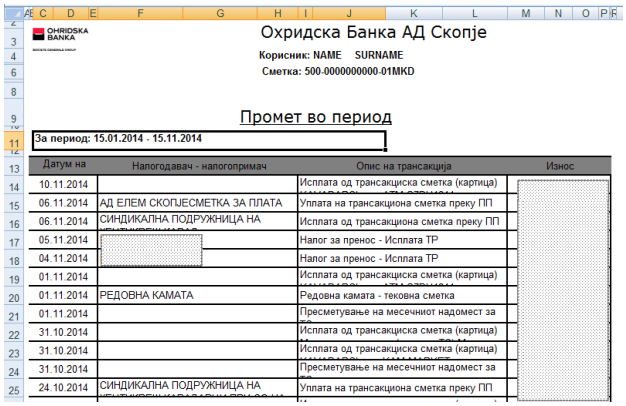
НЛБ Тутунска Банка
ул. Мајка Тереза, бр.1
МК - 1000 Скопје
T: 02/15 600
F: 02/3105 681
E: tbanka1@tb.com.mk
SWIFT:

Промет во период

За период: 15.12.2014 - 15.01.2015

Датум на валута	Датум на книжење	Опис	оригинална валута	Износ	Состојба
13.01.2015	14.01.2015	RAMSTORE MALL NLBTU			
08.01.2015	09.01.2015	SUPER TINEX NLBTU 3. MAK.			
08.01.2015	08.01.2015	Исплата од трансакциона			
05.01.2015	08.01.2015	VERO CENTAR NLBTU KOCO			
05.01.2015	05.01.2015	Исплата од трансакциона			
02.01.2015	02.01.2015	Електронско исплата од ТРС			
02.01.2015	02.01.2015	Исплата од трансакциона			
02.01.2015	02.01.2015	Уплата на трансакциона сметка			
02.01.2015	02.01.2015	Уплата на трансакциона сметка			
01.01.2015	31.12.2014	Пренос на камата од транс.			

Figure 1. Excel bank statement of "NLB Tutunska Banka"



OHRIDSKA BANKA

Охридска Банка АД Скопје

Корисник: NAME SURNAME
Сметка: 500-0000000000-01MKD

Промет во период

За период: 15.01.2014 - 15.11.2014

Датум на	Налогодавач - налогопримач	Опис на трансакција	Износ
10.11.2014		Исплата од трансакциска сметка (картица)	
06.11.2014	АД ЕЛЕМ СКОПЈЕСМЕТКА ЗА ПЛАТА	Уплата на трансакциона сметка преку ПП	
06.11.2014	СИНДИКАЛНА ПОДРУЖНИЦА НА	Исплата од трансакциона сметка преку ПП	
05.11.2014		Налог за пренос - Исплата ТР	
04.11.2014		Налог за пренос - Исплата ТР	
01.11.2014		Исплата од трансакциска сметка (картица)	
01.11.2014	РЕДОВНА КАМАТА	Редовна камата - тековна сметка	
01.11.2014		Пресметување на месечниот надомест за	
31.10.2014		Исплата од трансакциска сметка (картица)	
31.10.2014		Исплата од трансакциска сметка (картица)	
31.10.2014		Пресметување на месечниот надомест за	
24.10.2014	СИНДИКАЛНА ПОДРУЖНИЦА НА	Уплата на трансакциона сметка преку ПП	

Figure 2. Excel statement of "Ohridska Banka"

A. *Spreadsheet data format*

Most of the banks are using excel exports as a report with client transactions in a given period of time.

The first example is the bank statement provided by "NLB Tutunska Banka". Using the service for electronic banking, it is possible to make a report for a period of time. The user receives a document with sorted transactions, in html or excel format. An example of this excel document is presented in Figure 1. The following key data to be exploited by other applications for management of expenses and bank transaction:

- Field C6 contains information about user's name and surname
- Field C7 contains information about the user's transaction account
- All transactions are shown starting from the 20th row. Field B contains the transaction date, field G its description, and field I the transaction value.

The number of rows in the document is variable according to the given time range and the number of transactions that occurred in that period. The same layout of the fields is used for both the debit and credit cards.

Another example of use of Excel's reports as bank statement is the use of "Ohridska Banka", as presented in Figure 2. The used layout uses the following:

- Name and surname of the user account is recorded in the field H4
- Transaction account of the user is recorded in the field H6
- All transactions are shown starting from the 14th row. Field B contains the transaction date, field and its description, and field M the transaction value.

Similarly to the previous example, the number of rows is variable according to the given time range and the number of realized transactions.

One can conclude that although excel formatted documents give a relatively structured information each bank uses data in different positions. Structured information in this case means that the meaning of information is not described, so one needs to establish a mapping of the "meaning" and the value. For example C6 (or H4) identifies the user. This mapping of cells with their meaning is a file that has to be imported with each bank statements to enable information extracting from the corresponding statement.

Another problem is generated if the bank decides to change the layout, then all positions will be changed and the relevant information will be displaced.

B. *Swift related formats*

Some banks offer swift formatted bank statements using the MT940 format. An example is used by "Komercijalna Banka", which is shown in Figure 3. MT940 is Customer Statement Message international format used as digital equivalent of the official paper account statements. However, statement uses opening and closing balance on the book date, as well as specific information for all entries to the account. So the SWIFT MT940 structured format delivered by RCM is not fully compliant with SWIFT, i.e. the message starts with a specific value ':940:' and the allowed maximum length is more than the 2000 characters specified by SWIFT. The

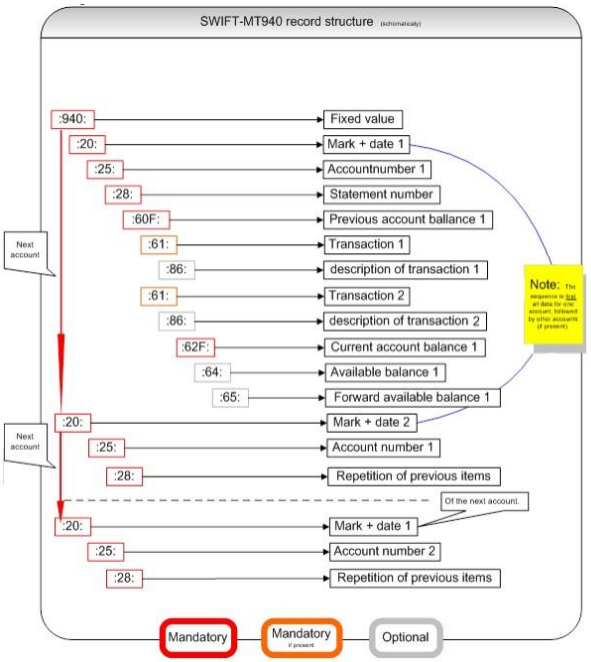


Figure 3. MT940 bank statement format of "Komercijalna Banka"

statements are empty if no entries are available for the specified entry date. The structured export format is adopted to provide additional information for SEPA orders. For example, Radobank will close any message with tag 62F, 64 or 65.

About format structure, each information block in an MT940 is preceded by a field (label, tag). The content of the field may consist of different elements. The file begins with a fixed value ":940:". Each statement in the file begins with field-20 and ends with the next field-20 or when the end of the file is reached. Within one statement the group consisting of field-61 and field-86 may occur several times. Field-65 may also occur more than once. Codes in field-86 are separated by a forward slash. These codes are used to provide structure in field-86 and to separate the various data elements [10].

This standard uses textual file with structured data and is better than the ordinary text based word documents. However, it does not offer flexible adaptation to new environments or processing of the files.

C. XML structured data

There are banks that use several document types for bank statements, including a XML format. An example is the use of "Stopanska Banka".

Figure 4 presents a typical example document. All information has explanation about its meaning and data values are described within the document using the following structure of XML objects:

- StatementRs is the main element in which are houses all XML objects from report
- StatementType, AccountInfo and ListOfTransactions are sub elements of the document, where in StatementType are given general information about the report, AccountInfo contains general information about the user's account and ListOfTransactions has data for all transactions made.
- In ListOfTransactions is described separately each transaction in element Transaction, and as the first element appears TranCount which containing information about how many transactions are described in the document.

The parser can easily analyze and further process the following XML data are:

- AccountInfo / AccountNumber - information about transaction account user
- AccountInfo / CustomerName - name and surname of the user
- ListOfTransactions / Transaction / ProcessDate - date when transaction happened
- ListOfTransactions / Transaction / Amount - value of the transaction
- ListOfTransactions / Transaction / OtherRefName – place where the transaction is made

```
<StatementRs>
    <StatementType>Retail
        <DateFrom>01.12.2014</DateFrom>
        <DateTo>31.12.2014</DateTo>
    </StatementType>
    <AccountInfo>
        <AccountNumber>1000000001</AccountNumber>
        <Currency>MKD</Currency>
        <CustomerID>2911</CustomerID>
        <CustomerName>IME i PREZIME</CustomerName>
        <CustomerAddress>ULICA 118-1</CustomerAddress>
    </AccountInfo>
    <ListOfTransactions>
        <TranCount>3</TranCount>
        <Transaction>
            <ProcessDate>2014-12-22T12:54:00</ProcessDate>
            <OrderId>-9655</OrderId>
            <ValueDate>2014-12-22T00:00:00</ValueDate>
            <Amount>-12.00</Amount>
            <Balance>1681.00</Balance>
            <RefTrans>FT14356580187257</RefTrans>
            <RefTime>1412221254</RefTime>
            <OtherRefName>СТОПАНСКА БАНКА АД СКОПЈЕ</OtherRefName>
            <OtherAccountNo>200000000099029</OtherAccountNo>
            <PurposeCode>666</PurposeCode>
            <Purpose>Провизија за трансакција: FT14356580187257</Purpose>
            <DebitReference></DebitReference>
            <CreditReference></CreditReference>
            <TransactionChannel>e-банк</TransactionChannel>
        </Transaction>
        <Transaction>
```

Figure 4. XML bank statement from "Stopanska Banka"

As a conclusion, XML bank statements offer the best suitable forms of documents that express the meaning of data and corresponding values. It allows easy further processing.

IV. INTEROPERABILITY FRAMEWORK

This section presents our solution of an adapter that can be used to communicate with different banks, presenting the software and hardware concepts.

A. Software concepts

The adapter analyzes reports collected by different banks and stores relevant data in the database using a pre-defined common format. Although various banks use different formats of bank statements, only common necessary data is extracted and inserted into the database. The output of the service is an XML file containing the pre-defined information. Our dedicated application form management of bank transactions uses these XML files to realize management of expenses and transactions.

The predefined common data extracted from all bank statements is using:

- Transaction Account
- Transaction date
- Transaction value
- Place / company where the transaction is made

These data should be read through for each transaction, respectively for any document submitted by a user. Furthermore, it should make connection between the user and the transaction. Accordingly, the display of table data is given at Figure 5. The special application for management of expenses and transactions is using an XML generated file and builds a database to store the information and generate various reports.

```
SELECT TOP 1000 [UserId]
      ,[TransactionId]
      ,[TDate]
      ,[Total]
      ,[MadeIn]
      ,[IsProcessed]
  FROM [MyMoney].[mm].[Transactions]
```

Results | Messages

	UserId	TransactionId	TDate	Total	MadeIn
121	8	3564400008800	2014-10-07 00:00:00.000	591	SKOPJE EVN MAKEDONIJA
122	8	3564400008800	2014-10-07 00:00:00.000	312	SKOPJE JP VODOVOD I KANALIZ
123	8	3564400008800	2014-10-07 00:00:00.000	131	SKOPJE ZEGIN LISICE
124	8	3564400008800	2014-10-05 00:00:00.000	348	KAVADARCI DIJANA 8
125	8	3564400008800	2014-10-04 00:00:00.000	2000	SKOPJE BDS IDRIZOVO P.E. 153
126	8	3564400008800	2014-10-04 00:00:00.000	900	KAVADARCI DIRA IMPEKS
127	8	3564400008800	2014-10-04 00:00:00.000	320	KAVADARCI ART DIZAJN
128	8	3564400008800	2014-10-03 00:00:00.000	170	SKOPJE SP MARKET 14
129	8	3564400008800	2014-10-02 00:00:00.000	850	STIP TRD ROBI DOOEL
130	8	3564400008800	2014-10-02 00:00:00.000	760	SKOPJE TINEKS DISKONT 33
131	8	3564400008800	2014-10-02 00:00:00.000	388	SKOPJE KAM MARKET 4
132	8	3564400008800	2014-10-02 00:00:00.000	226	SKOPJE TINEKS DISKONT 33

Figure 5. Display of the table in the database, in which are integrated necessary data from three different banks formats

B. *Architecture of solution*

The architecture of the interoperability process is presented in Figure 6. A user asks for an initial electronic bank statements transaction, the adapter transforms the statements into an XML recognizable file and send it to other applications (in our case to the application that manages the expenses). This application requires:

- Internet connectivity
- User's account for the electronic banking service
- User's account for the application that manages expenses and banking transactions

The adapter transforms all different types of bank statements into a pre-defined common XML statement that can be easily used by other software products.

Figure 6 presents the process of realized adapter and interoperability of bank statements. The adapter uses a configuration tool, where the user can configure the bank statement file specification and map relevant fields into pre-defined required data format.

V. DISCUSSION

A system that manages expenses and bank transactions faces a challenge to cope with various formats of bank statements. Although one can enter the relevant data by hand copying values from statements, we aim at automatic access to the bank statements. However, the main challenge is in recognition of various file formats.

Different approaches were used in similar systems, leading to several solutions of the problem.

In the first solution [1], it is a system that allows the integration of transactions made through credit cards in the system for financial management of a company, whose goal is to perform control over the budget of the company. The integration is done through the process of settlement of transactions. The bank captures every transaction for that company before it is processed. The recorded transactions and then sent to the company.

In the financial system of the company is done

A settlement between the recorded transactions from the bank and those who reported directly to the financial system of the company by employees.

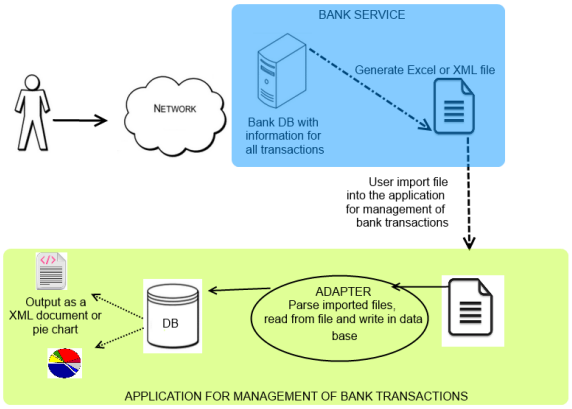


Figure 6. Interoperability process

Then based on the locked transactions, the system makes changes in the budget, plan, projects or books if such exist.

The second solution [2] is a system for monitoring and management of transactions in the process of using games in the casino. Namely, the system automatically monitors the transactions of gambling of the casino player. The system includes a casino database, which recorded the summary data for each of the many players, where each total field is associated with the identification number of the player, and includes the best rating of that player. One or more tables have multiple positions for the players and more readers' identification codes. Readers' codes initiate sessions betting in response to read the identification card of the player encoded with an identification code for that player. Communication network brings together a database table and casino gaming unit for automatic monitoring and management (ATMU). On this unit (ATMU) has a monitor and it displayed the related results. This unit (ATMU) collect data in real time of actual transactions, associated with casino database and makes settlement of transactions. Also with the data obtained from (ATMU) unit the player account is updated, according to transactions in real-time. In this way system monitoring the financial state of each player in real time and according to her system decides who has the right to play a particular game.

The third solution [3] is a system, which processes credit, debit and smart cards and related electronic transactions. The control system for the sale and the method used to connect the control unit for machine sales have input / output control lines that transmit control signals to control the operation of the machine for sale. Control devices include tools/ways to connect with the control lines of the machine for sale; means/ways for reading the control signals, tools/methods for determining an operative responses/outputs of the machine for sale, that corresponding to the control signals; means/ways for storage control signals and information for an operational responses; and means/ways for reading the stored information for control the machine for sale. The control and monitoring of transactions from the machine of sale include tools/ways for reading the magnetic data preserved by magnetic card; means / methods for keeping the approved code for specific authorized part of the transaction; means/ways for store the approved code;

means/ways for work with machines for sale without exceeding the authorized amount; and means/ways for store information about the transaction. Thus system manages transactions made between the bank and the corresponding cards issued by the bank.

In the fourth solution [4], there is a method for determining the terms of the transaction. The system and process that formed the function have a proper function identifier. This function is adapted to perform during the transaction, a point of sale transaction, which may include credit card. Systems and process include account identifier, the establishment of function that have function identifier and receipt of the request for a transaction involving account identifier, transaction amount and function identifier. Systems and process also processed transactions appropriate to the function that affects transactions, and perhaps the amount of the transaction. Also there is a consideration about the impact of the account. In addition, the system and the process are adapted to send messages to the user with the current state of the user card.

All analyzed solutions relating to projects with a specific problem, and all are doing some communication between two objects: a system of bank and system of company, the bank system with a device, functions, methods, etc. In none of these solutions there is no analysis of banking transactions per user from various bank documents.

XBRL (eXtensible Business Reporting Language) is a specially developed language to become a standard to exchange financial business information. It covers a much broader scope then the one we are analyzing in this paper. It is based on XML technologies to enable communication of business related information between business systems defined by appropriate metadata sets, taxonomies, and other relevant semantic meaning, concepts and relationships.

The scope of this paper is realization of a system that makes analysis of banking transactions per user from various bank documents. Our solution is based on building an adapter that is outside the application, and works as an independent service. It accepts different bank statement formats and outputs a pre-defined format that can be used for further processing by other applications, such as the one to manage expenses and transactions.

VI. CONCLUSION

The need for managing own finances is exposed a lot considering the dynamic lifestyle in the last decade. Banks in the digital society allow users to manage electronic transaction accounts, and control the finances. Users usually use several different bank accounts, which initiates the interoperability problem for a centralized solution and wish to control overall finances.

For this goal, we conducted research to analyze the existing solutions. We elaborated on the different approaches for the interoperability of bank statements. The existing solutions realize some analysis of the accounts, and perform connection with various devices, or produce confirmation reports of certain actions or data analysis. But none of these solutions makes data collection from different cards (debit or credit) of different banks, and provides summary reports based on the collected data.

Since our goal is to provide every user, regardless of that in which bank has his current account, to see his payments made in one place, grouped by certain criteria, we created a software solution that collects data from various formats from more banks, create a common format, then make different statements, according to which the user can see the analysis of his costs and implement a strategy for further management.

This paper analyzes the interoperability problem of different bank statement formats. The adapter solution we have developed, can understand various semi-structured formats, such as excel, MT-940 text or XML files, and also un-structured format files, such as textual files with pre-defined positions. The customization part enables the user to map the required fields onto specified formats. The Adapter is realized as a web service and can be used by other applications, such as those that manage expenses and bank transactions.

This application allows each user to analyze the transactions for a defined time period, and then to conclude a future spending strategy, which might set the proper and most efficient spending of analyzed finances.

REFERENCES

[1] R.Bissonette, C.R. Corrie, W.L.Wingate, "System integrating credit card transactions into a financial management system," US patent US6343279 B1, filling date 26.8.1998, www.google.com/patents/US6343279.

[2] John M. Strisower, "System for the tracking and management of transactions in a pit area of a gaming," US patent US5809482 A, filling date 01.9.1994, www.google.com/patents/US5809482.

[3] Brock Kolls, "Credit card, smart card and bank issued debit card operated system and method for processing electronic transactions US patent US6119934 A, filling date 30.12.1996, www.google.com/patents/US6119934.

[4] J.S. Walker, S.K. Jindal, D.E. Tedesco, "Method for determining the terms of a transaction," US patent US5945653 A, filling date 26.06.1997, url: www.google.com/patents/US5945653.

[5] IEEE web site, www.ieee.org, as seen on 4.2.2015

[6] J.A. O'Brien, G. Marakas, "Introduction to Information Systems(13th Edition)", Published 2005 by Mcgraw-Hill Higher Education, ISBN 0-07-304355-9

[7] Kubicek, H., Cimander, R., Scholl, H.J., "Organizational Interoperability in E-Government", Springer-Verlag, Berlin Heidelberg 2011, 2011.

[8] European Union, "European interoperability framework for pan-european e-government services", 2004.

[9] Kubicek, H., Cimander, R., "Three dimensions of organizational interoperability", European Journal of ePractice 6, 1-12, 2009.

[10] https://www.rabotransact.com/rtpubeu/ContentIntegrationPortlet/images/Format_description_SWIFT_MT940_Structured_tcm79-183712.pdfa

SOCIAL NETWORKS IN THE CASE OF TOURISM IN CROATIA

Ljiljana Zekanović-Korona
Department of Tourism and Communication Sciences
University of Zadar
Franje Tuđmana 24i
ljkorona@unizd.hr

Abstract - Social networks in terms of communication are bidirectional and provide the opportunity to participate in shaping the product. In e-business social networks have made progress that will forever change the way the companies (large and small) create their image and overall performance. Social networks is a rather broad term, but what defines it in the context of e-business is the ability to communicate within the company itself and with other users. A carefully planned campaign, adjusted to social networks, can bring manifold benefits to the company business. Social networks should be mastered as they will become the primary basis for creating (and closing) future gigantic corporations and smaller companies. This paper analyzes viral marketing and the marketing of social and mobile networks through a series of examples. The development of social and mobile networks is influenced by advanced information technologies. The analysis of the results is based on the online research of social networks, where the TRI index of their users was tested together with the extent of their usage in tourism business. Further data analysis included a variety of statistical methods aimed at analyzing time and manner of using social networks in e-business.

Keywords: social networks, digital economy, Internet, information technology, e-business, accommodation.

I. INTRODUCTION

In the beginning of the 21st century the Web 2.0 technology was developed. Web 2.0 technology concerns a wide spectre of new technologies, mostly composed of blogs, Wikipedia, podcasts, tagging, predictable markets and social networking. In 2004, Mark Zuckerberg launched a web page called Facebook, later proven as a brilliant idea, because he wanted to enable virtual communication and reconnection with friends previously separated due to life circumstances. Facebook is simple to use and technologically sophisticated. The fact that Facebook is free contributes to its huge popularity. The success of Facebook encouraged other business people to try to imitate it. Existing similar web sites like MySpace, Delicious, Linkedln and Flickr had been founded before Facebook, but later they have implemented some of Facebook's characteristics. Twitter, Bobo, Ozone and Scispace.net were founded after Facebook and in their offer of Internet services they have added some other technological benefits. With the growth of social networking since 2005 until today, a large part of Internet has rapidly adopted characteristics of social media. Internet is turning into a first class social media of the 21st century, which does not reduce its business role and value. On the contrary, from the business point of view the virtual space of Internet is becoming more interesting because of the growth of its plurality and diversity, while the interest of users for the things happening in it is getting increasingly larger. From the moment the members of network see information from other members that they find interesting, they start using them as business users. The users exchange information about the business subjects they are interested in, search for a job or look for a partner which will work with them on a mutual project.

"...The amount of information on the web is really exploding. They say that from the beginning of human history all the way up thru 2003 you took all the information that has ever been recorded, you take all the books that have been written, all the films that's been produced, that's about 40 egzabytes of information that was produced in all human history up to 2003. This year (**2011**) the World is producing about **800** egzabytes of information... and all that lines up on the web...", [1]

Social networks as a form of online communication connect acquaintances and strangers in the great maelstrom of information and content, which includes all social groups. If social networks are wisely used, they can be a place where brands are created or otherwise destroyed. In order to master social networks, it is important to understand their modes of operation, and choose the one that best fits one's projects and plans of action.

B2C is the company that markets the products for the female population and has chosen Pinterest as its homepage Social Network. However, if the target population prefers to receive information in a concise text format, then Twitter is the right choice. Facebook has been so far the most widespread network with a wide range of users and companies doing businesses all over the world. LinkedIn is a great place to connect with other professionals of the kind, especially appropriate for B2B Company.

Regardless of the type of company, the way we start using social networking in doing business is always the same – identification of the target by asking four questions:

1. Who are the target customers?
2. What social networks are used by clients?
3. How much time do customers spend on social networks?

4. How do customers use social networks?

II. THE MOST FAMOUS SOCIAL NETWORKS

Facebook ranks as the world's largest social service, but it is also a very popular platform for advertising. The initial idea of a Harvard group of students was to create a network that will facilitate communication within the community of people who already knew each other. The company has existed for seven years and brings together nearly half a billion of users.

"Twitter is a microblogging service. It may also publish a variety of messages and notifications, known as the statuses of up to 140 characters. The Following are the people you follow and see their tweets / status, and the Followers are the people you have chosen to follow you. It is obvious that Twitter differs from Facebook, among other things, in the fact that if someone is your "friend", it does not mean that you are his or hers. It is not surprising that the majority of Twitter users are from the United States and that the Americans are leading in the number of tweets, but interestingly right behind them are the Japanese. Why? For the simple reason that with 140 characters a lot more can be said in Japanese than in English or another language. In terms of the market size, there are over two million Internet users in Croatia, and 1.2 billion in the world. According to the latest data there are 1.116.92010 Facebook users in Croatia, or about 50% of the Internet users, accounting for a quarter of the population."[1] [2]

Facebook and Twitter are closely linked with each other without ever having intention to do so. After first four years, Twitter, following the model of Google, introduced promotional messages on its site. A corporate Twitter account is an ideal opportunity to gain real-time feedback from the market. When in February 2009, Dell announced that they would offer exclusive benefits and promotions throughTwitter, the number of users rose sharply from 8000 in February, to 500,000 in mid-March.[2] [2]

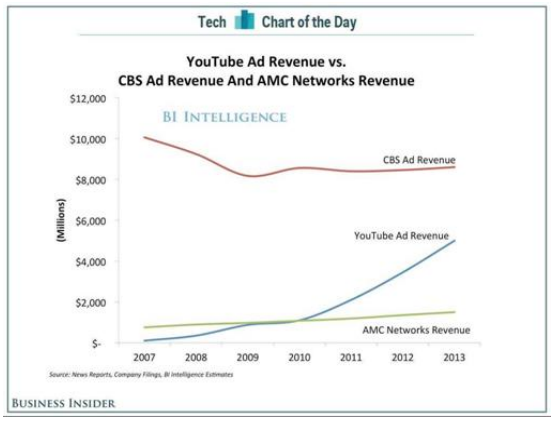
Pinterest is a relatively new social network compared to the already mentioned giants of the Internet marketing. It was launched in 2010, and promptly and significantly managed to change the structure of social networks. The emphasis is on the advertising photography and in such a way it initiated a new wave of advertising that focuses on textual minimalism. The principle of operation is a large advertising board, where users can pin the content they like and organize a themed collection.

The copies of Pinterest are appearing on daily basis. The most famous Chinese copies that generate turnover in millions for the companies being advertised on them are Mogujie and Meilishuo, and are directly associated with the giant of e-business: Taobao-m. According to the research of WhoIsHostingThis of 2014, Pinterest has more than 25, 000, 000 users and is one of the most popular social networks. 47% of online shoppers in the US have been influenced by the recommendations found on Pinterest. Each 'pin' generates approximately $ 0.78% on e-business pages - Pinterest generates four times more profits compared to Twitter does and 27% higher than Facebook.[3] [6]

One of the most popular social networks is YouTube, which daily publishes thousands of video clips. It was launched in 2005 by three former PayPal employees. Since its beginnings, the number of users, both individuals and multinational corporations doing campaigns on this platform, has been growing exponentially. "YouTube was acquired by Google in 2006 for $ 1.65 billion, which marked a new step in the evolution of the Internet. The first video 'uploaded' on YouTube in 2005 had only a few thousand views. Compare this with the most watched video of 2013 -PSY: Gangnam Style, which has had a staggering 1.3 billion views."[4] [3] The common feature of all social networks is non discrimination of users by the size of their income. YouTube is a platform that strives for constant development and integration with other social networks. In 2011 they confirmed the cooperation with the social network Google+ and moved into a new age of viral marketing.

Viral marketing is the attempt of creating an advertisement that will be orally spread online.

YouTube – ideal for viral marketing



Picture 1. YouTube and viral marketing

LinkedIn is a unique platform that has developed a special identity as a social network for professionals. But it has gone a long way since its beginnings (2003) as a forum that allowed employers to review one's resume and job seekers to see ads for vacancies.

„Engagement is rising fast on LinkedIn, as the social network becomes a content destination. The social network is also high-income and highly educated, and it has a big international presence. These factors will make LinkedIn increasing compelling to marketers. Already, a survey from Cogent Research finds that LinkedIn is the

[1] p.171.
[2] Ibid.
[3] http://whoishostingthis.com/, retrieved on 21 July 2014
[4] Zarella, D.; *The Social Marketing Book*; O'reilly, Canada, 2010, p. 7

preferred social network by a wide margin for building a brand identity. “[5]

Nowadays, LinkedIn has a pseudo-monopoly in the sector of connecting employees and employers, and in the future we can expect its further growth and gradual monopolizing. It is important to note that LinkedIn is not blocked in the People's Republic of China as Facebook, Twitter, Google and YouTube are. In addition to the network itself, also available are the tools which contribute to building better relationships among the contact members. Fivehundredplus enables the sorting of contacts by the frequency of communication. The Rapportive is the extension of Gmail and displays the details of the LinkedIn profiles of the persons you communicate with and allows you to connect with them.

"According to the Jobvite Social Recruitment survey roughly 80% of companies use social media to locate a new labour force; 95% of them use LinkedIn."[6] [4]

Picture 2. The best choices of social networks for B2B and B2C[7]

III. TRI - Technology Readiness Index

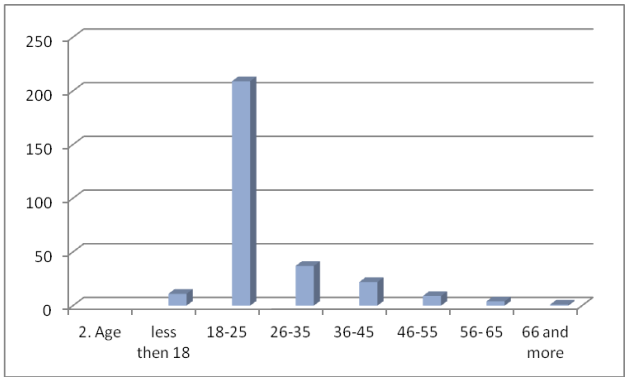
In the questionnaire there was a set of questions from which the TRI-index (*Technology Readiness Index*) was calculated for every respondent. Technology Readiness Index (TRI) is used to calculate the willingnes of respondents to adopt and use new technologies. This index is based on the scale of feelings that the respondents express towards new technologies. Components, i.e. feelings covered by the TRI-index are: optimism, innovativeness, discomfort and insecurity. The above feelings are measured by Likert scale as a series of statements in which the respondents can assess how much they agree or disagree with a given statement on a scale of 1 to 5. The obtained results can show us how much a particular individual or respondent is willing to accept new technologies. Readiness assessment includes the evaluation range of -20 to +20. If the above assessment scale is divided into five equal parts, it is possible to distinguish the following categories of respondents:

1) people reluctant to use new technologies (from -20 to -13)
2) paranoids (from -12 to -5)
3) skeptics (from -4 to +4)
4) pioneers (from +5 to +12)
5) researchers (from +13 to +20).

Therefore, if the level of TRI-index is higher, the tendency of respondents to the acceptance and use of new technologies is larger.[8] [5]

IV. The Questioning Method

The testing was conducted through an online questionnaire which was originally placed on the Facebook wall so that it could be accessible to the entire population of users. The online survey was opened during December 2014 and during that period a random sample of 293 respondents filled it out (245 females, i.e. 83,62% of the sample and 48 male respondents, i.e. 16,38% of the sample with the average age of 25 years). The age structure of respondents is showed in the picture 3. The online questionnaire consisted of 20 questions and all the questions in the survey were closed type with multiple choice answers. The limitation of this research is the fact that questionnaire was published on Facebook, and not on other social networks, as well as the fact that only the users of social networks were included.



Picture 3. Age structure of respondents

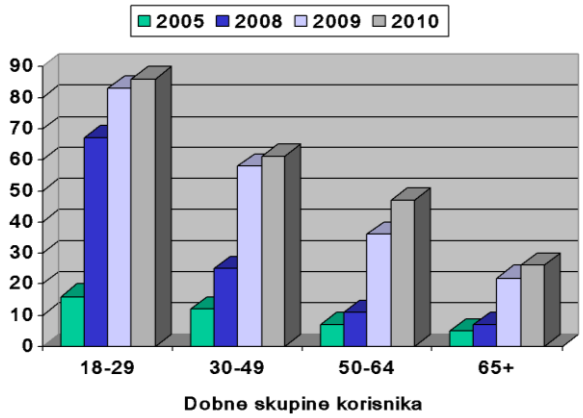
Picture 4. shows the age groups of Facebook users in the period between 2005 and 2010. Comparing these two pictures, it is obvious that our group of respondents, considering the size of sample, is correlated to the whole population of Facebook users.

[5] http://www.businessinsider.com/demographic-data-and-social-media-2014-2; retrieved on 13 August 2014

[6] Qualman, E.; *Socialnomics*, John Wiley & Sons, Canada, 2013, p. 470.

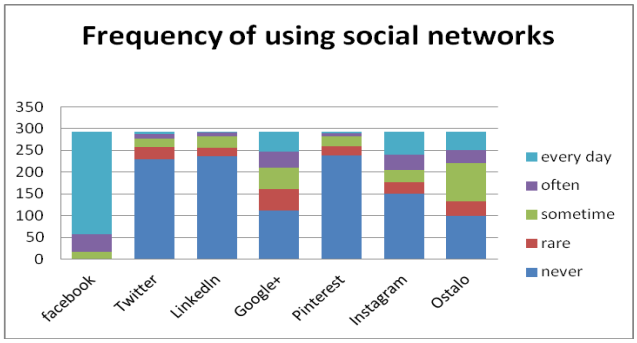
[7] ://marketingland.com/survey-facebook-best-for-b2c-sales-linkedin-for-b2b-20219, retrieved on 18 July 2014.

[8] Zekanović-Korona,Lj, Grzunov.J.: Evaluation of shared digital economy adoption: Case of Airbnb, MIPRO DE,Opatija 2014.



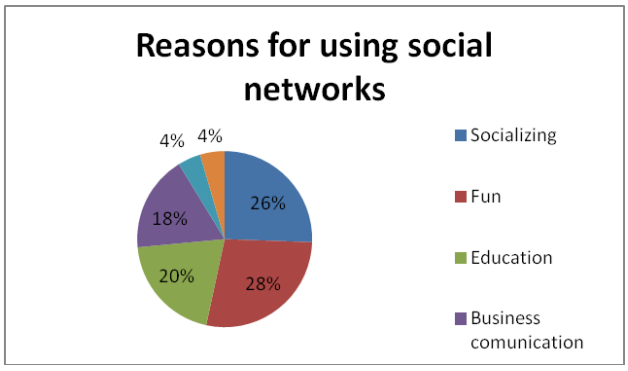
Picture 4. Age structure of Facebook users

The research results lead us to conclusion that the most popular social network is Facebook. In the tested sample there is not a single respondent that has never used Facebook, and 80% of the respondents use this social network on daily basis. The research results indicate that of all the networks the most common are Instagram and Google+, while the least preferable social networks are Pinterest and LinkedIn. (Picture 5.)



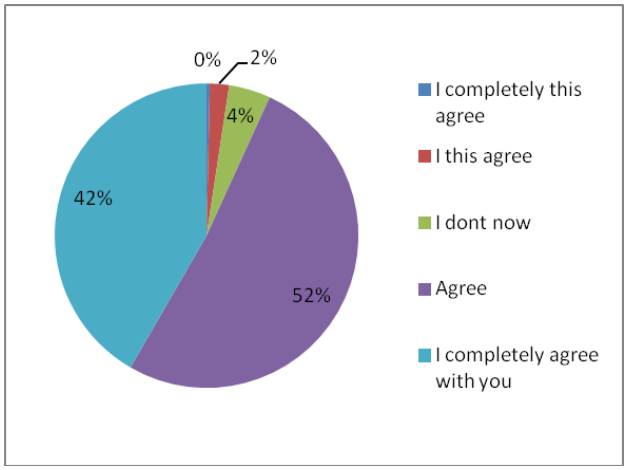
Picture 5. Frequency of using social networks

The Analysis of the survey results indicate that the respondents most often use social networks for fun and socializing, and they least use them without a clearly defined reason. As tourism covers travel, recreation and leisure, a logical conclusion would be that social networks are used largely for tourism. It can be said that tourism is one of the aspects of entertainment and socializing (Picture 6).



Picture 6. Reasons for using social networks

The survey was meant to determine the correlation between the use of social networks and tourism and it turns out that more than 50% of the respondents believe that they certainly facilitate access to information about people as well as tourist destinations. Only 2% of the respondents believe the opposite. (Picture 7)



Picture 7. Correlation between the use of social networks and tourism

The survey of the reasons, time spent on social networks and activities carried out on them was followed by measuring the satisfaction of the respondents with the quality of the content they consume on social networks. Although the largest proportion of the respondents (49%) are satisfied with the quality of the contents on social networks, a significant portion of them, slightly more than 40%, are only slightly satisfied.

Considering the age structure, we got the expected percentage of the respondents with respect to the TRI-index, i.e., sceptics 57,67%, paranoids 5,11 %, pioneers 33,79%, and researchers 3,4%.

The research results from this sample indicate that IT literate respondents use several different pages made for the purpose of accommodation renting and the review of the tourism offer. For example, the groups of Sceptics and Paranoids do not visit such sites, and the respondents from the group of Sceptics mostly use famous sites such as Booking.com and Airbnb for the purpose of accommodation renting while TripAdvisor, Booking.com, Futurama and Black egg are used for reviewing the tourism facilities. The respondents from the group of Pioneers, beside the above mentioned, also use some less known sites such as Couchsurfing, Hostelbook, Apartmanija, Galileo, Lonely Planet, Putovnica.net and the like. The least numerous group of Researchers, mostly use Booking.com in order to find accommodation and TripAdvisor for the reviews of hotels, restaurants, etc.

Today there are many social networking sites that offer a variety of activities to their users. One of the hypotheses of this study was to examine the association between the use of particular social networks and the basic activities performed on them. The questionnaire included a wider list of the most common and most popular social networks in the world today. The respondents in this study most commonly use the following social networks: Facebook, Google+ and Instagram. The testing of the hypothesis is based on three

most common social networks and the basic activities related to them.

The Chi-square test (χ 2) was 4.965 (p = 0.05, df = 6) and it found no association between the use of certain social networks and the basic activities that take place on them. Due to the multifunctionality of social networks it is very difficult to determine which network is suitable for the relevant activities (information sharing, publishing photos and videos, making friends, and communication and information about tourist destinations and travels).

With regard to the TRI indexes in this survey, the dominant groups are Sceptics and Pioneers. The testing hypotheses (χ 2) is the association between the level of computer literacy and the basic activities of respondents on social networks. The result is 9.267 (p = 0.05, df = 5), and we conclude that there is no mutual association, or regardless the level of computer literacy, the respondents use social networks for the same or similar reasons.

The level of computer literacy contributes greatly to respondents' views and attitudes towards social networks and the way they facilitate the availability of information about tourist destinations. It was also tested with (χ 2) test The result of the test was 9.49 (p = 0.05, df = 3), and we conclude that the higher level of computer literacy contributes to the wider use of social networks for getting information about tourist destinations. The higher degree of TRI means the openness of the respondents towards the acceptance of new technologies, and thus easier collecting of all kinds of information, including information on travel destinations.

The hypothesis on the association between the level of computer literacy and the time the respondents spend on social networks was also tested. The computer literacy variables and the time spent on social networks were used in calculating the chi-square. (Table 1)

Table 1: Time spent on social networks

	Less 2 ours	3-4 ours	more then 4 ours	∑
Sceptics	65	72	32	169
Pioneers	42	40	17	99
∑	107	112	49	268

The resulting value of the chi-square is 0.424 (p = 0.05, df = 2) and we can accept the null hypothesis that there is no association between the level of computer literacy and the amount of time spent on social networks. The lack of association between the level of computer literacy and the time spent on social networks can be interpreted by the fact that the time spent on social networks is more dependent on other variables such as free time and work status.

For most of the companies that conduct research in order to improve their business, an important part of the research is the satisfaction of users with their services. The study tested the hypothesis of association between the level of computer literacy and the satisfaction with the quality of the content the respondents consume on social networks. The resulting value of the chi-square is 9.49 (p = 0.05, df = 4), and we can conclude that the ICT educated people are more satisfied with the quality of the content they use on social networks. It is probably because they use particular social networks for the purpose they are meant for (eg. LinkedIn for business communication) and they know how to take advantage of all their possbilites while those with less computer literacy are more likely to use only a limited range of many possibilities offered by social networks.

IV. Conclusion

Social networks are increasingly gaining popularity among different age groups of users. A number of existing prejudices about the use of social networks were tested in this paper through different statistical methods.

The increased use of social networks for tourism purposes is proved by the fact that almost 50% of respondents answered that the information about tourist destinations and travel is one of the main activities they perfom on social networks.

Social networks made information about tourist destinations much more available to the largest part of the respondents. Considering a large number of social network users that is growing every day, social networks have beome an everyday source of first hand information about the whole offer of a tourist destination. All the users of social network will inform themeselves about accommodation and restaurants before the journey. The research results confirm that more ICT literate respondents frequently use online tourism offer. There is association between IT literacy of respondents and the use of technology for the puropose of getting information about tourist destinations. Tourism and e-business are strongly correlated and therefore we need ICT literate users. Population should be constantly educated so that information and communication technology can be efficiently used in tourism and other other related fields. In Web 2.0 environment and in business, various companies have their own on-line client communities that are showing tendencies toward social networking, which is understood as the improvement of e-business. For the purpose of tourism travelling, and especially for choosing a destination, it is important to know the experiences and impressions of the users that have already visited the same destination. Virtual client communities are groups of people in mutual interaction by means of ICT. The users of virtual communities share their concerns, problems, interests and passions toward particular issues. Numerous companies want to achieve active cooperation with customers and to collect data from them through the use of e-business and social networks. Computer literate users of social networks are more satisfied with the content offered by social networks because they know how to use all available opportunities they provide. The respondents mostly use Facebook, which was the expected result given the fact that the questionnaire was placed on it. The research should be repeated on a bigger sample and among older respondents, and the

questionnaire should be placed on other social networks in order to obtain a better dispersion of results. Almost 50 % of the respondents are satisfied with the quality of content on social networks (49%), 40% of the respondents want changes and improvements This research shows that there is still room for innovations in business communication, and they can be introduced in the domain of social networks through ICT.

LITERATURE

[1]].http://www.youtube.com/watch?v=pK02FzSKotc(Woodside,D., Google)

[2] Marija Stanojević, Marketing na društvenim mrežama, Medianali, Vol. 5 (2011), No. 10

[3] Zarella, D.; The Social Marketing Book; O'reilly, Canada, 2010.

[4] Qualman, E.; *Socialnomics*, John Wiley & Sons, Canada, 2013,

[5] Zekanović-Korona,Lj, Grzunov.J.: Evaluation of shared digital economy adoption: Case of Airbnb, MIPRO DE,Opatija 2014.

[6] http://whoishostingthis.com

[7] http://www.businessinsider.com/demographic-data-and-social-media-2014-2

[8] ://marketingland.com/survey-facebook-best-for-b2c-sales-linkedin-for-b2b-20219,

Financial Agency FINA and Online Auctions in Croatian Amended Distraint Act

Anton-Jan Klasinc
Vertex projekt d.o.o.
Institute for Public Administration, Zagreb, Croatia
e-mail adresa: ajklasinc@gmail.com; vertexprojektmail.com

Abstract – This paper presents an analysis of the newly amended Distraint Act in Croatia, concerning foreclosure and sales of properties. The idea behind the act is that making sales of foreclosed properties more transparent by the means of online auctions can create gains for both the foreclosed and the state. The banks may also profit from better repayment of defaulted loans. The unamended act provided that the properties are put on sale by the court in execution proceedings, and those sales were conducted by dealers that saw their interest in lowering the prices. This apparently meant that properties were sold below the market prices and that loans were still not payed off when the sale of property was concluded, which also presented a useful opportunity for some buyers of properties to make profits, as well as for dealers themselves, while the state did not collect as much as it could have and the process of selling was slow. With online sales of properties the process becomes more transparent and it can reach larger number of potential buyers, which could increase the price and profits for both state and the former owners of those properties that are in debt. However, a similar act in the Netherlands provided that the properties should be sold through private online auction resellers, while in Croatia this is given to the government financial agency FINA. The problem is analysed from the perspective of principal agent theory, and from the perspective of supply and demand.

I. Introduction

Recent changes to the Distraint Act have been criticised by some for allowing too much discretionary power to FINA – executive Financial Agency in Croatia, and tranferring the issue from Commercial courts. A similar problem has already occured with the issue of pre-bankruptcy proceedings, where the main power has been transferred from commercial courts to the executive (Ministry of Finance). We analyse and propose a model for new Distraint Act based on principal agent theory, demonstrating that private companies have lost a significant amount of bargaining power arising from assymetric information allowing them to obtain significant gains in under-the-value sales of property. The ammendments were introduced in order to obtain better values for sold properties, or as close to the market value as possible, and on the other hand for the sake of preserving debtors' from being deprived of their property for a scarce value of the original debt, leving them still very much indebted after the sale of the property with which they guaranteed for the debt.

II. Theoretical Background and methodology

A common problem with real estate sales is that if buyer and sellers have some special relationship, they may agree on a lower price than the real price of property should be if the buyer and seller were strangers. This is generally known as principal-agent problem, where conflicts of interest and moral hazard may arise. The solution for this is to employ arm's length agencies that are supposedly indifferent and can not be influenced in transactions, acting in their own best interest as well as in the interest of their principal. In reality, situations which present moral hazard are quite common, especially in financial industry and real estate. In addition to this, they are also very difficult to control and prevent, and things are even more complex when the government is involved as principal or one of the parties.

We would like to begin by describing a simple model of an oligopolistic market with firms acting as principals trying to distribute their products through distributors acting as agents. This could be compared to the government acting as principal, trying to sell foreclosed properties through commercial courts which are then able to offer the property on auction s to private resellers as agents that offer them to final buyers on the market. The commercial courts are independent from the executive government in their operations and may offer the foreclosed properties on auctions. So far, the volume of sales was too low and the prices achieved also much lower than market prices, as the low provided for two auctions of 2/3 and 1/3 of market value, respectively. One reason could also be that the budget for commercial courts is fixed, with no possibility of extra bonus or motivation for judges conducting sales, or rewards for better performance, which could also lead to moral hazard on their part (e.g. bribes by resellers and/or debtors).

In the ammended act, the courts are obliged to transfer all requests for sales to FINA financial agency which then conducts online auctions. The argument for using online auctions instead of conducting sales directly through commercial court is that online auctions make sales available to a larger number of potential buyers, thus enabling larger competition and higher prices of foreclosed properties (near or equal to prices on the market), which in turn benefits the both creditors and debtors, as the latter are left with less debt to repay. It would also mean that the government investment into an

executive agency conducting the sales would provide for a faster and more effective sale process, which would result in larger volume of sales and consequently more income would be generated. However, this would hurt the private sector resellers that form an oligopoly and profit from the original Distraint Act. They would now have to appear as „ordinary buyers“ on online auctions with information readily available to everyone, so they would not be as motivated, as they would not be able to make profits from the difference between auction prices and market prices, and the government as a whole may also incur losses due to excessive and unnecessary investments that are not controlled by the market forces, as it would now have to conduct operations through its executive agency that may prove to be more expensive.

Monopolistic government firms replacing free market are not uncommon in Croatia, as on local level a recent study found a number of government owned firms (usually registered as ltd. - d.o.o.) doing business for local governments that could have been better left to the free market competition, except for, perhaps, community services (waste disposal, water, etc.).

From the above also follows that, the information about auctions being completely transparent and online, the absence of asymmetric information would allow for better volume of online auctions and fairer prices. This, of course, would increase the profits for all sides except for a few private real estate agencies.

On the other hand, it can be shown that in oligopolistic industry that consists of identical firms, a subset of firms may find it optimal to commit to face asymmetric information about their agents' operations. Therefore some firms may choose to incur informational agency costs, even though information is avilable at no cost. The commitment to face asymmetric information is also a commitment on the part of the firm not to extract the entire agent's surplus and so agents have incentive to make a specific investment that increases firms' expected profits. The level of this investment increases with the proportion of firms that are not informed.1

Uninformed firms always produce lower quantities than informed firms and the expected quantity of output in the industry is lower when there are uninformed firms. Informed firms' agents do not make any specific investment whereas uninformed firns' agents make a strictly positive investment. Since informed firms can extract all the agent's surplus, the latter does not have any incentive to make the investment. By deciding not to acquire information firms commit themsleves not to extract their agents' surplus and thus give a credible incentive to them to make specific investments.1

Therefore, in the case of uninformed commercial courts and oligopolistic resellers, the volume of properties sold at auctions would be lower and the output of debt repayment would be lower than if the information about auctions was readily available to everyone.

The level of investment in a government controlled or owned agency, such as FINA, is determined by its budget and not by market laws. In general, the level of specific investment in such agencies may be shown to rise in general, while not increasing efficiency and effectiveness at the same time, which presents a specific type of moral hazard where investment may be extracted through corrupt means or contracts with preferred suppliers (e.g. of IT equipment and services) as well as hiring preferred staff (cronies, party members).

The problem of online auctions and sales through private agents is essentialy a principal-agent model, which is also well known in the context of a monopolistic upstream market. Acquiring information has the benefit of eliminating asymmetry of information between each firm and its agent (reseller). 8However, it has the shortcoming of eliminating incentives to make specific investments on the part of the agent.

In our case, „upstream firm“ or the principal is the executive government and the debtor (as „producer“), and the „downstream firm“ or the agent is either (i) the judicial branch of the government with its commercial courts selling almost certainly below market prices to a certain limited number of firms, allowing them to profit from the difference between auction and market prices; or (ii) the monopolistic government agency FINA selling to virtually all potential buyers via online auctions; or (iii) an oligopolistic market of resellers which may be entirely owned by government (FINA) in which case there is no information assymetry, or private resellers that benefit from information asymmetry and need to make specific investments in order to facilitate sales, acquire greater volume of sales and speed up the process.

It has been shown that in considering a fixed cost of acquiring information, circumstances may arise under which one firm decides to acquire information, and the other does not acquire the information even when both firms are equally efficient. 1Therefore, even if the financial agency FINA has made information available to every potential buyer for a fixed cost, it is still possible that it would not be more efficient than several private oligopolistic resellers operating under information asymmetry and same fixed cost.

This, in turn, would mean that in the case of government monopoly on online auctions the cost to the government could rise due to lack of market control of specific investments that it is obliged to make, and the volume of sales may not increase as fast as if the operation were left to private sector that could be taxed and would still have incentive to make specific investments on their own, in order to make profits. The taxation and contributions from the private sector, however, could be greater than the amount that government agency FINA could earn as a sole operator.

In fact, the level of investment in FINA could in the end be equal to the surplus that would have otherwise been extracted from the agents (private resellers) in case of commercial courts as prinicpals, but as in this case the principal is informed about the results and targets, but not about the costs and FINA is allowed to project its own out-of-budget costs, the largest portion of increase in volume of sales would go to the repayment of debtors' debt to the banks and covering FINA's operational costs

and specific investments into equipment and human resources, that may exceed by far necessary levels.

As FINA is a public agency, it could be argued that with proper management and controlling it would not be motivated to make unnecessary investments and that it would operate at optimum level, allowing the maximum repayment of debt with government costs well covered. 2 However, actual examples of Croatian ministries and so called arm's length agencies show that they overspend if they are legally allowed to do so, even on very expensive and unefficient premises, vehicles, and other products and services, as well as on employees' benefits and unnecessary hiring of staff.

III. New Croatian Distraint Act, online sales and comparison of with Netherlands

The new amended articles of Croatian Distraint Act stipulate that:

- Agency conducts the sale of property. The request for sale is submitted by the court. The court is obliged to submit to the Agency the foreclosure act, deeds of the estate and the sale decision.
- Agency has the right to request advance covering its costs related to the sale of the property. The distrainor is obliged to pay advance to teh Agency within eight days from the day of receipt of the payment request. If the distrainor does not advance the costs, the Agency shall not conduct the sale of the property, and shall notify the court instead. In that case, the court shall terminate the foreclosure.
- Sale of the property is conducted by online auction (OA). Online public auction begins with the publication of the call for participation in the OA.
- »Online auctions shall be carried out even if there is only one bidder.«. (Article 48 amending Article 100)
- On the first OA, the property shall not be sold under 4/5 of the assessed value. (Article 49 amending Art. 102)
- On the second OA, the property shall not be sold under 3/5 of the assessed value
- Bid collecting shall last 10 workdays. If on the fist OA no valid bit is collected, the second OA shall start with the publication of call for participation, first day after the end of the first OA
- If even on the second OA no valid bid is collected, Agency shall notify the court about it. In that case, the court shall terminate the foreclosure.6
-

The main difference between Croatin procedure and a similar procedure in the Netherlands is that there is no centralised government agency that performs online auctions at the request of the court, but instead there is a market of reselling private agencies organising online auctions.

IV. Available Statistical data on distraints and foreclosures in Croatia

According to the Government Statistics Bureau, the average price of an apartment in Zagreb was 1460 euros for square meter, but in case of foreclosed properties, the price was as low as 775 euros, for an apartment of 90 m2 that was estimated at 2280 euros per square meter, or only 1/3 of the real price. The evidence of properties (*očevidnik nekretnina*) of Croatian Chamber of Commerce shows 140 foreclosed properties in August and 6658 in whole 2014., and 3038 of those were apartments (1170) and houses (1868), most of which are located in the city of Zagreb (601), Zagreb county (600) and Istria county (1455). Among them is an apartment estimated at 1950 euros per m2 that can be purchased for only 1300 euros.45

Table 1

Foreclosed properties – Year/Number	
2008	2106
2009	2579
2010	3664
2011	4266
2012	4784
2013	6733
2014	6658
Total:	**30790**

The old act allowed for properties to be sold at 1/3 of the real price, but new rules provide that price can not be lower than ½ of the real price, which is not true for apartments sold in bankruptcy cases.

Type of property	Foreclosed in 2014
House	1868
Apartment	1170
Building	332
Arable land	1034
Meadow	208
Business premises	213
Forest	213
Pasture	366
Apartmental and business space	42
Hotel	6
Total	**6658**

The judge on the first auction lowers the price to 2/3 and if it remains unsold, to ½ of the real estimated price in the second auction. In case of bankruptcies, on the first auction the real price is offered, which lowered by 10% on each subsequent auiction, and ½ of the price is than reached in 5 auctions – however, there is no lower limit for the price in this case.3

V. The model of the problem

The problem which should be solved requires the level prices achieved by the sales to be as close to the

market value (optimum), while repaying the maximum amount of the outstanding debt and achieving the best possible value for firms competing for the property, while also trying not to increase the cost to the state to much. This also means eliminating as much as possible the profits of those firms, in terms of difference between market value and the value achieved by the sales, but also limiting the cost of the Financial Agency FINA. As this is essentially a social measure requiring state intervention, it differs from similar models in other (market oriented) countries, where firms compete openly on private online markets, whereas in Croatia, the government FINA – Financial Agency is a mnopolist. In the case of private firms, they may acting as resellers, they may be motivated (and generally, are) to invest into human resources and equipment, while in this case it is the government that has to invest into these things – namely in its agency, FINA, in order to achieve the proposed results. The cost to the government including loss of employment in private sector and taxes to private sector should not exceed or should be at least equal to the increased repayment of debt to the lending institutions and guarantors, additional employment in lending institutions and the gains in terms of reduction of social welfare to debtors that are left without property (foreclosed).

In this case, it is a principal – agent problem in which the efficiency of the monopolistic state agency FINA should be greater or equal to the efficiency of several oligopolistic private resellers (some of which may be owned by the government).

The demand function relates price and quantity. It tells how many units of a good will be purchased at different prices. In general, at higher prices, less will be purchased. Thus, the graphical representation of the demand function (often referred to as the demand curve) has a negative slope. The market demand function is calculated by adding up all of the individual consumers' demand functions.

Assume n debtors that are trying to sell the same type of property (e.g. a house) through government auctions and face a linear demand function given by

$$P = a - Q$$

Where P is the clearing market price, q_h is the output of debtor h and $Q = \Sigma^{n}_{h=0} q_h$ is the output of properties. We assume that debtors compete in quantities and we normalize the production costs of properties to zero. It is also assumed that a is large enough for the equilibrium quantities to be always positive.

If the total number of debtors q_h is high, Q will also be high, and the clearing market price will be lower than it could be achieved on the open real estate market. So, a is actually a function of the number of prospective buyers, N and likewise n_h for each respective property.. With a larger N, both P and Q may rise, or just P, which than gets closer to the proposed clearing price.

$$P = xN - Q$$

Solving for x on a sum of all properties with its prospective buyers yields:

$$\Sigma (p_h + q_h)/ n_h = x \quad ; \qquad \Sigma n_h = N$$

Where x obviously depends on the amount of available money that each buyer n in n_h is willing to invest in a property h. With the rise of n_h, x should become greater, as more buyers will compete and have more incentive to invest money in order to buy property and not just to resell it, until they reach the threshold at which they have virtually no utility. P will rise simultaneously with the volume V of sales of Q (V = QxP), and after that point, P will start to drop again, although the number of sold properties Q may remain at the same level. This can be determined on the basis of statistical data over a period of time, where $P = \Sigma p_h$ and $Q = \Sigma q_h$. As the number of buyers is lower in the case of an oligopoly of private resllers than in the case of ordinary buyers, registered on online auctions, it is certain that the price P in that case would rise, even if the number of sold properties rises as well, and the volume of sales shall grow as P x Q. If the average number of available buyers N is increased by online auctions from 10 to 100, or even 1000 for some attractive properties, x would drop, but the product xN would still be higher than the original number, thereby increasing the price P closer to the value of market value of the property (above 1/2 of the market value, i.e. 4/5 or 3/5 of the market value instead of 2/3 and 1/3 respectively for first and second auction).

The average debt of all debtors q_h is the amount of D = Q x a/2, where P = a/2 is the equilibrium price which gives the real market value of foreclosed properties. The actual amount of debt repayment that may be achieved on sales is lower. By data analysis, a in Croatia is rougly equal to 5000, and a/2 (or equilibrium price) is 2500.

As we know, the sales are usually governed by power law or Paretto principle that says that 20% of properties make 80% profits to private resellers and 80% of properties make 20% profits. In case of auctions organised by courts, the prices are set at 2/3 in in the first and 1/3 of the market price in the second auction. If market price is 2500 euros per m2, that means that 2/3 is 1667 and 1/3 is 833 euros, respectively. In case that we have 4000 properties on sale, 800 properties make 80% of the profits for resellers, and 3200 make 20% profits. In fact, it is almost certain that those 800 making the most profits are large apartments (e.g. 100 m2) located in city centers or large houses. The other 3200 may be smaller apartments, or smaller houses in villages. Supposing that we have only 2 types of property, an apartment of 100 m2 and an apartment of 12 m2, the first should be sold in 800 sales and make 80% of profits, and the second in 3200 sales and make 20% profits. It is clear that the first would be sold in the second auction (for 1/3 market price, or 833 euros), and the second in the first auction (for 2/3 market price, or 1667 euros). In the first case, unpayed debt would be 133,365 million euros, where each property was deemed to be worth at least the amount of credit at market price of the property, or 250000 euros, as only

83300 euros would be payed for each such property. In the second case, each property worth 30000 euros of credit would be repayed with 20000 euros, which would leave only 3,2 million euros of outstanding debt, but the profit for resellers would also be considerably smaller, as they would not be able to make a large difference between actual price and market price, for a larger number of much smaller properties (12 m2). If the government took 10% on each deal, it would earn 10 million euros in the first and 6,4 million in the second. However, if sold at ½ market price, the sales for private resellers would be only 100 million euros, but the state would cover 10 million instead of 6,664 million euros, which is 3,337 million euros more, whereas in the second case private resellers would take 48 million euros, or 16 million more, but the state would take only 4,8 million euros, which is 1,6 million euros less than original 6,4 million euros. In total, the state would gain 1,737 million euros, or about 14,02275 million kuna, that it could use for financing the operations and specific investments of financial agency FINA, instead of employing private resellers, that would lose a total of about 17 million euros. If 1/3 was their profit, 1/3 used for investment in equipment and 1/3 in employment (human resources), and 1/3 is 5,7 million euros, than 20% of the first and ½ of the third would return to the government as profit tax and contributions for employment to the state, and income taxes, which means that the government would cover at least 1,14 million euros and 2,8 million euros or a total of 3,94 million euros from the private sector if online auctions were left to the private sector. On the other hand, if FINA was the sole monopolistic operator on the market, it would make only about 1,74 million euros for the government, or, in fact, government would lose about 2,2 million euros, or about 16,5 million kunas, which is an amount for which it could set up a whole new agency! However, if we looked at the budegt of the agency, we may notice that the government has taken this in the account and that it has already invested in both amounts, or about 30 mililion kunas, instead of 14 million kunas that it can actually cover from the operation. Actual budget for the actual FINA agency confirms this hypothesis.

In case the first auction is set at 4/5 and second at 3/5, the figures are even worse, and show greater loss for the private sector, and greater gain for the government and the banks, in terms of taxes and debt repayment. In this case, government would gain about 5,6 million euros, but it would lose about 13 million euros contributions from the private sector in taxes and contributions, which would amount to a loss of 7,4 mil euro, or about 55,5 mil kuna.

At the beginning of the game, executive government on behalf of the debtor employs a commercial court and its judge acting as an independent agent for the provision of reselling services on the open market, or a monopolistic government agency FINA, or a number of private resellers acting as agents, supposing that debtors may choose their preferred agent. It is assumed that except in the case of a monopolistic reseller there is a competitive market for identical agents and the equilibrium contract in this market is such that the expected utility of agents equals their reservation utility U, which is normalized to zero. At this moment in the game, the executive government cannot write a complete contract specifying the remuneration of the agent and the level of output (volume of sales) that must be delivered.

Since agents anticipate that informed government representing debtors will expropriate all benefits of their investment, they do not have any incentive to invest. In case the government is not informed, the situation is rather different: the executive government representing debtors must pay a rent to their managers when the good state is realized in order to maintain incentive compatibility.

The commitment to face asymmetry of information is also a commitment of the part of the government not to extract the entire agent's surplus and so agents have incentive to make a specific investment that increases governments' (and debtors') expected profits. The level of this investment increases with the proportion of cases in which government is not informed (such as particular debtors in forclosure cases on commercial courts).

In case of a monopolistic government agency, however, it is still a question what the optimum initial investment on the part of executive government should be and how to calculate it on the basis of available information

If the increase in sales volume is Q, the government may be able to collect a certain money per sale for its services. The total sum of this should not be lower than the initial investment that was made into FINA in order to perform online auctions of foreclosed properties, and all the proper registrars, and to to this instead of commercial courts, that are already prepared for this job better than executive government.

If the monopolistic government agency is used, the government also has to write a contract specifying the remuneration of the agent with the agent although it is not really informed, and it cannot specify the level of output that must be delivered. It just assumes that the level of output must be greater if it makes the investment in first place, but it doesn't actually know by how much. On the other hand, it cannot simply write off the previous investment in commercial courts, as the courts still have to exist, so it could lose all its investment if it doesn't meet the target in which whole loss would be covered by the extra output reuturns. In addition to this, it cannot have returns in terms of taxes, because Croatia doesn't have a property tax, so the government has deprived itself of one important way of covering back what it has invested, and can only use guarantees for bidders. The only real return it can actually have for its investment is to reduce the burden of debtors and creditors, hoping that this will somehow materialise in lower social costs to the government and higher level of contributions and taxes. However, if the initial investment is not well suited (if it is too high) it will be difficult for government to motivate the agent to achive higher output targets. If the investment is too low, the agent may entirely miss the target, and the operation will be a failure. Therefore,

without the market mechanisms working, it is almost impossible for the government to estimate what the optimum level of investment should be – and that is probably the reason why other countries would never opt for a monopolistic government agency to do the job.

In this model it can be shown that FINA is not necessarily more efficient in terms of cost to the government, although wtih the changed rules it may be more efficient in terms of increasing debt repayment to the banks because of better value achieved for sold property. Therefore, it may be concluded that the model is generally profitable for financial institutions (banks), somewhat less so for the debtors, and detrimental to the private sector and contributions of private sector to the government. In order to improve the efficiency of FINA and cover these contributions, it would be necessary to allow competition between FINA and private online resellers that would have to meet the same standards of transparency in online sales, as the government monopoly in reselling is not the optimum solution for this market in terms of cost of human resources and equipment.

VI. Policy recommendations

It would therefore be advisable to reduce the scope of operations of FINA financial agency and perhaps to stimulate private sector to make specific investments in online auctions, as it is done in Netherlands and other countries, although it could be necessary to leave a certain scope of operations to FINA in order to determinme whether the market forces are sufficient to handle the demand and supply and to intervene on the market as necessary. However, squeezing private market out with a bloated government agency seems to be a less efficient solution.

The fact that Croatia doesn't have a property tax makes it difficult to assess the exact amount of profit made by private agencies. Therefore, it would be necessary to introduce a property tax of at least 1,5 % to 3% in order to achieve the goal of government „breaking even“ or perhaps achieveing a small level of profitability with respect to the situation where no property tax exists and private resellers are not allowed to hold online auctions.

As the investments in human resources (employment) are among most important investments that provide a lot of resources to the state budget in case of private employment, it is necessary for the state to stimulate private employment and investment of private entrepreneurs in human resources by providing adequate measures and subsidies, which are more suitable than further employment in the public sector.

VII. Conclusion

In this paper we have proposed an analysis of a recently amended Croatian Distraint Act within the framework of a demand and supply model for online auctions in which in the first case we have a monopolistic agency acting as provider of auctions, in the second case an oligopolistic market of resellers without auctions and in the third case a market of private online auction agents together with the government agency. It is cleary the third case which is the most efficient and effective, and this has been shown by our modelling. The model in which commercial courts act as principals conducting the sales of property through an oligopolistic market of bidders - resellers and final buyers is vulnerable to conflict of interest and moral hazard, which can reduce the price of the property and lead to „power law“ distribution of profits, reducing debt repayment and government finances, whereas the model with a monopolistic arm's length agency is economically not viable, as it is practically impossible to determine the proper initial level of investment in order to obtain the desired output (results). The market element increasing the efficiency of FINA and more favourable to state budget could be introduced by allowing for private online resellers in auctions, that would serve as a benchmark for the government agency, which would in turn prevent moral hazard of employing more staff or spending on unnecessary equipment, while the government agency on its part could reduce the moral hazard of oligopolistic private reselling firms interested in lowering the price of properties on auctions too much in order to make more profits. As to possible uncstitutionality of the Act, it should be noted that it does serve some better than the others, such as banks and debtors, but not citizens in general or the private sector.

References

[1] Fatima Barros, Assymetric information as a commitment in oligopoly, European Economic Review, Volume 41, No2, February 1997.pp207-225.

[2] Koprić, Musa, Đulabić: „Agencije u Hrvatskoj“, Institute for Public Administration, 2013

[3] Distraint Act http://narodne-novine.nn.hr/clanci/sluzbeni/2014_07_93_1877.html

[4] https://www.hgk.hr/odjel-za-poslovne-informacije/ocevidnik-nekretnina-i-pokretnina-koje-se-prodaju-u-ovrsnom-postupku

[5] http://ovrha.hgk.hr/ocevidnik-web/

[6] http://liderpress.hr/arhiva/85734/

[7] Accountability and principal agent models http://www.law.berkeley.edu/files/csls/Gailmard_-_Accountability_and_Principal-Agent_Models%281%29.pdf

[8] Trust and incentives in principal-agent negotiations http://jtp.sagepub.com/content/14/2/231.short

[9] https://www.nyu.edu/gsas/dept/politics/faculty/downs/conflict.pdf

[10] http://link.springer.com/article/10.1007%2FBF00240424

[11] http://www.academia.edu/2481040/Principal-_Agent_Theory_Is_an_effective_Tool_or_only_a_good_Hypothesis Principal Agent Theory is an effective tool or only a good hypothesis

[12] The political evolution of principal agent models. http://faculty.washington.edu/jwilker/571/571readings/Miller.pdf

[13] Patrick Keil: Principal Agent Theory and its Application to Analyze Outsourcing of Software Development, Technische Universitat Munchen, Institut for informatik

[14] Sanford T. Grossman and Oliver D. Hart, An Analysis of Principal-Agent Problem, Econometrica, Vol 51, No 1 (jan 19983), pp. 7-45

[15] Stephen A. Ross: The Economic Theory of Agency: The Principal's Problem, 1973

[16] K.J.Arrow, Essays in the Theory of Risk-Bearing, Chicago 1970

[17] H. Raiffa, Decision Analysis: Introductory Lectures on Chices Under Certainty, Reading, Mass., 1968

[18] Nico Groenenduk, A principal-agent model of corruption, in Crime, Law & Social Change, 27, 207-229, 1997, Kluwer Academic Publishers

Scalable Architecture of e-Ordering System in Cloud

Filip Dimitrievski, Sasko Ristov, Marjan Gusev, and Goce Armenski
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering,
Skopje, Macedonia
Email: dimitrievski@outlook.com, {sashko.ristov, marjan.gushev, goce.armenski}@finki.ukim.mk

***Abstract*—Todays dynamic economy enforces the companies to dismiss their minor-business departments and to turn the efforts towards the main business. This is emphasized especially to small and start-up companies, which would like to order the issues as a service from third parties as OPEX (Operational Expenditure), rather than investing in their own departments as CAPEX (Capital Expenditure). Even more, they expect to achieve their orders as faster as possible. Since the number of such companies is floating, as well as the number of their employees, these requirements demand a scalable system for e-ordering. Such a system will allow the consumer companies to order some services from the sellers. In this paper we present scalable architecture in the cloud, which can be offered as a service both to the consumers (buyers) and clients (sellers), along with two business models according to the market in a particular region or state.**

***Keywords*-Business, cloud computing, e-Ordering, SaaS, services.**

I. Introduction

Until about a decade ago, a lot of enterprises hosted their software on servers in their own data centers. Enterprises conducted detailed analysis and developed a strategy on how to design and maintain the datacenter. The building of such a complex data center was usually conducted in several phases in a period of several months to several years. The dynamic life and ICT growth could not have been predicted easily and thus lead towards huge losses or even bankruptcies. Also, the board of directors spent a lot of effort on the non-business ICT issues, rather than on their main goal.

Small and medium enterprises usually did not have such problems since they could not afford to build and manage their own data center, which forced them to use third party hosting [1]. This approach offers them to be more flexible in choosing the hosting provider, but still other issues and challenges arise: vendor lock in, third party dependence, security and privacy issues, lack of elasticity, etc. Recently, the new paradigm cloud computing has emerged. The cloud removes the need for the user to install any software on premises [2]. Running software remotely can be more cost effective [3], because of the reduced man hours and days for maintenance and administration, power consumption, etc. In order to provide these cost savings, the cloud providers charge their customers on a subscription basis. As a result, companies can spread their budget more efficiently and focus on growing the business [4].

Cloud computing paradigm matured over time, attracting enterprises that are seeking to make the most of todays technology at lower costs. That is, enterprises are not obliged to invest in any software or hardware. They only pay for what they use. In general, the cloud services are divided into three layers: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [5]. Each layer uses the services of the layers below. The IaaS layer is the lowest layer in the cloud architecture, while the SaaS layer is the highest. Going from IaaS to SaaS, the customers have less work and responsibility [6].

A typical deployment of a SaaS application does not require any hardware, because the vendors provide the entire cloud infrastructure. The SaaS model is designed to continuously and securely deliver applications anywhere and anytime. That is, the vendors need to have reserve capacity in order to handle any spikes in usage, outages or network mishaps. The cloud pricing model offers a company dynamic and elastic way to rent the resources where a SaaS application can be hosted. However, the SaaS application must have a scalable architecture, which can utilize the cloud elastic resources efficiently. Even more, the utilization will be more emphasized if the client requests are dynamic, but in some way predictive.

In this paper we present an architecture of an e-Ordering system, that is, a system that allows consumer companies to order some services from the sellers. We focus on a system for weekly orders catering. That is, the case where the orders are tightened in one or two days, while other periods the orders are minimized.

The rest of the paper is organized as follows. Section II elaborates on the motivation why the solution for e-Orders is selected and why a scalable architecture is proposed. Section III presents the details of the proposed scalable system concept for the e-Ordering system. In Section IV, the details of the system architecture and design are presented, along with detailed explanation of the systems modules and their agents. Section V presents the business models of the e-Ordering system. The system architecture challenges and benefits are discussed in Section VI. The last section, Section VII, concludes the paper and states future work.

II. Motivation

In this section we elaborate on the motivation why we analyze the system for e-Ordering. Also, some challenges are

analyzed about the performance and cost of such a system.

A. Main challenges

There are a lot of companies in the world that care about their employees, but only a few provide free food and other types of paid services to their workers. Every year, the Fortune magazine publishes a list of best companies to work for [7]. Last year, Fortune has once again ranked Google as number one. Google has always had a reputation for being the best employer and having a great office culture. As an employer who provides valuable benefits for its employees, Google has a profound influence on many small and start-up companies. The downside to this is that those companies dont have an approximate budget and resources as the tech giant. For example, they cant afford to have their own cafeteria and in order to satisfy their workers they order food from the local restaurants. Thus, the e-Ordering system comes to light, modernizing the ordering process of services and delivering them in one place. Because of this, companies can provide free food and other types of paid services to their workers in order to have higher employee satisfaction ratings, i.e. increased productivity and morale, healthier workforce, stronger and positive connections inside the company, etc.

The e-Ordering system must provide authorization and authentication for the users who are responsible for the administration of the profiles of both consumers (buyers) and clients (sellers), and those users who use the system to order some services from the sellers, i.e. the consumers employees. Their access rights are managed by the companys main user, who is also in charge of controlling and limiting the services they use. Those users are the most prevalent users in the system and also the ones who are responsible for slowing down the system.

The system performance depends on the total number of these active users and the most important factor that impacts the performance is their changing number. Thus, the system must be reliable and available during the e-Ordering periods in order to satisfy the increased number of orders that are tending to decrease the overall system performance. Security and data privacy are also important because the system stores sensitive user data, such as personal information of clients and consumers, and as well as their personal list of orders. All these requirements are essential to implementing a scalable and elastic system for e-Orders.

B. How to achieve a scalable system

SaaS vendors usually rent their virtual machines (VMs) on hourly basis [8]. The cost for a single VM is calculated by the hours the rented machine has been activated. On the other hand, the revenue that a given SaaS application receives from renting those machines, in our case that is the e-Ordering system, depends on the number of users and transactions made.

With exception of some sellers, most of the e-Ordering services are available for ordering just before the weekends start. That is, the system has increase in the number of orders every Thursday and Friday, as presented in Fig. 1. Since

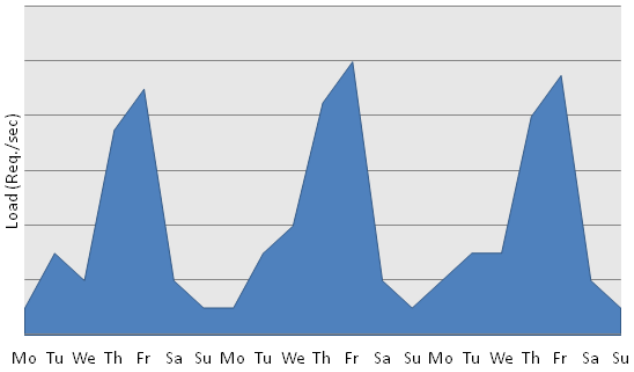


Fig. 1. System load in days per week.

both renting the VMs and consumers ordering most of the e-Ordering services from the sellers are happening at the exact time of the week, a business model needs to be created in order for the e-Ordering company to maximize the profit and grow its business. This will save the companys CAPEX and OPEX, and also resources needed for administration and management [9]. The details of the proposed business models are discussed in Section V.

As mentioned above, when the system receives the increased number of orders every Thursday and Friday, it should scale the resources in the cloud, either horizontally (utilize smaller VMs), or vertically (use greater VMs) [10]. A number of challenges of how to become a profitable provider by using the resources are discussed by Schwanengel and Hohenstein [11]. A few papers reported that the price for scaling the resources in a VM is linear [12], [13], i.e. using one VM with two cores is two times more expensive than one with one core, and so on.

In this paper we propose a scalable architecture of e-Ordering system in cloud which uses the cloud scalability and elasticity in order to achieve sustainable performance. The proposed solution will minimize the overall cost since it uses minimum resources utilized during the e-Ordering periods, thus better performance is expected to be achieved. The e-Ordering system requires huge amount of unpredictable resources only in small periods of time, while in most of the time it requires predictably much less resources. Using server-cluster architecture to create scalable and highly available solutions will only partially solve a performance peak problem since it is hard to be managed and administered, thus the proposed cloud hosted solution. However, although the peaks are unpredictable, the SaaS company still can predict the required resources in the peak period. That is, the SaaS company knows the total number of registered companies and the total number of their employees.

C. Service oriented architecture

A flexible and standardized architecture is needed to respond to the increasing demand for applications that support sharing resources and data. Service Oriented Architecture

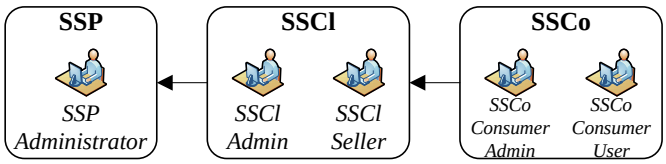


Fig. 2. Company roles interaction.

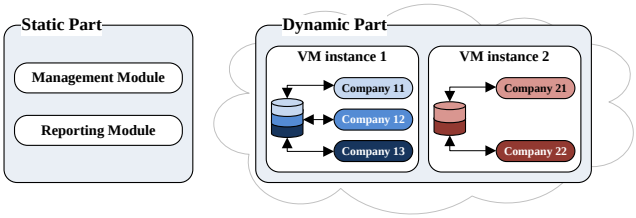


Fig. 3. Static and Dynamic parts of the system.

(SOA) is a flexible architecture that unifies business processes by structuring applications into services. SOA is being advocated as the next evolutionary step in software architecture [14], making a paradigm shift in architecture, i.e. from object to service oriented architecture.

SOA is based on the concept of a service [15]. Each SOA service is designed to perform one or more activities by implementing one or more operations. These services can be combined by other applications to provide the complete functionality of a larger application. Therefore, one service can rely on another to achieve its goals. These services communicate with each other by passing data or coordinating an activity between one or more services, allowing users to combine and reuse them in applications. Thus, the main benefit of SOA is the ability to mix together services from different systems.

Scalability feature of SOA is responsible for the wide adoption of this architecture. SOA enables interoperable communication among various systems, but has issues regarding the overall system cost and as well as resource utilization and provisioning. Therefore, on-demand infinite elastic resources are necessary to address these issues. Cloud providers are able to offer these resources using virtualization.

III. The Scalable System Concept

This section presents the details of the proposed scalable system architecture for e-Ordering system, along with various system roles and their interaction with the system modules.

A. System concept

The e-Ordering system is designed with three company roles: SaaS Service Provider (SSP), SaaS Service Client (SSCl) and SaaS Service Consumer (SSCo). The interaction among company roles, along with their users is presented in Fig. 2.

SSP offers the e-Ordering system as a service for SSCls to publish their services in a customized web site (service). SSCo interacts with one or several SSCls. That is, employees of SSCo can order services from different SSCls. The system roles of each company are elaborated in the next section.

B. System roles

SSP is represented with one user role: *SSP Administrator*, while SSCl is represented with two user roles: *SSCl Admin* and *SSCl Seller*. SSCo has also two user roles: *SSCo Consumer Admin* and *SSCo Consumer User*.

The *SSCo Consumer Admin* in the e-Ordering system is responsible for managing a given consumer company that uses the system to provide services from SSCls to its employees (*SSCo Consumer Users*). In order to provide these services, the SSCo needs to add or import its users into the system. *SSCo Consumer Admins* configure which SSCls are available to *SSCo Consumer Users*. Also, they can generate order reports and manage SSCl service rules, e.g. number of allowed orders per user, number of allowed items per order, periodical price limit for orders (daily, weekly, monthly), etc. On the other hand, the *SSCo Consumer Users* can only make orders from services per SSCl, predefined by their Consumer Admins. That is, they cannot order services that are not registered by their company administrators.

The *SSCl Admins* are responsible for managing the SSCl users - *SSCl Sellers*, which define the service menu items and their prices.

The *SSP Administrator* registers SSCl and SSCo companies and approves their administrators.

C. The scalable concept

Our main idea is to identify two subsystems of e-Ordering system, i.e. static and dynamic, as shown in Fig. 3. The first one will be active during a normal or a reduced system usage, while the other will be active at higher peaks of usage [16], [17].

The static part includes two modules: management and reporting modules, which do not require huge resources and can be predicted easily. This part can be hosted either on static physical server or VM with constant resources which can be predicted easily. The performance of this part depends directly on the number of SSCls only.

The dynamic part should be hosted in the cloud. The idea is to group, consolidate and host clients (those companies that offer) in separate VM instances according to the number of their consumers and their behavior. For example, one VM instance will host three companies and another VM instance two other companies, such that the total number of users per VM is balanced. The dynamic part depends on the total number of SSCo consumers, which is much greater than the number of SSCl clients. Additionally, the performance of this part is unpredictable because not only executes SELECT-queries,

but also performs a large number of INSERT, DELETE and UPDATE operations.

D. Architecture scalability

One of the main problems in this SOA architecture is lack of scalability, e.g. the administrator needs to upgrade the server infrastructure in order to satisfy the growing demand for orders. Thus, the administrator can either dedicate enough resources to host the system in order to handle the orders of all active employees in one moment or predict the average number of employees who are making orders in a specific moment and dedicate a required number of resources to host the system. The first solution is very expensive and the overall system will be underutilized most of the time. Also, the number of active employees varies from time to time, which introduces additional scaling problems. The second solution is less expensive, but it doesnt solve the problem, i.e. the system will be under or over utilized most of the time.

As the cloud computing paradigm emerged, the e-Ordering system will be hosted in cloud in exchange for better performance, optimal resource utilization and most importantly overall cost reduction. Because the cloud uses technologies that implement scalability, elasticity and interoperability between different systems, we propose a SOA based architecture that can be hosted on the cloud using all the benefits the cloud can offer in order to provide a working scalable solution as well as sustainable performance with minimal costs for hosting resources.

IV. System Architecture and Design

This section presents the details of the e-Ordering system architecture and design. Also, it explains the three modules of the system, as well as the agents with in these modules.

The e-Ordering system consists of three subsystems or modules, i.e. Management Module (MM), Reporting Module (RM) and Service Ordering Module (SOM), as presented in Fig. 4.

The latter is the dynamic part of the system, while other two modules belong to the static part. Each module has at least one agent. The e-Ordering system has total of four agents: Admin Agent, Infrastructure Agent, Reporting Agent and Consumer Agent. Each module and its agents are described in detail in the following subsections.

A. Management Module

This module is the core of the e-Ordering system, because it manages all resource provisioning in the cloud. The Management module can be hosted either on a bare metal server or VM instance in the cloud with constant resources. This module manages the system during inactive mode, providing user authentication and authorization, service menu items and categories, etc. Fig. 5 presents detailed design of this module.

Two agents are part of this module: Admin Agent and Infrastructure Agent.

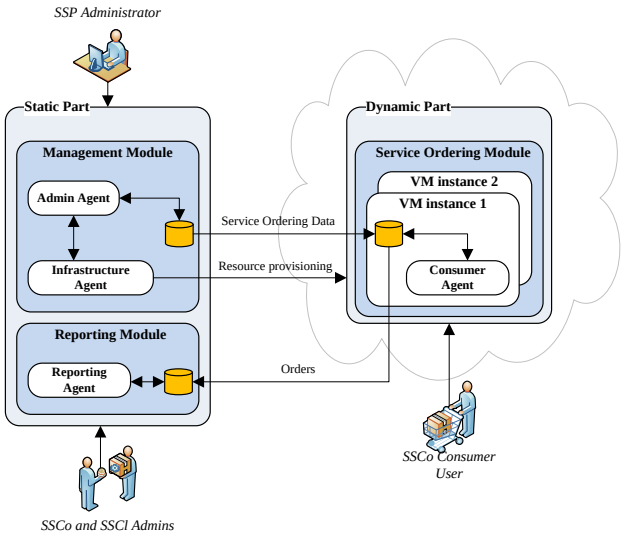


Fig. 4. System modules and roles interaction.

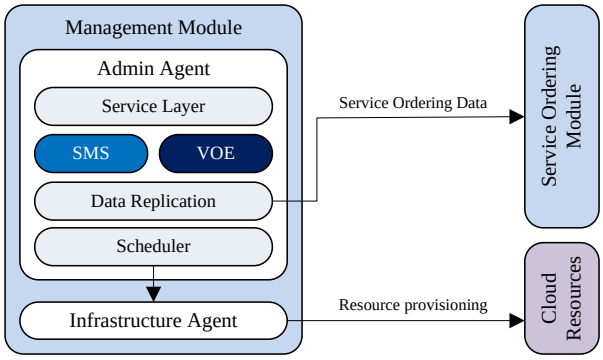


Fig. 5. Management Module design.

1) Admin Agent : The Admin Agent is responsible for authentication and authorization services of Service Layer, e-Ordering content administration, scheduling and data replication to the Consumer Agent in SOM. This agent is service oriented and contains Service Management System (SMS) and Virtual Ordering Environment (VOE). The former is used for SSCl management (menu and pricelists), while the latter is used for SSCo management (allowed SSCls for a particular SSCo).

The Scheduler communicates with the Infrastructure Agent for creating the SOM module. Through the VOE environment, SSCo Consumer Users can browse their personalized e-Ordering service menu. Based on the number of employees and the size of the allowed services from the menu, the Scheduler communicates with the Infrastructure Agent in order to supply the needed information for resource provisioning.

VOE as a subsystem communicates with the Data Replication subsystem in order to define the parts of data needed to be replicated to the SOM module.

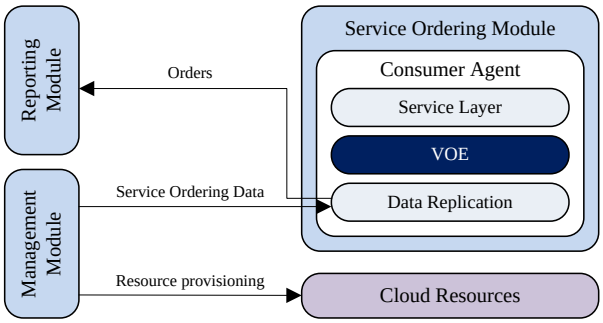


Fig. 6. Service Ordering Module design.

2) Infrastructure Agent: As we mentioned earlier, the Infrastructure Agent communicates with the Scheduler from the Admin Agent in order to create VMs in Thursday morning and shutdown instances in Friday evening, after the order data is replicated in the Reporting Module. Another important function is to choose an optimal algorithm how to organize SSCos in a single VM. This algorithm is out of scope of this paper.

B. Reporting Module

This module, similar to the Management Module, is the static part of the system. The main idea behind this module is dividing the purchased services from the e-Ordering system into a separate module for reporting and accounting. In Friday, after all orders have been purchased by SSCo Consumer Users, the order data is replicated for further management (accounting, reporting, business intelligence, etc.). Data privacy and security are very important for this part, because each order consists of a sensitive data.

This is a view-only module, i.e. only SELECT-queries are executed since both consumers and sellers view the executed orders. This module can be designed as OLAP part in order to speed up the report generation, where applicable.

C. Service Ordering Module

The Service Ordering Module is active only when the system is working in active mode (Thursday-Friday). Because its load varies depending on the number of orders and active users, we propose this module to have a dynamic resource allocation and therefore to be hosted in cloud in order for the system to be more efficient and cost effective. As a subsystem, SOM hosts only a small amount of data, i.e. SSCo Consumer Users can only order services from a small group of allowed SSCls. This makes the system more efficient and provides a better overall performance. Fig. 6 presents the detailed design of this module.

This module is specifically designed for the SSCo Consumer Users to order some services from the sellers. Because of its uniqueness, we can safely place the Consumer Agent in this module. The Consumer Agent receives the necessary data from the Admin Agent. It is a service oriented agent and contains only the VOE environment. VOE of Consumer Agent has significantly smaller database than VOE in Admin Agent. Each query in this module works on smaller data set improving the overall performance of the e-Ordering system, therefore making it more efficient. After all orders are committed, that is, after Thursday and Friday, the Consumer Agent copies the data of the orders in the Reporting Module as it can be viewed by the appropriate roles only.

V. Business model

In this section we propose two payment models for the e-Ordering system. That is, payment model per transaction (PPT) and payment model per consumer (PPC).

A. Payment model per transaction (PPT)

There are a number of SaaS applications where per consumer pricing makes sense, but it should never be the prime cause for choosing a payment model. For most applications it doesnt really matter how many users are using the application, but how many transactions they make. In order to grow the business we need to decide how to charge our clients. For example, if our clients are offering and selling food portions, then our payment metric will be - per food order transaction. As the consumers employees buy more food menus, our business (as SSP) will grow and expand.

We choose this payment model to be based on transactions, because the e-Ordering system does not only offer food but a various number of services. For example, when SSCo consumer orders some service from SSCl seller, the related seller is obliged to give a prearranged percentage to the SSP for every transaction that is made via the system. SSP helps SSCl clients to connect and sell their services to a countless number of SSCo consumers around the globe. Unlike SSCls, SSCo consumers in PPT are excused from paying a monthly fee to SSP for using the system.

B. Payment model per consumer (PPC)

Like most SaaS applications, this payment model is based on the number of active users a consumer company has. Unlike PPT model whose goal is to charge the SSCl clients for every transaction made via the e-Ordering system, the PPC as model intends to charge the SSCo consumers on a subscription basis, i.e. a monthly fee based on the number of their active employees. Thus, attracting more clients to come and use the e-Ordering platform because of its free SSCl access.

Depending on the market conditions and the business climate, the e-Ordering company can implement either one of these two models. Choosing an appropriate payment model for a particular region or state will help the e-Ordering company to maximize the profit and grow its business.

VI. Discussion

The e-Ordering system is designed to handle unpredictable peaks in the number of consumers and sellers. But, if the system starts spreading around the world as a proper SaaS

application, then a challenge for deploying additional VMs in different time zones arises. However, this is not a problem because the cloud service providers have several data centers around the world, which is enough to serve both the clients and consumers. Consumers usually order services from nearby warehouses, but they can also order services from sellers that are located in a different area, country or a continent. That is, a seller can offer services in different time zones, but its warehouse needs to be located closely to the consumers ordering location in order to distribute the ordered service to the consumer.

The e-Ordering solution is capable to instantiate as many VM instances as necessary. The communication with the cloud controllers is scalable and elastic. It creates and destroys VM instances during e-Ordering periods, with particular resources that can provide minimum system performance depending on the number of orders and active users.

VII. Conclusion and future work

In this paper we present a scalable architecture for e-Ordering system that is capable to retain its performance regardless of the peak periods. The architecture consists of two parts static and dynamic. The former is always active and charged for administration and reporting. The latter is active only during the period of active orders and it is charged for allowing the consumers to commit their orders. This part utilizes minimum resources to keep the performance at the required level.

The cost of this system is minimized since it uses the minimum resources only in active periods. Additionally, the performance is much better since each Consumer Agent works with its own small database, which has the data only for those companies (sellers and buyers) that are hosted within.

Our future work consists of developing such a system, with the described challenges and algorithms for consumer consolidation in a single VM.

Acknowledgment

This work was partially nanced by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University, Skopje, Macedonia.

References

[1] M. I. Williams, *A Quick Start Guide to Cloud Computing: Moving Your Business into the Cloud (New Tools for Business)*, 1st ed. Kogan Page Limited, 2009.

[2] K. Hwang, J. Dongarra, and G. C. Fox, *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things*, 1st ed. Morgan Kaufmann, 2011.

[3] J. Weinman, *Distributed and Cloud Computing: From Parallel Processing to the Internet of Things*, 1st ed. Wiley, 2012.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[5] P. Mell and T. Grance, "The nist definition of cloud computing sp 800-145," 2011. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[6] K. Kiroski, M. Gusev, M. Kostoska, and S. Ristov, "Modifications and improvements on cen/bii profiles," in *ICT Innovations 2011*, ser. Advances in Intelligent and Soft Computing, L. Kocarev, Ed. Springer Berlin / Heidelberg, 2012, vol. 150, pp. 395–404.

[7] Fortune, "100 best companies to work for," Feb. 2015. [Online]. Available: http://fortune.com/best-companies/

[8] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," in *High Performance Computing and Communications, 2008. HPCC '08. 10th IEEE International Conference on*, Sept 2008, pp. 5–13.

[9] O. Rana, "The costs of cloud migration," *Cloud Computing, IEEE*, vol. 1, no. 1, pp. 62–65, May 2014.

[10] M. Gusev, S. Ristov, B. Koteska, and G. Velkoski, "Windows azure: Resource organization performance analysis," in *Service-Oriented and Cloud Computing*, ser. Lecture Notes in Computer Science, M. Villari, W. Zimmermann, and K.-K. Lau, Eds. Springer Berlin Heidelberg, 2014, vol. 8745, pp. 17–31. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-44879-3_2

[11] A. Schwanengel and U. Hohenstein, "Challenges with tenant-specific cost determination in multi-tenant applications," in *Proceedings of CLOUD COMPUTING 2013 : The Fourth International Conference on Cloud Computing, GRIDs, and Virtualization*, 2013, pp. 85–95.

[12] M. Simjanoska, M. Gusev, S. Ristov, and G. Velkoski, "Scaling the performance and cost for elastic cloud web services," *CIT. Journal of Computing and Information Technology*, vol. 21, no. 2, pp. 85–95, Oct. 2013.

[13] M. Simjanoska, S. Ristov, G. Velkoski, and M. Gusev, "Scaling the performance and cost while scaling the load and resources in the cloud," in *Information Communication Technology Electronics Microelectronics (MIPRO), 2013 36th International Convention on*. Opatija, Croatia: IEEE, May 2013, pp. 151–156.

[14] K. Channabasavaiah, E. Tuggle, and K. Holley, "Migrating to a service-oriented architecture," Dec. 2003. [Online]. Available: http://www.ibm.com/developerworks/library/ws-migratesoa/

[15] M. Endrei, J. Ang, A. Arsanjani, S. Chua, P. Comte, P. Krogdahl, M. Luo, and T. Newling, "Patterns: Service-oriented architecture and web services," Jul. 2004. [Online]. Available: http://www.redbooks.ibm.com/abstracts/sg246303.html

[16] M. Gusev, S. Ristov, G. Arminski, G. Velkoski, and K. Bozinoski, "e-assessment cloud solution: Architecture, organization and cost model," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 8, no. S2, pp. 55–64, Jun. 2013, special issue: EDUCON2013. [Online]. Available: http://online-journals.org/i-jet/article/view/2783

[17] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," in *IEEE Global Engineering Education Conference (EDUCON), 2013 IEEE*. Berlin, Germany: IEEE, Mar. 2013, pp. 736–743, best Paper Award.

The potential of SEPA Credit Transfer implementation in Croatia

M. Ptiček*, B. Vrdoljak *, L. Humski*, Z. Skočir*, G. Bolanča** and Ž. Gašparić**
* University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
** Financial Agency, Ulica grada Vukovara 70, 10000 Zagreb, Croatia
{marina.pticek, boris.vrdoljak, luka.humski, zoran.skocir}@fer.hr, {gordana.bolanca, zeljka.gasparic}@fina.hr

Abstract - After joining the European Union, the Republic of Croatia became obliged to consolidate its laws and regulations with the EU's. European Commission's Regulation No 260/2012, obligatory for Croatia too, has set the 31st of October 2016 as the deadline for replacement of national euro credit transfer and direct debit schemes by SEPA Credit Transfer (SCT) and SEPA Direct Debit for the EU member countries that do not use euro as their currency. From that date, use of SCT will be obligatory for the credit transfers in euro, but it has been decided that it will be obligatory for national currency kuna, too. Since SEPA is based on ISO 20022 norm, which is built upon XML language, it is not realistic to expect from business subjects, i.e. future SEPA users, the understanding of XML technologies and ISO 20022 norm necessary for creating SCT electronic payment order. For that reason, we have built a web application, which provides for users a simple web interface with forms to fill and facilitate their transit towards SCT. This paper explains SEPA and describes its implementation in Croatia, structure of SCT electronic payment initiation order format and the web application we designed for its creation.

I. INTRODUCTION

Single Euro Payments Area (SEPA) is a project initiated by European Union and its main goal is creating an area in EU where the consumers, business subjects and government bodies can conduct their payments in euro currency under equal terms, regardless of their location. In the past few years SEPA payment tools have been slowly becoming mandatory for the EU member countries. Since it has joined the EU, Croatia is also obliged to implement SEPA until the deadline defined by the European Commission, which means that Croatian national euro credit transfer and direct debit system will have to migrate to SEPA until no later than November of 2016. This paper will describe the general purposes of SEPA, its brief history and standards it uses in the background, and SEPA activity called the SEPA Credit Transfer (SCT) along with its *pain* message format. This paper will represent a brief history of implementation of SEPA in Croatia and the main functionalities of an application developed in cooperation of Croatian Financial Agency (FINA) and Faculty of Electrical Engineering and Computing (FER) from the University of Zagreb, in order to facilitate a part of SEPA implementation in Croatia related to SEPA Credit Transfer.

Second chapter describes SEPA project into more details and the third chapter describes the norms on which SEPA is based. The fourth chapter explains SEPA Credit Transfer activity and the main document (payment order) by which a payment can be initiated. The fifth chapter brings the brief history of SEPA implementation project in Croatia and describes the application related to SCT made in cooperation of FINA and FER. Sixth chapter brings the conclusion of this paper.

II. ABOUT SEPA

Project SEPA is a part of a bigger project initiated by European Commission, which aims to unite national markets of the EU member countries into a singular European market. Creation of the united market was proposed in 1990 in the report *Making payments in the Internal Market*. The united market would have euro as a common currency, same payment operation costs regardless of the location within the area, same legal rules for national and cross-border payments, etc. The key prerequisite for such united market were cheap, fast and reliable money transfers across the area [1] and SEPA was meant to meet those requirements; SEPA payment tools would enable the users to conduct cashless payments to anyone within the SEPA area by using only their bank account and SEPA payment tools.

At the end of 2000 EU has defined SEPA as one of the main political goals and in the following years many legal acts have been brought in order to increase the integration of European market. One of the most important directives was *Payment Services Directive*, which has set the legal foundations for SEPA [2]. European Payments Council (EPC), which is consisted of European banks, created the necessary trading and technical schemes and frameworks. EPC's duty was also the creation of SEPA payment tools (schemes): SEPA Credit Transfer (SCT), SEPA Direct Debit (SDD) and SEPA Cards Framework, along with the belonging rulebooks and implementation guidelines. SEPA payment instruments became available for the first time in January 2008.

In 2010 SEPA payments became the most dominant payment type in EU and in 2011 it replaced national payment schemes of the Eurozone countries. SEPA includes 34 countries - 28 EU member countries, 4 EFTA members, Monaco and (as of February 2014) San Marino. On February 1, 2014, national payment schemes have been cancelled in all SEPA countries that have euro as the

main currency. On that date, all of the credit transfer and direct debit procedures have expired and afterwards all of the transactions can be conducted exclusively through SEPA instruments and using the IBAN bank account numbering system. For the rest of SEPA countries, which do not have euro as their main currency, the deadline for that same scenario is October 31, 2016.

III. NORMS USED IN SEPA

Any process that aims to integrate heterogeneous systems and processes has standardization as its key method for achieving that goal. SEPA was meant to be the standard that would unite heterogeneous national payment schemes of EU member countries. However, a new standard was not created, but a decision has been brought to build SEPA upon existing standards - ISO 20022, IBAN and BIC, which are described in the following subchapters.

A. ISO 20022

ISO 20022 is an industrial ISO standard for Financial Services Messaging that was decided to be the most suitable for SEPA [3][4]. This standard ensures consistency in the exchange of messages concerning the financial business. The messages of this standard suit the needs of the entire supply chain; from the clients to the banks, between banks, price and transaction reporting, money transfers, etc. Conceptually, this standard is organized in three layers. The upper layer provides key business processes and concepts, the middle layer provides logical models of the messages and the lower layer deals with the syntax and the actual implementation. The methodology of ISO 20022 is special because it differs business information and the way it is interpreted in the message [5].

The first step of the message modeling begins with modeling of the business process, activities and actors it includes. The business data are then organized into components, which contain the business elements. For instance, in the processes of credit transfer, the key terms are debtor, creditor, debtor agent, creditor agent and payment. For each of these terms, a data type is defined. However, these data types are not necessarily simple, but can be composed of more details (e.g. creditor data type contains name, address, identifier, etc.).

In the second step of message modeling, the defined types (components) model the message logically, while the third step is related to the lowest conceptual layer and includes physical implementation of the message model. For the purpose of the implementation, ISO 20022 is built upon XML language (Extensible Markup Language), even though using other language is also possible within special business domains and with mutual agreement. Consequentially, ISO 20022 business messages are XML documents and the term of message exchange means the exchange of actual XML files, whose structure complies the standardized format. In ISO 20022, the actual standardization mediators are XSD (XML Schema Definition) schemes that define the structure of the XML files and thus standardize their format. XSD files can be found in the ISO 20022 catalogue of messages, which is available online.

B. IBAN

IBAN stands for International Bank Account Number. It is an internationally agreed standard system for numeration and identification of bank accounts. It was developed by European Committee for Banking Standards and later as a part of ISO 13616:1997 standard. The current standard is ISO 13616:2007. IBAN was initially created to facilitate money transfers within the European Union and automate processing of payment orders, but due to its flexibility it has been globally applied. As of September 2014, 66 counties are using the IBAN system.

IBAN consists of 34 alphanumeric characters at most. First two characters have to be letters that denote the country code according to ISO 3166-1, a standard for representation of countries, dependent territories and special areas of geographical interest. The following two characters have to be digits, which have the role of control digits; they ensure the overall accuracy of the IBAN and thus minimize the possibility of error during transcriptions. Use of IBANs has reduced trans-national money transfer errors to under 0.1% of total payments. After these four characters come 30 alphanumeric characters at most and they are called Basic Bank Account Number (BBAN). BBANs are country-specific; each country can decide on its own how many characters will construct the BBAN and the only limitation is that all IBANs across a country must have the same length. For instance, Croatian IBANs must be 21 characters long. First two characters are letters "HR", denoting Croatia, and the following two characters are control digits. Of the rest 17 alphanumerical characters, the following 7 characters denote the leading bank number, and 10 remaining characters denote the account number.

Use of IBAN is mandatory within the European Economic Area and its usage is one of prerequisites for implementation of SEPA payment tools. Since February 2014 usage of IBAN is mandatory for all national systems for credit transfer and direct debit.

C. BIC

BIC is an ISO 9362 standard and stands for Business Identifier Code. BICs were originally introduced by the SWIFT Standard and they are still called SWIFT addresses of codes. It is a standard format for internationally unique business identifier applicable for both financial and non-financial institutions and is approved by the ISO organization. When it identifies a non-financial institution, it is also known as Business Entity Identifier (BEI), and when it identifies banks as financial institution, BIC can be interpreted as Bank Identifier Code.

BIC information is used for routing of the money transfers and exchange of messages between banks. The latest BIC version is ISO 9326:2009. It is made of 8 or 11 characters. In case of identification of primary bank office, BIC is an 8-character code; otherwise it is 11 characters long. The first four characters have the role of Institution Code (e.g. bank code). The next two characters are letters

denoting the country code according to ISO 3166-1, just like first two characters of IBAN. The following two characters can be either letters or digits and they denote the location code, and the last three characters are alphanumerical optional characters denoting the branch code.

At the moment, SEPA is based on IBAN account numbering system and SWIFT-BIC bank identification. However, an overlapping issue between ISO 9362 (BIC) and ISO 13616 (IBAN) has been noticed and discussed recently. Within the Eurozone, SWIFT network is not required for payment transmission because BIC information can be derived from IBANs. Consequentially, usage of BIC information will gradually become optional within the Eurozone (and thus the SEPA area). From the February 2014 BIC usage information became optional for the national transactions, and from February 2016 it will become optional for the cross border transactions, too.

IV. SEPA CREDIT TRANSFER

SCT is SEPA payment tool for non-urgent payment in euro currency within the SEPA area. SCT are used for Payment Initiation activities. They can only be sent in euro currency and its user, i.e. payment initiator, has to know the recipient's IBAN.

In order to achieve "straight-through processing" of the data, EPC has developed a subset of ISO 20022 XML standard messages that were suitable because they defined the structure of the messages exchanged in payment services. Consequentially, all SCT-related messages are exchanged as XML documents whose structure corresponds with the XSD schemas defined by ISO 20022 standard.

ISO 20022 provides packages of standardized messages for various bank communications performed during the conduction of SCT activities. For instance, in the processes of Clearing and Settlement, usage of message package *pacs* is mandatory. Package *camt*, designed for Reporting, is not mandatory, but the banks have to be able to offer it on demand. Package *pain*, used for Payment Initialization, is also not mandatory, but its use is recommended. Even though the mentioned packages define the structure of the messages, this structure is not ultimate. The basic message structure can be further restricted according to the specific needs. For instance, it is expected that several *pain* specifications will be developed among the SEPA members.

A. SCT PAIN

SCT Payment Initiation payments are national or cross-border euro payments from the clients' transactional accounts they have in their national banks toward transactional accounts of the recipient, which can be either national or within Europe Economic Area or EU. The client initiates the payment by giving the payment order to his bank.

SCT payment order is an XML document, whose structure is defined by a XSD schema found in the *pain* package of the ISO 20022 standard. The *pain* schemas can be found in various versions in the ISO 20022 archive, but they do not differ drastically. For the purposes of SEPA, commonly used version of the *pain* XSD schema is *pain.001*, i.e. *pain.001.001.03*, which is recommended by the European Payment Council and was also selected to be the basis of Croatian national *pain* schema. After creating the XML document, a good practice is to validate the created document against that schema definition.

SCT payment order (XML document) has two main parts - a group header and one or multiple payment instruction information elements [6].

Group header refers to all of these payment instruction information elements. It contains the general information about the payments, such as message identifier, number of transactions, control sum and information about the payment initiator. Payment initiator can either be a person or a business entity (organization), so it contains initiator's name and identification element, which is optional. For instance, organizations can be identified by BIC/BEI and a person can be identified by date, city and country of birth.

Payment instruction information is the key element for conduction of the transactions. It also contains some general information such as control sum, number of the transactions, identifier, required execution date and various codes which determine the way the payment will be conducted, e.g. urgency of the transaction, which SEPA instrument will be used, payment purpose, etc. Payment instruction information also contains information about the debtor, but it differs the debtor from the ultimate debtor. This comes practically in use in scenario when a parent is making a payment for the child's school trip, i.e. in its child's name. In that case, the parent is the debtor, but the ultimate debtor is the child. Debtor's and ultimate debtor's information set is very similar to the one previously described for the payment initiator in the group header. Since the debtor is the side that will ultimately transfer money, payment instruction information set contains information about debtor's account and debtor's agent (usually a bank). The account is identified by IBAN and the agent can be identified by BIC code.

It is important to emphasize that payment instruction information can refer to one or more actual transactions. Information for a transaction contains similar data set as the general payment information, e.g. codes which determine the way the payment will be conducted and an identifier. The actual transaction amount is also defined within this data set, as well as the currency and optional remittance information. In the transaction information part, creditor (recipient), creditor's account and agent, and ultimate creditor are defined, while ultimate debtor can be redefined in case that ultimate debtor in the payment instruction information data set is not defined. When all of the transactions have the same ultimate debtor, it is enough to define it in the payment instruction information and leave it empty in the transaction information. Creditor, ultimate creditor and ultimate debtor in this part are defined as either a person or a business entity, similarly to initiator in group header, or debtor and ultimate debtor in the payment instruction information part. Creditor's account and agent are defined similarly to debtor's account and agent in the payment instruction information part.

V. SEPA PROJECT IN CROATIA

After joining the European Union in July 2013, Croatian national payment system was incorporated into SEPA area. From that moment, the option to join SEPA Credit Transfer Scheme became administratively available to Croatian banks. Even though SEPA became available after Croatia joined the EU, it has still remained optional for the upcoming period which will last until October 31, 2016. This date is the deadline because Croatian main currency is not euro and for that reason the later deadline - defined by the European Commission - is applied to Croatia. Until that date, the national euro credit transfer and direct debit schemes will have to be replaced with SEPA Credit Transfer and SEPA Direct Debit. However, a decision has been brought that SEPA will be implemented not only for national euro transactions, but also for the national currency kuna and thus will be one of the first EU countries to implement ISO 20022 payments in non-euro currency [7]. Regarding the deadlines, Croatian banking community has decided that an earlier deadline will be applied for SCT, April 1, 2016.

In order to fulfill the obligation of implementation of SEPA in Croatia, Croatian National Payment System Committee initiated the SEPA project in Croatia [8] in 2012 and still governs it as the supervisor. In May, 2013 they formed the project's working bodies - Croatian SEPA Coordination Committee and Croatian SEPA Forum [9], whose main mission in the implementation of SEPA project in Croatia0. Both the Committee and the Forum are consisted of the representatives of Croatian Banking Association, Croatian National Bank, Croatian Chamber of Economy, Ministry of Finance and Fina. Croatian SEPA Coordination Committee deals with the migration of the national payment system to SEPA, including the credit transfers and direct debits, creates and conducts the solutions necessary for credit transfers and direct debits consolidation with the Regulation No 260/2012, etc. Croatian Banking Association leads the Committee, whose members are elected as its chairmen. Croatian SEPA Forum is oriented towards informing and educating the payment services users about SEPA and general promotion of SEPA payment services, and is presided by the Croatian Chamber of Economy. The Committee and the Forum can also found special workgroups (task forces) for the purpose of fulfillment of their objectives, whose goal will be proposing and finding the practical solutions for the set objectives and tasks. Figure 1. displays the general organization of the working bodies.

In the countries that have adopted SEPA, creating another national body called National Adherence Support Organization has been common practice. Its main purpose was giving support to the payment services providers during the process of transition to SEPA payment schemes. This body establishes contact between the providers and the European Payments Council and facilitates their procedures of collecting the documentation necessary for accessing SEPA schemes. On May 15, 2014 National Payment System Committee has brought the decision that Croatian Financial Agency (FINA) will perform the role of Croatian National Adherence Support Organization (NASO), as mediating body between EPC and payment service providers. As a part of the Croatian SEPA project, FINA will support the implementation of SEPA in Croatia, i.e. migration of the Croatian national payment system to SEPA payment tools, schemes and formats.

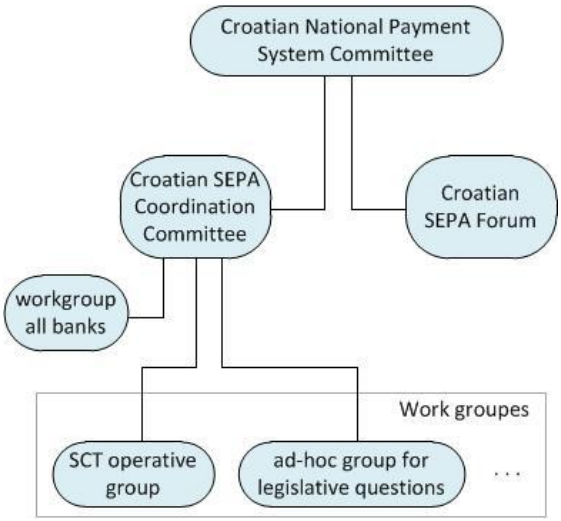


Figure 1. Organization of the SEPA working bodies [7][9]

In order to be able to participate in SEPA area, Croatian monetary institutions had to transit to IBAN account numbering system. Banks had to enable its usage until June 1, 2012. From that date started a two-year transitional period toward exclusive use of IBAN, during which the usage of IBAN was optional for either creditor's or recipient's account number field of the payment orders, and its purpose was to introduce the users to using IBAN. One year later, on June 1, 2013, IBANs became mandatory for the creditor's account number field. One year after that, on June 1, 2014, IBANs became mandatory for the recipient creditor's account number field too and that was the end of transition to IBAN account numbering system in Croatia. Since June 1, 2014 all of the transactions in Croatia can be conducted exclusively by using IBAN account numbers of the creditor and recipient.

A. FINA's Cooperation with University of Zagreb

As a part of SEPA project in Croatia, FINA and Faculty of Electrical Engineering (FER) from University of Zagreb have developed an application, whose purpose is creating one or more SEPA electronic payment orders (XML documents). Its goal is preparing business subjects for SEPA payments and using SCT formats and thus facilitate this transition which will become mandatory in 2016 for both currency of euro and kuna. FER helped FINA in specifying the user requests and modeling the application. The main functionalities of this application will be described in the following subchapter. In the nearer future, it is possible to expect that FINA will develop this web application further.

B. Web Application for Creation of SEPA Electronic Payment Order

The web application created by FINA and FER helps users to create a SEPA electronic payment order (XML

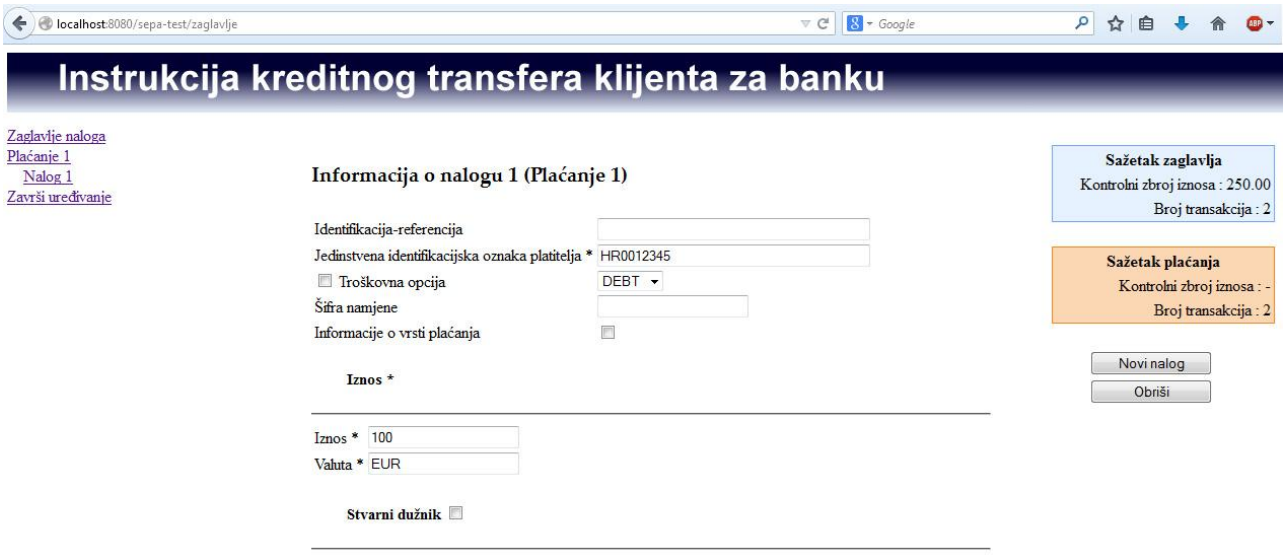


Figure 2. Screenshot of the web application for creation for SEPA electronic payment order

file) through filling and editing of simple web forms. Along with this main functionality, application enables user registration and authentication with basic built-in security mechanisms, validates user's inputs and generated XML file, and offers the generation and download of payment report in a PDF file.

The application has been developed in Java programming language using JavaServer Pages (JSP) and Java servlet web technologies. The servlets process HTTP requests and render the HTML code, acting as intermediary between the client and the actual JSP documents on the web server. This way, clients do not access the JSP documents directly and do not have to worry about the documents' actual location - which is hidden from them - but they access just the URL address, which is mapped to the corresponding servlet. The application provides bilingual interface - the users can choose whether they want the interface in Croatian or English language. The user's choice of language is stored in HTTP session attribute, and based on it the labels (displayed text) in the JSP page are built during the page rendering. The labels are dynamically fetched from a corresponding *property* file; for each of the offered languages one *property* file has been created. This way the application remains easily extensible with more interface languages by adding more *property* files and making few minor changes in the application code.

In order to be able to work with the application's web forms, users will have to be registered and logged in. Users register by their names and e-mails and they must define their password. The application stores the data necessary for the login in the database; the passwords are stored encrypted to keep them safe in case of database breach. The application communicates with the database by *Prepared statements* in order to avoid the possibility of *SQL Injection* attack. After the registration, the users will receive an e-mail from the application with the activation link and will be able to log in after opening that link in the web browser. However, the activation has to take place within the 20 minutes from the registration; otherwise the registration will be annulled. In the moment of registration, the application stores the timestamp so it could compute the time difference in the moment of activation attempt. If the time difference is greater than 20 minutes, the database record will be deleted and user will have to register again. Application will also enable redefinition of the password in case the user has forgotten it. The process is similar to registration process, but the user needs to provide the correct name and e-mail that were used for registration. Pages for login, registration and password recovery (i.e. publicly available pages) also implement CAPTCHA security mechanism so that application could be safe from *botnet* attacks.

After they log in, users will be able to access the electronic payment order form pages and fill the web forms for payment order group header, payment instruction information and the belonging credit transaction information. They will be able to manipulate the number of total payment instructions and the number of their credit transactions, i.e. add or delete them, and navigate through the filled data in order to check or update them. Figure 2. displays a web form for data input for credit transfer information. Navigation is on the left of the screen and helpful summary data, along with the buttons for adding or deleting credit transfer information is displayed on the right. During any kind of data manipulation, application will perform validation of the entered data; e.g. if the mandatory data are filled, if the order contains at least one payment instruction information and a payment instruction information contains at least one credit transfer, if the predefined codes are entered correctly, if IBANs and BICs have the correct form, etc. A part of user's input validation is conducted client-side (i.e. in the user's web browser) and is implemented by JavaScript functions. However, a part validation of the data policies concerning the elements' cardinality, such as number of payment instruction information elements in the group header element, can not be implemented by JavaScript functions and has to be conducted server-side, after the data are submitted from web browser to the application server. The submission

occurs when the users fill the form and want to save the data. After the reception of the data, the application performs server-side data validation and tries to store the data into the data model objects in the session attributes.

The application's main data model is derived from the XML Schema (*pain* XSD schema), meaning that the relations of the data model classes correspond to the relations of elements defined by the XML Schema. The model was generated by JiBX, a framework for XML data binding. After the users have finished their work with the web forms, application will offer them the download of the XML file or the PDF payment report. In both cases, an XML file will be generated based on the data stored in data model objects in the session attributes. In order to map Java objects to XML and generate the XML file, application uses JiBX marshalling methods that generate the contents of the XML file. Internally, once the XML file has been generated, application validates it against *pain* XSD schema by using common Java libraries for XML processing. If user chooses to download the payment report in the PDF, the application will use the generated XML file as a source to create a PDF report using JasperReports, a Java reporting tool.

After generating the requested file, the application sends it and the users can save it locally on their computer. Afterwards, they can either send the XML file by e-mail to their bank or agent, or bring it on a memory stick.

VI. CONCLUSION

SEPA offers potentially ultimate payment tools that could unite all of the EU member countries one step further. Deadlines for scheme implementation favor its success, especially with the fact that it was built upon existing standards which were partially used before they became imperative. This paper has explained the use of these standards and benefits SEPA would bring for each EU member country and SEPA area in general. Croatia has entered the SEPA area only recently and has a lot to achieve in the following two years to migrate to SEPA. However, Croatia's recent entrance is not such a bad thing, because in the meantime European SEPA project has developed to a decent state of maturity. FINA and FER have developed a web application for creation of SEPA electronic payment order in successful collaboration. This application will in future be provided as a service to business subjects and facilitate their transition to using SEPA Credit Transfer for the currencies of euro and kuna. This way SEPA will become one technical standard for payments both in currency of euro and kuna. The application will facilitate a part of the implementation of SEPA Credit Transfer in Croatia, which will be implemented by the beginning of April 2016.

REFERENCES

[1] European Payments Council, SEPA Legal and Regulatory Framework, http://www.europeanpaymentscouncil.eu/index.cfm/about-sepa/sepa-legal-and-regulatory-framework/, accessed April 2014.

[2] European Payments Council, What is a payment scheme?, http://www.europeanpaymentscouncil.eu/index.cfm/sepa-credit-transfer/what-is-a-payment-scheme/, accessed April 2014.

[3] European Payments Council, ISO 20022 Message Standards, http://www.europeanpaymentscouncil.eu/index.cfm/sepa-credit-transfer/iso-20022-message-standards/, accessed February 2015.

[4] European Payments Council, Shortcut to the SEPA Data Format, http://www.europeanpaymentscouncil.eu/index.cfm/knowledge-bank/epc-documents/epc-publication-shortcut-to-the-sepa-data-format/epc346-08-v60-shortcut-to-the-sepa-data-format/, accessed February 2014.

[5] The SWIFT Standards Team, ISO 2022 For Dummies, http://www.sepahungary.hu/uploads/ISO20022_fordummies.pdf, November 2013.

[6] European Payments Council, SEPA Credit Transfer Scheme Customer-to-Bank Implementation Guidelines Version 7.0 Approved, November 2012.

[7] Bohaček Z., Gašpert M. Pregled statusa Projekta SEPA u Hrvatskoj (*National SEPA Migration Plan - project status*). E-biz2015, Zagreb, March 2015.

[8] Official Croatian SEPA page, http://www.sepa.hr/, accessed February 2015.

[9] Croatian SEPA Coordination Committee, Republic of Croatia National SEPA Migration Plan, April 2014.

Business process orientation construct analysis in companies in the Republic of Macedonia

M.Mijoska* and N.Levkov*
* Faculty of economics – Skopje, Ss. Cyril and Methodius in Skopje, Republic of Macedonia
e-mail address: marina@eccf.ukim.edu.mk; nikolal@eccf.ukim.edu.mk

Abstract - In today's global and continuously changing environment, companies are pressured to become more flexible, fast responding, customer oriented and to follow the fast technological development. To achieve these goals they should become business process oriented. Business process orientation (BPO) and process thinking help management to leverage the use of information and communication technologies (ICT) and the Internet, and hence to implement e-business more successfully. E-business strategies have made the need of streamlining, integrating and automating business processes even more pressing. Hence, efficient business processes and higher levels of process orientation are become elementary for organization's success in e-business environment.

Due to the potential of e-business technologies and organizational transformation towards process oriented organization in dynamic environment, the goal of this paper is to analyze business process orientation of Macedonian companies in selected industries and to estimate their level of business process maturity. For this purpose an empirical research was conducted based on the previously prepared questionnaire interviewing managers in Macedonian companies in the banking, telecommunication and IT sector.

I. INTRODUCTION

Today, e-business strategies have made the need to streamline, integrate and automate business processes even more pressing for organizations. Business processes are become elementary for organization's success. In order to maximize its competitiveness in the global, dynamic environment companies need to have business processes which are both efficient and effective. This requires process orientation and achievement of higher levels of process maturity.

In the e-business environment managing a business, means managing its business processes. In that sense, business process orientation – BPO has become one of the most important management paradigms in the new millennium.

But, being aware of its business processes is not enough. Managing mature business processes is the new challenge for contemporary management, since higher levels of process maturity promises better organizational performance. In this sense, maturity models have been proposed to gradually assess and improve business processes in organizations.

II. BUSINESS PROCESS ORIENTATION AND BUSINESS PROCESS MATURITY

A. *Business porcess orientation - BPO*

In the extensive literature on business process management, there are numerous definitions of business process orientation - BPO that vary. One that is more generic and comprises all of the basic elements of the concept is McCormack's and Johnson's [6] definition of process orientation. According to them, an organization that, in all its thinking, emphasizes process as opposed to hierarchies with a special emphasis on outcomes and customer satisfaction is considered as process oriented [6]. The concept of BPO is based on the assumption that the value to customers is delivered by streamlining and accelerating work patterns [6]. Placing the focus directly towards customer and managing end-to-end processes provides a strategic approach in achieving a competitive advantage in the current customer-centric e-business environment.

Organizational transformation towards process orientation for companies means numerous benefits like more efficient execution of work resulting in cost savings, improved customer focus, better integration across the organization, increased flexibility of the company accompanied with improved customer satisfaction etc. Another benefit from process focus is improved hand-offs between functions that leads to cycle time reduction. Process orientation reflected in processes that are broadly defined eliminates redundant activities, verifying inputs one time for all functions within organization. Further, it is indicated that investments in business processes and their management creates competitive advantage for companies and provides significant improvement to the overall system [5], [6].

The concept of BPO in the literature is considered as multidimensional and holistic. Hence, different authors have identified different elements that constitute BPO, each of them having different impact on BPO maturity level. As most relevant in the literature are models and frameworks defined by Kohlbacher, (2010), Škrinjar et al.,

(2010), Willaert et.al, (2007), Reijers (2006), Rosemann et.al, (2006), McCormack, and Johnson (2001). These models in general differ in the number of elements i.e. critical success factors influencing the level of BPO in organizations. Most common are the three elements present in the model of McCormack, and Johnson (2001).

According to McCormack and Johnson (2001), the BPO construct is defined through the following three basic elements: process view, process jobs and process management and measurement [6]. In addition to the basic components, they also discuss about the importance of two supporting components: organizational (process) structure and organizational culture referring to the process values and believes. These two elements enable the basic components to operate, interactively [3], [6]. The above mentioned BPO elements refer to the following:

Process view (PV) refers to the importance of the definition and documentation of business processes and their understanding from top to bottom and from the beginning to end of a process in the organization. w. Documentation of process steps, activities and tasks comes in both visual and written formats that allow people in different job functions and companies to communicate using the same vocabulary. This component includes broad understanding of the processes across the organization, not just documentation.

Process Jobs (PJ) refer to the definition of process related tasks and roles in the organization. These jobs include horizontal rather than vertical responsibility. People participate and take ownership of the entire process.

Process Management and Measurement systems (PMM). This component includes process measurement systems, rewards for process improvement, outcome measurements, customer-driven and team-driven measures, and rewards.

These individual dimensions of the BPO construct i.e. measured variables - PV, PJ and PMM in the model are measured by 3, 3 and 5 items respectively, as given in the table below.

TABLE I. OPERATIONALIZATION OF THE BPO CONSTRUCT

Process view (PV)	
PV1	The average employee views the business as a series of linked processes.
PV2	Process terms such as input, output, process and process owners are used in conversation in the organization.
PV3	The business processes are sufficiently defined so that most employees know how they work.
Process jobs (PJ)	
PJ1	Jobs are usually multidimensional and not just simple tasks.
PJ2	Jobs include frequent problem solving.
PJ3	Employees are constantly learning new things on the job.
Process management and measurement system (PMM)	
PMM1	Process performance is measured.
PMM2	Process measurements are defined.
PMM3	Resources are allocated based on process.
PMM4	Specific process performance goals are in place.
PMM5	Process outcomes are measured.

B. Business porcess maturity model

Extensive literature on business process management suggests that organizations could enhance their overall performance by adopting a process view of business. [5], [6], [8]. But, being aware of its business processes is not enough. Managing mature business processes is the new challenge for contemporary management. In this sense, maturity models have been proposed to gradually assess and improve business processes in organizations. Hence, business process management maturity models are becoming the new promising tool for managing successful companies in e-business environment. They become practical tools that assist organizations in achieving business (process) excellence, since organizational transformation towards business process oriented organization is an evolutionary pattern and as such is based on the concept of maturity.

As the topic became very popular, a proliferation of business process maturity models can be witnessed. A wide research on the topic reviled that most of them are based on the Capability Maturity Model developed by Software engineering Institute (SEI CMM) and differ among four or five maturity levels. Generally, in the literature are familiar maturity models that refer to three major process categories: general (core) business processes, supply chains and collaborative processes [1]. The first category of process maturity models further differ maturity models that refer to a specific process (CMM for example) and maturity models that refer to the overall BPM practice and process orientation of the organization.

Besides their popularity, maturity models have been a subject to a criticism as well. Mainly they are criticized for being a simple consultancy speech or step-by-step recipes that simplify the very complex reality of organizational transformation [4], [8], [11]. In spite they represent a basic road map for organizations in their attempt at moving toward process orientation.

For the purpose of this research the model of McCormack and Johnson (2001) comprising three BPO construct elements (as discusses previously) will be used that differ four levels of maturity: ad hoc, defined, linked and integrated processes [6]. Each step builds on the work of the previous steps to apply improvement strategies that are appropriate to the current maturity level. The following definitions for the stages that an organization goes through when becoming BPO are provided [4], [6]:

[1] *Ad hoc.* The processes are unstructured and ill defined. Process measures are not in place and the jobs and organizational structures are based upon the traditional functions, not horizontal processes.

[2] *Defined.* The basic processes are defined, documented and available in flow charts. Changes to these processes must now go through a formal procedure. Jobs and organizational structures include a process aspect, but remain basically functional. Representatives from functional areas (sales, manufacturing, etc.) meet regularly

to coordinate with each other, but only as representatives of their traditional functions.

[3] *Linked.* The breakthrough level. Managers employ process management with strategic intent and results. Broad process jobs and structures are put in place outside of traditional functions.

[4] *Integrated.* The company, its vendors and suppliers, take cooperation to the process level. Organizational structures and jobs are based on processes, and traditional functions begin to be equal or sometimes subordinate to process. Process measures and management systems are deeply imbedded in the organization.

The first level is estimated with the average score of 1,5; the second 2,5; the third 3,5 and the forth level with average score of 4,5. This measurement tool will help us to estimate on which maturity level are analyzed companies in our survey [6].

III. BUSINESS PROCESS ORIENTATION AND MATURITY OF COMPANIES IN SELECTED SECTROS IN THE REPUBLIC OF MACEDONIA

A. General characteristic of the sample

Competitive advantage of companies today is reflected in aligned and innovated business processes. This implies that successful implementation of e-business strategies requires companies to achieve higher levels of maturity in process orientation.

BPO is becoming promising concept for managing organizations and improving their organizational performance, not only in manufacturing sector, which is traditionally considered as process oriented sector, but in other industries like some high technology industries, telecommunications, finance and insurance etc. BPO is becoming common methodology for improving organizational performance in service industries as well [7].

Due to that, for the purpose of this research companies from sectors like telecommunication, IT sector and banking in the Republic of Macedonia which at the same time show higher level of e-business penetration in the country were selected for further analysis. The structure of the sample by industry type is given in Table II.

TABLE II. SAMPLE BY INDUSTRY TYPE

INDYSTRY TYPE	**Frequency**	**%**
Banking sector	37	54,41
IT sector	18	26,47
Telecommunications	13	19,12
Total	**68**	**100**

Although, the sample size is not sufficient to declare the sample as representative for complex statistical analysis, but it can certainly be considered informative for gaining insight into presence of process orientation in companies from selected sectors in the Republic of Macedonia, having in mind the specific structure of these sectors in the country.

Namely, the banking sector in the Republic of Macedonia comprises 16 banks (12 of them are in foreign ownership) and 4 saving houses (as at the end of 2013) in total. The Banks represent the most significant financial mediator, whereas, savings houses held an insignificant share. The concentration of the sector remains high, with 68, 8% share of the three largest banks in the total assets of the sector [15].

According AEC [2], the telecommunication market in the Republic of Macedonia, as in the third quarter of 2014, is represented by four companies which are FDI (foreign direct investments) (T-Mobile, VIP, ONE and Albafone). Three of them (T-Mobile, VIP and ONE) hold the 99% of the market.

The structure of the domestic IT market is represented by IT companies (which are mainly medium and small companies) including: software and IT services companies, hardware companies and distributors, as well as training providers and ICT consulting companies [14]. Most of the world's largest IT companies such as Microsoft, CISCO, IBM, Compaq, Hewlett Packard, Dell, Siemens, Sun Microsystems, Apple, Lotus, ORACLE are present in the Republic of Macedonia via branch offices, distributors, dealers, resellers, solution providers and business partners. In addition, there are a number of companies oriented towards assembling, sale and maintenance of their own computer systems [12]. Among these companies, of our interest were the ones that are engaged in software development i.e. software and IT services companies.

Due to this the sample of 68 responses from companies in the selected sectors, is considered representative for the task of the paper i.e. gaining the first insights of the process management maturity practices in these sectors in the Republic of Macedonia.

During the period from May to September 2014, an interview with 68 managers from companies in the selected sectors in the Republic of Macedonia was performed based on the previously prepared questionnaire. The questionnaire was consisted of three parts. The questions in the first part refer to the demographic characteristics of companies and respondents (managers). The second part contains question regarding the e-business strategy implementation, as well as major process characteristics of companies. The third part contains questions regarding the BPO construct which was based on the McCormack and Johnson (2001) model. Despite the questions in the first part and several questions in the second part, which were multiple choice questions, for questions in the third part a five point Likert scale was used.

Data gained from part I and part III from the questionnaire were basis for the analysis presented in this paper. The results show the following independent characteristics of the sample (Table III):

TABLE III. THE INDEPENDENT CHARACTERISTIC OF THE EXAMINED COMPANIES AND RESPONDENTS

A. Independent characteristics of the companies	**Frequency**	**%**
Number of employees		
10 - 49	13	19,12
50-249	14	20,59
250+	41	60,29
Market share		
up to 5%	13	19,12
6-10%	10	14,70
10 - 25%	18	26,47
more than 25%	27	39,71
B. Independent characteristics of respondents	**Frequency**	**Percentage**
Hierarchical position of the respondents		
Top management	5	7,35
Middle management	48	70,59
Lower management	9	13,24
Other	6	8,82

B. Data analysis

Regarding the BPO construct, the results of the survey are given in Table IV. For the measurement of the level of BPO, three measurement variables were taken into consideration: process view, process jobs and process management and measurement system which represent critical success factors of BPO in organizations. In the questionnaire they were represented by 3, 3 and 5 items respectively as in the original model [6].

TABLE IV. THE BPO CONSTRUCT – DESCRIPTIVE STATISTICS

BPO construct	**Mean**	**Std. deviation**
Process view - PV		
PV1	3,62	0,90
PV2	3,44	0,98
PV3	3,65	0,94
Process jobs - PJ		
PJ1	3,81	1,12
PJ2	3,43	1,06
PJ3	3,60	1,08
Process management and measurement - PMM		
PMM1	3,69	0,92
PMM2	3,57	0,85
PMM3	3,19	0,87
PMM4	3,29	0,88
PMM5	3,28	0,99

Based on the mean of analyzed critical success factors of the defined BPO construct (PV, PJ and PMM), an average process maturity level of the analyzed companies in the Republic of Macedonia can be estimated. The data showed that the average value for BPO is 3.51. This means that companies in analyzed sample are among second and third level of BPO maturity according the measurement tool of McCormack and Johnson (2001). This result cannot be generalized for all companies in the Republic of Macedonia since the sample was targeted based on the authors previous knowledge that this companies have some awareness about business process orientation and hence a certain degree of BPO. These companies as well have higher level of e-business penetration which is another indicator why a certain degree of BPO in these companies was expected, since e-business is recognized as one of the main driving forces of BPO [6]. Therefore it can be assumed that the overall process maturity of companies in the Republic of Macedonia will be different.

If we look deeper in the given descriptive statistics, based on the data given in Table IV, the lowest average value has PMM3. This means that in the analyzed companies in the Republic of Macedonia, resources are still allocated more on functional basis instead on processes. The highest average value has PJ1, which means that the jobs are more perceived like multidimensional and not simple tasks in most of the analyzed companies.

At the same time PJ1 has highest standard deviation i.e. variance in the answers. This means that there are companies where jobs are very multidimensional, but also there are cases where this statement is not true at all. The lowest standard deviation i.e. variance in the answers is evident for PMM2. This item refers to the definition of process performance measures, which means that most of the respondents recognize importance to this issue.

In the literature there is no significant indicator that there are differences in the level of BPO among companies from different sectors and different size. Even though large companies are making massive efforts to expand their BPO, in contrary, there is evidence that smaller companies tended to score better than larger ones [5]. Another question posed in the literature is the one of "natural BPO" concept which requires further investigation [5]. This means to test whether service companies, which by their nature have muchmore day-to-day contact with their customers, have naturally higher BPO scores than for example manufacturers, who tend to work further down the value chain

Regarding the tested sample that was carefully selected based on the authors expectations that a certain level of BPO exist in companies from selected sectors and size in the Republic of Macedonia, authors do not expect the data obtained from this survey to reveal differences based on these two criteria: industry type and company size regarding the level of BPO.

For that purpose, authors performed one way analysis of variance (ANOVA) to determine whether company size or industry type plays significant role in maturity of business process orientation construct. All assumptions for performing ANOVA like homogeneity of variances, absence of significant outliers and approximately normal distribution of the dependent variables for each group of the independent variable have been previously satisfied.

TABLE V. THE BPO CONSTRUCT AND COMPANY SIZE

Grouping variable - Company size	Small (10-49)	Medium (50+249)	Large (250+)	F ratio	F probability
PV construct	3.30	3.69	3.61	1.116	.334
PJ construct	3.74	3.71	3.53	.354	.703
PMM construct	3.15	3.52	3.44	1.193	.310

TABLE VI. THE BPO CONSTRUCT AND INDUSTRY TYPE

Grouping variable - Industry type	Banking sector	Telecommunications	IT sector	F ratio	F probability
PV construct	3.62	3.71	3.35	1.181	.313
PJ construct	3.47	3.53	3.94	1.672	.196
PMM construct	3.49	3.24	3.33	.791	.458

Based on the F probability statistics, no difference was found in business process orientation maturity constructs between the groups when the grouping variable is company size or industry type as given in Table V and Table VI.

Thus we conclude that company size or industry type does not play significant role in business process orientation maturity on our sample data, since we do not have enough evidence to reject the null hypothesis.

IV. CONLUSION

In order to take the advantage of business process orientation when implementing integrated e-business solutions i.e. automating business processes, companies and their managers have to understand that it is very important to start thinking and acting on a process way meaning to change the complete mindset of an organization.

Regardless the industry type and size, managers in all companies should focus on business processes and streamline their effort in achieving higher level of business process maturity since they promise higher organizational performance and hence better competitive position on the global market.

This research gave us important initial information on process maturity in companies in the Republic of Macedonia. As well, it showed that company size or industry type does not play significant role in business process orientation maturity on our sample data.

But in order to perform profound analysis, authors plan to extend the research in the following directions:

- to increase the number of companies in the sample by including companies from various sectors and from the overall territory of the country;
- to expand the original model by including more critical success factors in the analysis that have impact on BPO, since it is considered as multidimensional and holistic concept;
- to test the impact of BPO construct on organizational performance in companies in the Republic of Macedonia.

Business process orientation and business process management are strategic decisions and hence, require top management support, financial investment, time and commitment of all human resources in the company. BPO requires changes in the complete mindset in organizations.

The organizational transformation towards process focused organization is not a revolutionary but rather an evolutionary effort. This means that companies should join all of its resources and forces in order to gain the promised benefits from these undertakes. This counts for Macedonian companies as well.

REFERENCES

[1] A. Van Looy, M. de Backer and G. Poels, "Defining business process maturity. A journey towards excellence", *Total Quality Management and Business Excellence*, Vol.22, No.11, pp.1119-1137, 2011

[2] AEC, Report on development of the Electronic communication market Quarter 2, 2014

[3] K. McCormack, Business process orientation, supply chain management, and the e-corporation, IIE Solut. Vol. 33, No. 10, pp. 33-35, 2001

[4] K. McCormack, et al., "A global investigation of key turning points in business process maturity", *Business Process Management Journal*, Vol. 15, No. 5, pp. 792-815, 2009

[5] K. McCormack, "Business process orientation: do you have it?", *Quality Progress*, Vol. 34, No.1, pp.51-58, 2001.

[6] K. McCormack and W. Johnson, "Business process orientation – Gaining the e-business competitive advantage", Florida: St. Lucie Press, 2001

[7] M. George, "Lean Six Sigma for Service: How to Use Lean Speed and Six Sigma Quality to Improve Services and Transactions", McGraw-Hill, New York, 2003.

[8] M. Röglinger, J. Pöppelbuß and J. Becker, "Maturity models in business process management", *Business Process Management Journal*, Vol. 18, No.2, pp. 328 – 346, 2012

[9] P. Harmon, "Business Process Change: A Guide for Business Managers and BPM and Six Sigma Professionals," 2nd edition, Morgan Kaufman Publishers, San Francisco, 2007

[10] S. L. Mansar and H. A. Reijers, "Best practices in business process redesign: use and impact", Business Process Management Journal, Vol.13, No.3, pp 193-213, 2007

[11] T, De Bruin et.al., "Understanding the Main Phases of Developing a Maturity Assessment Model", *16th Australasian Conference on Information Systems (ACIS)*, November 30 – December 2, Australia, New South Wales, Sydney, 2005

[12] USAID, ICT Country profile - Macedonia, Regional Competiveness Initiative, 2013
[13] www.gartner.com
[14] www.masit.org.mk
[15] www.nbrm.mk
[16] www.stat.gov.mk

The role of the public sector in defining the communication strategies in the Croatian tourism industry

Ph.D. Đani Bunja
University of Zadar/Department of Tourism and Communication Studies, Zadar, Croatia
gianni@unizd.hr

The paper describes communication strategies in the tourism industry. After an overview in the paper are discussed issues in tourist destination communication, i.e. the larger social, political, economic, and cultural climate in which tourist destinations creates their images and projects their voices, the "opportune moment" for communications, the need to understand and capitalize on the psychology of constituencies, the best mix of communication channels and the ethical dimension of tourism industry communication. Tourist destination needs to pay attention to communication strategies because missed opportunities and potentially dangerous threats to industry performance are the alternatives. In the paper then are described the possibilities for applying communication strategies on Croatian tourism industry. It is emphasized that the main characteristics of Croatian tourism industry are the inadequate quality level of services, unfavorable structure and low level of utilization. However, unclear identity can easily hurt tourism industry success and in order to bring about change (improvement the quality of Croatian tourism industry) the application of the communication strategies will be needed.

I. INTRODUCTION

Croatia earlier this year (2015.) chose a new tourism slogan, i.e. a new key message that will be used for its promotion on the international tourism market. The slogan reads: "Croatia, full of life" and will replace the previous one used for the last fifteen years, "The Mediterranean as it once was". Another slogan that Croatia also used for the purpose of tourism promotion was a "Small country for a big vacation", as the previous and the first slogan since Croatian independence and its appearance on the international tourism market under its own name "Croatia, a paradise on earth" practically had not been used.

The slogan is, according to the authors and the client, i.e. the Croatian Tourist Board and the Ministry of Tourism, applicable to all languages and in all communication channels, and summarizes all Croatian advantages - of natural beauty, cultural and historical sites, gastronomy, love, secrets and more.

The selection of a new key message will be followed by the design of a new logo, visual and other promotional materials accompanying it, as well as the elaboration of new concepts of advertising and overall marketing campaign of Croatian tourism for the next five to ten years, which is adjusted to the requirements of the last year adopted Strategic Marketing Plan for Croatian tourism in the period until 2020.

However, the adoption of this slogan, as most of other decisions related to the Croatian tourism, has caused many disputes.

The performance evaluation of tourism policy in a specific tourist destination is often a political issue, resulting in conflicts among political options, which is especially the case in Croatia.

II. ISSUES IN TOURISM INDUSTRY COMMUNICATION

Tourism industry communication is the voices and the images it projects of itself on the world tourism market by its various audiences. Included in this field are areas such as tourism industry identity, image and reputation, tourism industry marketing and advertising, investor relations, government relations, media management, change management, crisis communications, etc.

Tourism industry communication can be possibly observed from several vantage points. It is most visible as a function [19]. In addition, tourism industry communication is also the processes used to communicate all its messages to key constituencies. Finally, it is possible to see ethics as a necessary component of tourism industry communication.

In order to get the best results from communication there is a need to develop a communication strategy. An effective communication strategy involves process of setting objectives, analysis and synthesis, strategy implementation and its evaluation [7].

A developed communication strategy reduces the barriers to communication. The four main elements in an effective communication strategy are communicator strategy, audience strategy, channel strategy and message strategy [18].

Tourism industry communication can be one of the major ways to counteract the persistent scrutiny and negative attention that business has to face [19]. Secondly, a good strategy can set up the technological advances that enhance the scope of the publicity. Thirdly, numerous lessons from worldwide tourism industry demonstrate the harsh consequences of avoiding tourism

industry communication. And, finally, tourism industry communication can serve as a differential advantage on the international market.

As a result, the decision-making process should take into account which strategic direction a tourism industry will choose to take and ultimately how various tourism industry constituencies will react to new strategic initiatives.

The most relevant issues in the field of public relations [19] in general, but also in tourism industry, are the following:

- The larger social, political, economic, and cultural climate in which tourism industry creates their images and projects their voices;
- The "opportune moment" for communications, i.e. sensing the right moment to communicate messages;
- The need to understand the psychology of the market. Clients' constituencies seemed to be capable of rational decision making and so should be provided the facts on which to base decisions. The public was driven by primitive emotions. As a result, the public could be persuaded by symbols. Symbols, by their nature, appeal to the emotions, not to the judgment or intellect. The public was susceptible to the entertainment value of news events, even when they were contrived;
- The best mix of communication channels;
- The ethical dimension of tourism industry communication.

Tourism industry has to pay attention to communication strategies or otherwise missed opportunities and potentially serious risks to efficient and profit are rather costly alternatives.

III. TOURISM INDUSTRY COMMUNICATION STRATEGIES

The idea of communication strategy is taken from strategic management. However, while tourism industry strategic management deals with issues of productivity, profitability, efficiency and effectiveness, a communication strategies objective is development of tourism industry identity, image and reputation.

Generally, numerous examples from worldwide tourism industry demonstrate that tourism industry communication strategies by its nature are either offensive or defensive.

The different parts of a contemporary tourism industry communication function perform worldwide, activities such as tourism industry advertising, and the management of tourism industry identity, image and reputation.

Tourism industry identity is the various attributes, including names, brands, and symbols. A related concept, tourism industry image, is a reflection of that tourism industry identity. In other words, tourism industry image is how each constituency views specific tourism industry. Reputation represents the cumulative impact of the images all constituencies have of specific tourism industry. Both reputation management and tourism industry advertising should be handled in the tourism industry communication function because they have a direct bearing on how the tourism industry is perceived by its most important constituency's [19].

Activities of the tourism industry communication function are [19]:

- Strategy formulation and implementation
- Tourism industry identity, image, and reputation
- Tourism industry advertising and advocacy
- Customer relations
- Employee communications
- Investor relations
- Government affairs
- Community relations
- Media relations
- Crisis communication
- Management of new communication technologies

Not every tourism industry will be able to include all of the activities previously identified as parts of one strategy, but a majority of these functions must be included for the best operation. Indeed, the activities must be connected in some way for a tourism industry to speak with a single voice.

Along with a role in strategy formulation and implementation, tourism industry communication has recently expanded its scope of responsibilities in order to include the use of communication technologies.

The access to data is a crucial issue and therefore very important for managing worldwide communications. Through contemporary technology there is information available through the communication networks of a quality and a degree of sophistication that only, a few decades ago, was solely held by the largest multinational corporations.

With the technology ubiquity in tourism industry the functions of communication enabled by the given technology become ever more important.

Not only has the tourism industry communication function expanded to include responsibility for communication technology as it affects tourism industry communication, but the expansion of technology use in the broader society has given greater visibility and importance to the function of tourism industry communication.

IV. THE POSSIBILITIES FOR APPLYING COMMUNICATION STRATEGIES ON CROATIAN TOURISM INDUSTRY

The Republic of Croatia considers tourism as one of its main strategic guidelines of development. The tourism sector is undoubtedly one of the most significant sectors of the Croatian economy. However, regarding the

available resources, Croatia has not developed a product competitive enough. Croatia is in general unrecognizable as a destination. Croatia disposes of inadequate capacities, the inland, and the historical and cultural heritage are not evaluated enough [7].

Unfortunately, in spite of the country's hundred years of tourism tradition, we have to still speak of it as a new destination, primarily as its name is concerned [4].

Croatia on the international tourism market shares the faith of all Mediterranean destinations. Therefore tourism industry should also respect demand characteristics of customers attracted to this region. More frequently tourists' chief interest is in the quality of the product purchased. Croatia is facing the peril of market defeat in the case of untimely or inadequate inclusion into competition on Mediterranean tourist market. Therefore the immediate and most important aim of the Croatian tourism is qualitative transformation.

Tourism industry facilities in Croatia are mostly appropriate for mass-tourism, which nowadays is not an up-to-date option. The insufficient quality level of service is caused by the lack of standard in projecting, constructing and equipping its buildings. Unprofitable facilities prevail – those that cannot create income. Supply structure in this way cannot be efficient. With all this, the capacities of the Croatian tourism industry are becoming more and more uncompetitive on the international market. The consequence of inadequate tourism industry facilities is deficient utilization level in compare to competitors.

It is one of the reasons for a low occupancy rate. Other reasons can be searched in the context of the lack of identity of Croatia as a tourist destination, i.e. its poor offer unable to attract tourists and other consumers during the low season.

Neither the rate of capacity exploitation, nor the profit itself adequately indicates the successfulness in the modern tourism industry. A pleased customer should be the ultimate result of a well-done tourism-related enterprise.

An accelerated adoption of new strategy requires changes on the macro-level. The provision of additional investments in entertainment facilities and equipment in order to prolong tourist season is of great significance for Croatia. It is necessary to build small accommodation facilities adapted to the requirements of individual tourists. Investments may play an important role by investing in family and smaller tourist and hotel facilities. Innovative forms of investments involve long-term cooperation of partners' i.e. direct investments, joint investments, management controls and franchising. Foreign direct investments include know-how, and development of human resources. Franchising is particularly important for Croatia because it provides faster and easier internationalization of business.

The main tourism industry objective now is desirable identity configuration, and the way the potential customers perceive it [13]. In that respect, tourism industry should constantly expand their offer with new facilities and in such a way they can be more competitive and attractive to their clients.

Therefore, the turning point for Croatian tourism industry is not only in ownership and operative restructuring, but also in creation of a completely new identity and creation of destination's own image. The Croatian strategic communication goal has to be to develop a tourism promotion that will help Croatia create its own tourism identity and the image of a desirable tourism destination. Successful promotional activities are long-term and expensive, so penetration into international tourism market asks for a strategic decision, with clearly objectives.

Simply, success of tourist product and tourist destinations will not only depend on development of products in line with consumer demands and new trends but also on the more intensive marketing of their product and careful targeting of relevant market segments.

V. THE IMPORTANCE OF MEASURING TOURISM

Measuring tourism helps to understand some of the problems of those in charge of decisions related to plans for tourism growth. Statistics are needed for gathering the data which will enable an estimation of tourism importance to researchers, managers and other users, especially to public sector. The reasons for statistical measuring of tourism are as follows [26]:

- to estimate importance of tourism for tourist destinations;
- to make a quantitative analysis of contribution of tourism to economy, i.e. to society as a whole, and especially to define the impact of tourism on balance of payments;
- to help tourist industry and public sector in their process of planning and development of tourism infrastructure;
- to help in implementation of promotional activities;
- to help tourism industry in making decision related to future business activities.

However, there is an information gap between types of statistical data being offered to users and the real users' needs. [20] Actually, most of tourism statistics present measuring of tourist arrivals, overnight stays and spending by the use of questionnaires, the results of which are then interpreted to become representative for the real situation. Therefore, the questionnaires are liable to significant mistakes, depending on the size of the sample. While tourism sector requires reliable statistical data, which managers or decision-makers will be informed about, the public sector in Croatia often manipulates the data [9].

VI. THE ROLE OF THE PUBLIC SECTOR IN TOURISM DESTINATIONS DEVELOPMENT

Tourist destinations take part in a global tourism market competition. The market is characterized with a continually growing competitiveness. In this context, it is possible to speak about the so-called hyper competition [25] in tourism. According to Page, it is possible to define the following four processes, typical of modern tourism and contributing to hyper competition [25]:

- tourist require an improved quality at lower prices;
- implications of IT development;
- occurrence of very aggressive business entities ready to enter the market for a longer period with the so-called loss leader products aiming to destroy their competition;
- barriers of authorities towards competitions are progressively being removed worldwide.

Tourism management (elected by the public sector) requires certain measures to correct any deviations in accomplishing the objectives determined [9]. Unpredictability and taking risk are here unavoidable. In this context, managers of a certain tourist destination must ensure the destinations being able to adjust, which will allow for the destination's survival.

Authors of papers on tourism agree trends in tourism are influenced by a series of exogenous and endogenous factors not related with nor controlled by tourism. Managers must forecast the future to reduce risks and to increase possibilities for success. Therefore, tourism managers cannot avoid the need for a kind of forecasting the future trends in tourism. Those forecasts are results of guessing, teamwork or use of complex models and have a significant impact on managerial decisions. Jefferson and Lickorish in Marketing Tourism: A Practical Guide [2] point out the fact that forecasting is not a precise science as it tries to estimate the future potential and a series of possible scenarios related to possible level of changes in a certain category. According to Bull [1] there are two basic types of forecasting: forecasts based on qualitative techniques and quantitative forecasting methods using techniques deriving from statistics and economics.

Forecasting tries to determine the way the tourist flows influenced former trends and the way they can change in the future. Although forecasting will never be a precise science, it is one of the foundations for making plans about future changes in tourism.

Management of tourism destinations development is a complex question for the following reasons [26]:

- tourism sector is not a homogenous sector of economy as it consists of various activities included in tourism flows;
- certain business activities included in tourism sector are not always seen as an integral part of this sector;
- tourist destinations, i.e. the areas where tourism activities take place, are not controlled only by management of tourism sector but also of public sector, which coordinates the interests of the local community and business entities in terms of preservation of the resource basis they use;
- public sector is responsible for promoting development and marketing of tourism.

To formulate a successful developmental policy, all those connected with a tourist destination should be involved in different stages of this process. Developmental policy of a certain destination presents a smaller or a larger segment in formulating tourism policy within an integrated regional or national policy. The destination's surrounding is constantly changing; therefore the politics itself should be dynamic.

VII. CONCLUSION

The development of tourism needs to be managed and not left uncontrolled. The public sector approaches tourism with special interest due to its positive impact on the economy. However, in order to realize the benefits of the effects of tourism, the public sector must achieve certain prerequisites for the development of tourism. Ultimately, the development of tourism must stimulate the development of the economic activities that do not belong to the tourism sector [9].

Achieving a balance between the needs of each interested business and a sustainable development of tourist destinations is a real challenge. Therefore, the management of tourism is a complex task and there are good reasons why the public sector is usually in charge of its good functioning in society.

The fundamental role of the public sector in tourism lies in planning its development. The public sector by its intervening in tourism development seeks to achieve its long-term sustainable development, i.e. the maximization of benefits from tourism activities while preserving the attractiveness of tourist destinations.

The achieved level of development and utilization of capacities of Croatian tourism industry beyond any doubt require change. The one of possible options to achieve determined objective, i.e. qualitative improvement of Croatian tourism industry is defining the communication strategy. The communication strategy result, desirable identity and positive image in social and ecological environment, also require for an analytical research of competition.

Successful management of tourism as much as tourism planning and regulation of activities related to tourism, requires coordination between government bodies, interest groups, private sector and the involvement of local and regional government. The public sector, on the one hand, generally has a reactive role in terms of the regulation and control of tourist activities on the other hand, has a proactive role in planning and managing tourism.

LITERATURE

[1] A. Bull, The Economics of Travel and Tourism, Pitman, London, 1991.

[2] A. Jefferson & L. Lickorish, Marketing Tourism: A Practical Guide, Longman, Harlow, 1991.

[3] A. Mathieson & G. Wall, Tourism: Economic, Physical and Social Impacts, Longman, Harlow, 1982.

[4] B. Pirjevec, "Survey and assessment of real potentials for development of Croatian tourism", Acta Turistica, 1997, pp. 5-45.

[5] B. Wheeller, Tourism's troubled times – Responsible tourism is not the answer, Tourism Management, vol. 12, 1991, pp. 91-96.

[6] Đ. Bunja, "Communication Strategy and Culture in Tourism and Hospitality Industry", Hotel & Tourism 2002, pp. 301-308.

[7] Đ. Bunja, "Management and Organizational Communication in Cross-Cultural Environment in Tourism and Hospitality", Informatologia, 2002, vol. 35. n. 4, pp. 303-308.

[8] Đ. Bunja, "Modernizing the Croatian Tourism Industry", International Journal of Contemporary Hospitality Management, 2003, vol. 15, n. 2, pp. 126-128.

[9] Đ. Bunja, "The public sector interventionism regarding tourism development – the example of Croatia", SGEM Economics & Tourism, Conference Proceedings Volume IV., 2014, pp. 775-782.

[10] D. Hickson & D. Pugh, Management Worldwide: The Impact of Societal Culture on Organizations around the Globe, Penguin, London, 1995.

[11] D. Magaš, Turistička destinacija, Faculty of Tourism and Hospitality Management, University of Rijeka, Opatija, 1997.

[12] E. Cater & G. Lowman, Ecotourism: A Sustainable Option? Wiley, 1994.

[13] F. P. Drucker, New Realities, Novi Liber, Zagreb, 1992.

[14] J. Conger, "The necessary art of persuasion", Harvard Business Review, May–June 1998, pp. 84–95.

[15] J. Krippendorf, The Holidaymakers Understanding the Impact of Leisure and Travel, Butterworth-Heinemann, Oxford, 1987.

[16] J. Swarbrooke, Sustainable Tourism Management, CAB International, Wallingford, 1999.

[17] L. L. Byars, W. L. Rue & S. A. Zahra, Strategic Management, Irwin, Chicago, 1996.

[18] M. Rouse & S. Rouse, Business Communication, Thomson Learning, London, 2001.

[19] P. A. Argenti & J. Forman, The Power of Corporate Communication, McGraw Hill, New York, 2002.

[20] P. R. Brunt, Market Research in Travel and Tourism, Butterworth-Heinemann, Oxford, 1997.

[21] R. Davidson & R. Maitland, Tourism Destinations, Hodder and Stoughton, Bath, 1997.

[22] R. Mitchell, B. Agle & D. Wood, "Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts", Academy of Management Review, 1997, 22, pp. 853–886.

[23] R. Ohanian, "Construction and validation of a scale to measure celebrity endorser's perceived expertise, trustworthiness and attractiveness", Journal of Advertising, 1990, 19(3), pp. 39–52.

[24] R. W. Griffin, Management 3rd ed., Houghton Mifflin Company, Boston, 1990.

[25] S. J. Page & co., Tourism: A Modern Synthesis, Thomson Learning, London, 2001.

[26] S. J. Page, Tourism Management, Managing for change, Butterworth-Heinemann, Oxford, 2003.

[27] V. Jadrešić, Turizam u interdisiplinarnoj teoriji i primjeni, Research papers collection, Školska knjiga, Zagreb, 2001.

[28] W. C Frederick, J. E. Post & K. Davis, Business and Society – Corporate Strategy, Public Policy, Ethics, 7th ed., McGraw Hill, New York, 1992.

[29] W. David, Managing Company - wide Communication, Chapman & Hall, London, 1995.

[30] World Commission on Enviroment and Development, Our Common Future, Oxford University Press, Oxford, 1987.

[31] WTO, Guidelines for the Collection and Presentation of Domestic and International Tourism Statistics, Madrid, 1981.

Accounting Information Systems and their use in Regional and Local Governments sector: Quality, Efficiency, Security and Control procedures as (present) Challenges

Marija Boban, PhD*, Toni Šušak, MA**
*University of Split Faculty of Law
Department for Economic and Financial Sciences, Split, Croatia
marija.boban@pravst.hr
** University Department of Professional Studies
Department of Finance and Accounting, Split, Croatia
tsusak@oss.unist.hr

Abstract – The basic function of accounting information system in local government is production of information on the basis of data processing as a result of financial transactions. This kind of information usually is the foundation and support of information systems management accounting and is conceived as a system for identifying, collecting, processing and exchange of financial and other data used to support decision making.
The main object of this paper is to analyse the importance and the impact of use of Accounting Information System (AIS), as well as use of information technology in accounting in whole. Also the authors will analyse the modelling of AIS, establishing quality of financial information, security of model of AIS in local government units, especially in decision making process in AIS and in internal control procedures.
The author's will also give and overview of standards in financial reporting with emphaisis on ISO/IEC 2510: 2011 and XBRL global standard as well as their assessment in use in local and regional governments.

Key words: accounting information system, control, quality, efficiency, financial reports, control, regional and local governments

I. INTRODUCTION

The standards of the European Union, in terms of local government, are clearly defined with European administrative space based on the principle of good governance and management, harmonization, development and codifying with it's convergence.

In most European countries, local authorities rely in an appropriate (usually higher) rate on financial grants from the central government and by budget planned accounting. [1] Although the constitutional basis for financial grants is not the same in all European countries is the central government grants to local budgets vary from 30% to 70% (United Kingdom, Netherlands, Ireland, etc.).[2]

In Croatia, the current issue is the issue of decentralization of the fiscal system and the transfer of authorities, responsibilities and funds from higher levels of government to the local level. Also this process has inevitably a political dimension which can be reduced to different ways of organizing economic and political institutions, which then define the financial strength of the local community. Local and regional governments in Croatia are characterized by a relatively weak financial capacity - share in GDP in budgetary spending is negligible; efficient use of public resources and effectiveness in achieving the objectives of relevant public policies are still not at a satisfactory level. In this context, reform of the part of the fiscal system, which regulates the relationship between central and local authorities is a necessary part of both economic and political reforms. Specifics of regional and local self-government in the Republic of Croatia speak in favor of greater decentralization, especially in order to stop the trend of depopulation of certain areas, but also to relieve the pressure of population on urban areas and attracting foreign investment. In this direction, in order to achieve the highest possible degree of fiscal decentralization, local governments must ensure that generous funding that will allow the provision of public services within their purview.

The next step in this direction is the establishment of a computerized accounting system which would enable a more efficient overview and insight into the current indicators of budgetary accounting. In favour of giving

[1]For more on this subject see McLeod and, R., H., Harun, H., Public Sector Accounting Reform at Local Government Level in Indonesia , Volume 30, Issue 2, John Wiley & Sons Ltd. p 238–258, May 2014.

[2] AULFIELD, J., „Local govermnent finance in OECD countries", University of New south Wales, 2008., p 4

the bigger efficiency and transparency as well as higher level of control and security – implementation of Accounting Information Systems (furthermore AIS) in the actual suggested process of decentralization which is discussed at this moment – can present the step forward in Croatia.

II. DECISION-MAKING PROCESS USING AIS

The goals of AIS differ depending on characteristics of organization which implements them. When speaking of companies, efficient usage of information through AIS can have various goals. However, in the end, entire set of goals converge to the fundamental goal every company has – profit maximization for the owners.

The situation with the local government units is somewhat different because they aim to provide the best possible service to the community and, accordingly, public benefit maximization tendency exists unlike companies which aim to maximize owner's profit.

One of the most important functions of AIS is to support decision-making at all organizational levels within a local government unit.

In addition to the function of providing periodical information to the external users through formalized financial statements, AIS also has function of providing relevant information at all organizational levels within an organization (in this case local government unit), especially to the highest management levels in process of decision-making. AIS handles both financial and non - financial transactions that directly affect the processing of financial transactions. [3] Besides traditional financial information (the amount of cash available, cost structure etc.), non-financial information is also needed for the more efficient decision-making process because certain information that can be very significant aren't contained in financial statements and attention should be paid to them.[4] When changes are made to customers' data, for instance change in next of kin and/or address, these changes provide vital information for processing of future data to such users/customers..[5]

Nowadays, „information overabundance" [6] is additional problem which can further complicate the decision-making process and because of which „computational procedures for decision-makers have been developed when using such information for specific decisions.". The same attention should bei also given to "data quality". The main understanding of this term is shown on Figure 1.

Dimensions	Definitions
Accessibility	The extent to which data is available, or easily and quickly retrievable
Appropriate Amount of Data	The extent to which the volume of data is appropriate for the task at hand
Believability	The extent to which data is regarded as true and credible
Completeness	The extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	The extent to which data is compactly represented
Consistent Representation	The extent to which data is presented in the same format
Ease of Manipulation	The extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	The extent to which data is correct and reliable
Interpretability	The extent to which data is correct and reliable
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial
Relevancy	The extent to which data is applicable and helpful for the task at hand
Reputation	The extent to which data is highly regarded in terms of its source or content
Security	The extent to which access to data is restricted appropriately to maintain its security
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand
Understandability	The extent to which data is easily comprehended
Value-Added	The extent to which data is beneficial and provides advantages from its use

Figure 1. Data quality dimensions (Source: Wang & Strong 1996.)[7]

In the view of before mentioned information overabundance there is the question of utilization of certain information. Key characteristics when considering degree of information usability are:[8]

- Relevance,
- Reliability,
- Completeness,
- Timeliness,
- Intelligibility,
- Verifiability.

Accountability of leaders and groups that manage modern local government units is significantly increasing which indicates the requirement for efficient use of scarce resources and achieving planned goals.[9]

It is important to highlight that the manner of functioning of local government units in Republic of Croatia is significantly different than in the past. Political parties gain the trust of citizens on local level and, accordingly, legitimacy within a specified period to improve the local government unit in the different segments. If they fail, it is most likely that they will be „punished" at the following elections and their possibility of further local government unit management denied.

[3] Muhrtala, T. O., Ogundej, M., Computerized Accounting Information Systems and Perceived Security Threats in Developing Economies: The Nigerian Case, Universal Journal of Accounting and Finance, p 9

[4] Ibid. p 12

[5] Vitasović M.: „Analiza stanja i doprinosa računovodstvenog informacijskog sustava i sustava upravljanja financijama u segmentu razvoja proračunskog sustava lokalnih jedinica, Ekonomska misao i praksa, Dubrovnik, god. XXI, 2012.,vol .2, p 570-572

[6] Meter M.:„Potreba intenzivnijeg korištenja i mogućnosti unapređenja računovodstveno informacijskih sustava u hrvatskim poduzećima", Ekonomski pregled, 57(7-8), 2006., p 519.

[7] Wang R. Y., Strong, D.M., Beyond accuracy: what data quality means to data consumers, Journal of Management Information Systems archive, Volume 12 Issue 4, March 1996., pp 5-33

[8] Ibid., p 520.

[9] Nuryanto M., Nur Afiah N., [7] Nuryanto M., Nur Afiah N.: „The Impact of Apparatus Competence, Information Technology Utilization and Internal Control on Financial Statement Quality (Study on Local Government of Jakarta Province - Indonesia)", World Review of Business Research, Vol. 3. No. 4., November 2013 Issue., 2013., p 157.

Therefore, it is in the interest of local government unit leaders to ensure the quality information which enables them to bring optimal decisions and enhance the probability of retaining their political positions.

The Republic of Croatia belongs to the countries that experienced the transition and thoroughly changed economic system in the late 20th century. Accordingly, we can talk about „progressive environment“ and „the need of a more informative accounting information system“ in Republic of Croatia which went through significant changes similar to Poland in the same period.[10] Unfortunately, the crisis and the following experience of GDP falling (continuously) on the national level in Croatia didn't go in favour of getting the AIS on the higher level so in this matter the authors will disccuss on the best way to implement the AIS and AIS reporting in the public, reginal and local self-governments in order to harmonise their functioning and their level of use of modern information technologies.

The good example, in that area presents the reform conducted by Italian local authorities' also emphasized performance control of public managers with evaluation of „their ability to achieve stated objectives, both financial and non-financial, with the set of resources available and within their degree of autonomy and responsibility.“11

Speaking of local government units reforms, the case of Portugal can also be mentioned because Portuguese local governments improved their information system (and consequentially the quality of information and decision making) in a way that besides „traditional cash-based budgetary accounting, the system now includes accrual-based financial accounting and reporting, as well as cost accounting.“[12]

The research conducted in Indonesia showed that „information technology utilization has positive and significant impact on the quality of financial statements of the provincial government.“[13]

III. QUALITY CHARACTERISTICS IN ACCOUNTING INFORMATION SYSTEMS

The basic function of accounting information system is production of information on the basis of data processing as a result of financial transactions. This kind of information usually is the foundation and support of information systems management accounting and is conceived as a system for identifying, collecting, processing and exchange of financial and other data used to support decision making. There is no doubt that the function of AIS same in this case with local governments and with businesses as the quality of accounting information to ensure the achievement of the targets (as shown on Figure 1.) set by the idea that someone always has to be responsible for certain costs, otherwise it will cost to grow out of control.[14]

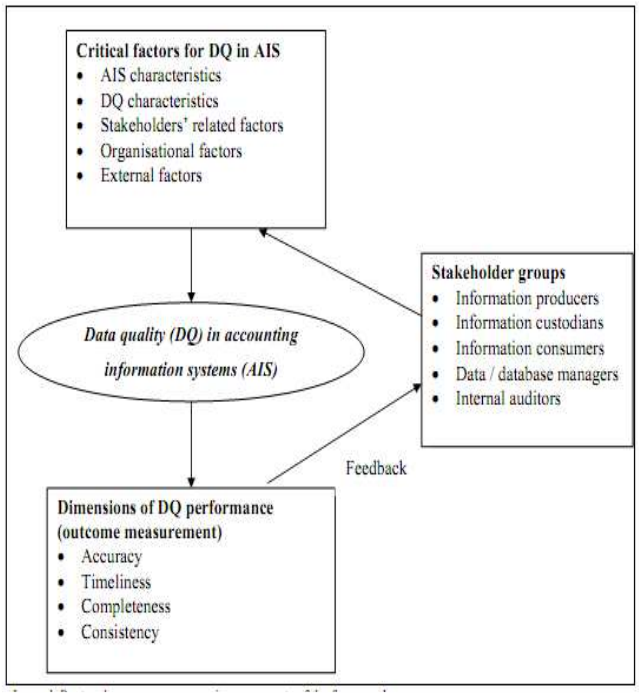


Figure 2. Data quality in Acounting information systems (AIS)[15]

If we are addressing to the quality characteristics in accounting information systems, several high-level approaches to modelling of the quality of software products have been proposed: these models are based on the idea that there are a number of important high-level quality factors of software products that should be measured. These factors are determined by lower-level criteria which are supposed to be much easier to measure than factors; for this reason, metrics are proposed as criteria. As a consequence, all the approaches proposed so far are based on the decomposition of a set of quality attributes in order to define a hierarchic structure of characteristics and a way to relate them to some technical aspects of the product.

These and other efforts have led to the definition of a number of draft standards which directly address software measurement issues. The main and most popular one, ISO/IEC 2510: 2011, gives standard definitions of quality attributes and proposed measures, together with a set of application guidelines.

[10] Faustino Jorge, S.M.: „Local Government Accounting in Portugal in Comparative-International Perspective“, Thesis submitted to the University of Birmingham, 2003., p 151.
[11] Bracci, E. Menagerialism, Accounting and Accountability int he Italian Local Governments: An Empirical Analysis“, 2008., available at SSRN: http://ssrn.com/abstract=1184621 , 2008., p.13. (15. 02. 2015.),
[12] Faustino Jorge S.M. (2003.), p ii.
[13] Nuryanto M., Nur Afiah N. (2013.), p 168.
[14] Peršić, M., Accounting inforamtion in management quality system, Praksa i daljnji razvoj urpavljanja kvalitetom / Bešker, Marko (ur.). - Zagreb : OSKAR, Centar za razvoj i kvalitetu , 2001. , p 259-267
[15] Availbale at Hongjiang Xu, M., Critical Success Factors for Accounting Information Systems Data Quality, https://eprints.usq.edu.au/1526/2/Xu_2003_whole.pdf p 29, (23. 02. 2015.)

The fundamental objective of the ISO/IEC 2510 standard is to address some of the well known human biases that can adversely affect the delivery and perception of a software development project. These biases include changing priorities after the start of a project or not having any clear definitions of "success."

By clarifying, then agreeing on the project priorities and subsequently converting abstract priorities (compliance) to measurable values (output data can be validated against schema X with zero intervention), ISO/IEC 2510 tries to develop a common understanding of the project's objectives and goals – as shown on Figure 2.[16] Also, it is important to mention that ISO/IEC 25012 contains a model for data quality that is complementary to this model .

In whole ISO/IEC 25010:2011 defines:

- A quality in use of model composed of five characteristics (some of which are further subdivided into subcharacteristics) that relate to the outcome of interaction when a product is used in a particular context of use. This system model is applicable to the complete human-computer system, including both computer systems in use and software products in use.
- A product quality model composed of eight characteristics (which are further subdivided into subcharacteristics) that relate to static properties of software and dynamic properties of the computer system. The model is applicable to both computer systems and software products.

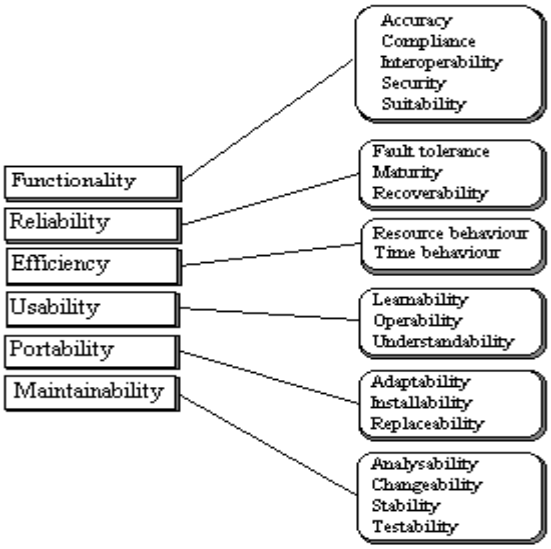


Figure 3. Software Quality Model (ISO/IEC 2510).[17]

[16] ISO/IEC 25010:2011Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality modelsat http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=35733 (21. 02. 2015.),

[17] Ibidem.

The characteristics defined by both models are relevant to all software products and computer systems. The characteristics and subcharacteristics provide consistent terminology for specifying, measuring and evaluating system and software product quality. They also provide a set of quality characteristics against which stated quality requirements can be compared for completeness.

Although the scope of the product quality model is intended to be software and computer systems, many of the characteristics are also relevant to wider systems and services as well at this example on accounting information system.

The main scope of the proposed model excludes purely functional properties, but it does include functional suitability. The scope of application of the quality models includes supporting specification and evaluation of software and software-intensive computer systems from different perspectives by those associated with their acquisition, requirements, development, use, evaluation, support, maintenance, quality assurance and control, and audit.

The importance of this model is that for example, it can, be used by developers, acquirers, quality assurance and control staff and independent evaluators, particularly those responsible for specifying and evaluating software product quality. Activities during product development that can benefit from the use of the quality models include:

- identifying software and system requirements;
- validating the comprehensiveness of a requirements definition;
- identifying software and system design objectives;
- identifying software and system testing objectives;
- identifying quality control criteria as part of quality assurance;
- identifying acceptance criteria for a software product and/or software-intensive computer system;
- establishing measures of quality characteristics in support of these activities.

These activities are of great importance for accounting information system so the suggestion will be that when modelling the AIS in local governments units it is proposed to be applied by the ISO/IEC 2510 procedures.

IV. INTERNAL CONTROL IN AIS

Internal control implies procedures „performed by management, executives and employees to ensure the provision of moderate guarantees with regard to the

achievement of objectives that can be grouped in the following categories:

- Efficiency and effectiveness of business,
- Reliability of financial statements,
- Compliance with applicable laws and other regulations."[18]

Control procedures imply methods for „recognition, prevention and elimination of irregularities that impede the fulfilment of the objectives of the enterprise", and the classification is as follows:[19]

- Preventive controls – they have the function of preventing irregularities,
- Detective controls – they are carried out with the purpose of analyzing deviations,
- Corrective controls – they are conducted after the irregularity has already emerged, with the aim of its removal and improving system functioning.

Control procedures are useful in all phases of accounting process:[20]

- Input – preventive control application when entering data,
- Processing – the control of the data and compliance with the relevant regulations,
- Output – checking the information through financial reporting.

Research has shown that, unlike positive impact of information technology on quality of financial statements of local government units, „internal control has no positive significant effect on the quality of financial reporting in work units of provincial government."[21]

In that area, the most famous global standard for exchanging business information is XBRL (eXtensible Business Reporting Language)[22]. It is freely available allows the expression of semantic meaning commonly required in business reporting. The language is XML-based and uses the XML syntax and related XML technologies such as XML Schema, XLink, XPath, and Namespaces. One use of XBRL is to define and exchange financial information, such as a financial statement.

As a proposed model, the positive aspects of XBRL is firstly the wiki repository of XBRL projects is available to be freely explored and updated.[23] Within the last fifteen years, the Securities and Exchange Commission (SEC) in the US, the United Kingdom's HM Revenue & Customs (HMRC), and Companies House in Singapore had begun to require companies to use it, and other regulators were following suit. The SEC's deployment was launched in 2008 in phases, with the largest filers going first: foreign companies which use International Financial Reporting Standards (IFRS) are expected to submit their financial returns to the SEC using XBRL once the IFRS taxonomy has been accepted by the SEC. In the UK in 2013, both HMRC and Companies House accepted XBRL in the iXBRL format. XBRL was adopted by the Ministry of Corporate Affairs (MCA) of India for filing financial and costing information with the Central Government. The only complaint on XBRL, is it's lack of accuracy. The study from April 2009, conducted by North Carolina State University Department of Accounting College of Management, which valuated the accuracy of XBRL filings for 22 companies participating in the SEC's voluntary filing program in 2006 showed it's lack of accuracy. Results of a comparison of XBRL filings to Forms 10-K revealed multiple errors in signage, amounts, labeling, and classification. The study considers that these errors are serious since XBRL data is computer-readable and users will not visually recognize the errors, especially when using XBRL analysis software. There was also a different conclusion, reached in a study in 2013 who argued that companies are going through a learning curve and are steadily improving.[24] However, the ocnlusion based on this studies, is that XBRL still needs improvements to be used as a recmmended model of financial data exchange used in the public, regional or local governments in the future.

CONCLUSION

Author's research and analysis security, control procedures and efficiency of accounting information systems and their use in regional and local Governments sector, has shown firstly that unlike positive impact of information technology on quality of financial statements of local government units, internal control has no positive significant effect on the quality of financial reporting in work units of provincial government. Also if we are addressing to the quality characteristics in accounting information systems, several high-level approaches to modelling of the quality of software products have been proposed: these models are based on the idea that there are a number of important high-level quality factors of software products that should be measured.

On the national level the future prospects for Croatia's strong recovery from recession and return to the pre-recession growth rates are slim unless the country vigorously moves forward in addressing the structural

[18] Mamić Sačer, I., Žager, K. (2007.), p.75.
[19] Ibidem, p.79.
[20] Ramljak, B.: „Interne računovodstvene kontrole u javnom sektoru", 15. Savjetovanje Interna kontrola i revizija, Zagreb, Hrvatska zajednica računovođa i financijskih djelatika, 2012., p 150
[21] Nuryanto, Nur Afiah, 2013., p169.
[22] XBRL International (XII) is the international standards organization which develops and maintains the XBRL Standard and related specifications[1] in order to to improve business reporting for the public good. It is a global not-for-profit consortium of approximately 600 public and private organisations working together to support the collection, sharing and use of structured data for data reporting and analysis. More inofrmation available at http://specifications.xbrl.org/ (15. 01. 2015.)
[23] XBRL projects available at http://www.xbrlwiki.info/index.php?title=Main_Page (15. 01. 2015.)
[24] Du, H., Vasarhelyi, M. A. & Zheng, X. 2013. XBRL Mandate: Thousands of Filing Errors and So What? Journal of Information Systems, 27(1), pp 61-78. doi: http://dx.doi.org/10.2308/isys-50399

weaknesses of its economy, according to the World Bank's new Public Finance Review: Restructuring Spending for Stability and Growth report. The global financial crisis, coupled with a delay in the implementation of structural reforms, exposed the weaknesses of the Croatian economy which was especially emphesized on the process of fiscal decentralization. The most important role in this area should be on the Croatian Ministry of Finance which should be give directives and standards in the Public Finances by analyzing the key issues that are standing in the way of strengthening macroeconomic stability and of laying the foundations for a robust economic recovery. However, on streamlining the fact is that government of Croatia spends more on public administration than the new EU member states. Also, it has relatively weak performance in terms of effectiveness, rule of law, and administrative barriers to doing business. The public wage bill is close to 12 percent of GDP, with a tendency to increase further due to further decentralization, creation of new agencies, and the need to build additional capacity to absorb large EU funds. So giving the directives in the area of fiscal reporting and making accounting information systems as a standard on regional and local level should help in process of fiscal decentralization as well as funding based on EU funds.

Until the unique model on national public level in the process of fiscal decentralization is formed (or proposed) the regional and local governments, after the author's analysis of the most important procedures in AIS based on security, quality assurance and control, the suggestion of the author's is modelling the AIS in regional and local governments units based on ISO/IEC 2510:2011 procedures. Also the situation with the regional and local government units is somewhat different because they aim to provide the best possible service to the community and, accordingly, public benefit maximization tendency exists unlike companies which aim to maximize owner's profit, the model of AIS based on procedures of ISO 2510 meets all of the needed requests especially the most important function of AIS: to support decision-making at all organizational levels within a local government unit s described in paper.

LITERATURE

[1] Bracci, E.: „Menagerialism, Accounting and Accountability int he Italian Local Governments: An Empirical Analysis", 2008., available at SSRN: http://ssrn.com/abstract=1184621

[2] Public finance review (Croatia) : restructuring spending for stability and growth (English) - document available at http://documents.worldbank.org/curated/en/2014/10/23014557/croatia-public-finance-review-restructuring-spending-stability-growth

[3] Du, H., Vasarhelyi, M. A. & Zheng, X. 2013. XBRL Mandate: Thousands of Filing Errors and So What? Journal of Information Systems, 27(1), pp 61-78. doi: http://dx.doi.org/10.2308/isys-50399

[4] Faustino Jorge S.M.: „Local Government Accounting in Portugal in Comparative-International Perspective", Thesis submitted to the University of Birmingham, 2003.

[5] Hongjiang Xu, M., Critical Success Factors for Accounting Information Systems Data Quality, https://eprints.usq.edu.au/1526/2/Xu_2003_whole.pdf

[6] ISO/IEC 9126-1:2001, Software engineering -- Product quality -- Part 1: Quality model, available at http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22749

[7] Mamić Sačer, I., Žager, K.: „Računovodstveni informacijski sustavi, Rif, Zagreb, 2007.

[8] McLeod and, R., H., Harun, H., Public Sector Accounting Reform at Local Government Level in Indonesia , Volume 30, Issue 2, John Wiley & Sons Ltd. p 238–258, May 2014.

[9] Meter M.: „Potreba intenzivnijeg korištenja i mogućnosti unapređenja računovodstveno informacijskih sustava u hrvatskim poduzećima", Ekonomski pregled, 57(7-8), 2006., 518-545,.

[10] •Muhrtala, T. O., Ogundej, M., Computerized Accounting Information Systems and Perceived Security Threats in Developing Economies: The Nigerian Case, Universal Journal of Accounting and Finance 1(1): 9 - 18, 2013.

[11] Nuryanto M., Nur Afiah N.: „The Impact of Apparatus Competence, Information Technology Utilization and Internal Control on Financial Statement Quality (Study on Local Government of Jakarta Province - Indonesia)", World Review of Business Research, Vol. 3. No. 4., November 2013 Issue., 2013. Pp. 157 – 171.

[12] Peršić, M., Accounting inforamtion in management quality system, Praksa i daljnji razvoj urpavljanja kvalitetom / Bešker, Marko (ur.). - Zagreb : OSKAR, Centar za razvoj i kvalitetu , 2001. , p 259-267

[13] Ramljak, B.: „Interne računovodstvene kontrole u javnom sektoru", 15. Savjetovanje Interna kontrola i revizija, Zagreb, Hrvatska zajednica računovođa i financijskih djelatika, 2012., str. 141. – 153.

[14] Vitasović M.: „Analiza stanja i doprinosa računovodstvenog informacijskog sustava i sustava upravljanja financijama u segmentu razvoja proračunskog sustava lokalnih jedinica, Ekonomska misao i praksa, Dubrovnik, god. XXI, 2012.,vol .2, pp 563. – 594., .

[15] Zenzerović, R.: „Računovodstveni informacijski sustavi", Sveučilište Jurja Dobrile, Odjel za ekonomiju i turizam, Pula, 2007.

[16] Wang R. Y., Strong, D.M., Beyond accuracy: what data quality means to data consumers, Journal of Management Information Systems archive, Volume 12 Issue 4, March 1996., pp 5-3

Citybus mobile application

Jasmin Abou Aldan, Mario Lončar, Marina Ivašić – Kos
Department of Informatics University of Rijeka, Rijeka, Croatia
{jasmin.abou.aldanm, mario.loncar, marina}@uniri.hr

Abstract – Students, tourists and citizens are often confused by bus lines and timetables or get tired of running to catch their bus or looking for bus stops. This paper describes the mobile application that conveniently provides information about city bus lines. For a selected line, it can gives an overview of timetables, bus stop locations and estimated departures. It also calculates the shortest route from user location to a certain bus stop and can navigate the user to that bus stop. The prototype application is made for local bus lines in Rijeka.

Keyword: iOS application, navigations, maps

I. INTRODUCTION

The idea for this mobile application came in 2012 when we noticed a need for a better access to information about public bus transportation among university freshmen who came to the city for the first time, as well as among the citizens of Rijeka. Until then, available information about the bus lines were shown only on bus stations and on official website of the transporter, but information about exact locations of all stations was not available anywhere.

That kind of arrangement was rather impractical because to know when a bus would arrive to a particular station, one would have to be at that exact station and the problem gets bigger if one does not even know where the station is. We were encouraged to start this project by a frequent sight when people run after or miss a bus because they do not know the exact time when it will arrive at a station. Therefore, we started making various sketches and ideas to solve this problem.

In earlier versions, the application was reading a simple JSON format from Parse [1] because it was easier way to manipulate with data about local bus lines and stations than with the PDF document that was available on the official websites. The next functionality added in the first version of the application were coordinates of all the bus stations and the possibility to start a GPS navigation to a selected station.

II. APPLICATION DESCRIPTION

The main functionality of this application is to give information about local bus lines e.g., directions, locations of bus stations and timetables. To be as intuitive and easy to use as possible, the main applications screen contains a menu in which user chooses whether he needs information about locations of the bus stations or about the timetables (Fig. 1).

Figure 1: Main Screen

If a user selects the Stations button, a screen with all available bus stations will appear (Fig. 2). At first, all of the stations are shown on the map in form of pins, and the user can zoom the map to have a closer look at bus stations in the area of interest. When a particular station is selected, the information box appears on the screen and the user can see which busses pass through that station.

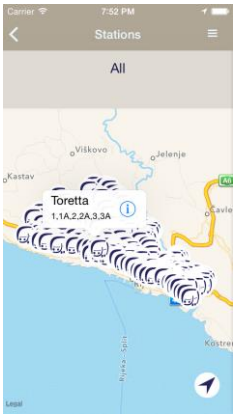
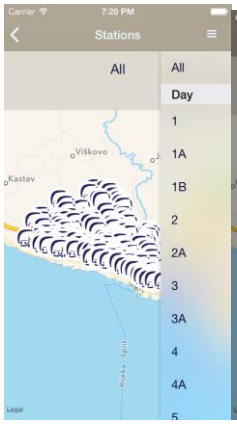


Figure 2: All Bus Stations

User can also select a side menu item to view all available local bus lines, e.g. 1, 1A, 1B, 2, 2A, etc. (Fig. 3). If the user needs all information about a particular bus line, he can select that line on a given list. Then, available directions will be shown in an alert view (Fig. 4). For each bus line, there are two directions: A direction, from the first to the last station and B in the opposite direction, from the last station to the first one.



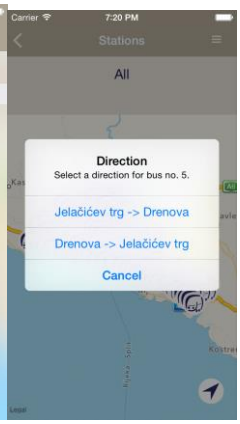


Figure 3: Side Menu Figure 4: Direction Alert

Then, for a given bus line and a selected direction, all stations will be indicated on the map (Fig. 5).

Additionally, by selecting a specific station, the user can start a GPS navigation by Apple Maps application (Fig. 6). Only navigation on foot is relevant, so for that navigation type several different routes are offered and time necessary to reach the station, from the current user location, is calculated.

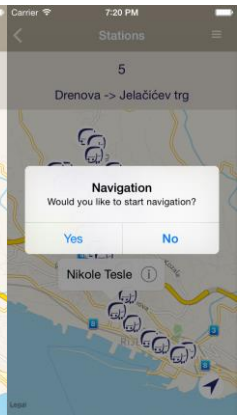


Figure 5: Single Line Figure 6: Navigation Alert

Figure 7 shows the selected route from the user location to the location of the selected station. By default, the walking route is selected, but the user can choose whether he/she wants the pedestrian or car route.

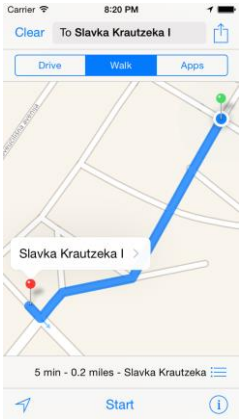


Figure 7: Navigation

For a selected route in the Apple Maps application, user has the ability to assign that route to any contact or to send it via SMS, e-mail, Facebook, etc. (Fig. 8).

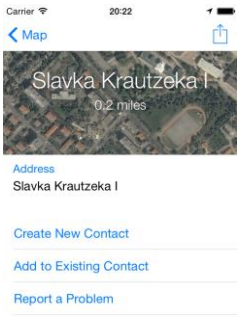


Figure 8: Navigation options

III. IMPLEMENTATION

A. General

The platform used to build the application is Apple iOS. The key feature of this application is manipulation with the maps service that provides a "turn-by-turn" navigation option for cars as well as pedestrians [2].

Both, Apple maps and Google maps can be embedded into an iOS application. Although Google maps where until recently the main maps used in iOS applications, we have decided to use Apple maps because of up-to-date application programming interface (API) and good support for iOS platform. Furthermore, it was shown that Apple maps work faster on the iOS platform than Google maps. However, as Apple maps were recently introduced, Google maps have better and more developed cartography. Since the performance was more important than accurate cartography, Apple maps were embedded in this application.

For the applications backend, Parse iOS software development kit (SDK) [1] is used. Parse iOS SDK is primary used to store data about the bus lines, directions and stations. It has proven to be a stable data storage that

can be easy accessed. Available functions for sending push notifications and statistics are implemented too.

The programming language used to develop this application is Swift [3]. It is a relatively new programming language introduced by Apple in 2014 [4] as a replacement to the former Objective-C language.

The Classes used within the application are part of the Cocoa Touch [5], a frameworks collection for creating native iOS applications. Used frameworks in this application are FoundationKit [6], UIKit [7] and MapKit [8].

Framework FoundationKit is a core framework in which classes for manipulation with strings (Class NSString), time (Class NSTime), numbers (Class NSNumber), arrays (Class NSArray), etc. are defined. All these classes are subclasses of the abstract class NSObject, and inherit their properties. Hierarchy of mentioned classes is shown on the Fig. 9.

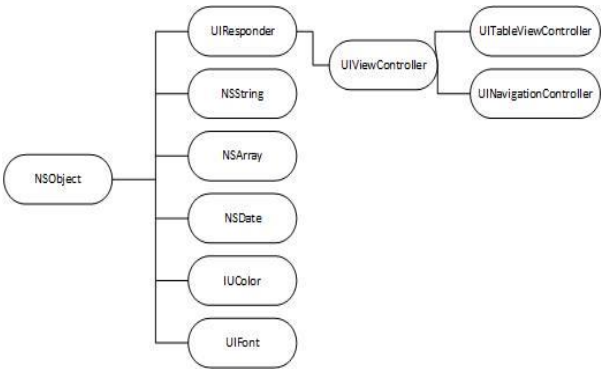


Figure 9: Class flowchart

Framework UIKit [7] contains all classes that are used to build user interface. This framework provides the window and view architecture with a view controller model to be able to encapsulate the contents of a user interface, event-handling infrastructure needed to respond to user inputs, and the application model needed to interact with the system.

To embed a fully functional map interface into an application and for working with maps, classes within MapKit framework [8] are used. The map support provided by this framework includes standard street-level map information or satellite imagery. In addition, automatic support for the touch events that let users zoom and pan the map display is provided. CoreLocation API [9] is used for locating current user position and for GPS navigation.

B. Maps

To represent distribution of bus station on the map and for side menu creation and filling with bus lines three classes are instantiated and overridden: *MapsViewController*, *SideMenu* and *SideMenuTableViewController*.

The *MapViewController* is the interface used to access the basic map services. It creates a map instance that can be displayed using methods inherited from its superclass *UIViewController*[10].

To manipulate with the map user can use finger gestures on the screen but map can also be manipulated by using methods for setting the display type, zoom level or a specific angle.

For example, the code snippet to set a standard map type with the possibility of 3D display and pitch is:

```
mapView.mapType = MKMapType.Standard
mapView.showsBuildings = true
mapView.pitchEnabled = true
```

First zoom is set to the Rijeka city. To specify a map data points that correspond to Rijeka, a latitude and the longitude coordinates of Rijeka on the spherical representation of the Earth is used to set values of the structure *CLLocationCoordinate2DMake*:

```
let zoomLocation = CLLocationCoordinate2DMake (45.337596, 14.426216)
```

Current zoom level is defined with structure MKCoordinateSpan where a smaller delta values correspond to a higher zoom level. For example, the amount of north-to-south distance and east-to-west distance (measured in degrees) that will be display on the map are set to 0.15 with a following code line:

```
let span = MKCoordinateSpanMake(0.15, 0.15)
```

To define and represent areas of interest around a given location (here the *zoomLocation*), with desired zoom level (here the *span*) a data in structures *MKCoordinateRegion* is set as follows:

```
let region = MKCoordinateRegion(center: zoomLocation, span: span)
mapView.setRegion(region, animated: true)
```

The pin that represents bus station is changed from standard red pin to a custom one, e.g. an image of a small bus:

```
pinView.image = UIImage(named:"pinstation.png")
```

Two functions of a *MapsViewController* class are used for customising the map. That is a *sideMenuDidSelectButtonAtIndex* that recognizesrecognizes which button is selected at the side menu and uses its index to call a function *makePins* responsible for drawing pins on the map for a selected bus line. Signature of function *makePins* is:

```
func makePins (coordinates1: Array <String>, coordinates2: Array <String>, linNumb: String, direction: Array<String>) { //code}
```

A *makePins* function opens a pop-up alert view *alertView* in which user can choose the direction of the selected bus line. For example if a direction A is selected, then, for every list item *el* in array *self.direction1*, function reads data and draws pins on the map. Data about bus lines and corresponding directions and bus stops are stored in an external database .plist [11] (a database similar to XML) in form of key – value data.

```
alert.addAction (UIAlertAction (title: dirA,
style:.Default, handler:
{action in
    for el in self.direction1 {
      let lat = tmp[0] as NSString
      let long = tmp[1] as NSString
      let name = tmp[2]
      let stations = MKPointAnnotation()
      let station_coordinates =
    CLLocationCoordinate2DMake(lat,long)

stations.setCoordinate(station_coordinates)
       stations.title = name

    self.mapView.addAnnotation(station1)
      }}))
```

Similar code is defined for the direction B, only difference is array self.direction2 used for data about stations in directions B.

```
alert.addAction(UIAlertAction(title: dirB,
style:.Default, handler: {action in
   for el in self.direction2 {
       //the same code
   }
}))
```

C. Navigation

After pins are created and shown on the map, the user can activate GPS navigation to the selected station.

For navigation, application uses MKDirections API within the Map Kit framework. MKDirections API provides route-based directions data from Apple servers and contains step-by-step walking or driving directions, travel time estimation, alternate routes etc.

When navigation is called, the application calls *provideDirections* function within which the part of the code responsible for navigation is executed. This function takes the user position and the position of the destination (selected bus station), generates direction request and sends it to the Apple servers where the route will be calculated.

```
let placemark = placemarks[0] as CLPlacemark
let destinationCoordinates =
placemark.location.coordinate
directions.calculateDirectionsWithCompletion
Handler {(response: MKDirectionsResponse!,
error: NSError!) in //code for starting the
navigation}
```

App server makes the calculations and sends them back to the application. The application should be able to accept directions response and should declare a document type to handle incoming routing requests. Then, Apple maps application opens and the navigation starts.

```
MKMapItem.openMapsWithItems([response.destin
ation], launchOptions: launchOptions)
```

D. Menus

Data about bus lines are shown on the side menu. To define menu, its size, position, colour and functionality Class *SideMenu* is used. The width of the menu is set to 100 pt (point). If the menu is opened, then the starting point of the menu is the width of the screen reduced by the width of the menu. If it is closed, then the initial point is on the edge of the screen increased by 1 pt to be sure that it does not appear on the screen.

SideMenu is a subclass of the *NSObject* [12] class. It uses *SideMenuTableViewControllerDelegate* that communicates with *SideMenuTableViewController* in order to fill out the menu with data. Data used to fill out the menu were forwarded with *MapsViewController* from .plist database.

The *SideMenuTableViewController* forwards the chosen index to the *SideMenu* class, which forwards it later to the main *MapsViewController* class.

MapsViewController class is a subclass of the *UITableViewController*[13] class that is usually used to display table data in different views.

For communication with a *SideMenu* class a *SideMenuTableViewControllerDelegate* protocol is used (Fig. 10).

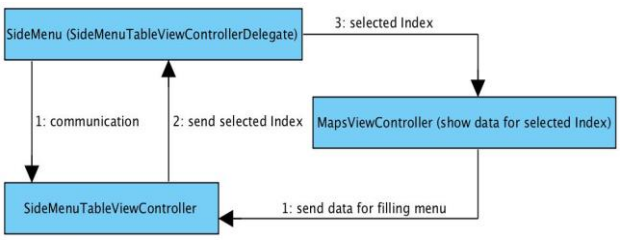


Figure 10: Communication flowchart

IV. CONCLUSION

The mobile application that conveniently provides information about city bus lines is displayed.

The application is available in beta test version for iOS, and it will soon be available for Android platform.

In the first upgrade, the number of bus lines will be expanded with nightlines and with suburb lines. In addition, for a chosen bus station estimated time needed to reach the station on foot from the user current location will be calculated as well as the calculation of closest bus departure from that station. As a new functionality, the application will deliver the latest news from the public transportation web site services that will be available in form of push notifications.

The application can be modified to provide useful information for other means of public transport and other transportation service providers in Croatia, and beyond.

ACKNOWLEDGMENT

On this occasion, I would like to thank my brother Damjan Abou Aldan for his contribution in creating design for the User Interface of the application.

REFERENCES

[1] Parse.com; Parse Core; https://www.parse.com/products/core; (Accessed 2014-06)

[2] Apple Inc.; About Location Services and Maps; Updated: 2014-03-10; https://developer.apple.com/library/ios/documentation/UserExperience/Conceptual/LocationAwarenessPG/Introduction/Introduction.html; (Accessed 2014-06)

[3] Apple Inc.; The Swift Programming Language (Book 1,Swift Programming Series);2014-06-02; https://itunes.apple.com/us/book/swift-programming-language/id881256329?mt=11; (Accessed 2014-06)

[4] Apple Inc.; Introduction to Swift; https://developer.apple.com/videos/wwdc/2014/#402; (Accessed 2014-06)

[5] Apple Inc.; Cocoa Touch Framworks; https://developer.apple.com/technologies/ios/cocoa-touch.html; (Accessed 2014-06)

[6] Apple Inc.; The Foundation Framework; 2013-10-22; https://developer.apple.com/library/mac/documentation/Cocoa/Reference/Foundation/ObjC_classic/index.html; (Accessed 2014-07)

[7] Apple Inc.; The UIKit Framework Reference; 2014-09-17; https://developer.apple.com/library/ios/documentation/UIKit/Reference/UIKit_Framework/; (Accessed 2014-07)

[8] Apple Inc.; The MapKit Framework Reference; 2013-09-18; https://developer.apple.com/library/ios/documentation/MapKit/Reference/MapKit_Framework_Reference/; (Accessed 2014-08)

[9] Apple Inc.; The Core Location Framework Reference;2013-09-18; https://developer.apple.com/library/ios/documentation/CoreLocation/Reference/CoreLocation_Framework/; (Accessed 2014-08)

[10] Apple Inc.;UIViewController Class Reference; 2014-09-17; https://developer.apple.com/library/ios/documentation/UIKit/Reference/UIViewController_Class/ ; (Accessed 2014-08);

[11] Joe Hoffman; Using Propery List (.plist); 2013-02-14; http://nscookbook.com/2013/02/ios-programming-recipe-13-using-property-lists-plists/; (Accessed 2014-07)

[12] Apple Inc.;NSObject Class Reference;2013-12-16; https://developer.apple.com/library/mac/documentation/Cocoa/Reference/Foundation/Classes/NSObject_Class/index.html; (Accessed 2014-06)

[13] Apple Inc.;UITableViewController;2015-01-10; https://developer.apple.com/library/prerelease/ios/documentation/UIKit/Reference/UITableViewController_Class/index.html; (Accessed 2015-01)

Analysis and recommendations for e-business development in Croatia

Ranko Smokvina [1], Sasa Aksentijevic [2], Edvard Tijan [3]
[1] InfoExpert, Ltd.
Lorenzov prolaz 13, Rijeka, Croatia
Tel. +385 98 293660 E-mail: infoexpert@ri.t-com.hr
2 Aksentijević Forensics and Consulting, Ltd.
Gornji Sroki 125a, Viškovo, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr
[3] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr

Abstract – In this paper, the development of e-business in Croatia in the last decade has been researched, with special emphasis on legislation, organization and state influence, along with the influence of other relevant and potential stakeholders. The influence of Croatian inclusion into the European Union has been analyzed, both before and after the inclusion in 2013. The main directions of e-business development in Croatia are presented. Main segments of e-business have been analyzed (b2b - business-to-business, b2g - business-to-government and c2g - citizen-to-government). EU DESI, UN e-Government Survey, e-Participation Index (EPI), World Economic Forum (WEF) data, UN Open Government Data (OGD) and other relevant indices have been studied and interpreted. In the closing part of the paper, guidelines for development of e-business in Croatia are presented along with author conclusions, specifying e-business as a key element for the evolution of digital economy and digital society.

I. INTRODUCTION

The research started with three different relevant resources for analysis and recommendations of e-business development in Croatia:

- e-Domains observed in the paper: B2B & B2G, as being the most socially relevant,
- Nicholas Negroponte‘s vision outlined in the 1995 book: Being Digital, and
- EU 2010-2020 Digital Agenda for Europe.

Croatia has joined EU in the summer of 2013 and the object of our research is the achieved level of adoption of e-business and e-government paradigms one and a half years after the accession. To understand the current state of e-business in Croatia, available and published indices from relevant organizations have been identified, studied thoroughly, explained and conclusions were drawn based on them.

Finally, recommendations are given on how to improve the current situation in form of exact steps to be taken after identifying the main involved stakeholders.

II. ASSESSMENT METHODOLOGY OF E-BUSINESS DEVELOPMENT IN CROATIA

Digital economy and society results presented in the paper are assessed by using the latest indices and reports from reputable and up to date sources. These sources include the following:

- EU: Digital Economy & Society Index 2015 [1],
- UN: E-Government results assessed according to UN E-Government Survey 2014 [2],
- World Economic Forum: Global Competitiveness Report 2014-2015 [3],
- World Economic Forum: Global Information Technology Report 2014 [4].

III. RECENT DEVELOPMENTS IN E-BUSINESS AND E-GOVERNMENT AREAS IN CROATIA

In order to analyze various facets of achieved level of implementation of e-Business and e-Government systems in Croatia, selected aspects are presented including the most recent developments related to the implementation of such systems.

National PKI system provided by Financial Agency – FINA [5] is being used by public and private users. FINA is the only issuer of digital certificates in Croatia. Citizens are generally not using FINA certificates to access e-services. The eID project, which was promised by Ministry of Interior to include certificates for digital identification and signature is still not on the horizon. FINA is also serving some important applications related to e-Government and b2g, among them most notably HITRO.HR (Croatian Government’s service for quick communication between individuals and business entities and the state administration), Central Registry of Insured Persons (REGOS), Public Sector Employees Registry (RegZap), Blue Diesel Consumption Supervisory System, Registry of Court and Notary Public Securities, HITRONet end e-File.

The e-Citizen system [6] promoted by Ministry of Administration is in its initial phase, offering several services to citizens and business. Since there is no widespread national certification mechanism available, e-Citizen system is trying to temporarily fill the gap by using its own hybrid NIAS (*National Identification and Authentication System*) [7] consisting of combination of several local and partial identification and authentication solutions. E-Citizen is not expected to show its full potential until the 'real' NIAS is in operation and when a fully expanded set of useful and practical services is offered to citizens and businesses.

GS1 organization in the Republic of Croatia [8] is, among other activities, active in providing its members with electronic cataloguing GlobeCat service which is, thru built-in global synchronization network, offering far reaching possibilities for Croatian GS1 member companies.

The Eurostat eInvoicing data for Croatia's eInvoicing penetration in 2014 is about 10% which is, according to authors' opinion, also a statistically optimistic result. Slow take-up of eInvoicing in Croatia, in both public and private sector is a good indicator of the whole eBusiness implementation along both the supply and financial chains. Some good examples in supply-chain automation [9] clearly show what results could be reached on the larger scale if more companies in Croatia adopt similar development path. Croatia is lagging in adoption of eInvoice, being included into the lowest <15% penetration group [10].

IV. OVERVIEW OF EU DESI 2015. (DIGITAL ECONOMY AND SOCIETY INDEX)

On 24th of February 2015, EU has published the results of the new Digital Economy and Society survey with the appropriate scoring and rankings among 28 EU member states. All data contained in this paragraph is derived from the data presented in that survey [11]. Here the authors reproduce some of data presented by the EU. It should be noted that Croatia has been analyzed by the EU as a member of the low-performance cluster of countries.

DESI survey scores range from 0 to 1, the higher the score the better the country performance. Structure of DESI indices and composition factors related to principal policy area, sub dimension weights, indicators, units and ranges are shown in Table 1.

TABLE I. STRUCTURE OF DESI INDICES AND COMPOSITION FACTORS

Principal policy area (weight %)	Sub-dimension (weight %)	Indicator	Unit	Min	Max
Connectivity (25%)	Fixed BB (33%)	Fixed broadband coverage	% households	80%	100%
		Fixed broadband take-up	% households	50%	100%
	Mobile BB (22%)	Mobile broadband take-up	Subscribers per 100 people	25%	150%
		Spectrum	% of the target	25%	100%
	Speed (33%)	Next Generation Access coverage	% households	0%	100%
		Subscriptions to fast broadb.	% of subscriptions >=30Mbps	0%	75%
	Affordability (11%)	Fixed broadband price	% of individual gross income (EUR PPP)	0%	4%
Human Capital (25%)	Basic skills and usage (50%)	Internet users	% individuals	40%	100%
		Basic digital skills	% individuals	0%	100%
	Advanced skills and development (50%)	ICT specialists	% individuals	0%	5%
		Science, Technology, Engineering, Math graduates	Graduates in STEM per 1000 individuals (20-29 year old)	0	40
Use of internet (15%)	Content (33%)	News	% individuals who used internet in the last 3 months	33%	100%
		Music, videos and games	% individuals who used internet in the last 3 months	20%	100%
		Video on demand	% households with a TV	0%	100%
		Internet Protocol TV	% households with a TV	0%	75%
	Communication (33%)	Video calls	% individuals who used internet in the last 3 months	20%	100%
		Social networks	% individuals who used internet in the last 3 months	40%	100%
	Transactions (33%)	Banking	% individuals who used internet in the last 3 months	0%	100%
		Shopping	% internet users (last year)	0%	100%
Integration of digital technology (20%)	Business digitalization (60%)	Electronic information sharing	% enterprises	0%	50%
		RFID (Radio Freq. IDentif.)	% enterprises	0%	15%
		Social media	% enterprises	0%	50%
		eInvoices	% enterprises	0%	50%
		Cloud	% enterprises	0%	50%
	eCommerce (40%)	SMEs selling online	% of SMEs	0%	33%
		eCommerce Turnover	% turnover	0%	33%
		Selling online cross-border	% of SMEs	0%	25%
Digital public services (15%)	eGovernment (67%)	eGovernment users	% of internet users (last year)	0%	80%
		Pre-filled forms	score (0 to 100)	0	100
		Online Services completion	score (0 to 100)	0	100
		Open data	score (0 to 700)	0	700
	eHealth (33%)	Medical data exchange	% general practitioners	0%	100%
		ePrescription	% general practitioners	0%	100%

In the DESI 2015 survey, low-performing cluster of countries comprises of Bulgaria, Cyprus, Greece, Croatia, Hungary, Italy, Poland, Romania, Slovenia and Slovakia. As shown in Fig. 1, Croatia ranks better than EU's average in integration of digital technology. Croatia ranks better than low-performance cluster of countries in digital public services and human capital.

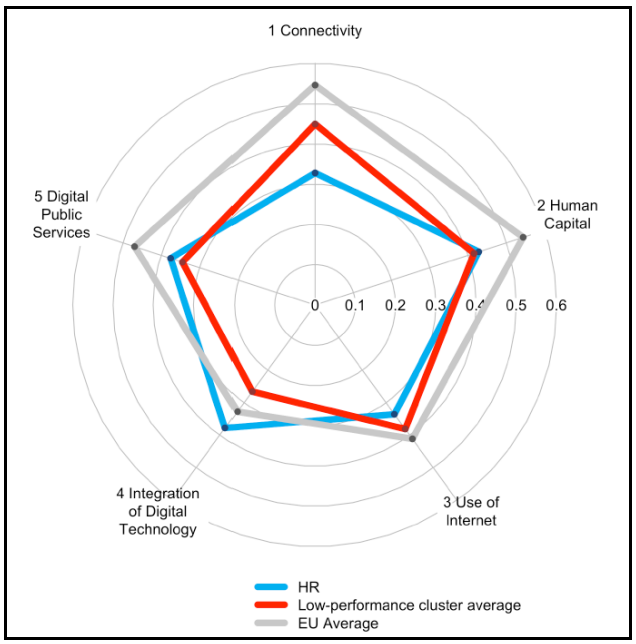


Figure 1. Croatia's DESI 2015 overall score

Croatia has a weaker rank than the representative cluster in use of Internet and connectivity, where it has the worst rank among all EU member states.

In the DESI 2015 survey, Croatia has achieved an overall score of **0.37** (the same as the previous year) and ranks **24th** (three places down compared to the previous year) out of the 28 EU Member States. Croatia falls into the low-performance cluster, ranking below EU's average (Fig. 2).

	Croatia		Cluster	EU
	Rank	Score	score	score
DESI 2015	24	0.37	0.38	0.47
DESI 2014	21	0.37	0.34	0.44

Figure 2. Croatia's DESI 2015 ranking within the EU

Data, ranks and scores related to different domains (Connectivity, Human Capital, Use of Internet, Integration of Digital Technology, and Digital Public Services) are shown in Table II.

TABLE II. CROATIA'S DESI 2015 RANKING COMPARED TO PREVIOUS YEAR, CLUSTER SCORE AND EU SCORE

		Croatia		Cluster	EU
		rank	score	score	score
1. Connectivity	**2015**	**28** ↓	**0,33** ↓	**0,45**	**0,55**
	2014	26	0,36	0,39	0,51
2. Human Capital	**2015**	**21** ↓	**0,43** ↓	**0,42**	**0,54**
	2014	20	0,45	0,38	0,52
3. Use of Internet	**2015**	**25** ↓	**0,34** ↓	**0,38**	**0,41**
	2014	22	0,35	0,33	0,39
4. Integration of Digital Technology	**2015**	**10** ↑	**0,38** ↑	**0,27**	**0,33**
	2014	17	0,31	0,23	0,30
5. Digital Public Services	**2015**	**19** ±	**0,38** ↓	**0,34**	**0,47**
	2014	19	0,36	0,33	0,45

In the year 2015 Croatia has achieved better score than in the previous year in Integration of digital technology and Digital public services, but other three DESI scores are lower than in the previous year. Better rank in the year 2015 than in the previous year Croatia obtained only in Integration of digital technology (+7), mainly because the e-Prescription service (ranking 3rd in EU, only behind Estonia and Denmark) entered in full operation and received maximal scores. All other rankings are lower than previous year, except for Digital public services where the score remained the same.

EU DESI data for the year 2015 are published accompanied with a very detailed Methodological note (68 pages) [1]. In addition to that, EU also provides interested researchers with raw data which can be used for further deep analysis of digital phenomena in EU member states, and EU as a whole.

As shown in Figure 2, in DESI 2015 Croatia ranks at 24th place among 28 EU member states. Denmark is at the first place with the DESI score of 0,68 and Romania is at the last place with the DESI score of 0,31.

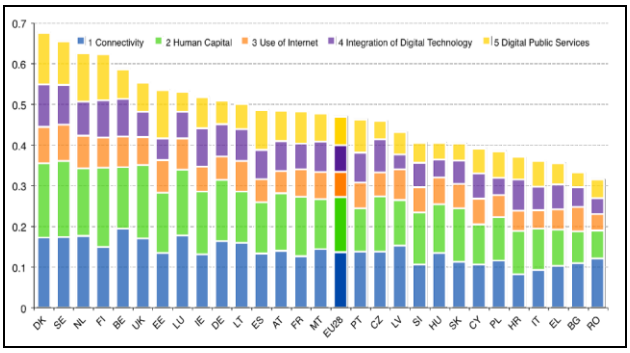


Figure 2. DESI 2015 EU main rankings

V. UNITED NATIONS E-GOVERNMENT SURVEY 2014

The UN has after two years published its new eGovernment Survey in New York in 2014 [2], so it is possible to compare similar data for the year 2012. The survey includes the Annexes with detailed survey methodological manual describing formulas used to calculate all relevant indexes. Croatia was ranked 47th in the survey, with EGDI (E-Government Development

Index) of 0,6282 (17 places lower than two years before when Croatia was ranked 30th). The survey, among other things, states:

"The small proportion of high scores in the Online Service Index (OSI), for example in the range of 0.7 and above, suggests that some governments with access to more advanced telecommunication infrastructures may be able to leverage existing technology more fully in public service delivery, especially in cases where human capital is high. To illustrate, consider on the one hand the case of Republic of Croatia that is well advanced in human capital and telecommunication infrastructure, but has an Online Service Index that is less than 0.5. On the other hand, consider Uruguay which has the same size and similar income levels to those of Croatia, as well as similar Human Capital Index (HCI) and Telecommunication Infrastructure Index (TII), but has invested more in online services, which is reflected in its OSI (0.8504) as shown in Figure. This shows that Republic of Croatia has great potential to improve its online services."

According to that, the comparison between two quite similar states, Uruguay and Croatia, shows some interesting results, as shown in Figure 3.

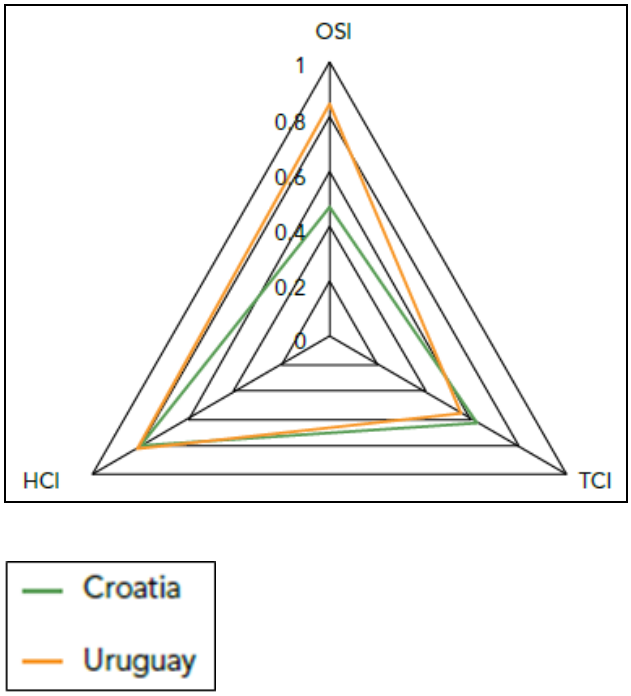


Figure 3. OSI, HCI and TCI comparison between Croatia and Uruguay

It is important to note that eGovernment eCentral State Office for eCroatia was three year ago transferred into a Ministry of Public Administration by a newly appointed government, becoming the Department for eCroatia.

Some more information from the UN E-Government 2014 survey can be extracted from component indices shown in the following figures.

	1.Online Service Component (OSI)	2. Telecomm. Infrastructure Component (TII)	3. Human Capital Component (HCI)
EGDI = 0,6282	**0,4646**	**0,6271**	**0,7928**

Figure 4. E-Government Development Index (EGDI)

	Stage 1 Emerging presence	Stage 2 Enhanced presence	Stage 3 Transactio-nal pres.	Stage 4 Networked (connected) presence
OSI = 0,4646	**59%**	**68%**	**26%**	**18%**

Figure 5. Online Service Index (OSI)

	% of Internet Users	Fixed telephone Subscr. per 100 inhab.	Mobile-Cellular Subscr. per 100 inhab.	Fixed BB Subscr. per 100 inhab.	WIFI BB Subscr. per 100 inhab.
TII = 0,6271	**63**	**38,07**	**115,41**	**20,67**	**53,9**

Figure 6. Telecommunication Infrastructure Index (TII)

	Adult literacy (%)	Gross enrolment ratio (%)	Expected years of schooling	Mean years of schooling
HCI = 0,7928	**98,99%**	**81,8%**	**14,11**	**9,8**

Figure 7. Human Capital Index (HCI)

Survey methodology for 2014 states: *„For the 2014 Survey questionnaire, the thematic subthemes identified are:*

- *Whole-of-government;*
- *Multichannel service delivery;*
- *Bridging the digital divide;*
- *Increasing usage;*
- *Open Government;*
- *E-participation."*

E-Participation Index (EPI) is derived as an important supplementary index to the UN E-Government Survey. It extends the dimension of the Survey by focusing on the use of online services to facilitate provision of information by governments to citizens ("*e-information*

sharing"), interaction with stakeholders ("*e-consultation*") and engagement in decision-making processes ("*e-decision making*"). Explanation of the principle of *E-Participation Framework* is the following*:*

• **E-information:** Enabling participation by providing citizens with public information and access to information without or upon demand.

• **E-consultation:** Engaging citizens in contributions to and deliberation on public policies and services.

• **E-decision-making:** Empowering citizens through co-design of policy options and co-production of service components and delivery modalities.

	Total %	Stage 1 % e-Information	Stage 2 % e-Consultation	Stage 3 % e-Decision-making
EPI = 0,333	**31,03**	**44,44**	**27,27**	**0**

Figure 8. E-Participation Index (EPI)

Regarding e-Participation, Croatia was not ranked among the top 50 global performers (ranked 97th) by the UN E-Government Survey 2014. Finally, the same is also valid for UN Open Government Data (OGD). According to UN's definition of OGD, *"data is considered open when it is shared with an open license in a way that permits commercial and non-commercial use and reuse without restrictions."* Furthermore, Croatia was not ranked among the top 50 global performers regarding the usage of Open Government Data by the UN E-Government Survey 2014. However, a large step towards data opening by the Croatian government was done on March 19th 2015 when *data.gov.hr* public portal was open. This portal will serve as a data hub for collecting, categorizing and distributing open data of the public sector in Croatia.

VI. WORLD ECONOMIC FORUM DATA

World Economic Forum data are sourced from two WEF reports: Global Information Technology Report 2014 [4], and Global Competitiveness Report, 2014-2015 [3]. As shown in Fig. 9, Croatia's Networked Readiness Index (NRI) for 2014 was **4,3** and it was ranked 46th out of 148 countries. For comparison, Croatia's NRI for 2013 was **4,2**, ranking it 51st out of 144 surveyed countries.

sub index	pillar	value (1-7)	rank (out of 148)
Environment	Political and regulatory	3,5	88
	Business and innovation	4,7	44
Readiness	Infrastructure and digital content	4,6	54
	Affordability	6,0	36
	Skills	5,6	36
Usage	Individual	5,0	39
	Business	3,5	81
	Government	4,1	65
Impact	Economic	3,6	40
	Social	3,8	66

Figure 9. WEF Networked Readiness Index (NRI) data

WEF Global Competitiveness Index (GCI) is shown in the Fig 10. According to GCI 2014-2015 index, Croatia's value is **4,1**, ranking it 77th out of 144 surveyed countries.

		score (1-7)	rank (out of 144)
Basic requirements	Institutions	3,6	87
	Infrastructure	4,7	44
	Macroeconomic environment	4,4	91
	Health and primary education	5,9	60
Efficiency enhancers	Higher education and training	4,8	53
	Goods market efficiency	4,1	105
	Labor market efficiency	3,9	106
	Financial market development	3,9	74
	Technological readiness	4,6	44
	Market size	3,6	79
Innovation and sophistication factors	Business sophistication	3,8	83
	Innovation	3,1	93

Figure 10. WEF Global Competitiveness Index (GCI) 2014-2015

For comparison, the data for the previous three year periods are shown here:

1. GCI 2013-2014 = 4,1 - ranking 75th (out of 148)
2. GCI 2012-2013 = 4,0 - ranking 81st (out of 144)
3. GCI 2011-2012 = 4,1 - ranking 76th (out of 142)

VII. CONCLUSION

Being late is not a good strategy in e-Business and e-Government, and it seems that Croatia is at least on the brink of being late according to researched and described indices.

On the whole, at the current level of the e-Business adoption Croatia is suffering from the following issues:

- Poor eRecords,
- Lacking and late e-Legislation (no national initiatives),
- Inadequate organization and management of national e-Bodies,
- Poor e-Governance,
- Automation of business processes is not included in the National Industrial Strategy,
- There are no national standards for e-business.

Development of e-business and its components, especially Internet and communications are stagnating in Croatia, which places Croatia at the very bottom of the EU and at the lower end of the first third of surveyed countries according to UN's methodology. There is no significant advance in ranking in the past years.

There are several very noticeable advances in implementation of some e-government services that are contributing to slight elevation of ranking (notable example is e-prescription, where Croatia is among European leaders, and data openness promoted by *data.gov.hr* portal).

However, the authors recognize the lack of government involvement to coordinate all e-business and e-Government effort. Creation of a new central governmental body (independent of specific ministerial domains, responsible for Digital Croatia strategic planning), and co-ordination between Government agencies and Industry (both public and private) is highly recommended and seems to be a prerequisite for further advances in Croatia's digital position.

The person appointed as head of the body for Digital Croatia should be a high positioned official at the level of State Minister (as in UK, for example) or minister for digital economy (as in France) or even two persons (like commissioners in EC), each responsible for digital economy and society and separately for digital market. No more 'silo' organizations should be allowed.

New and comprehensive national digital strategy should be proposed by the government and approved in Croatian Parliament, with clear goals to be reached by 2020, according to national and EU priorities.

Financial push should be given for research and development of digital competencies and knowledge through grant initiatives. Tax-reductions are highly advised for enterprises adopting new e-business practices. Available EU funding should be used in preparations, and maximal use of digital projects and programs should be endorsed in addition to own funds and grants.

In order to enhance the penetration of e-technologies in general population, national digital literacy program for "everybody to be included everywhere" should be devised, including but not restricted to the new European Computer Driving License.

REFERENCES

[1] European Commission, Directorate-General for Communications Networks, Content and Technology, DESI 2015 Digital Economy and Society Index, Methodological note (accessed March 19th 2015)

[2] unpan3.un.org/egovkb/Reports/UN-E-Government-Survey-2014 (accessed March 19th 2015)

[3] Global Competitiveness Report, 2014-2015, World Economic Forum, http://www.weforum.org/reports/global-competitiveness-report-2014-2015 (accessed March 19th 2015)

[4] The Global Information Technology report 2014, World Economic Forum, http://www3.weforum.org/docs/WEF_GlobalInformationTechnology_Report_2014.pdf (accessed March 19th 2015)

[5] http://www.fina.hr/Default.aspx?sec=1152 (accessed March 19th 2015)

[6] https://gov.hr/e-gradjani/23 (accessed March 19th 2015)

[7] https://nias.gov.hr/Home/TermsOfUse (accessed March 19th 2015)

[8] http://www.gs1hr.org/djelatnosti/razmjena/gdsn (accessed March 19th 2015)

[9] http://www.agrokor.hr/hr/ (accessed March 19th 2015)

[10] Koch, B: "The impact of B2G E-Invoicing to the whole economy", 13th conference eBiz 2014, Zagreb, Croatia, March 24th 2014

[11] http://ec.europa.eu/digital-agenda/en/digital-agenda-scoreboard(accessed March 19th 2015)

Analiza sadržaja internetskih stranica županija, gradova i općina u Republici Hrvatskoj u prvoj polovici 2014. god.

Tomislav Lamut i Agneš Sarić
Digitalni informacijsko-dokumentacijski ured Vlade Republike Hrvatske
Zagreb, Republika Hrvatska
Tomislav.Lamut@digured.hr, Agnes.Saric@digured.hr

Sažetak - Nakon što je uočeno da na njihovim internetskim stranicama često ne postoje informacije koje bi po Zakonu o pravu na pristup informacijama trebale biti objavljene, Digitalni informacijsko-dokumentacijski ured proveo je analizu sadržaja stranica svih hrvatskih jedinica lokalne i područne samouprave. Na temelju odredaba čl. 10. Zakona, određeno je 20 kategorija informacija i dokumenata, čija se (ne)objava provjeravala na stranicama svake od 576 jedinica. Uz to, istražilo se i (ne)postojanje dodatnih sedam kategorija koje se smatraju bitnima za kvalitetu stranica. Nakon prve faze istraživanja, svim jedinicama je poslana obavijest s rezultatima analize njihovih stranica i molbom za povratnom informacijom u slučaju uočenih pogrešaka ili u međuvremenu izvršenih promjena. Nakon što su, na temelju naknadno izvršenih objava i uvaženih prigovora, izmijenjeni pojedinačni rezultati, prišlo se njihovoj analizi i interpretaciji. Rezultati su izračunati po pojedinim kategorijama i interpretirani na bazi međusobne usporedbe, te sveukupno, a potvrdili su pretpostavljane značajne manjkavosti u proaktivnoj objavi informacija. Kako su pri svemu tome uočene i velike razlike u principima strukturiranja sadržaja stranica, gdje su često mnoge informacije i dokumenti bili objavljeni na nedovoljno pregledan način, zaključeno je da bi napredak u proaktivnoj objavi informacija trebao biti potaknut putem prijedloga konkretnog standarda i procedure njihova objavljivanja.

I. Uvod

Digitalni informacijsko-dokumentacijski ured Vlade Republike Hrvatske (dalje: Digitalni ured), koji je nastavio djelatnost bivše Hrvatske informacijsko-dokumentacijske referalne agencije (HIDRA) koja je djelovala od 1994. godine, je stručna služba Vlade Republike Hrvatske nadležna za obavljanje informacijsko-dokumentacijskih poslova sa zadaćom osiguranja dostupnosti javnih službenih dokumenata i informacija Republike Hrvatske radi informiranja i ponovne uporabe.

Nakon što je u sklopu obavljanja svojih redovnih djelatnosti, što uključuje i rješavanje pojedinačnih izravnih zahtjeva, uočeno da na nekim internetskim stranicama županija, gradova i općina ne postoje informacije i dokumenti, koji bi po Zakonu o pravu na pristup informacijama [1] (dalje: Zakon) trebali biti objavljeni, Digitalni ured [2] proveo je analizu sadržaja stranica svih hrvatskih jedinica lokalne (općine i gradovi) i područne (županije) samouprave.

II. Analiza Sadržaja

Čl. 10. Zakona glasi:“ (1) Tijela javne vlasti obvezna su na internetskim stranicama objaviti na lako pretraživ način:

1) zakone i ostale propise koji se odnose na njihovo područje rada;
2) opće akte i odluke koje donose, kojima se utječe na interese korisnika, s razlozima za njihovo donošenje;
3) nacrte zakona i drugih propisa te općih akata koje donose, sukladno odredbama članka 11. ovog Zakona;
4) godišnje planove, programe, strategije, upute, izvještaje o radu, financijska izvješća i druge odgovarajuće dokumente koji se odnose na područje rada tijela javne vlasti;
5) podatke o izvoru financiranja, proračunu i izvršenju proračuna;
6) informacije o dodijeljenim potporama, bespovratnim sredstvima ili donacijama uključujući popis korisnika i visinu iznosa;
7) informacije o svom unutarnjem ustrojstvu, s imenima čelnika tijela i voditelja ustrojstvenih jedinica i njihovim podacima za kontakt;
8) zapisnike i zaključke sa službenih sjednica tijela javne vlasti i službene dokumente usvojene na tim sjednicama, te informacije o radu formalnih radnih tijela iz njihove nadležnosti;
9) informacije o postupcima javne nabave i dokumentaciji za nadmetanje te informacije o izvršavanju ugovora;
10) obavijesti o raspisanim natječajima te natječajnu dokumentaciju;
11) registre i baze podataka ili informacije o registrima i bazama podataka iz njihove nadležnosti i načinu pristupa;
12) obavijesti o načinu ostvarivanja prava na pristup informacijama i ponovnoj uporabi informacija s podacima za kontakt službenika za informiranje;
13) visinu naknade za pristup informacijama i ponovnu uporabu informacija, prema kriterijima iz članka 19. stavka 3. ovog Zakona;
14) najčešće tražene informacije;

15) ostale informacije (vijesti, priopćenja za javnost, podaci o drugim aktivnostima).

U skladu s time, određeno je 20 kategorija, odnosno vrsta informacija i dokumenata, čije se postojanje odnosno nepostojanje utvrdilo za svaku pojedinu jedinicu. To su sljedeće kategorije: predstavničko tijelo, izvršno tijelo, upravna tijela, kontakt, pravne osobe (ustanove), službenik za informiranje, službeno glasilo, opći pravni akti i druge odluke, statut, poslovnik, proračun, izvršenje proračuna, prostorni plan, dnevni redovi sjednica, prijedlozi i nacrti općih akata, zapisnici sa sjednica, programi, projekti, planovi i sl., izvješće o izvršenju programa, javna nabava te natječaji i potpore.

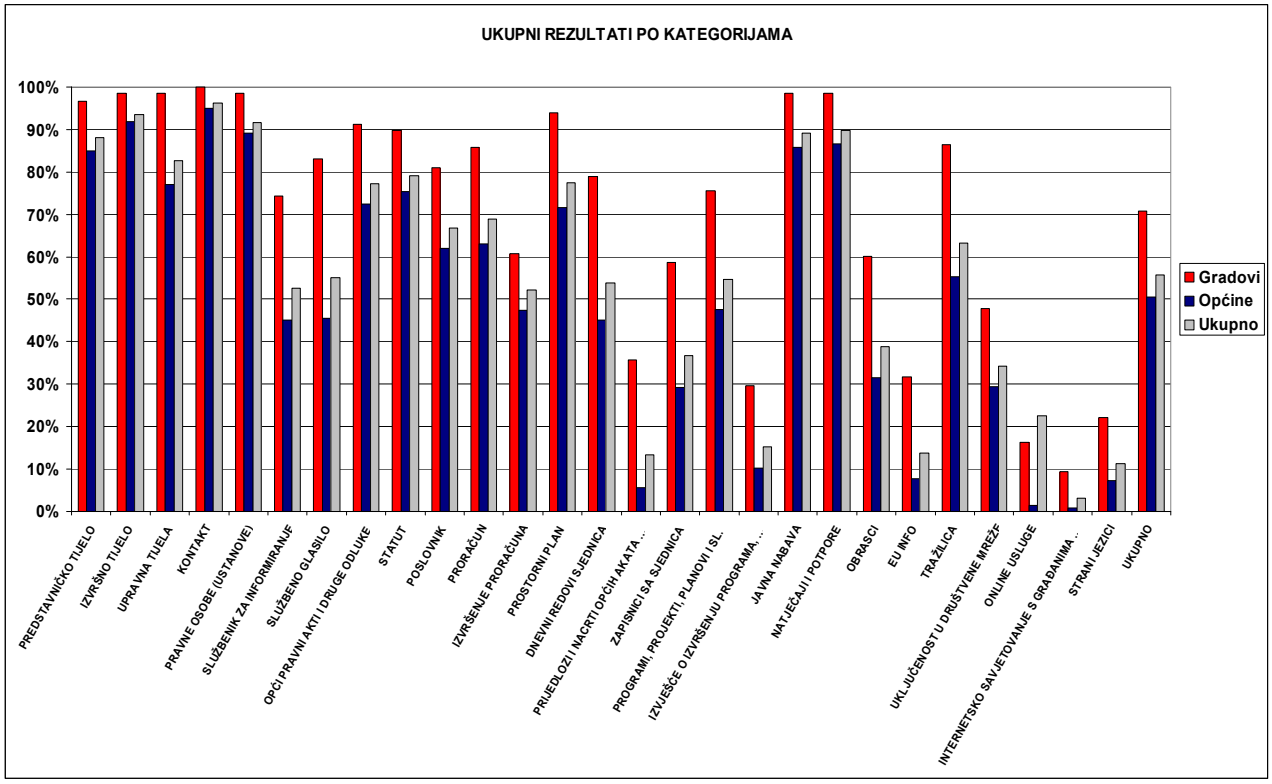
Ovdje treba spomenuti da čl. 62. st. 1. Zakona propisuje novčane kazne od 2000 do 4000 kuna za tijela javne vlasti u slučaju neobjavljivanja informacija u skladu s čl. 10., st. 1. i čl. 11. Zakona. Prema čl. 35. i čl. 41. Zakona, tijelo nadležno za nadzor i inspekcijski nadzor nad provedbom ovog Zakona jest Povjerenik za informiranje, koji također „podnosi optužni prijedlog i izdaje prekršajni nalog za utvrđene prekršaje“. U sklopu ovog istraživanja, nisu se prikupljale informacije o eventualnim kaznama izdanim od strane Povjerenika za informiranje.

Uz navedeno, istražilo se i (ne)postojanje dodatnih sedam elemenata, koji ne spadaju u ove izričito propisane Zakonom, ali se mogu smatrati bitnim za kvalitetu sadržaja internetskih stranica jedinica lokalne i područne samouprave. To su sljedeće kategorije: obrasci, EU info, tražilica, uključenost u društvene mreže, online usluge, internetsko savjetovanje s građanima i strani jezici. One su određene samoinicijativno od strane djelatnika Ureda, na temelju višegodišnjeg iskustva u pretraživanju internetskih stranica domaćih i stranih lokalnih jedinica.

Ukupni broj općina, gradova i županija u Republici Hrvatskoj je 576, od čega je 428 općina, 127 gradova, 20 županija i Grad Zagreb sa statusom županije. 15 jedinica, tj. općina nije u razdoblju istraživanja uopće imalo internetske stranice.

Nakon prve faze istraživanja, koju je dvoje djelatnika Ureda obavljalo kada je to bilo moguće s obzirom na osnovnu djelatnost i koja je trajala od siječnja do lipnja 2014. godine, svim jedinicama je poslana obavijest s rezultatima analize i molbom za povratnom informacijom u slučaju uočenih pogrešaka ili u međuvremenu izvršenih promjena. Odgovori su poslani iz 151 jedinice.

Najčešći prigovori su bili oni u kojima se navodilo kako je neki dokument objavljen u elektroničkom izdanju službenog glasila, što se nije uvažavalo jer se smatralo da time nije zadovoljen uvjet iz čl. 10. Zakona o objavljivanju „na lako pretraživ način“. Prethodno navedeno je ukazalo da u značajnom broju jedinica ne postoji spoznaja kako o nužnosti objavljivanja službenih dokumenata i informacija, tako i o omogućavanju preglednosti u njihovom pretraživanju. S druge strane, u značajnom broju dopisa kao i u nekoliko izravnih telefonskih razgovora, navodi se kako je nešto odmah po dopisu iz Digitalnog ureda bilo objavljeno i/ili će biti u „skorom roku“, što se zaista u većini takvih slučajeva pokazalo točnim, pa su takve, već obavljene nadopune, pribilježene u tablicama. Smatramo da je ovo jasno pokazalo kako je spoznaja da se kvaliteta sadržaja stranica prati od strane vanjske institucije, pozitivno utjecala na podizanje kvalitete stranica, pa čak i usprkos



Grafikon 1. Ukupni rezultati po kategorijama

tome što je Digitalni ured naveo, kako je ovo tek istraživanje u informativne svrhe. Nakon što su, na temelju u međuvremenu izvršenih objava i uvaženih prigovora, izmijenjeni pojedinačni rezultati, prišlo se njihovom zbrajanju i analizi.

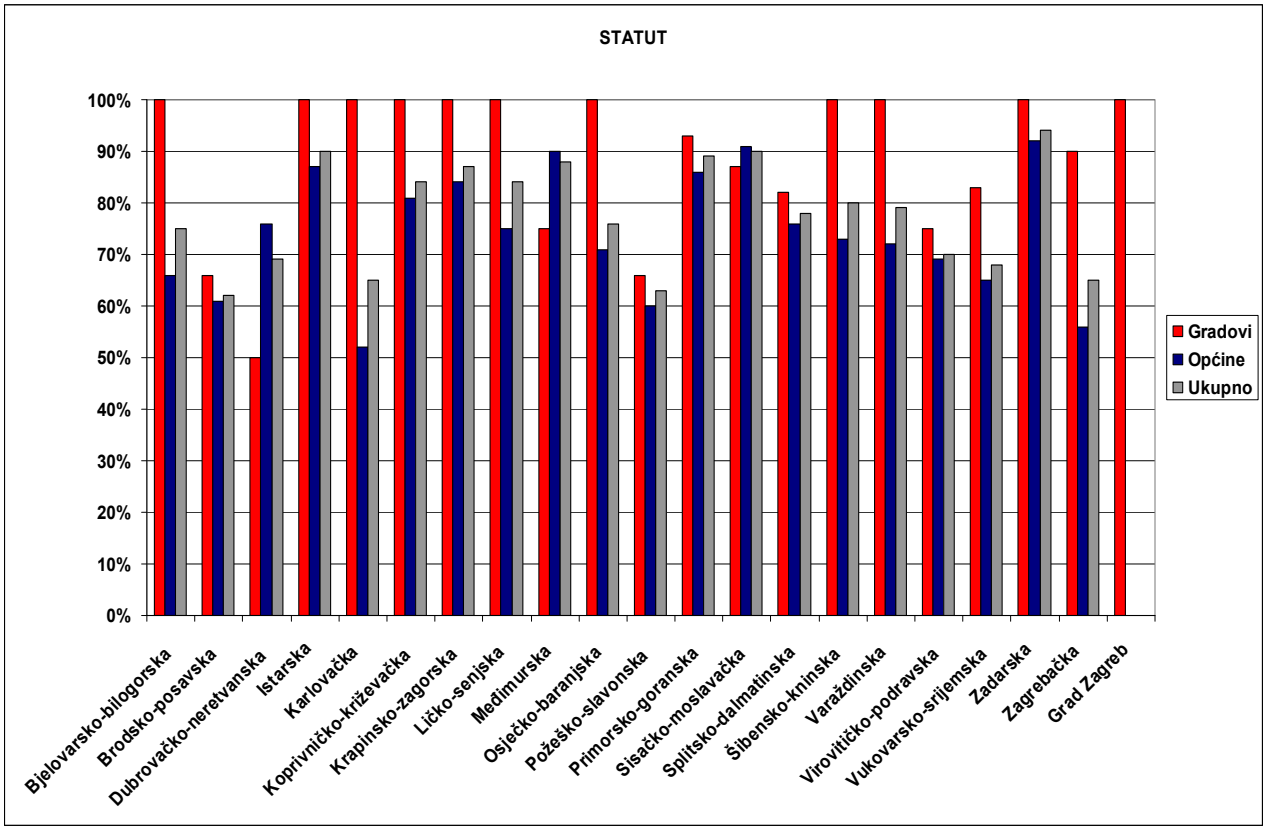
Rezultati su interpretirani na bazi međusobne usporedbe pri čemu su izračunati sveukupni rezultati te rezultati posebno za gradove i posebno za općine, uz napomenu da su rezultati županija pribrojani rezultatima gradova. Bitno je navesti da rezultati nisu vrijednosno interpretirani jer kada se radi o ispunjavanju zakonskih obveza, jedini zadovoljavajući rezultat je onaj stopostotni (takav je zabilježen za tri grada: Zagreb, Rijeku i Pulu). Po očekivanju, s obzirom na puno veći ljudski i tehnički potencijal gradskih i županijskih uprava, rezultati gradova su bili viši jer npr. u tijelima najmanjih općina, nije rijetko da su općinski načelnik i tajnik jedini profesionalni djelatnici.

Svi rezultati su prikazani u grafikonima, ali kako istih ima 28 (po jedan za svaku kategoriju i jedan sa sveukupnim rezultatima, koji su prikazani u grafikonu 1), nije ih bilo moguće sve uvrstiti u rad od maksimalno 6 stranica. Stoga su ovdje prikazani samo grafikon s ukupnim rezultatima i pet odabranih grafikona kao primjeri.

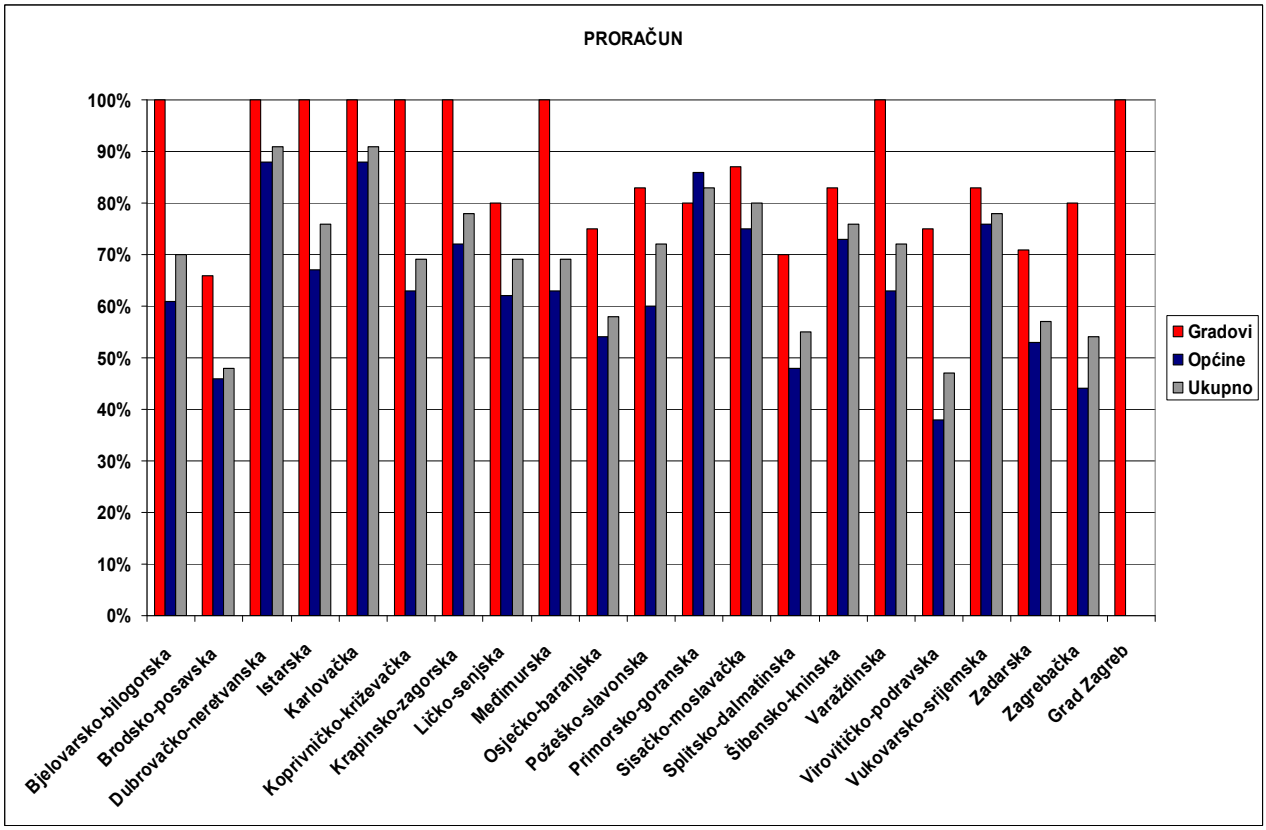
Što se tiče rezultata po pojedinim elementima, najviši rezultati su zabilježeni pri objavljivanju informacija vezanih uz kontakt podatke te ustrojstvo vlasti u jedinicama, ali već i kod toga su zabilježene prve manjkavosti. Tako je podatke o članovima predstavničkog tijela objavilo 96% gradova i županija, odnosno 85% općina, dok su podaci o sastavu izvršnog tijela bili objavljeni na 98% gradskih tj. županijskih, odnosno 91% općinskih stranica. Podatke o upravnim tijelima, odnosno djelatnicima, objavilo je 98% gradova tj. županija, ali u općinama je taj rezultat iznosio 77%. Podatke o raznim ostalim ustanovama, tvrtkama, organizacijama i udrugama u svom sastavu ili sa svog područja objavilo je 98% većih jedinica i 89% općina. Potpune informacije o ostvarivanju prava na pristup informacijama objavilo je 74% gradova i županija dok je kod općina rezultat bio puno niži i to 45%. Slični su rezultati (83% i 45%) vezani uz objavljivanje službenog glasila, pri čemu treba napomenuti da općine nisu obvezne imati svoje vlastito službeno glasilo, pa se u tom slučaju tražilo tek da na općinskim stranicama bude jasno objavljena poveznica na županijsko službeno glasilo što je velikom broju općinskih stranica nedostajalo. Dio prigovora iz pojedinih općina se odnosio upravo na negativni znak u rubrici „službeno glasilo“, uz obrazloženje kako imati ga, nije njihova zakonska obveza pri čemu se očito nije razumjelo da je u tom slučaju bilo potrebno isto navesti i korisnike uputiti na odnosno županijsko glasilo i osnovne općinske akte, koji su u njemu objavljeni. Osnovni akti poput statuta jedinice (grafikon 2) i poslovnika vijeća tj. skupštine su objavljeni u preko 80% slučajeva za gradove i županije (statut – 89%, poslovnik predstavničkog tijela – 81%), odnosno 75% (statut) i 62 % (poslovnik) za općine. Kriterije za ocjenu dostatnog objavljivanja tj. navođenja općih pravnih akata nije bilo moguće točno utvrditi pa se pozitivni znak upisivao tamo gdje je procijenjeno da je objavljen barem minimalan broj najvažnijih akata pri čemu je rezultat na taj način iznosio 91% za gradove i županije i 72% za općine. Zatim, zanimljivi su bili rezultati za objavljivanje proračuna (grafikon 3) i izvješća o njegovom izvršenju (grafikon 4); dok je proračun za tekuću godinu objavljen u 85% (gradovi i županije) odnosno 63% (općine) slučajeva, izvješće o njegovu izvršenju za prethodne godine je objavljeno na značajno manjem broju stranica (60% gradovi i županije, 47% općine). Prostorni plan (grafikon 5) je objavljen na 93% (gradovi i županije), tj. 71% (općine) stranica. Kod informacija vezanih uz sjednice predstavničkih tijela, dnevni redovi sjednica (grafikon 6) su objavljivani na 79%, tj. 45%, a zapisnici na 59%, tj. 29% stranica. Značajno niži rezultati su zabilježeni u objavljivanju prijedloga i nacrta akata (35% i 5%). Slična situacija kao s proračunom je uočena i kod objave raznih planova i programa rada, odnosno izvješća o njihovu izvršenju: planovi i programi su bili objavljeni na 75%, tj. 47% stranica, dok je rezultat vezan uz izvješća o njihovom izvršenju bio značajno niži: 29% i 10 %. U slučaju objavljivanja informacija o javnoj nabavi i raznim natječajima, rezultati su bili iznad 98% za gradove i županije, odnosno 86% za općine.

U drugoj grupi od sedam ostalih vrsta informacija i sadržaja, rezultati su po očekivanju, bili još niži pa su tako npr. informacije o EU objavljene na 31% stranica gradova i županija odnosno 7% onih općinskih. Obrasci i to uglavnom ne svi, objavljeni su na 60% stranica gradova i županija, tj. 31% općinskih stranica, a bilo kakve administrativne online usluge omogućuje tek 16% gradova i županija i nešto više od 1% općina. Internetsko savjetovanje s građanima vezano uz prijedloge akata je omogućeno u 9% gradova i županija i manje od 1% općina dok je pregled osnovnog dijela sadržaja na stranim jezicima (engleski i talijanski u Istri), omogućen kod 22% gradova i županija i 7% općina. Očekivano je bilo da mnoge, posebno manje jedinice nemaju informacije na stranim jezicima, ali se pokazao i neočekivano veliki broj jedinica iz turističkih područja, koje nisu imale informacije za strane turiste pa čak ni poveznice na neke druge takve stranice, poput onih npr. turističke zajednice i sl. Tražilicu na svojim stranicama ima 86% gradova i županija, te 55% općina, ali u mnogim slučajevima je njena funkcionalnost upitna. I na kraju, podatke o uključenosti u društvene mreže objavilo je 71% gradova i županija i 29% općina iako je broj u društvene mreže uključenih jedinica nešto veći, ali u nekim slučajevima podaci o tome jednostavno nisu bili objavljeni.

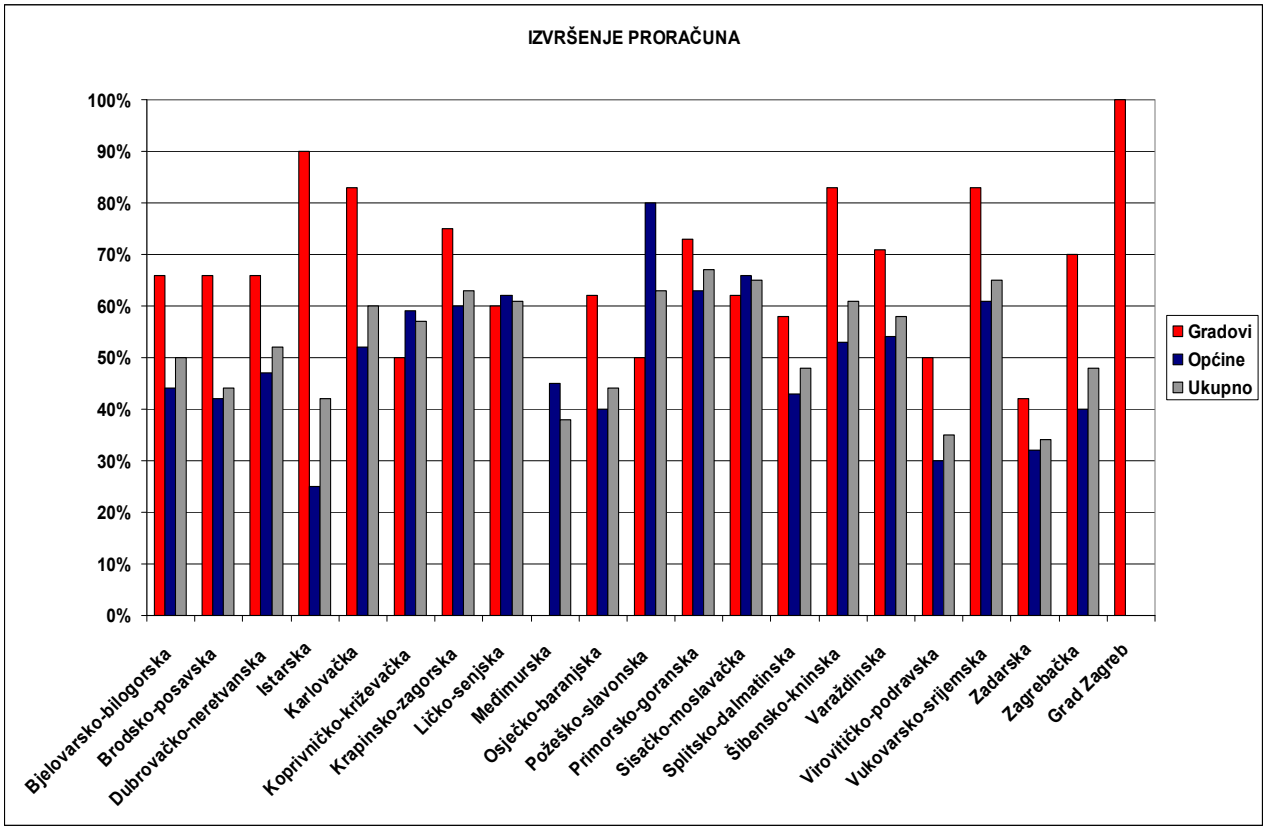
Sveukupni rezultati su grafički prikazani u grafikonu 1. Iz svega navedenog je jasno da su rezultati, s već navedenim razlikama između gradova i županija naspram općina, bili znatno niži kod onih vrsta dokumenata koji su mnogobrojniji, a aktualnost im je kraća, dok su viši bili kod pojedinačnih dokumenata, s dužim rokom važenja i pretpostavljano većom „važnošću“ poput npr. statuta ili prostornog plana.



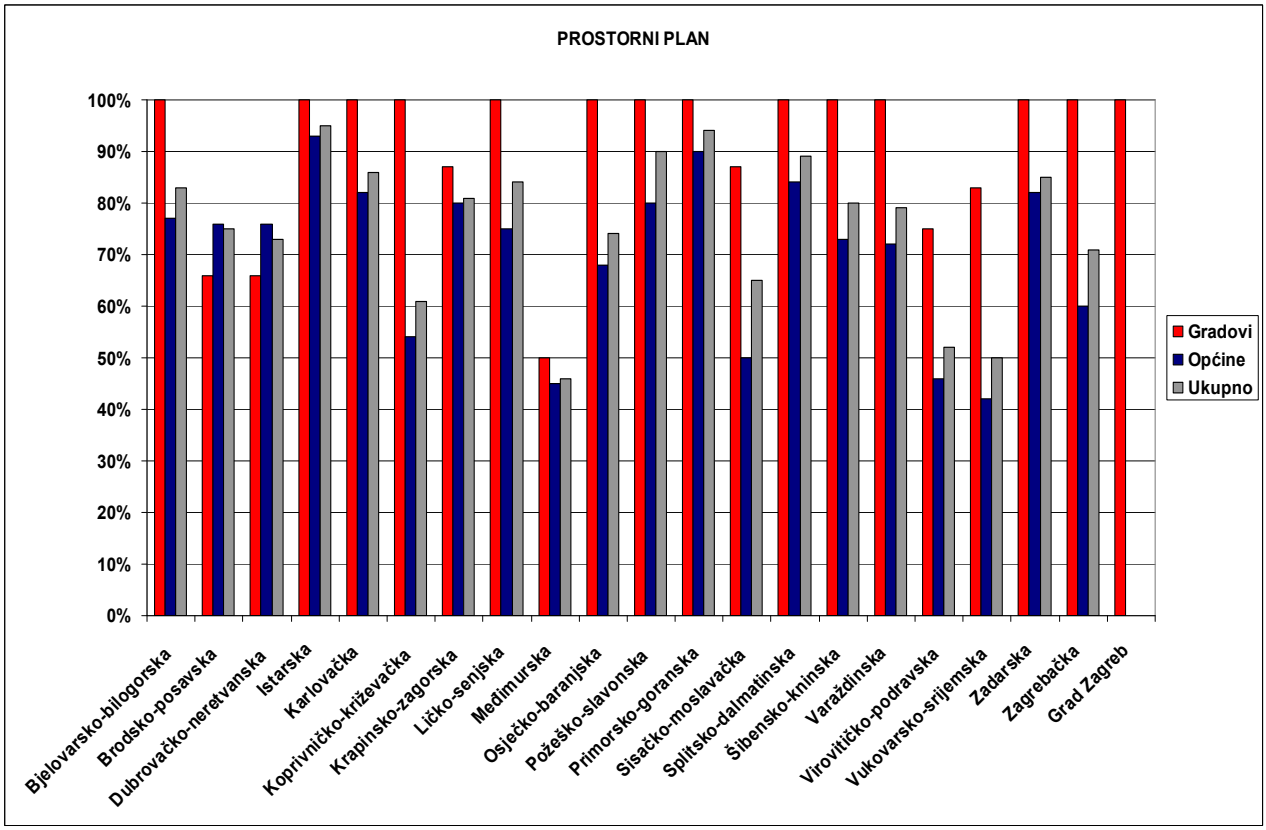
Grafikon 2. Statut



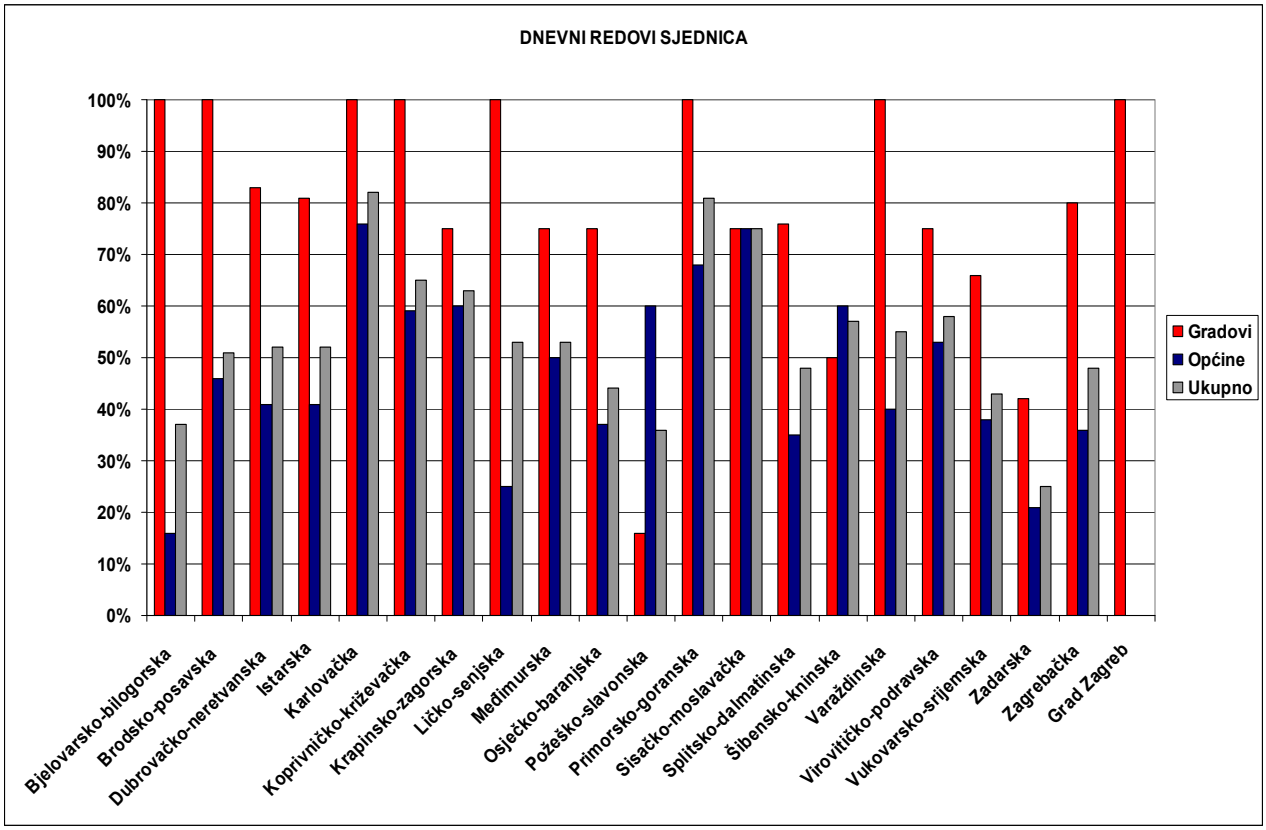
Grafikon 3. Proračun



Grafikon 4. Izvršenje proračuna



Grafikon 5. Prostorni plan



Grafikon 6. Dnevni redovi sjednica

Uza sve brojčano izraženo i vidljivo iz priloženih grafikona, uočeno je i da su mnoge informacije i dokumenti često bili objavljeni na neočekivanim dijelovima stranice i/ili pod upitnim naslovima. Takvi slučajevi jesu označavani znakom plusa, iako se može pretpostaviti da je kod dijela korisnika „laka pretraživost“ time bila više ili manje otežana. U svakom slučaju, također su uočene velike razlike u principima strukturiranja sadržaja stranica.

III. ZAKLJUČAK

S obzirom na sve navedeno što je ovo istraživanje pokazalo vezano uz nedostatno objavljivanje dokumenata i informacija u skladu sa Zakonom o pravu na pristup informacijama, mišljenja smo da bi napredak u proaktivnoj objavi informacija trebao biti potaknut putem prijedloga konkretnog standarda i procedure njihova objavljivanja, zamišljenog kao svojevrsnog jednoobraznog elektroničkog pretinca za unošenje odnosno objavu službenih dokumenata, pri čemu bi sa velikom vjerojatnošću, a nakon eventualnog ustrojavanja takvog modela, isti ne samo bio od izravne pomoći u lokalnim jedinicama, već bi dodatni pozitivni indirektni poticaj kod njihovih djelatnika bila i spoznaja da su „rezultati rada“ lako usporedivi u odnosu na ostale jedinice, tj. da je (ne)ispunjavanje zakonskih obveza na taj način puno lakše uočljivo. Naš zaključak se nalazi na tragu točke 3.5. Akcijskog plana za provedbu inicijative Partnerstvo za otvorenu vlast u Republici Hrvatskoj za razdoblje 2014. do 2016. godine [3] koja predviđa izradu uputa i priručnika za proaktivnu objavu informacija, čiji je nositelj Povjerenik za informiranje, sunositelj Ministarstvo uprave, a predviđeni rok srpanj 2015. god.

LITERATURA

[1] Zakon o pravu na pristup informacijama, NN 025/2013

[2] www.digured.hr, pristupljeno 07.01.2014.

[3] Akcijski plan za provedbu inicijative Partnerstvo za otvorenu vlast u Republici Hrvatskoj za razdoblje 2014.-2016. - http://digured.srce.hr/arhiva/1571/109347/Akcijski plan za provedbu inicijative Partnerstvo za otvorenu vlast u Republici Hrvatskoj za razdoblje 2014-2016..pdf

Implementacija informacijskog sustava za praćenje aktivnosti unutar jedinice javne uprave

O. Jukić, I. Heđi, I. Špeh
Visoka škola za menadžment u turizmu i informatici, Virovitica, Hrvatska
{oliver.jukic | ivan.hedi | ivan.speh}@vsmti.hr

Sažetak - U radu je opisana arhitektura informacijskog sustava čija osnovna zadaća je pružanje potpore prilikom odlučivanja, praćenja prijedloga i pritužbi građana te praćenja aktivnosti i stanja na određenim projektima jedinice javne uprave koja je implementirana na primjeru jedinice javne uprave. Funkcije sustava distribuirane su u tri sloja. Sloj medijacije daje potporu procesima prikupljanja podataka ili primanja aktivnosti iz različitih izvora. To mogu biti: postojeće baze podataka, baze podataka vezanih poduzeća, dopisi, telefonski kontakti ili kontakti elektronskom poštom, kontakti ostvareni putem web stranice i sl. Tako prikupljeni podaci se pohranjuju na sloju baze podataka, u više različitih baza podataka. Tipični primjeri su skladište podataka o radu jedinice samouprave, baza podataka o pritužbama građana, baza podataka o procesima koji su u tijeku (npr. zahtjevi za službena putovanja koji čekaju na odobrenje), baza podataka o projektima i sl. Iznad sloja baze podataka nalazi se prezentacijski sloj, čija osnova zadaća je kreiranje izvješća, njihov prikaz na grafičkim sučeljima te distribucija izvješća. Osim toga, arhitektura uključuje i *northbound* sučelje prema drugim informacijskim sustavima i tehnologijama, čija zadaća je omogućiti dostup podacima i izvješćima drugim sustavima i aplikacijama (npr. aplikacijama na pametnim telefonima). Implementacija ovako predložene arhitekture u tijeku je u jednoj jedinici javne uprave.

Ključne riječi: informacijski sustav, javna uprava, medijacija, baza podataka, izvješće

I. UVOD

Informacijski sustavi u jedinicama javne uprave s služe kao alati za povećanje učinkovitosti prilikom izvršavanja procedura javne administracije i upravljanju iste. Istraživanja su pokazala da informacijski sustavi pokreću modernizaciju javne uprave i dovode do organizacijskih promjena. Organizacije bez potpore automatskog procesiranja informacija imaju smanjenu učinkovitost zbog spore obrade tradicionalne "papirologije". Povećanje učinkovitosti je nemoguće bez automatskog procesiranja informacija [1].

Implementacija informacijskih sustava u javnoj upravi dovodi do poboljšanja kvalitete usluga prema građanima i poslovima. Prednosti dobivene implementacijom informacijskog sustava u rad javne uprave moraju se na neki način povezati sa ciljevima usluga koje se pružaju.

Ovaj rad je rezultat uvođenja informacijskih sustava u jedinicama javne uprave u Virovitičko – podravskoj županiji.

Te prednosti se postižu korištenjem usluga kako od strane građana, tako i od strane jedinica javne uprave. Prednosti prema građanima uključuju (1) pristup informacijama za javnost, (2) online usluge i (3) pristup pravovremenim, relevantnim i točnim informacijama. Prednosti za jedinice javne uprave su (1) standardizacija i redizajn procesa, (2) poboljšanje pri upravljanju projektima, (3) povećanje kvalitete te (4) povećana podrška i uključenost svih razina osoblja. Indirektna prednost implementacije informacijskog sustava u jedinice javne uprave je poboljšanje otvorenosti, transparentnosti i odgovornosti administracije te osvještavanje građana o lakšem pristupu uslugama javne uprave.

Implementacija informacijskog sustava u jedinice javne uprave rezultira sljedećim prednostima:

- Učinkovitu i logički smislenu cirkulaciju informacija među jedinicama javne uprave
- izvođenje operacija jedinica javne uprave koje su predefinirane procesima
- integraciju podataka
- lakši pristup podacima i informacijama za jedinice javne uprave
- pristup informacijama i uslugama putem različitih medija (računalo, tablet, pametni telefon) za građene.

Klasifikacija sustava implementirana u jedinicama javne uprave prema [2] je sljedeća: (1) Workflow Management System (WfMS), (2) Enterprise Information System (EIS), (3) Enterprise Application Integration (EAI), (4) Business Process Management System (BPMS), Business intelligence (BI) i (6) Enterprise Portals.

Sustavi za upravljanje tokovima su jedni od najvažnijih sustava koji pružaju podršku procesima javne uprave. Ovakvi sustavi djelomično ili u potpunosti automatiziraju definiranje, stvaranje, izvršavanje i upravljanje radnim procedurama koje se ne mijenjaju prečesto [3]. *Workflow* sustavi automatiziraju poslovne procese, djelomično ili u cijelosti pri čemu se informacije, dokumenti i zadaci prenose od jednog do drugog sudionika toka prema određenim pravilima [2].

Poslovni procesi, radne procedure i radni zadaci koji se mogu opisati *workflow* sustavom unutar jedinice javne uprave u radu su predstavljeni kao aktivnosti. Namjena

informacijskog sustava opisanog u radu je potpora za praćenje aktivnosti unutar jedinice javne uprave pri čemu su aktivnosti kategorizirane prema prirodi radnih procedura koje se izvršavaju. Prema tome razlikujemo (1) aktivnosti vezane za radne procedure unutar jedinice javne uprave, (2) aktivnosti vezane za pritužbe građana i (3) aktivnosti na projektima koje izvodi jedinica javne uprave.

Svaka aktivnost sastoji se od radnih zadataka, sudionika, dokumenata i informacija. Aktivnost može imati jedan ili više radnih zadataka koji se slijedno izvode. Svaki radni zadatak unutar pojedine aktivnosti u radu je definiran kao faza aktivnosti. Izvršavanjem svih faza aktivnost se završava.

II. ARHITEKTURA INFORMACIJSKOG SUSTAVA

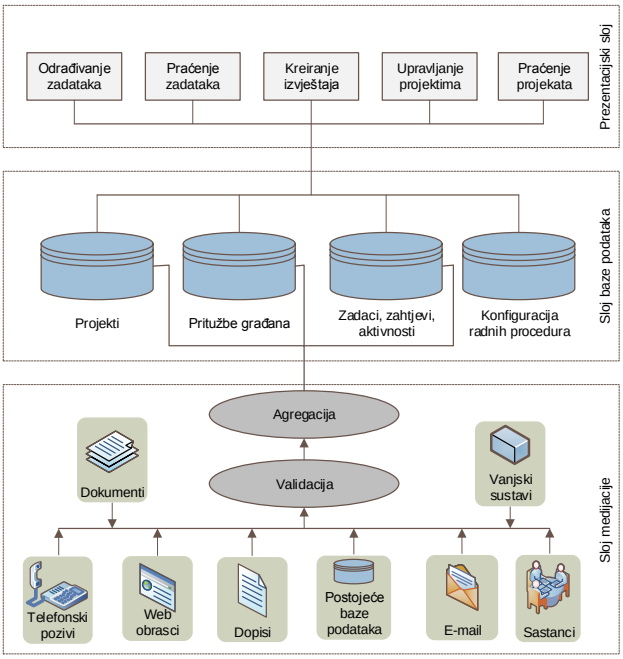
Za implementaciju informacijskog sustava potrebno je koristititi logičku konstrukciju (arhitekturu) sustava za definiranje sučelja i integraciju svih komponenti sustava [4]. Na slici 1 prikazana je arhitektura informacijskog sustava jedinice javne uprave čija je osnovna zadaća pružanje potpore prilikom odlučivanja, praćenja prijedloga i pritužbi građana te praćenja aktivnosti i stanja na određenim projektima. Funkcije sustava podijeljene su na tri sloja: sloj medijacije daje potporu procesima prikupljanja podataka ili primanja aktivnosti iz različitih izvora, sloj baze podataka te prezentacijski sloj.

Glavne funkcije sloja za prikupljanje podataka su:

- prikupljanje podataka od građana, djelatnika javne uprave ili drugih jedinica uprave
- validacija podataka
- agregacija podataka

Upravljanje poslovnim sustavom uključuje planiranje, organiziranje i kontroliranje aktivnosti poslovnog sustava, a u velikoj mjeri temelji se na donošenju odluke. Za donošenje dobre poslovne odluke potrebne su kvalitetne informacije (točne, potpune, primjerene i pravovremene). Zbog toga dobar poslovni sustav mora imati informacijski sustav koji će ga opskrbljivati kvalitetnim informacijama [6]. Za kvalitetno rješavanje problema potrebno je prikupiti informacije od građana na njima najprihvatljivi način. Na slici 1. prikazani su neki od načina prikupljanja podataka: web obrasci, telefonski pozivi, dopisi ili postojeće baze podataka. Prikupljanje podataka od građana putem web obrazaca odvija se putem modula web grafičkog sučelja otvorenog prema građanima ili putem web stranice javne uprave. U trenutnoj implementaciji građani ne koriste autenikaciju pošto pristupaju javnom sučelju i nisu dio sustava već su samo izvor informacija. Putem e-mail pošte građani pritužbe šalju na određenu adresu putem koje se prikupljaju informacije potrebne za rješavanje određenog problema. Postojeće baze podataka koriste se u slučaju da jedinica javne uprave već koristi određeni informacijski sustav. Tako prikupljeni prolaze proces validacije i agregacije (nepravovaljani i neprimjereni zahtjevi građana se odbacuju) i šalju se na pohranu i obradu u više slojeve.

Raznoliki unutarnji i vanjski izvori podataka "pune" skladišta podataka. Većina unutarnjih podataka (nastalih unutar javne uprave) rezultat je poslovnih aktivnosti. U



Slika 1. Arhitektura informacijskog sustava

sloju baze podataka (slika 1) prikazana su skladišta koja služe za pohranu pritužbi građana zaprimljenih na slojumedijacije, informacija o stanju na projektima, zadatke i rezultate pojedinih radnih procedura. Osim skladišta za prikupljanje podataka prikazana je i konfiguracijska baza podataka u kojoj su zapisani postupci izvršavanja određenih radnih procedura.

Korisničko sučelje obuhvaća mehanizme potrebne za unos podatka u sustav, izvođenje modela te izlaz podataka (sadržaj i grafički oblik ekrana). Ono je važna komponenta sustava za potporu praćenja aktivnosti jer je sučelje za korisnika zapravo sam sustav [7]. Grafičko sučelje implementiranog sustava sastoji se od modula koji su podijeljeni prema vrstama aktivnosti. Administrator sustava određuje koji moduli će se prikazati djelatniku na njegovom grafičkom sučelju. Npr. osoba zadužena za rješavanje pritužbi građana ne mora nužno imati modul za praćenje aktivnosti na projektima.

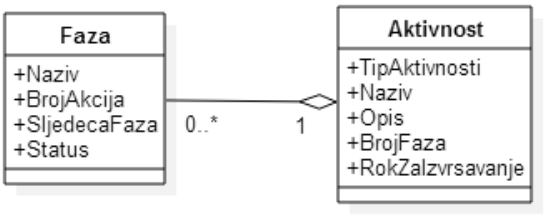
Djelatnici javne uprave pristupaju svom profilu putem web grafičkog sučelja koji se sastoji od dva glavna modula:

- grafičko sučelje za praćenje zadanih i primljenih zadataka vezanih za radne procedure unutar javne uprave, zadatke vezane za pritužbe i zahtjeve građana
- grafičko sučelje za praćenje stanja i aktivnosti na projektima jedinice javne uprave

III. PRAĆENJE AKTIVNOSTI

A. *Praćenje aktivnosti unutar jedinice javne uprave*

Radne procedure koje imaju isti slijed izvršavanja definiraju jedinstveni tip određene aktivnosti unutar sustava. Svaki tip aktivnosti, odnosno radne procedure karakteriziraju parametri zapisani u konfiguracijskim tablicama od kojih su najvažniji prikazani na UML dijagramu klase (Slika 2).



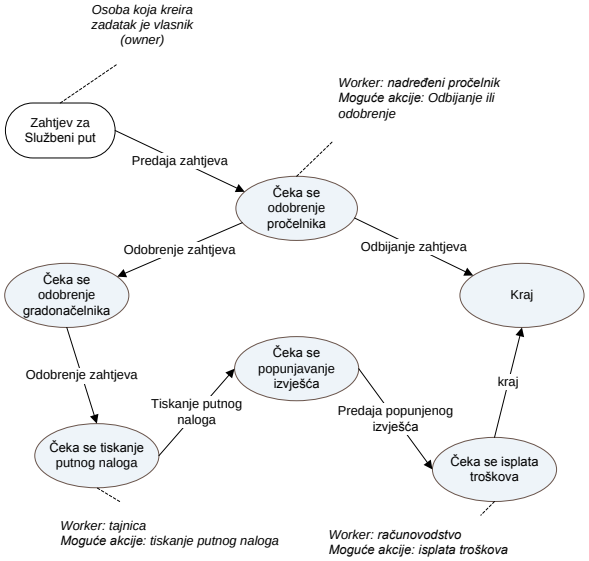
Slika 2. UML dijagram klase aktivnosti

Svaki korak izvršavanja radne procedure zapisan je u obliku faze. Radna procedura mora imati minimalno jednu fazu a ukupan broj faza određen je atributom "*BrojFaza*". U svakoj fazi definiran je djelatnik koji izvršava fazu - izvršitelj/radnik (*worker*). Izvršitelj može biti konkretan djelatnik u javnoj upravi ili određena funkcija (npr. pročelnik odjela). Rok za izvršavanje - datum do kojeg određena aktivnost mora biti odrađena. Ovisno o tipu aktivnosti rok može biti predefiniran (npr. izrada izvješća krajem mjeseca) ili se može birati prilikom kreiranja aktivnosti. Vlasnik aktivnosti može mijenjati rok za izvršavanje na grafičkom sučelju. Unutar svake faze definirane su akcije koje mogu izvršiti vlasnik zadatka ili izvršitelj zadatka. Ovisno o izvršenoj akciji, zadatak se završava ili prelazi u sljedeću fazu. Tipovi akcija mogu biti: (1) prijelaz aktivnosti u sljedeću fazu, (2) završetak aktivnosti, (3) završetak aktivosti sa negativnim ishodom (npr. odbijanje zahtjeva za odlazak na službeni put), (4) proslijeđivanje aktivne faze drugom djelatniku javne uprave, (5) otkazivanje aktivnog zadatka i zatvaranje njegove aktivne faze (akcija vlasnika aktivnosti koja rezultira zatvaranje aktivnosti) te (6) promjena roka za izvršavanje. Prilikom implementacije sustava akcije se prilagođavaju tipovima aktivnosti koje se nalaze u javnoj upravi. Tipični primjeri akcija za izvršitelja su "*Potvrdi*", "*Odbaci*", "*Proslijedi*" a za vlasnika aktivnosti su "*Otkaži*" i "*Promijeni rok za izvršavanje*".

Tijekom korištenja sustava korisnici kreiraju instance pojedine vrste predefiniranog zadatka, zahtjeva ili radne procedure (aktivnosti). Djelatnik javne uprave koji kreira novu instancu ujedno je i vlasnik zadatka (*owner*). Nakon kreiranja instance, zadatak postaje aktivan i spreman za izvođenje. Sve aktivnosti spremljene su u bazu podataka nad kojom se mogu kreirati izvješća o radu pojedinog djelatnika ili odjela, brzina reagiranja na zadane poslove, rezultati završenih zadataka i sl. U aktivnom stanju, zadatak prolazi kroz svoje faze izvršavanja pri čemu je u svakom trenutku aktivna samo jedna faza pojedine instance aktivnosti.

Vlasnik na grafičkom sučelju u svakom trenutku može vidjeti koje su faze zadatka izvršene te koja je faza trenutno aktivna. Svaki zadatak ima definiran rok za izvršavanje koji se postavlja prilikom kreiranja zadatka, a vlasnik ga u svakom trenutku može promijeniti (ukoliko tip zadatka dopušta promjenu roka za izvršavanje). Za svaku promjenu na zadatku (npr. istek roka za izvršavanje, promjena faze zadataka, završetak zadatka) obavještavaju se svi aktivni sudionici u zadatku putem notifikacije na grafičkom sučelju i/ili putem e-mail poruke.

Procedura izvršavanja jednog tipa radne procedure opisana je na slici 3. Djelatnik putem grafičkog sučelja odabire jedan od ponuđenih zadataka (u ovom slučaju



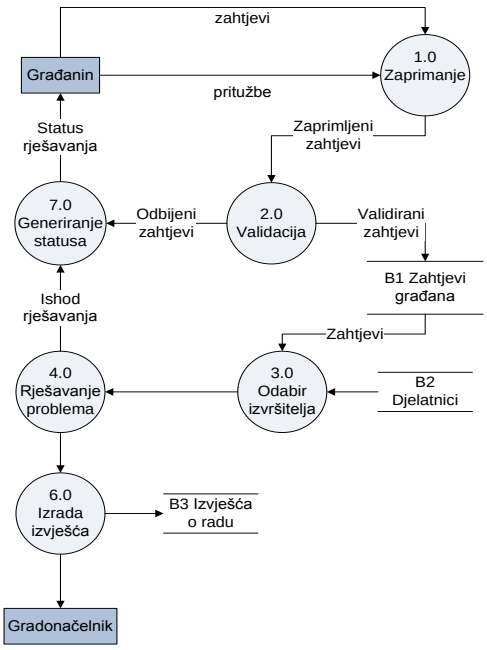
Slika 3. Workflow radne procedure službenog putovanja

"*Zahtjev za službeni put*"). Nakon popunjavanja web obrasca (sloj prikupljanja podataka) podaci o službenom putu spremaju se u bazu podataka službenih putovanja (sloj baze podataka) a nadređeni pročelnik na svom profilu dobiva notifikaciju sa novim primljenim zadatkom (prezentacijski sloj). Uz osnovne informacije o zadatku (tko, što, gdje, kada) pročelnik ima predložene akcije u obliku ikonica. Odabirom akcije za odobravanje završava se trenutno aktivna faza i zahtjev prelazi u novu fazu u kojoj izvršitelj postaje gradonačelnik nakon čijeg odobravanja zahtjev dalje prolazi kroz faze tiskanja putnog naloga, popunjavanja izvješća i isplate troškova. Svaka faza prikazuje se na profilu izvršitelja kojemu je zadano odrađivanje te faze Nakon što je zahtjev prošao kroz sve faze postaje neaktivan i sprema se u arhivu. Arhiva aktivnosti je podijeljana na aktivnosti u kojima je korisnik vlasnik i na arhivu u kojima je korisnik izvršitelj. Na temelju arhive moguće je izraditi izvješća na tjednoj mjesečnoj ili godišnjoj bazi (npr. koliko je djelatnik imao službenih putovanja u godini dana). Ovime je prikazana jedna od radnih procedura koja se može sresti u svakodnevnom djelovanju javne uprave.

B. Potpora za praćenje aktivnosti vezanih uz pritužbe građana

Rješavanje pritužbi i zahtjeva građana dio su svakodnevnih poslova djelatnika javne uprave. Dijagram toka podataka [5] prikazan na slici 4 prikazuje proces rješavanja zahtjeva i pritužbi građana putem informacijskog sustava. Građanin, kao vanjski entiteti u ovom procesu sustavu daje svoju pritužbu ili zahtjev na jedan od načina prikazanih na slici 1. Nakon zaprimanja pritužbe građanina i validacije ulaznih informacija (procesi 1.0 i 2.0), sustav dodjeljuje aktivnost određenom djelatniku na rješavanje koji će na svom profilu zaprimiti novi zahtjev sa opisom problema, rokom za izvršavanje i akcijama koje će odabrati nakon rješavanja problema (proces 3.0). Nakon što djelatnik odradi aktivnost vezanu za pritužbu građana generira se status u procesu 7.0, a on ovisi o ulaznim tokovima podataka koji mogu biti ishodi rješavanja problema ili rezultati validacije određenih

podataka. Gradonačelnik kao vanjski entitet zaprima izvješća o pritužbama građana i ishodima rješavanja istih.

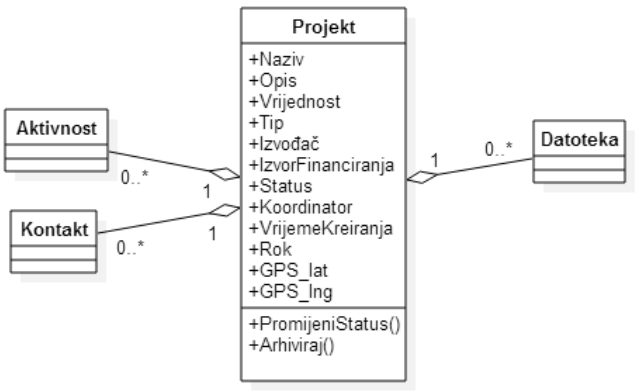


Slika 4. Dijagram toka podataka zaprimanja pritužbe građana

C. *Praćenje aktivnosti nad projektima*

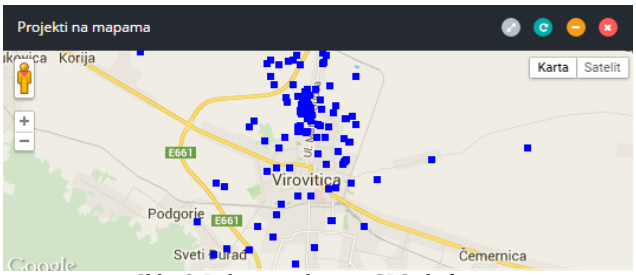
Jedna od glavnih zadaća jedinica javne uprave je provođenje projekata. Za projekte sa niskom dinamikom promjena i velikim brojem sudionika teško je uvijek znati trenutno stanje na projektu. Putem web grafičkog sučelja djelatnici javne uprave dobivaju brze i kvalitetne informacije o trenutnom stanju na pojedinom projektu kao i praćenje tijeka odvijanja projekta.

Svaki projekt karakterizira određeni broj atributa koji se mogu definirati tijekom prilagodbe sustava, na primjer: vrijednost projekta, tip (gradski, državni, privatni...), izvor financiranja, nositelj projekta, rok za izvršavanje, geografska lokacija itd. Njih unosi vlasnik projekta prilikom kreiranja novog projekta i vidljivi su na grafičkom sučelju. Za svaki projekt može se vezati i niz drugih atributa koji se ažuriraju kroz izvođenje projekta - to su aktivnosti na projektu, kontakti, prilozi, galerija fotografija i podsjetnici.



Slika 5. UML dijagram klase projekta

Projekt je logički okvir unutar kojeg se provode određene aktivnosti. Na primjer: projekt može biti 'Organiziranje proslave Dana Grada', na kojem će biti određeni broj aktivnosti (organiziranje koncerta, priprema domjenka, organizacija svečane sjednice i sl.). Premda te aktivnosti ne moraju imati direktne poveznice, one su logički povezane jer pripadaju istom projektu. Nakon što se odrade sve aktivnosti, projekt se zatvara. Kako na svakom projektu postoji čitav niz aktivnosti koje iziskuju kontakte sa djelatnicima jave uprave, ali i sa osobama iz drugih institucija (druge gradske tvrtke, ministarstva, državne tvrtke, dobavljači....) potrebno je za svaki projekt osigurati i rezime svih relevantnih kontakata koji su prikazani na grafičkom sučelju prilikom pregleda projekta. Uz sve informacije, projekt može imati i određene priloge u vidu različitih dokumenata: npr. ugovori, računi, skice, fotografije i sl. Nad projektom se mogu dodavati podsjetnici koji ukazuju na određene akcije koje se moraju izvršiti do određenog roka. Na grafičkom sučelju bit će prikazane u obliku led indikatora. Svaki projekt može se povezati sa geografskom lokacijom koja se na grafičkom sučelju prikazuje na GPS platformi (Slika 6).



Slika 6. Prikaz projekata na GPS platformi

IV. IMPLEMENTACIJA SUSTAVA

Svaka jedinica javne uprave (grad, općina. županija) predstavlja jednu domenu unutar koje se izvode aktivnosti vezane za pritužbe građana, unutarnje poslove ili aktivnosti vezane na projekt. Domene imaju jedinstveni identifikator po kojemu se međusobno razlikuju unutar sustava. Korisnici unutar jedne domene nemaju prava pristupa aktivnostima unutar druge domene osim ako politika praćenja drugačije ne nalaže, npr. administrator domene Virovitičko-podravske županije može imati pristup svojim *child* domenama. Cjelokupan sustav sa skladištima podataka izveden je u *cloud* okruženju.

Slika 7. Domene spojene na sustav u cloud okruženju

V. ZAKLJUČAK

U radu su opisane prednosti korištenja informacijkih sustava u jedinicama javne uprave sa naglaskom na workflow sustave. Predstavljena je arhitektura informacijskog sustava podijeljena na tri sloja: sloj medijacije koji daje potporu procesima prikupljanja podataka ili primanja aktivnosti iz različitih izvora, sloj baze podataka koji sadri više skladišta podataka o radu jedinice samouprave, baza podataka o pritužbama građana, baza podataka o procesima koji su u tijeku, baza podataka o projektima i sl. Iznad sloja baze podataka prikazan je prezentacijski sloj, čija osnova zadaća je kreiranje izvješća, njihov prikaz na grafičkim sučeljima te distribucija izvješća. Na primjerima je opisan workflow sustav za izvršavanje aktivnosti u jedinicama javne uprave. Opisana su grafička sučelja temeljena na web tehnologijama za prikazivanje workflow aktivnosti i projekata jedinice javne uprave.

VI. LITERATURA

[1] A. Pawlowska, "Information Technology in Service of Public Administration", neobjavljen rad

[2] E. Zembia i I. Obłąk, "The Survey of Information System in Public Administration in Poland,", Interciplinary Journal of Information, Knowledge, and Management, vol 9, 2014., pp. 31-56

[3] W. van der Aalst, "Generic Workflow Models", Coorporative Information Systems, 1999, pp. 115-126

[4] J.A. Zachman, "A framework for information system architecture, IBM System Journal, vol 26, 1987, pp. 276-292

[5] A. M. Langer, Analysis and Design of Information Systems, 2008, pp. 51-58

[6] Ž. Panian, K. Ćurko, Poslovni informacijski sustavi, Zagreb, 2010.

[7] V. Bosilj Vukšić, M. Pejić Bach i ostali, Poslovna informatika, Zagreb, 2009.

.

Analiza sustava za nadzor potrošnje posebnih goriva

S. Bojko*, V. Kumpar-Tasevski **, M. Mateljan * i A. Topić *
* Financijska agencija, Vukovarska 70, Zagreb, Hrvatska
** Ministarstvo financija, Carinska uprava, Alexandera von Humboldta 4a, Zagreb, Hrvatska
slaven.bojko@fina.hr, Vesna.Kumpar-Tasevski@carina.hr, marino.mateljan@fina.hr, ana.topic@fina.hr

Sažetak - Ovaj rad analizira sustav za nadzor potrošnje posebnih goriva za namjene u poljoprivredi, ribarstvu i akvakulturi i za javne i gospodarske namjene u plovidbi kako su ga uspostavile Carinska uprava Ministarstva financija Republike Hrvatske i Financijska agencija. Pravni temelj ovakvog sustava nadzora potrošnje je europska Direktiva 2003/96 EZ o restrukturiranju sustava zajednice o oporezivanju energenata i el. energije, Zakon o trošarinama, Pravilnik o primjeni Zakona o trošarinama što se odnosi na plinsko ulje obojano plavom bojom za namjene u poljoprivredi, ribolovu, ribogojstvu i akvakulturi i Pravilnik o uvjetima i načinu oslobođenja od plaćanja trošarine na energente koji se koriste kao pogonsko gorivo za plovidbu. Sadašnjem sustavu nadzora prethodio je sustav posebnog poreza i knjižica goriva s neučinkovitim nadzorom.

Svrha uspostave sadašnjeg sustava je učinkovitiji nadzor i dostupnost podataka u realnom vremenu. Kako sustav mora omogućiti jednostavnu i brzu realizaciju korištenja prava na kupnju posebnih goriva velikom broju korisnika na velikom broju prodajnih mjesta, implementiran je pomoću tehnologija koje se koriste u sustavima plaćanja putem platnih kartica izdavanjem posebnih kartica goriva korisnicima, označavanjem plinskog ulja propisanim markerima i nadzorom samog korištenja plinskog ulja kod korisnika.

Ovaj rad će detaljno opisati način realizacije i učinke propisanog cjelovitog sustava nadzora.

I. UVOD

Predmet ovog rada je analiza utjecaja uvođenja naprednijeg sustava za nadzor potrošnje posebnih goriva. Novi sustav je uveden jer se pretpostavljalo da stari sustav ne sprječava adekvatno eventualne zloupotrebe posebnih goriva. Cilj ovog rada je, osim da opiše implementaciju novog sustava, kvantificirati efekt primjene sustava kroz promatranje relevantnih statističkih vrijednosti pojedinih pokazatelja kroz određeni niz godina prije i poslije uvođenja novog sustava.

II. ZAKONSKA REGULATIVA SUSTAVA

Nadzor potrošnje plinskog ulja obojanog plavom bojom korisnika u poljoprivredi, ribolovu i akvakulturi, kao i plovidbi uspostavljen je u svrhu kontrole potrošnje i sprečavanja zloporaba korištenja. Prethodni je sustav nadzora potrošnje plinskog ulja obojanog plavom bojom bio uređen putem Knjižica goriva svakog pojedinog korisnika, s obvezom popunjavanja od strane prodavatelja prilikom prodaje, a kontrola takve evidencije bila je uvelike otežana zbog pojedinačnog nadzora, s ograničenim opsegom uslijed nedostatnosti ljudskih resursa i relativne dostupnosti podataka. Današnji način kontrole potrošnje putem Kartica goriva uspostavljen je 1. srpnja 2010. godine.

Pravni temelj sadašnjeg nadzora potrošnje je pravna stečevina Europske unije koja između ostalih kategorija trošarinskih roba regulira i oporezivanje trošarinom energenata. Nadzor potrošnje plinskog ulja sa sniženom trošarinom do nule propisuje europska Direktiva 2003/96 EZ o restrukturiranju sustava zajednice o oporezivanju energenata i el. energije. Kako se sniženjem trošarine zadire u fiskalne učinke državnog proračuna svake države članice, plinska ulja koja imaju sniženu trošarinu moraju biti označena u svrhu smanjenja zloporaba i povećane učinkovitosti kontrole. Način i vrsta označavanja plinskih ulja propisan je Direktivom Vijeća 95/60 EZ o fiskalnom označavanju plinskog ulja i kerozina.

Nacionalni temelji ovog nadzora potrošnje su Zakon o trošarinama („Narodne novine“ br. 22/2013, 32/2013, 81/2013), Pravilnik o primjeni Zakona o trošarinama što se odnosi na plinsko ulje obojano plavom bojom za namjene u poljoprivredi, ribolovu, ribogojstvu i akvakulturi („Narodne novine“ br. 156/2013), te Pravilnik o uvjetima i načinu oslobođenja od plaćanja trošarine na energente koji se koriste kao pogonsko gorivo za plovidbu („Narodne novine“ br. 58/2013, 72/2013,69/2014, 157/2014).

Navedenim pravnim okvirom uspostavljen je sustav nadzora potrošnje plinskog ulja obojanog plavom bojom za namjene u poljoprivredi ribolovu, ribogojstvu i akvakulturi, te plovidbi, pomoću kartice goriva za poljoprivredu i ribarstvo, te plovidbu. Uređen je i izgled te sadržaj i oblik kartice goriva. Uređen je i način korištenja kartice goriva, te sukladno tome i programska podrška za sustav kontrole pri prodaji plinskog ulja obojanog plavom bojom. Uređena je mogućnost usporedbe zaduženja i odobrenih količina, te povezivanja odobrenih količina sa kvotama potrošnje plinskog ulja za te propisane namjene.

Osnove sustava implementirane iz europske pravne stečevine propisane u Zakonu o trošarinama sadržavaju odredbe koje se odnose na sniženje trošarine do 0,00 kuna, za plinsko ulje u tu svrhu označeno i obojano plavom bojom. Pravo korištenja navedenog ulja propisano je korisnicima prava u poljoprivredi, ribogojstvu, akvakulturi i ribolovu, te plovidbi. Da bi određeni subjekt postao korisnikom prava za namjene u poljoprivredi mora biti upisan u odgovarajući upisnik poljoprivrednih gospodarstava pri ministarstvu nadležnom za poljoprivredu. Korisnici prava u ribogojstvu, akvakulturi i ribolovu moraju posjedovati povlasticu sukladno posebnim propisima. Korisnicima prava propisani su pokazatelji prava na godišnju potrošnju. Pokazatelje utvrđuje ministarstvo nadležno za poljoprivredu uz zadovoljenje određenih uvjeta. Spomenuto ministarstvo prava na potrošnju goriva posebnih namjena, iskazuje putem kartica goriva, koje su javna i zaštićena isprava i obrazac namijenjeni za kontrolu. Na Kartici goriva, vidljiva je oznaka nadležnog ministarstva, naziv kartice goriva za poljoprivredu i ribarstvo, broj kartice goriva, OIB korisnika prava ili matični broj poljoprivrednog gospodarstva ili CFR broj plovila, kratki naziv korisnika prava i oznaka ispostave Agencije za plaćanja odnosno nadležnog tijela koje je izdalo povlasticu za ribolov ili za uzgoj ribe odnosno za akvakulturu. Kartica goriva za poljoprivredu i ribarstvo u vlasništvu je Ministarstva poljoprivrede.

Uvjet, propisan u naznačenim propisima, za stjecanje prava korisnika u plovidbi je upis fizičkih i pravnih osoba u odgovarajući upisnik odnosno očevidnik kao vlasnika, brodara, ili korisnika financijskog leasinga plovnih objekata ili plovila, registriranih za gospodarske svrhe sukladno posebnom propisu, a djelatnost je komercijalni prijevoz putnika ili tereta. Pravo na karticu goriva za plovidbu stječe se na temelju rješenja koje izdaje ministarstvo nadležno za plovidbu po službenoj dužnosti za svaki pojedinačni brod osim za brod koji plovi u međunarodnoj plovidbi i međunarodnoj unutarnjoj plovidbi. Ministarstvo je ujedno i vlasnik Kartice te utvrđuje i kriterije za određivanje najveće dopuštene količine godišnje potrošnje. Kartica gorivo za plovidbu sadrži podatke o nazivu ministarstva nadležnog za plovidbu, naziv kartice gorivo za plovidbu, OIB korisnika prava, ime odnosno oznaku plovnog objekta odnosno plovila, broj kartice goriva, ime ili naziv korisnika prava.

III. METODOLOGIJE NADZORA KORIŠTENJA PRAVA NA POSEBNA GORIVA

Cjelovit sustav nadzora mora omogućiti ograničavanje korištenja posebnih goriva kod pojedinog korisnika samo do odobrene količine i dokazivanje zloupotrebe posebnih goriva kada se koriste u neodobrene svrhe.

Prvi cilj nadzora može se ostvariti nadzorom proizvodnje posebnih goriva uspostavom odgovarajuće evidencije i usporedbom evidencije proizvodnje s evidencijom distribucije posebnih goriva krajnjim korisnicima. Stoga je [1] propisao vođenje evidencija kojima se evidentira ukupna proizvedena ili uvezena količina posebnog goriva, te distribucija posebnih goriva krajnjim prodavateljima. Posebno gorivo, već prilikom proizvodnje ili uvoza, sukladno [1] mora biti označeno propisanim indikatorom i bojom što se izvršava u posebnim trošarinskim skladištima ovlaštenog držatelja trošarinskog skladišta. Kontrolom navedene evidencije moguće je utvrditi koliko je svako prodajno mjesto zadužilo posebnog goriva u određenom periodu.

Kako bi se spriječilo da prodajna mjesta distribuiraju posebno gorivo neovlaštenim korisnicima, dužna su gorivo razdužiti evidentiranjem svake prodaje posebnog goriva kroz sustav nadzora potrošnje. Uspoređivanjem evidencije prodaje posebnog goriva krajnjim korisnicima s evidencijom distribucije opisane u prethodnom odlomku, moguće je ustanoviti da li je pojedino prodajno mjesto svu zaduženu količinu posebnog goriva prodalo ovlaštenim krajnjim korisnicima do pojedinačnih razina dozvoljenih količina, ili postoje razlike.

Mada je u [1] propisano da prodavatelji posebnih goriva prodaju moraju izvršiti izravnim utakanjem u poljoprivredne strojeve, ribarice ili plovila, postoji mogućnost da će zlonamjerne osobe posebno gorivo koristiti u strojevima za svrhe drugačije od onih propisanim zakonom. Odgovarajućim inspekcijskim nadzorom, moguće je analizirati pojavnost propisanih indikatora u spremnicima strojeva čija namjena nije propisana kao namjena za koju je moguće koristiti posebna goriva. Ovlaštene analize postojanja indikatora mogu se koristiti kao dokaznice u odgovarajućim prekršajnim ili kaznenim postupcima, te se time provode mjere sprječavanja zloupotrebe posebnih goriva.

U svrhu omogućavanja što učinkovitijeg nadzora, slijedeći postupci provode se sukladno opisanim pravilima:

- za namjene u poljoprivredi pravo na potrošnju plinskog ulja obojanog plavom bojom korisnici ostvaruju na temelju podnesenih zahtjeva za izravna plaćanja u prethodnoj godini odgovarajućem tijelu sukladno [1]. Pravo na karticu goriva za plovidbu stječe se na temelju rješenja kojeg izdaje ministarstvo nadležno na plovidbu.

- na temelju prethodno spomenutih zahtjeva korisnicima se izrađuju kartice. Karticu goriva korisnicima prava na potrošnju plinskog ulja obojanog plavom bojom u poljoprivredi izdaje Agencija za plaćanja u poljoprivredi, ribarstvu i ruralnom razvoju (u daljnjem tekstu: APPRRR), a korisnicima prava u ribarstvu nadležno tijelo koje je izdalo povlasticu za ribolov ili povlasticu za uzgoj ribe i drugih morskih organizama (u daljnjem tekstu: povlasticu za uzgoj) odnosno povlasticu za akvakulturu. Kartica goriva smatra se javnom i zaštićenom ispravom i obrascem kao kartica za kontrolu potrošnje goriva posebnih namjena čiji se sadržaj, oblik, način zaštite i obvezna primjena posebno uređuje. Karticu goriva za plovidbu izdaje ministarstvo nadležno za plovidbu. Kartica goriva za plovidbu smatra se javnom i zaštićenom ispravom i obrascem kao kartica za kontrolu potrošnje goriva posebnih namjena čiji se sadržaj, oblik, način zaštite i obvezna primjena posebno uređuje.

- svaki korisnik dobiva određenu količinu goriva za pojedinu godinu. Podatke o odobrenoj količini plinskog ulja obojanog plavom bojom za namjene u ribolovu, ribogojstvu i akvakulturi obračunava ministarstvo nadležno za ribarstvo koje je izdalo povlasticu za ribolov ili povlasticu za uzgoj odnosno povlasticu za akvakulturu, jednom godišnje. Podatke o odobrenoj količini plinskog ulja obojanog plavom bojom za namjene u poljoprivredi obračunava Agencija za plaćanja u poljoprivredi, ribarstvu i ruralnom razvoju jednom godišnje.

- korisnik prava koji je preuzeo karticu goriva dužan je prije kupovine plinskog ulja obojanog plavom bojom za namjene u poljoprivredi, ribarstvu, ribogojstvu i akvakulturi, prodavatelju predočiti:

1. karticu goriva i

2. identifikacijsku ispravu korisnika prava na potrošnju plinskog ulja obojanog plavom bojom propisanu posebnim propisima koja obuhvaća iskaznicu obiteljskog poljoprivrednog gospodarstva, ili rješenje o upisu u upisnik poljoprivrednih gospodarstava, ili povlasticu za ribolov, ili povlasticu za uzgoj, ili povlasticu za akvakulturu.

- plinsko ulje se može koristiti kao pogonsko gorivo za plovidbu ako:

1. se neposredno nabavlja iz trošarinskog skladišta te ako

2. korisnik prava pri nabavi podnese isprave kojima dokazuje komercijalno obavljanje plovidbe u međunarodnoj pomorskoj i međunarodnoj unutarnjoj plovidbi.

- prodavatelj koji prodaje plinsko ulje obojano plavom bojom korisnicima prava dužan je prije prodaje izvijestiti korisnika prava o raspoloživoj količini te izvršiti prodaju izravnim utakanjem do iznosa raspoložive količine.

- trošarinski obveznik i druga pravna ili fizička osoba – prodavatelj, koji prodaje plinsko ulje obojano plavom bojom za namjene u poljoprivredi, ribolovu, ribogojstvu i akvakulturi, dužan je voditi Evidenciju o kupcima plinskog ulja obojanog plavom bojom koji gorivo nabavljaju radi daljnje prodaje.

IV. TEHNOLOŠKA IMPLEMENTACIJA SUSTAVA

Za uspostavu učinkovitijeg sustava evidentiranja i nadzora potrošnje posebnih goriva za namjene u poljoprivredi, ribarstvu, akvakulturi i za javne i gospodarske namjene u plovidbi od onog koji je postojao do 2010. godine bilo je potrebno omogućiti kupnju posebnog goriva velikom broju korisnika na prodajnim mjestima diljem teritorija Republike Hrvatske, mogućnost za zloporabu prava korisnika svesti na najmanju moguću, osigurati da godišnja potrošnja korisnika ne premašuje odobrenu godišnju količinu te osigurati da proces provjere prava i evidentiranja kupnje traje što je moguće kraće.

Iz opisa zahtjeva može se zaključiti da je riječ o tipičnom slučaju dvostranog tržišta, odnosno, o usluzi koja mora zadovoljiti potrebe dviju potpuno različitih grupa korisnika usluge, u ovom slučaju korisnika prava na potrošnju posebnih goriva i distributera posebnih goriva.

U bankarskom sektoru slične zahtjeve već godinama ispunjavaju kartični sustavi koji se temelje na magnetskim i čip karticama izdanim korisnicima usluge plaćanja putem platnih kartica te EFT POS uređajima postavljenih kod trgovaca koji su voljni prihvatiti platne kartice kao sredstvo plaćanja. Kako je tehnologija kartičnih sustava dokazana u praksi, bilo je logično primijeniti istu za uspostavu učinkovitog sustava nadzora potrošnje posebnih goriva za namjene u poljoprivredi, ribarstvu, akvakulturi i za javne i gospodarske namjene u plovidbi.

A. Sudionici u kartičnoj shemi

Svako kartično poslovanje se temelji na reguliranim odnosima među sudionicima kartične sheme u kojoj svaki od sudionika imaju svoju ulogu, a informacijska i komunikacijska tehnologija je u službi kartičnog poslovanja gdje ima za cilj osigurati interakciju među sudionicima kartične sheme. U platnim kartičnim shemama prepoznajemo slijedeće ključne sudionike: izdavatelj (engl. Issuer), stjecatelj (engl. Acquirer), trgovac (engl. Merchant) i korisnik kartice (engl. Cardholder) kao shemu s četiri sudionika koja je primijenjena kod većine bankarskih kartica. Postoje i platne sheme s tri sudionika, gdje su izdavatelj i stjecatelj isti sudionik, te se ista primjenjuje kod kartica koje ne izdaju banke, već specijalizirani poslovni subjekti.

U kartičnoj shemi sustava za nadzor potrošnje posebnih goriva prepoznajemo Ministarstvo poljoprivrede i Ministarstvo pomorstva prometa i infrastrukture u ulozi izdavatelja i stjecatelja, distributere ili prodavatelje posebnog goriva koje prepoznajemo u ulozi trgovaca te poljoprivrednike, ribare i pomorce kao korisnike kartica. Ova shema spada u tzv. „zatvorene“ kartične sheme ili sheme sa tri sudionika. Dio poslova izdavatelja i stjecatelja, a temeljem odgovarajućih odluka i ugovora, u njihovo ime izvršava Financijska agencija (u daljnjem tekstu Fina) kao tehnički pružatelj usluge.

B. Izdavanje kartica

Izdavateljima kartica zainteresirani subjekti predaju zahtjeve s potrebnim podacima i temeljem kriterija utvrđenih u [2], [3] izdavatelji odobravaju korisnicima godišnju količinu posebnog goriva - kvotu. Podatke o korisnicima i godišnjim kvotama izdavatelji šalju na siguran način Fini, gdje se isti učitavaju u kartični sustav. Temeljem zaprimljenih podataka o korisnicima u kartičnom sustavu Fine kreiraju se računi korisnika s podacima o dozvoljenoj godišnjoj količini posebnog goriva – kvoti te se kreiraju podaci o kartici korisnika. U kartičnom sustavu Fine se također odrede podaci potrebni za izradu kartica i na siguran način dostavljaju ovlaštenoj ustanovi za izradu zaštićenih isprava Republike Hrvatske, tj. Agenciji za komercijalnu djelatnost (u daljnjem tekstu AKD).

Za potrebe izrade kartica, izdavatelji moraju prikupiti osobne podatke korisnika kartica. Stoga su kao voditelji zbirki osobnih podataka imenovane Agencija za plaćanje u poljoprivredi, ribarstvu i ruralnom razvoju, Uprava ribarstva Ministarstva poljoprivrede i Ministarstvo pomorstva, prometa i infrastrukture koji su zbirke prijavili u odgovarajući registar zbirki osobnih podataka koji vodi Agencija za zaštitu osobnih podataka.

AKD izrađuje kartice „Gorivo za poljoprivredu i ribarstvo“ (u daljnjem tekstu GPR kartica) i „Gorivo za plovidbu“ (u daljnjem tekstu GP kartica) koje su prema [4] standardu registrirane od ovlaštenog tijela „Bank of America“ i dodijeljen im je IIN broj – „Issuer Identification Numbers“. IIN broj GPR kartice je 639732, a GP kartice 637162. Kartice su izrađene sukladno [5], [6] i [7], s odgovarajućim prilagodbama u strukturi podataka kako bi se moglo zadovoljiti propisane zahtjeve. Izrađene kartice se distribuiraju korisnicima, a kako bi kartice bile spremne za korištenje iste je potrebno aktivirati.

Slika 1. Izgled prednje strane kartice „Gorivo za poljoprivredu i ribarstvo“

C. Prihvat transakcija

U ulozi stjecatelja Fina je dužna omogućiti prihvat kartica posebnih goriva na skoro 600 prodajnih mjesta diljem teritorija Republike Hrvatske. Kao i u svakoj kartičnoj shemi trgovci su dužni formalizirati svoju ulogu trgovca sklapanjem ugovora o prihvatu kartica te dostavljanjem podataka o prodajnim mjestima na kojima žele prihvaćati kartice posebnih goriva. Temeljem ugovornih odnosa s trgovcima podaci o svim prodajnim mjestima se unose u sustav nadzora potrošnje goriva te im se definiraju identifikacijski brojevi u sustavu, a sva prodajna mjesta potrebno je opremiti EFT POS uređajima na kojima je omogućen prihvat GPR i GP kartica.

Ukoliko trgovci imaju svoje vlastite EFT POS uređaje i kartičarske sustave, te su ih voljni prilagoditi za prihvat GPR i GP kartica, moguće je povezati takve sustave s kartičnim sustavom Fine. Komunikacija koja se odvija između takva dva sustava temelji se na „Host to host“ protokolu razmjene poruka kartičnih transakcija definiranom prema [8].

Za prodajna mjesta malih distributera posebnog goriva koji nemaju vlastite EFT POS uređaj, Fina je osigurala odgovarajuće EFT POS uređaje za prihvat GPR i GP kartica. Svaki EFT POS uređaj prije isporuke na prodajno mjesto prolazi kroz proces inicijalizacije u kojem je EFT POS uređaju na siguran način izdan i učitan X509 v3 certifikat kojeg koristi za ostvarivanje sigurne komunikacije sa sustavom u Fini putem Interneta. Pritom se koristi obostrana SSL/TLS autentikacija, tj. prilikom uspostave sigurne veze svojim certifikatima predstavljaju se klijent (EFT POS uređaj) i poslužitelj te se izmjenjuje jednokratni simetrični ključ kojim se štiti (enkriptira) sva komunikacija unutar uspostavljene sesije. Svi korišteni digitalni certifikati uređaja kreirani su sukladno X.509 v3 standardu, u kojem su definirane informacije koje se pohranjuju u certifikatu i sam format zapisa tih informacija.

Terminalska EFT POS aplikacija omogućuje trgovcu provođenje transakcija upita u stanje, prodaje, strona i naknadnog unosa te razmjenu kartičnih poruka sa sustavom u Fini koja se temelji na [8]. Prije samog istakanja željene količine goriva korisniku kartice posebnih goriva trgovac je dužan napraviti transakciju upita u stanje kako bi provjerio raspoloživu količinu goriva na računu korisnika. Nakon istakanja dozvoljene količine goriva trgovac koristeći karticu korisnika zadaje transakciju prodaje na točno istočenu količinu goriva. Ukoliko transakcija zadovolji sve formalne provjere i količina goriva za autorizaciju ne premašuje raspoloživu količinu na računu korisnika, transakcija biva odobrena. Kako bi se osigurala dvofaktorska autentikacija korisnika kartice, uz karticu goriva propisano je i predočenje osobnih dokumenata korisnika kartice te se je isti dužan potpisati na potvrdu o odobrenoj transakciji. Podaci o svim odobrenim i odbijenim transakcijama bilježe se u kartičnom sustavu Fine i na EFT POS uređaju te se na kraju svakog dana provodi poravnanje evidencija EFT POS uređaja s kartičnim sustavom Fine uspoređivanjem podataka o broju zabilježenih transakcija kao i o ukupnim količinama prodanog posebnog goriva. Tek se po provedenom postupku poravnavanja evidencija između EFT POS uređaja i kartičnog sustava u Fini transakcije smatraju konačnima i proknjiženima. U slučaju ustanovljenih razlika, posebice razlika između evidencija u blagajni distributera posebnog goriva i evidencija kartičnom sustavu, iste se rješavaju putem propisanih reklamacijskih postupaka koje može pokrenuti bilo koji od triju sudionika kartične sheme.

D. Izvještavanje nadzornih tijela

Podaci u sustavu nadzora potrošnje posebnih goriva omogućavaju praćenje potrošnje korisnika kartica, ali isto tako i prodavatelja goriva. Svi podaci o zabilježenim transakcijama se jednom dnevno iz transakcijske baze šalju u izvještajno-analitički sustav Carinske uprave za potrebe provođenja nadzora. Za potrebe provođenja izdavateljskih i stjecateljskih poslova, Fina je uspostavila izvještajni sustav koji omogućuje izdavateljima pregled podataka registriranih u sustavu.

V. REZULTATI PRIMJENE SUSTAVA

Učinci primjene sustava promatrani su samo na slučaju primjene u poljoprivrednoj proizvodnji radi dostupnosti odgovarajućih podataka. Naime, podaci na temelju kojih se dodjeljuju kvote za namjene u plovidbi, ribolovu i akvakulturi nisu javno dostupni. S druge strane, podaci o ukupno obrađenim poljoprivrednim površinama, što je prema [2] temeljni kriterij za određivanje kvote, dostupni su kroz [9]. Potrošnja posebnih goriva, odnosno, plinskog ulja za namjene u poljoprivredi čini većinu ukupne potrošnje, te je time najznačajnija za promatranje i analizu.

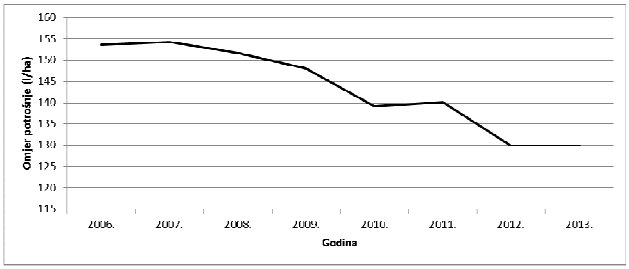
U [10], [11] i [12], u tablicama „3.3.6 Dizelsko gorivo" tih dokumenata, objavljeni su podaci o potrošnji dizel goriva, odnosno, plinskog ulja po pojedinim sektorima. Iskazana je potrošnja dizel goriva u milijunima litara za sektor poljoprivrede po godinama za razdoblje od 2006. do 2013. godine. Iz [9] uzeti su podaci o ukupnoj površini korištenog poljoprivrednog zemljišta za isto razdoblje i uspoređeni u Tablici 1 i Slici 2.

Iz podataka je vidljivo veliko smanjenje potrošnje dizel goriva po jedinici obrađene poljoprivredne površine 2010. godine te 2012. godine sa zadržavanjem gotovo jednake razine omjera potrošnje po jedinici poljoprivredne površine u 2013. godini u odnosu na prethodnu godinu. Navedene promjene u podacima, prema [2], koincidiraju s uvođenjem novog sustava nadzora potrošnje posebnih goriva u srpnju 2010. godine, odnosno, činjenicom da je 2011. godine kvota povećana za polovicu neiskorištene kvote iz 2010. godine te da je time 2012. godina prva godina u kojoj je dosljedno primjenjena kvota izračunata prema propisanim pravilima u [2].

Zbog navedenih utjecaja, iz detaljnih promatranja izuzete su 2010. i 2011. godina, kao godine izuzetaka nastalih polovičnom primjenom novog sustava i dodatnim povećanjem kvota.

Tablica 1 Korištena poljoprivredna površina i potrošnja plinskog ulja po godinama

	Korištena poljoprivredna površina u hektarima (ha)	Potrošnja plinskog ulja za namjene u poljoprivredi u milijunima litara (Ml)	Omjer Potrošnje plinskog ulja i Korištene poljoprivredne površine (l/ha)
2006.	1.230.183	189,0	153,64
2007.	1.201.756	185,5	154,36
2008.	1.289.091	195,6	151,73
2009.	1.299.582	192,4	148,05
2010.	1.333.835	185,7	139,22
2011.	1.326.083	185,8	140,11
2012.	1.330.973	173,0	129,98
2013.	1.301.985	169,1	129,88



Slika 2 Prikaz omjera potrošnji plavog dizela po jedinici poljoprivredne površine od 2006. do 2012. godine

Statistička analiza za radzoblje od 2006. do 2009. godine pokazuje da je godišnje prosječno potrošeno 151,94 litara dizel goriva po hektaru korištene poljoprivredne količine. Pearsonov koeficijent korelacije veličina potrošnje i površine poljoprivrednog zemljišta za razdoblje od 2006. do 2009. godine od 0,92 samo dosljedno pokazuje snažnu, propisima određenu, ovisnost tih veličina. Kako je prosjek omjera potrošnje dizelskog goriva po jedinici poljoprivredne površine za 2012. i 2013. godinu samo 129,92 litara dizel goriva po hektaru obrađenih poljoprivrednih površina, može se zaključiti da je 2012. i 2013. godine potrošnja dizelskih goriva smanjena za 14% u odnosu na prosjek godišnje potrošnje dizel goriva po jedinici površine prije uvođenja sustava, odnosno, u razdoblju od 2006. do 2009. godine.

Kako Pearsonov koeficijent nema previše smisla promatrati na skupu parova veličine 2, odnosno, samo za 2012. i 2013. godinu, nužno je skrenuti pažnju na pad vrijednosti standardne devijacije omjera potrošnje dizelskog goriva po jedinici poljoprivredne površine sa 2,82 za razdoblje od 2006. do 2009. godine na samo 0,07 za 2012. i 2013. godinu, odnosno, čak za dva reda veličine.

VI. ZAKLJUČAK

U ovom radu prezentirali smo način primjene i učinke sustava za nadzor potrošnje posebnih goriva. Prezentirani rezultati ukazuju na snažnu povezanost primjene sustava na smanjenje potrošnje posebnih goriva u poljoprivrednoj proizvodnji, te time i smanjenje upotrebe posebnih goriva mimo propisanih namjena. Vrijednost standardne devijacije omjera potrošnje dizelskog goriva po jedinici poljoprivredne površine za 2012. i 2013. godinu koja je za 4 reda veličine manja od promatrane veličine ukazuje da je razlika vjerojatno manja od statističke pogreške. Ukoliko podaci iz slijedećih godina pokažu ovakvu konzistentnost, svakako će se nametati zaključak da je primijenjeni sustav nadzor uklonio gotovo sve anomalije u korištenju plavog dizela u poljoprivredi. Ovakav sustav moguće je efikasno primijeniti i za nadzor bilo koje druge vrste robe ili usluga koje imaju smanjenu cijenu za krajnje korisnike ali samo za određene namjene. Kako autorima nije poznat sličan sustav nadzora potrošnje goriva s smanjenim trošarinama niti u EU niti bilo gdje drugdje u svijetu, već se isti većinom svode na markiranje goriva ili povrat iznosa trošarine prema plaćenim računima, svakako treba promovirati njegovu primjenu u

zemljama članicama i kandidatkinjama za pristup EU koje imaju zajedničku politiku trošarina na goriva za namjene u poljoprivredi.

LITERATURA

[1] Zakon o trošarinama, Narodne novine Republike Hrvatske

[2] Pravilnik o primjeni Zakona o trošarinama što se odnosi na plinsko ulje obojano plavom bojom za namjene u poljoprivredi, ribolovu i akvakulturi, Narodne novine Republike Hrvatske

[3] Pravilnik o uvjetima i načinu oslobođenja od plaćanja trošarine na energente koji se koriste kao pogonsko gorivo za plovidbu, , Narodne novine Republike Hrvatske

[4] ISO 7812 - Identification cards, Identification of issuers

[5] ISO 7810 - Identification cards - Physical characteristics

[6] ISO 7811 - Identification cards - Recording technique

[7] ISO 7813 - Identification cards - Financial transaction cards

[8] ISO 8583:1993 - Financial transaction card originated messages

[9] Državni zavod za statistiku, www.dzs.hr, PC-Axis baze podataka, Tablica 1.1.1. Površina korištenoga poljoprivrednog zemljišta po kategorijama u hektarima, Republika Hrvatska

[10] Dr. sc. Branko Vuk i ostali, „Godišnji energetski pregled – energija u Hrvatskoj 2011," Energetski institut Hrvoje Požar Zagreb, Zagreb: Ministarstvo gospodarstva

[11] Dr. sc. Branko Vuk i ostali, „Godišnji energetski pregled – energija u Hrvatskoj 2012," Energetski institut Hrvoje Požar Zagreb, Zagreb: Ministarstvo gospodarstva

[12] Dr. sc. Branko Vuk i ostali, „Godišnji energetski pregled – energija u Hrvatskoj 2013," Energetski institut Hrvoje Požar Zagreb, Zagreb: Ministarstvo gospodarstva

Social Media Marketing Activity in Western Balkan Banking Industry

Nikola Levkov*, Marina Mijoska**, Dimitar Jovevski***
* University Ss. Cyril and Methodius – Skopje, Faculty of Economics – Skopje, Department of Management, Skopje, R. of Macedonia
** University Ss. Cyril and Methodius – Skopje, Faculty of Economics – Skopje, Department of E – business, Skopje, R. of Macedonia
*** University Ss. Cyril and Methodius – Skopje, Faculty of Economics – Skopje, Department of E – business, Skopje, R. of Macedonia

Abstract - The number of people who are using social media websites is constantly growing. Customers increasingly expect banks to offer more services via social media platforms. Recent studies confirmed that customer retention is likely to increase at banks that offer a strong social media experience. This paper explores the maturity of social media use in banking service sector with particular focus on Facebook, Twitter, YouTube and LinkedIn. Empirical data have been collected mainly through visual inspection of social media accounts of major banks in Western Balkan Countries and the results have been compared and presented. Various types of indicators for measurement of social media activity have been used: number of fans, subscribers and followers, number of video views, number of posts in last 15 days, visibility of the social media channel on the official web page, social engagement and others. Based on these and other data we made quantitative comparison and qualitative analysis of the intensity of social media use in banking industry in the region of Western Balkan Countries.

I. INTRODUCTION

According to Digital Insights blog and its social media 2014 statistics, Facebook has 1.28 billion registered monthly active users, Twitter has 255 million, and LinkedIn has 187 million, while Youtube has 1 billion total users globally. Looking closely to this statistic an interesting question is placed in front of business practitioners: Where else can you find such a large number of potential buyers at one place and having the opportunity for instant direct communication with all of them? Businesses must understand the role of information technology for their marketing strategy and its impact on shaping the global marketplace in order to survive and compete. Today technology plays vital role in peoples everyday life activities and marketers need to reshape their marketing strategies for communication and interaction with their customers. The global financial crises and economic recession from 2008 affected firms' marketing strategies and budgets. Usually companies in a period of financial crises reduce their marketing and IT budgets. But marketing function plays key role during and after financial crises for companies to remain profitable and customer responsive. Scholars (Kirtiş and Karahan, 2011) have confirmed that for both big and small firms, social media marketing as strategy compared to traditional media help to advertise products and services during and after economic crises with lower cost.

In the last two decades marketing as academic discipline and marketing activities as business practices have been changed dramatically. The main driver for such a change in the domain of marketing management is the emergence of the internet and the era of web 2.0. Several important changes occurred in the domain of contemporary marketing activities: 1. broader access to marketing media – everyone who have access to internet can distribute content with low cost not only corporations with large marketing budgets (Tuten, 2008) 2. Power shift – empowered and intelligent customers versus weaker competitive position of corporations (Porter, 2001) (Urban, 2006) 3. Higher level of interaction – old fashioned one way of communication is changed to two way direct communication (Kim and Ko, 2012).

Answering the question what can be included under term social media is not easy task. Kaplan & Haenlein (2010) indicate that often there is confusion among managers and academic researchers regarding the meaning of social media. The main question appointed by the same authors is: How social media differ from Web 2.0 and User Generated Content. They explain that Web 2.0 is representation of ideological and technological foundation while User Generated Content can be seen as sum of all ways in which people use Social Media. On the other hand social media is defined by these authors as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content (Kaplan & Haenlein, 2010). In their article Kaplan and Haenlein have classified social media in the following six categories: 1) blogs, 2) collaborative projects 3) social networking sites 4) content communities 5) virtual social worlds and 6) virtual game worlds.

Banking industry cannot stay aside just watching the enormous growth of social media activity. Banks like no other organizations rely on reputation. The positive brand

awareness and banking reputation is on stake. Since the global economic and financial crises client sensitivity regarding banking service reliability and trustworthiness has become even more important. Vemuri (2010) stresses the need for bridging the gap between banking and social media use. He advocates for stronger activity of banks on social media. Dissatisfied clients spread very fast bad news on the internet through number of tweets, retweets and viral videos with million views. These interactions create negative awareness and influence the way clients make decisions.

However a research study published by Teller Vision in 2013 showed that retail banks in USA, United Kingdom, Germany, France and Australia will spend considerably less comparing to other industries on social media marketing. This study confirms manifestation of vigilance behavior by banks when it comes to marketing through social media. Banks' caution regarding social media use is influenced also by result reported that only a quarter of 3000 quizzed consumers use social media to follow and keep up-to-date with certain companies. More than three quarters of consumers are predominantly on the various platforms to keep in touch with friends and family.

The results from Global Retail Banking Digital Marketing Report in 2013 reported that amount being spent by retail banks on social media is very small and as a final remark at the end of the report a notion is mentioned that social media use by banks has reached a "plateau". Hence we can say that banks are facing with a key challenge in the use of social media in terms of opportunities and threats. This challenging situation in social media use by banks was the main driver for the authors of this paper to delve more deeply into this problem. Second main driver for the research conducted in this paper was the lack of data about social media use by banking industry in the region of Western Balkan Countries. The authors expect to contribute to the research literature by covering this research gap.

II. RESEARCH METHODOLOGY

Large amount of data have been collected in order to achieve the research goal of the study related to the main research question: How much banks from Western Balkan countries are active in use of social media platforms? The region of western Balkan countries was chosen for this research because all countries have similar level of economic development and political environment (not EU members) which makes comparison between countries more reliable. Also most of the countries (Macedonia, Serbia, Kosovo, Bosnia and Herzegovina and Montenegro) in the past were belonging to one state and all have passed through more or less similar reforms in the banking sector. Hence, banks from six western Balkan countries were included in this research study: Macedonia, Kosovo, Serbia, Bosnia and Herzegovina, Montenegro and Albania.

Regarding the methodology the authors were following consistent procedure with several steps in the process of data collection. Firstly, a list was created with all licensed banks operating in the western Balkan. To compile the list authors have used official data from national banks of all countries in the region. At the beginning the total number of banks included in the list was 107 banks. Two banks, one from Serbia and one from Montenegro have been removed from the sample because they recently received the license and they still did not start to operate. Thus a list of 105 active banks was used to conduct the research. Also it is important to mention that only privately owned retail banks were included in the sample.

Secondly, four main social media platforms have been used to collect data about banks social media activity: Facebook, YouTube, Twitter and LinkedIn. These four platforms have been chosen as a result of checking the leading platforms by use in all countries from this region. We were using Alexa as web analytics tool and its rankings of top websites in each country from western Balkan region. According to Alexa (http://www.alexa.com/topsites/countries) four social media platforms Facebook, YouTube, Twitter and LinkedIn were always rated as top websites in front of all other social media platforms in all countries, except Kosovo and Montenegro for which there were no data by Alexa rankings. Also we were following the results from two industry reports. First was the Social Media Marketing Industry Report for 2014 ranking the same four platforms as most often used from marketers (Stelzner, 2014). The second report was the First Global Retail Banking Digital Marketing Report for 2013 who was also ranking these four platforms as most often used by banks in all regions in the world.

Thirdly we made visual inspection of all banks' official websites from the sample to determine social media visibility. If we could find on the official web page link to the social media platform we were using the link to navigate our search towards the respective platform and to collect data such as: number of Facebook fans, Twitter followers, YouTube subscribers etc. If there was no any link confirming presence on some of the four social media platforms, than we were checking existence of social media account by using the official name of the bank or their official brand abbreviation in a direct search on all four selected platforms. All these data have been collected in a period of one month somewhere from the middle of November to the middle of December in year 2014.

The research goals of the study were twofold: First, this study is mainly exploratory and descriptive in its nature with a research endeavor mainly focused to depict the situation with social media use in banking sector in one region, Western Balkan Region. Second the authors are trying to make comparison between the social media presence and activity in banking industry between the countries in order to identify interesting trends and substantial differences.

III. DATA ANALYSIS AND RESULTS

In the following section we present and analyze the collected data and discuss the results. Fig. 1 shows social media presence of banks from Western Balkan Countries measured by existence of official account on a particular platform.

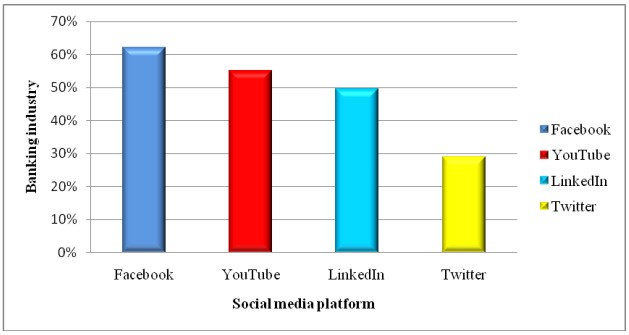


Figure 1. Social media platform presence by banks

The analysis of the results given in fig.1 shows that Facebook is the most often used platform by banks measured by existence of official bank fan page. Further are following YouTube, LinkedIn and at the end are the Twitter accounts. These results are consistent with both industry reports used in this study regarding the Facebook as most often used platform by retail banks. Slight difference exist when we compare the rest of the three platforms meaning that in the Western Balkan Countries banks are more present on YouTube comparing to Twitter and LinkedIn reported as more often used by retail banks globally.

In fig. 2 data are given about number of social media accounts, expressed as relative proportion to the number of banks in whole banking sector in a particular country. From the data on this figure we can notice that the presence of banks on certain social media platforms differs by country. Facebook is the platform most often used by banks from all countries except Macedonia. In Macedonia LinkedIn is the most often used social media platform based on number of official accounts. The second important trend easily noticeable is that YouTube and LinkdIn in all countries are more often used by banks than Twitter. But the presence of banks on YouTube and LinkedIn differs by country indicating that in Serbia, Kosovo and Montenegro, YouTube is more often used unlike in Albania, Bosnia and Herzegovina and Macedonia where LinkedIn is more often used.

Having this broad picture about banking industry presence on social media platforms in western Balkan countries, it is interesting to delve more deeply and to investigate social media activity on particular platforms beside only presence. First interesting social media metric used to compare social media activity of banking industry was number of Facebook fans, YouTube subscribers, Twitter and LinkedIn followers. In order to be able to compare the data from countries with different population and different potential for subscriptions to a particular platform we were using relative proportion of actual users to the total population of the country.

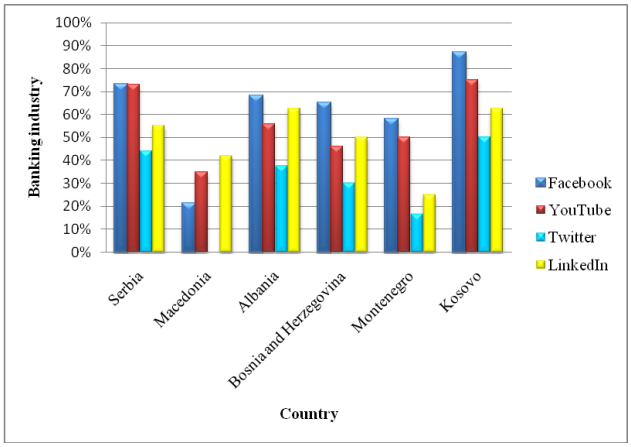


Figure 2. Social media platform presence by banks per country

Table 1 shows the percentile share of platform members in form of fans, subscribers or followers in relation to the total population of the country. The data from table 1 indicate that banks from all countries are most active on Facebook when we use Facebook fans as metric comparing to YouTube subscriptions, Twitter and LinkedIn followers. We also made visual inspection of Facebook timeline of all bank profiles for the whole region to determine the average time present on platform. The results indicate that until present moment on average banks have been present on Facebook for 1.9 years, on YouTube 1.3 years and on Twitter 0.9 years. For LinkedIn we could not determine when a profile was created and thus we have not included data for this social media platform. However data which show the length of time being on a certain social media platform are quite consistent with the number of active fans, subscribers or followers as a social media activity metric.

TABLE I. RELATIVE PROPORTION OF PLATFORM USERS TO THE TOTAL POPULATION OF THE COUNTRY

		Facebook fans	***YouTube subscribers***	***Twitter followers***	***LinkedIn followers***
1	Macedonia	2.34%	0.01%	0	0.28%
2	Serbia	9.06%	1.21%	0.4%	0.22%
3	Albania	5.50%	0.001%	0.011%	0.33%
4	Bosnia and Herzegovina	4.97%	0.004%	0.014%	0.13%
5	Montenegro	7.50%	0.007%	0.489%	0.06%
6	Kosovo	11.52%	0.019%	0.022%	0.25%

Second interesting remark can be derived by data comparison from table 1 and fig. 2, related to consistency in social media activity by country and platform. Namely, Kosovo and Serbia are the countries with the highest percentage of banks having official fan page on Facebook or official channel on YouTube. At the same time they are the two countries having largest number of Facebook fans and YouTube subscribers calculated as relative proportion to the total population of the country. On the other hand

Macedonian banking sector has the lowest number of official fun pages, measured as relative proportion to the total number of banks. Also Macedonia has the lowest average number of Facebook fans expressed relatively to the total population of the country. But this is not consistent trend for all social media platforms because data showed that although in some countries there are less official accounts on some platforms (for example YouTube, Twitter and LinkedIn) there are more subscriptions or followers. For example Montenegro banking sector has relatively lower number of Twitter accounts than Serbia but relatively more followers. Also Macedonian banking sector have lower number of YouTube official channels comparing to Albania and Bosnia and Herzegovina but relatively more subscribers. This indicates that we cannot derive straight correlation between presence on account on a particular platform and number of subscriptions for each country. The difference is a result of the various intensity with which the banks are using particular platforms in different countries. There is big disproportion in active use of social media platforms by particular banks in each country which adds value to the average scores per country.

We were also using social media visibility on the web page as metric to determine the level of platform visibility in banking sector. For this purpose we have manually inspected all bank web pages to check whether we could identify visible sign of the social platform on which the particular bank is active. The fig. 3 shows the data related to social media visibility.

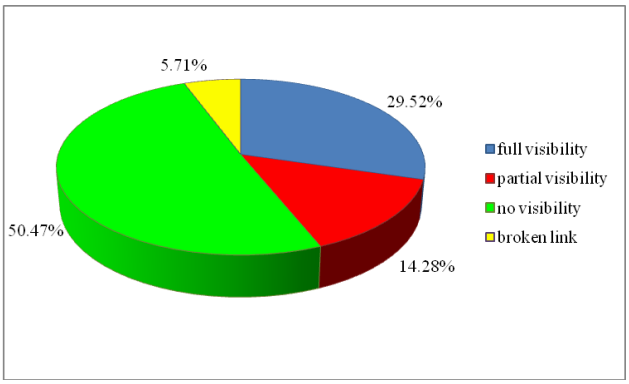


Figure 3. Social media visibility in the banking industry in the region of Western Balkan countries

To check the visibility of social media platforms on the web pages in banking industry we developed four categories: full visibility, partial visibility, no visibility and broken link. The first category full visibility means that the bank has visibly marked all social platforms on which is officially active on its own web page. The second category partial visibility means that the bank has visible links to some of the social media platforms but not to all. This means for example that the bank have official Facebook fun page but does not have visible link on its official web page which navigates the clients towards that particular platform. Third category no visibility means that the bank does not have any visible sign of any social media platform on its official web page. And the fourth category broken link represents situation when we could identify visible sign of some social media platform but the link was not navigating towards that social media platform.

If we analyze the data presented in fig.3 we could easily notice that 50.47% of the banks have no visibility at all of social media presence on its own official websites. Around 30% (or only 29.52%) of the banks have full visibility which is closely to only one third of the whole banking sector in the region of western Balkan countries. Even if we would add the numbers of partial visibility and broken link (which summed together are closely to 20%) to the number of full visibility we would come up with the number of 50% of the banks who have been involved in providing some visibility of its own social media platform on the official web page. It is also interesting to note that from 50% of the banks who did not have visibility on the web page half (or 25%) did not have presence on any social media platform. This is also important because one fourth (or 25%) of all banks from all countries from the western Balkan region did not have presence on any of the four social media platforms included in the research of this paper. If we compare the data by country we could easily notice that Kosovo and Serbia again are two countries leading in full social media visibility on the official banking web pages. Only small number, several banks from Albania and Bosnia and Herzegovina has been classified to the category broken link (18.75% or 11.55% of the total banks in the country).

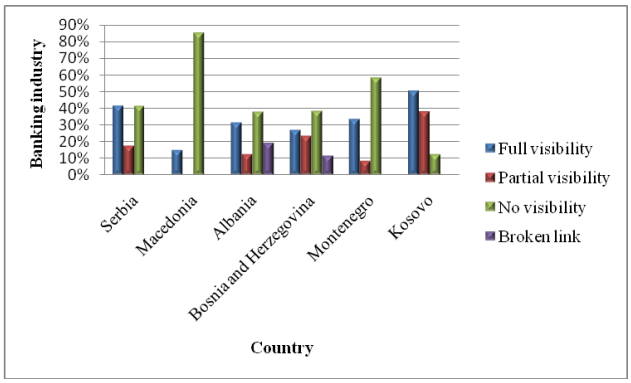


Figure 4. Social media visibility in banking industry per country

The data confirmed again that highest full visibility is present in Kosovo and Serbia, in the two countries who were scoring in the top also on the previous metrics presented in table 1 and figure 2. Macedonian banking industry was scoring very low on previous indicators of social media activity and again is at the very bottom regarding the social media visibility measurement.

In order to acquire deeper understanding of social media activity of banking industry in the region of our research interest, we also collected data for social engagement of

banks and their members on the social media platforms. For this purpose we were collecting data for:
1. Facebook – in form of Facebook posts, fans and bank comments in last 15 days;
2. YouTube – video uploads in last 3 months, total views and views of last video;
3. Twitter - number of total tweets, tweets and retweets in last 15 days;
4. LinkedIn – updates and comments in last 15 days. We have split the data in two general categories: 1) engagement of banks and 2) engagement of their social platform members. The table 2 presents the engagement of banks on four social media platforms with metrics as average number to the total number of banks in each country.

TABLE II. SOCIAL MEDIA ENGAGEMENT OF BANKS AS RELATIVE PROPORTION TO THE TOTAL NUMBER OF BANKS IN COUNTRY

		Facebook posts (last 15 days)	***YouTube video uploads (last 3 months)***	***Twitter total tweets***	***LinkedIn updates (last 15 days)***
1	Macedonia	1	0.14	0	0
2	Serbia	7.48	2.13	517	0.48
3	Albania	0.68	0.81	18.1	0.125
4	Bosnia and Herzegovina	4.8	4.03	53.62	0.14
5	Montenegro	3.3	0.5	118	0.16
6	Kosovo	5.87	3.75	64.25	0.625

Having the data for social engagement by banks we can easily notice again that two banking sectors from two countries Kosovo and Serbia are scoring again very high. This is again confirmation for the banking sector of these two countries as leading in using social media platforms for promoting own services. We can also confirm that Macedonian banking sector is again at the bottom of social media engagement comparing to banks from other countries in the region. Some metrics related to bank engagement (such as bank comments) have not been presented because they have been very low close to zero for all countries comparing to members comments.

TABLE III. SOCIAL ENGAGEMENT OF PLATFORM MEMBERS AS RELATIVE PROPORTION TO THE TOTAL NUMBER OF BANKS BY COUNTRY

		Facebook fans comments (last 15 days)	***Total YouTube video views***	***Retweets by followers***	***LinkedIn Comments (last 15 days)***
1	Macedonia	2	3898	0	0
2	Serbia	15	87933	4	0.03
3	Albania	0.625	2267	0	0
4	Bosnia and Herzegovina	63	13107	0.03	0
5	Montenegro	1.58	1992	0.58	0
6	Kosovo	33.5	29371	0	0

In table 3 data are presented for social media engagement by platform members. Interesting change for the first time appeared regarding the top two countries ranked by social media activity. Bosnia and Herzegovina appeared as most active on Facebook and YouTube regarding the platform members engagement measured by fans comments and total video views. Right behind Bosnia and Herzegovina are the two countries Kosovo and Serbia who were scoring very high in most of the metrics especially for these two platforms taking into account previously presented data.

We have also compared the data for social media activity of bank members of larger group operating under same brand in more than one country. We compared their social media metrics between banks within the group expecting to identify some common behavior regarding social media bank group policy. But we did not identify any consistency within the groups in different countries. This means that the banks which belongs to the same group and operate in different countries have presence on very different social platforms with diverse intensity. Hence they manifest different maturity in social media activity. This situation give us the knowledge that social media activity is highly decentralized between the banks within the same group and management of the group does not have social media strategy for the entire group.

In order to identify trends between certain social media metrics used in this paper we performed several bivariate Pearson correlations between social media metrics. The most significant correlation coefficients are presented bellow in the table 4.

TABLE IV. RESULTS FROM BIVARIATE CORRELATIONS

		Facebook fans	***Twitter followers***	***LinkedIn followers***	***YouTube subscriptions***
1	Total Assets	0.636**	0.415**	0.505**	0.269**
2	Time being on Facebook	0.365**	/	/	/
3	Time being on Twitter	/	0.548**	/	/
4	Time being on Youtube	/	/	0.056	/

** Correlation is significant at the level 0.01 level (2 - tailed)

* The data for total assets of banks are for year 2013 collected by publicly available financial reports from official bank web pages.

The results in table 4 shows that we have identified positive correlation between the sizes of banks measured as total assets and number of members of different social media platforms. This means that larger banks have more Facebook fans, Twitter followers, LinkedIn followers and YouTube subscriptions comparing to smaller banks. It is interesting to notice that smaller banks are not using more intensively the opportunity of social media in their competitive battle with larger banks taking into account

their smaller marketing budgets for promotion on traditional marketing media. When we analyze the data related to time having an official account on some social media platform and number of members in form of fans or followers we identified significant positive relationship for Facebook and Twitter but not for YouTube. This means that the banks that have been longer period present in the past on Facebook and Twitter have more fans and followers. This proposition was not confirmed for YouTube.

IV. RESEARCH LIMITATIONS

The research endeavor of the authors of this article was to depict the situation of social media use by banks in the region of Western Balkan Countries. Although this study has contributions, it has inherent limitations that warrant caution in interpretation of the results. First, the whole study was based on a data collected by desk research which means that data from primary sources such as surveys or interviews with banks officials were not included in the analysis. It would be interesting in future research to compare some field data with the results presented in this study. Second, the data have been collected in a single point in time which means that longitudinal trends in social media use cannot be derived from the study results. Third, subjective interpretation of the study results from the authors can be also strong limitation of the study.

V. CONCLUSION

The purpose of this research was to analyze and explore how banks from Western Balkan Countries are currently using social media for their marketing purposes, and to compare the results between the countries in order to derive some general conclusion for the whole banking sector in the region. Through the results that we have obtained from data manually collected by visual inspection of official websites and social media platforms social media marketing in banking sector in this region is still in emerging phase. The results confirmed that banks are not exploiting fully the potential of social media. The results showed that in countries with the highest participation in terms of presence on social media platforms, one-third of all banks did not have at all presence on any social medium. We also found that size of the market and the number of banks that constitute the banking sector does not play a role in boosting the social media activity of the banks. Maybe some other market characteristics play significant role in social media activity in banking industry which further needs to be investigated. In most of the rankings Serbia and Kosovo were scoring very high on social media activities and at the same time these two countries lay on the total opposite side on a continuum regarding market size. Another evidence of immature use of social media by banking industry can be found in the data showing social media visibility. More than 50% of the banks do not have any visible link for navigation towards the social media platforms, in around 5% of the cases the links were broken and in 15% of the cases not all media platforms on which the bank was active were included on the official webpage. The results were even worse when we inspected social engagement of banks and their social media members. We concluded that most of the banks are not continuously involved in communication with their clients with exception of only few of them. The results regarding the most often used platforms converged with other studies for other regions that most often used platforms by banks have been Facebook, YouTube, Twitter and LinkedIn. When we were inspecting social media visibility we could very rarely notice some other social media platform such as Google + or Instagram. We identify significant positive relationship between size of the bank and number of subscriptions to certain social media platform. This means that smaller banks do not outperform larger banks by increasing its' own population on the social media platforms. Although this is expected to a certain extent because larger banks normally would have larger number of clients, having more strong social media marketing strategy by smaller banks can help them to have influence on wider market space. Also the data showed positive relationship between time being present on a certain social media platform and number of subscriptions. This means that the banks that started earlier to be present on a particular social media platform have larger audience to influence. Taking into account the research limitations mentioned above we can conclude that maturity of social media marketing of banking sector from Western Balkan Countries is still on modest level.

References

[1] A.M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media" Business Horizons, vol. 53 (1) pp. 59–68, January 2010.

[2] A.J. Kim and E. Ko, "Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand", Journal of Business Research, Series Fashion Marketing and Consumption of Luxury Brands, vol. 65 (10) pp. 1480–1486, October 2012.

[3] A.K. Kirtiş and F. Karahan, "To be or not to be in social media arena as the most cost-efficient marketing strategy after the global recession", Procedia-Social and Behavioral Sciences vol. 24, pp. 260–268, September 2011..

[4] M.E. Porter, "Strategy and the Internet", Harvard Business Review vol. 79 (3), pp. 62–78 March 2001.

[5] T.L. Tuten, "Advertising 2.0: Social media marketing in a web 2.0 world",Greenwood Publishing Group 2008.

[6] G. Urban, "Don't Just Relate-Advocate", Retail Technology Quarterly, 30A, January 2006.

[7] A.Vemuri, Getting Social: "Bridging the gap between Banking and Social Media", Global Finance, vol. 4 (5), May 2010.

[8] "Social Media Brings Benefits, Pitfalls to Retail Banks", Teller Vision, ISSN: 08951039 issue 1430 pp.1-2, 2013.

[9] M.A. Stelzner, "Social media marketing indsutry report", 2014. Downloaded from: http://www.socialmediaexaminer.com/SocialMediaMarketingIndustryReport2014.pdf

[10] "The Global Retail Banking Digital Marketing Report", WIPRO, May 2013. Downloaded from:https://www.wipro.com/documents/the-global-retail-banking-digital-marketing-report-2013.pdf.

[10] "Retail banks warned of social media marketing pitfalls", January 2013. Downloaded from:http://www.finextra.com/News/FullStory.aspx?NewsitemID=24459.

S P

International Conference on

MIPRO Junior - STUDENT PAPERS

Steering Committee

Chairs: Andrej Sokolić, University of Zagreb, Croatia
Dino Lukman, University of Zagreb, Croatia

PAPERS

Optical Add/Drop Multiplexers for All-optical Fiber Communication

J. Ruzbarsky*, J. Turan* and L. Ovsenik*
* Technical University of Kosice, Dept. of Electronics and Multimedia Communications, Slovak Republic
jan.ruzbarsky@tuke.sk, jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstract - Demands on transmission systems increased due to the growing demand for services supported by the Internet. It has consequences in the continued development network in communication systems. This work is focused on the current state of Wavelength Division Multiplex (WDM) systems, on the recommendation of ITU for Coarse WDM (CWDM) and Dense WDM (DWDM) and on the possibilities of using different types of add/drop multiplexers (OADM) in optical transmission paths. In the present there exist many solutions of OADM and some of them are shown in this article. In the environment of optical systems with increasing amount of transmission information and power undesirable effects occur, such as nonlinear effects, but this subject is out of scope of this article.

I. INTRODUCTION

With the coming of optical fibers there were also new opportunities within transmission systems for short as well as long distances. In addition to the many benefits of new technology as for example increased transmission capacity, higher transmission speed, greater security etc. also new challenges that need to be solved come up. The main challenges include linear and nonlinear effects. Restrictions occasioned by linear effects are not examined in this work. Attention is primarily devoted to nonlinear phenomena and their effects in an optical fiber.

The work is focused also on introducing of current solutions of transmission systems which include dense wavelength division multiplexing (DWDM). For each approach standard that include some recommends is used. The same applies for DWDM, its way of working is formed by the International Telecommunication Unit (ITU). Specifically, there is the recommendation identified as - Recommendation ITU-T G.694.1. With the increasing amount of information there is an ambition to use a larger number of transmission channels, and this happens even when using DWDM technology in optical communication networks. In other words the traffic channels are narrower and they are proposed in greater numbers on the same size of bandwidth.

On the transmission path there is a large number of devices which are designed to receive, modify, amplify signals, etc. Devices receiving and working with optical signal also include optical add/drop multiplexers (OADM). This device has many similarities with optical cross connectors (OXC). Currently types of reconfigurable OADM (ROADM) are used but in early development fixed OADM (FOADM) was preferred.

II. ALL OPTICAL DWDM SYSTEMS

Wavelength division multiplexing (WDM) is presently the most innovated technology in optical fiber communications. In the past two decades, optical communications has totally changed the way we communicate. It is a revolution that has fundamentally transformed the core of telecommunications, its basic science, its enabling technology, and its industry. By this method, capacity can be increased by using more than one optical carrier (wavelength) on a single fiber. Therefore, adding a second transmitter and receiver to an optical fiber can double the bandwidth of that communications system. This method of increasing the capacity of an optical system has appeal for a variety of reasons. For transmission systems using a 1310 nm laser, a second laser at 1550 nm is usually added. The reason for choosing these wavelengths is that they lie in the "windows" or ranges of least attenuation. This allows the signal to travel a longer distance. Fig. 1 shows specification of wavelengths.

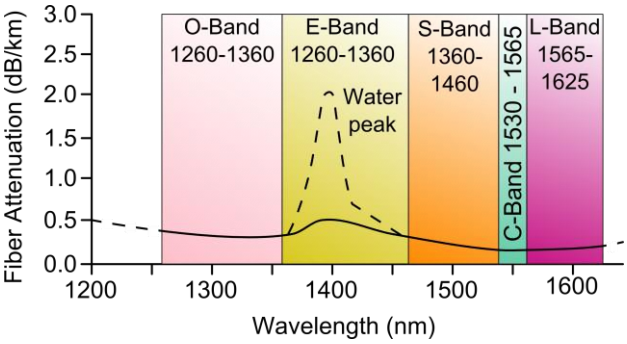


Figure 1. Graph of dependencies between fiber attenuation and wavelength for individuals bands.

To meet the rapid increase in the demand for multiterabit/s transmission capacities, 40 Gbit/s based dense wavelength division multiplexing (DWDM) is becoming the next generation in large-capacity systems. Expansion of WDM channels results in an increase in the complexity of optical network node [1].

The cost-difference between CWDM and DWDM is known to be large, favoring CWDM as the low-cost alternative. For channel counts beyond eight channels, DWDM may be a preferred technology because:

- Flexible upgrade: DWDM are more flexible and robust with respect to fiber types. DWDM

upgrade to 16 channels is viable on both G.652 and G.652.C fibers. This stems from the fact that DWDM always employs the low-loss region of the fiber, while 16 channel CWDM systems involves transmission in the 1300-1400 nm region, where attenuation is remarkable higher.

- Scalability: DWDM solutions allow upgrade number of channels, allowing a much higher total capacity on the fiber than a CWDM solution.
- Cost: CWDM still shows to be cost-efficient for 8 channel systems. However, for higher channel counts than 8, the higher attenuation in the fiber may prove DWDM to be a better alternative than CWDM.
- Long spans: DWDM employs the 1550 nm wavelength band which can be amplified using conventional optical amplifiers, Erbium Doped Fiber Amplifiers (EDFA's), enhancing transmission distance to hundreds of kilometers [2].

Dense wavelength division multiplexing (DWDM), a wavelength division multiplexing (WDM) technology, is characterized by narrower channel spacing than coarse WDM (CWDM), as defined in [ITU-T G.671]. In general, the transmitters employed in DWDM applications require a control mechanism to enable them to meet the application's frequency stability requirements, in contrast to CWDM transmitters, which are generally uncontrolled in this respect. DWDM uses frequency grid. It is a reference set of frequencies used to denote allowed nominal central frequencies that may be used for defining applications [3].

The frequency grid defined by ITU G.694.1 supports a variety of fixed channel spacing ranging from 12.5 GHz to 100 GHz and wider (integer multiples of 100 GHz) as well as a flexible grid. Uneven channel spacing using the fixed grids are also allowed. The current steps in channel spacing for the fixed grids have historically evolved by sub-dividing the initial 100 GHz grid [4, 5, 6].

Today, wavelength transport uses a 100 Gbit/s infrastructure, but with the introduction of super-channels and higher order modulation rates, 400 Gbit/s or even 1 Tbit/s transport will be realized. An IP or MPLS core network may require 100 Gbit/s links between locations, but for many applications finer bandwidth granularity can improve network efficiency. To that end, integrating OTN or multi-protocol label switching (MPLS) switching with the transport network to support sub-wavelength services is important.

Table I illustrates some nominal central frequencies within the C and L bands based on the 12.5 GHz minimum channel spacing anchored to the 193.1 THz reference. Table I also illustrates the 25, 50 and 100 GHz grid frequencies within the same region. The endpoints shown are illustrative, not normative. Fig. 2 shows basic scheme of DWDM transmission system [4, 7].

TABLE I. EXAMPLE OF NOMINAL CENTRAL FREQUENCIES OF THE DWDM GRID

Approximate nominal central wavelength (nm)	**Nominal central frequencies (THz) for spacing of:**			
	12.5 GHz	***25 GHz***	***50 GHz***	***100 GHz***
1530.0413	195.9375	—	—	—
⋮	⋮	⋮	⋮	⋮
1530.3341	195.9000	195.9000	195.9000	195.9000
1530.4318	195.8875	—	—	—
1530.5295	195.8750	195.8750	—	—
1530.6271	195.8625	—	—	—
1530.7248	195.8500	195.8500	195.8500	—
⋮	⋮	⋮	⋮	⋮
1624.8914	184.5000	184.5000	184.5000	184.5000

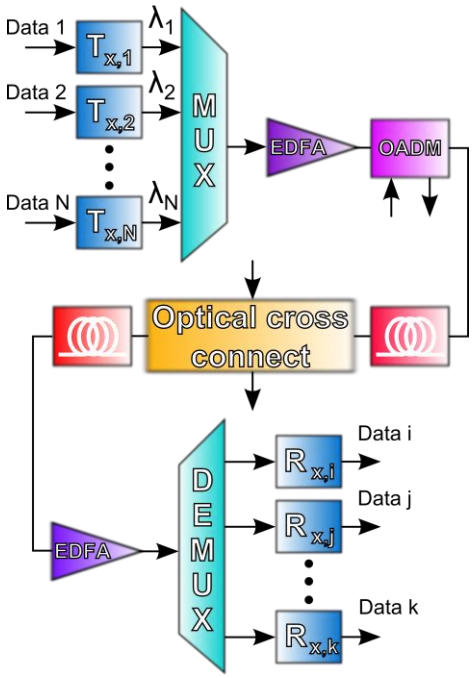


Figure 2. Model of optical transmission DWDM system using Optical Add/Drop Multiplexers.

All efforts are targeted at more efficient transmission. Future of optical transmissions is joined with reduction crosstalk among channels traffic, superior source of light, higher spectral efficiency, more efficient use of fiber spectrum, and corresponding scaling in WDM line-card capacity of WDM systems. Smarter ways to utilize optical capacity are critical since transmission costs have been decreasing slower than grooming and switching costs. Integrated transport and switching can improve the efficiency of the client network using techniques such as port virtualization and transit traffic reduction. Trends in transmission systems are showed in the Fig. 3.

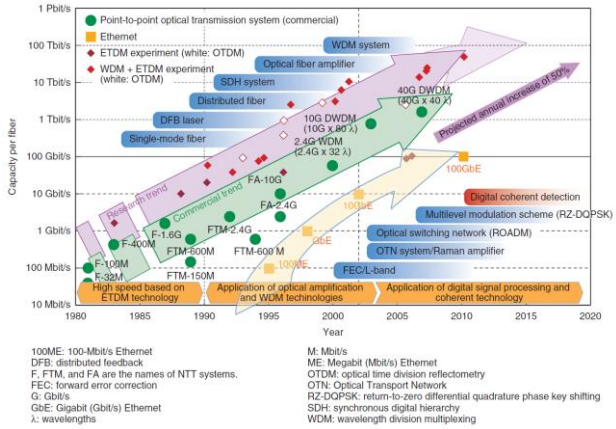


Figure 3. Trends in optical transport technologies [15].

Optical transport network capacity requirements continue to grow in order to meet worldwide IP traffic demands. Today's networks can provide 10 Tbit/s per link, and doubling or even tripling that capacity will be possible in the next several years. Technologies that include higher order modulation (e.g. 16-QAM) will increase spectral efficiency. They can be combined with Nyquist or OFDM super-channels to provide the necessary capacity. Various optical technologies, such as higher order modulation, coherent detection, soft-decision FEC, and super-channels will help transport networks provide the needed capacity for the near future. Long term, increasing a link's capacity can be supported by using wider bandwidth optical amplifiers or utilizing fibers with multiple cores and multiple propagation modes. These innovations are not yet mature, and significant development is still required. Maximizing network efficiency through smart bandwidth management can reduce required transmission capacity, resulting in lower cost and power consumption. Grooming can take the form of either OTN or MPLS switching with OTN preferred for private line applications and MPLS optimized for data applications. It is more efficient to integrate the transport and switching functions into a single platform, but low-cost optical interconnect will support separate transport and switching implementations. Joint optimization of the transport and client networks is important to maximize efficiency, and techniques such as port virtualization and transit traffic reduction can significantly reduce client network bandwidth requirements [8, 9].

III. Optical Add/Drop Multiplexer

Add/drop multiplexers are needed for wide-area and metro-area networks (similar to SONET ADMs) in which one or more channels need to be dropped or added, while preserving the integrity of other channels. The demultiplexer separates all channels, optical switches drop, add, or pass individual channels, and the multiplexer combines the entire signal back again [10].

Most multiwavelength-ring networks being deployed today employ fixed (or static) OADMs (FOADMs) because FOADM technology is quite mature today. However, reconfigurable OADMs (ROADMs) are more powerful because a network based on tunable ROADMs can easily adapt to fluctuating traffic demands.

A. Fixed Optical Add/Drop Multiplexers

Fixed OADMs are very often called also Configurable AODM (COADM). We can tell that their basic role is dropping and adding data signals on dedicated channels. FOADMs have allocated channels at non-dynamic nodes and enabled dropping and adding of single or multiple wavelength channels from DWDM. Configurable OADM allow the producer to customize the number of channels, specifications of filters, and power equalization. For dense channel plans with 50 GHz channel spacing, the FOADM typically covers an optical frequency interleaver. Most configurations have low loss and are temperature-insensitive. Fixed OADMs are less expensive than ROADMs, the efficiency of this architecture is dependent upon having a network with predictable and very stable traffic demand.

B. Reconfigurable Optical Add Drop Multiplexers

With approach a lot of video applications on the Internet is growing demands to the larger bandwidth. It is possible the link capacity increase by migration from lower speeds to 40 Gbit/s or up to 100 Gbit/s and using WDM technology. This is good result only for point-to-point communication but we need to change ever requirements for the bandwidth demand. The optical networks have to be able dynamically add, drop, and route channels at individual nodes. This type of management in optical networks is supported by ROADMs [11].

ROADM provides adaptivity in re-routing optical stream, avoiding defective connections, looks for the stability of the services, it can be configured remotely. They enable new wavelength channels to be installed by simply placing transponders at the end point. The rapid network reconfiguration provided by ROADMs could also become an enabler of dynamic network services, such as switched video for IPTV. For all of these reasons, ROADMs will continue to have a significant effect on the design of optical networks [12, 13].

Reconfigurable OADMs include three different types:

- Wavelength Blocking (WB).
- Planar Lightwave Circuit (PLC).
- Wavelength Selective Switch (WSS).

ROADMs are capable of enabling an advanced dynamic power/gain control for each channel and enable the remote management functionalities such as remote monitoring and remote configuration management. ROADMs not only can effectively increase the wavelength efficiency and network capacity of to day's optical network, but also are a promising technology for future high-bit-rate optical networks such as grid-less optical networks and all optical networks.

The most used type of ROADM is wavelength selective switch. This technology have become the established heart of the modern DWDM reconfigurable agile optical network. It serves as the optical component which is responsible for dynamically routing, blocking and attenuating all DWDM wavelengths within a network node. Therefore, the performance of this device is

generally the one of the dominant attributes which define the performance of the overall network system.

WSSs are more versatile, smaller, consume less power, and are less expensive than 2nd generation ROADMs. WSSs are typically designed to handle 40 incoming wavelengths on a fiber, and drop any combination of wavelengths out any of 8 ports (some systems can drop more). The WSS is an N-degree ROADM, which allows reconfiguration beyond typical 2-degree (usually east-west) configurations (WSS system can be used to interconnect rings or mesh points). This means that wavelengths can be routed from any node to any other node across multiple rings and networks [14].

IV. EXPERIMENTS

Experiments with OADM show basic properties of optical add/drop multiplexer in DWDM system. Fig. 4 shows block scheme of used system. Simple WSS-based ROADM architecture was used in this scheme.

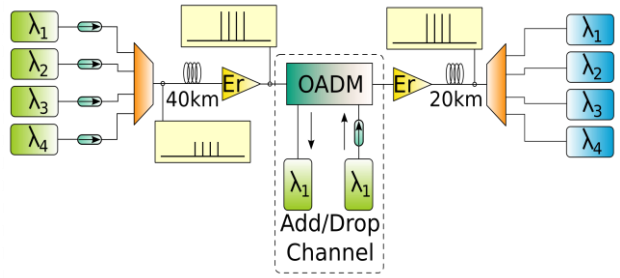


Figure 4. Block diagram of experimental system.

In this model was used four wavelengths (tx_1 = 193.35 THz, tx_2 = 193.45 THz, tx_3 = 193.55 THz, tx_4 = 193.65 THz). Fig. 4 shows spectrum transmitted signals. Wavelengths in the figure are before amplification.

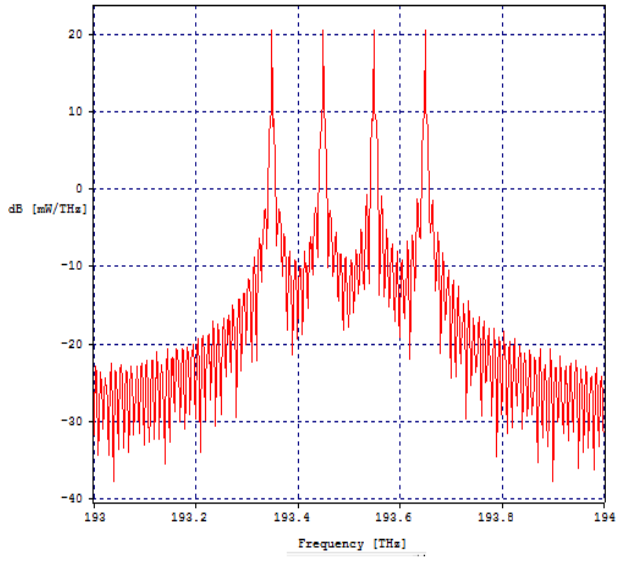


Figure 5. Spectrum of transmitted signals.

Fig. 5 shows signal after amplification. It is possible to see that levels of signals are more intensive.

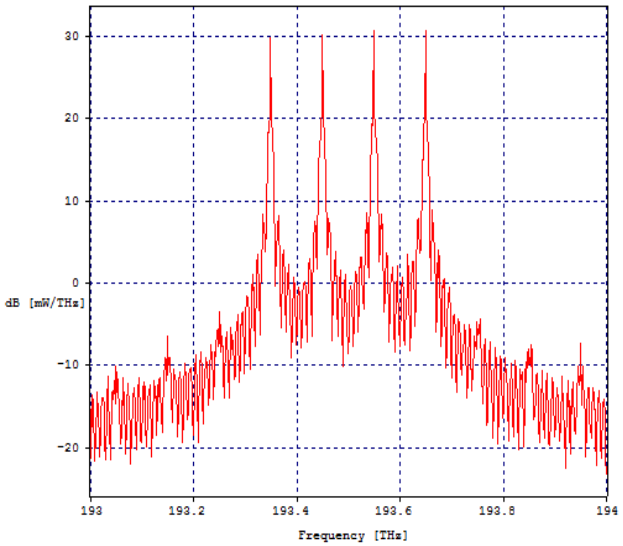


Figure 6. Spectrum of signals after amplification.

Bit Error Rate (BER) of signal with frequency 193.35 THz (dropped signal) is 7.67299×10^{-20}. This value is still on the value which can consider as sufficient for well recognized at the receiving end. Fig. 6. shows Eye diagram of dropped signal (193.35 THz). The signal was influenced by different attenuation effects on the 40 km transmission path.

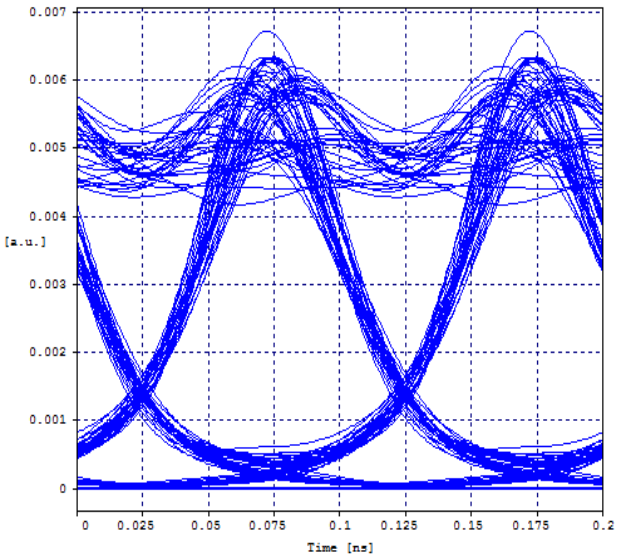


Figure 7. Eye diagram of dropped signal.

In the next step is frequency which had dropped, it is again used for transmitted another information. Eye diagram is almost ideal, after passing the signal through the OADM (Fig. 8.).

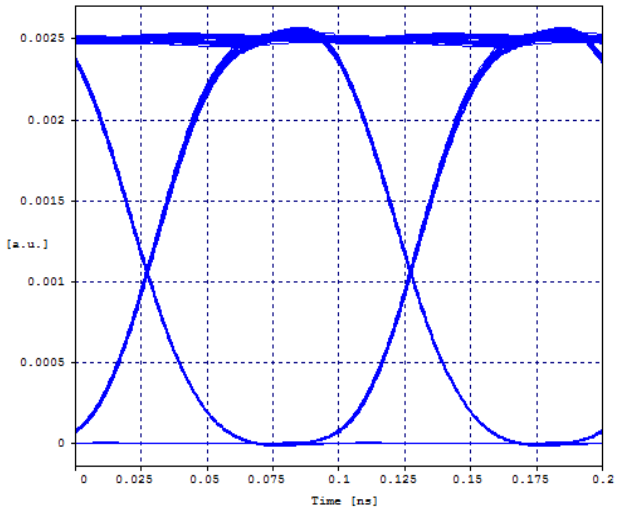


Figure 8. Eye diagram of inserted signal.

Signals on the received side are in the good condition and their BERs are at the border $1.10163x10^{-16}$(Fig. 9).

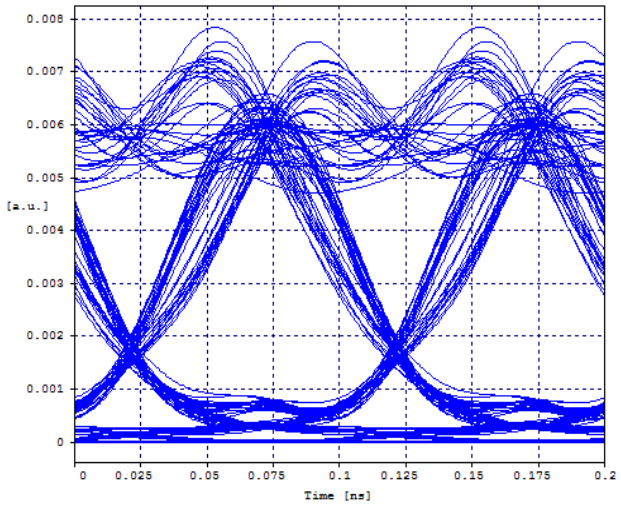


Figure 9. Eye diagram of received signal at the end transmittion path.

V. CONLUSION

Workplace Research of optoelectronic systems on Department of Electronics and Multimedia Communications at the Technical University in Kosice within the specialization is focused on the investigation of optical multiplexers and demultiplexers. This paper constitutes basic information about Optical Add/Drop Multiplexers. Initial experiments are designed to present behavior of transmission systems operating in accordance with the recommendations of ITU standards and manufacturers of optical devices. The measurements showed what effect does the distance on the transmitted signal and the fundamental property OADM, ergo withdrawal (attenuation) certain wavelength and wavelength addition to the transmission path. The experiment uses 100 GHz frequency grid spacing for different transmission channels in DWDM system. The quality of the transmitted signal is showed in the form of BER and eye diagram.

ACKNOWLEDGMENT

This work was supported by Cultural and Educational Grant Agency (KEGA) of the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project no. „006TUKE-4/2014 - The Use of TUKE PON Experimental Model in Teaching. This work is also the result of the contract no. “APVV-0025-12 - Mitigation of Stochastic Effects in High-Bitrate All Optical Networks” supported by the Slovak Research and Development Agency.

REFERENCES

[1] J. N. DOWNING, "Fiber Optic Communications", Clifton Park, NY,USA: Thomson Delmar Learning, 2005. ISBN: 1-4018-6635-2.

[2] TRANSPACKET "CWDM and DWDM networking", TransPacket whitepaper, 2011.

[3] ITU-T "Transmission characteristics of optical components and subsystems", Recommendation ITU-T G.671., International Telecommunication Union, Series G: Transmission Systems and Media, Digital Systems and Networks, 6th ed., February 2012.

[4] ITU-T "Spectral grids for WDM applications: DWDM frequency grid", Recommendation ITU-T G.694.1., International Telecommunication Union, Series G: Transmission Systems and Media, Digital Systems and Networks, 2nd ed., February 2012

[5] E. IANNONE, "Telecommunication Networks", USA: CRC Press Taylor & Francis Group, 2012. ISBN: 978-1-4398-4636-0.

[6] R. PASCHOTA, "Encyclopedia of Laser Physics and Technology", 1st.ed. Vol. 2, Wiley-VCH, Berlin, 2008. ISBN: 978-3-527-40828-3.

[7] A. GUMASTER, T. ANTONY, "DWDM Network Designs and Engineering Solutions", Indianapolis, IN USA: Cisco Press, 2003. ISBN: 1-58705-074-9.

[8] T. J. XIA, S. GRINGERI, "High-Capacity Optical Transport Networks", IEEE Communications Magazine, November 2012.

[9] J. RAHN, S. KUMAR, M. MITCHELL, H. SUN, K. T. WU, G. GOLDFARB, "Super-Channels: DWDM Transmission Beyond 100 Gb/s", Photonics Conference (IPC), IEEE, September 2012 ISBN: 978-1-4577-0731-5.

[10] G. P. AGRAWAL, "Fiber-Optic Communication Systems. 4th. ed", Rochester, NY : Academic Press, 2007. ISBN: 978-0-12-369516-1.

[11] P. WHITE, "ROADM Evolves: Should You Be Paying Attention?", Infonetics Research, Inc., San Jose California , February 2006.

[12] V. ABEDIFAR, S. A. SHAHKOOH, A. EMAMI, A. POURESLAMI S. A. AYOUGHI, "Design and Simulation of a ROADM-based DWDM Network", IEEE, ISBN: 978-1-4673-5634-3/13.

[13] W. I. WAY,"Optimum Architecture for MxN Multicast Switch-Based Colorless, Directionless, Contentionless, and Flexible-Grid ROADM", Optical Fib er Communication Conference and Exp osition (OFC/NFOEC), 2012 and the National Fib er Optic Engineers Conference, ISBN: 978-1-4673-0262-3.

[14] S. LI, Z. WAN, J. XU, S. ZHONG, Y. WU, "Wavelength-Selective Switch Based on a Polarization-Independent Transmission Grating and a High Fill-Factor Micromirror Array", IEEE Photonics Technology Letters, Vol. 23, No. 17, September 2011.

[15] S. MATSUOKA, "Ultrahigh-speed Ultrahigh-capacity Transport Network Technology for Cost-effctive Core and Metro Networks". NTT Technical Review, Japan, Vol. 9 No. 8 August 2011

Korelacijska analiza dotoka u akumulacijske bazene

E. Sutlović, K. Zoričić, T. Modrić
Sveučilište u Splitu, Fakultet elektrotehnike, strojarstva i brodogradnje, Split, R. Hrvatska
Elis.Sutlovic@fesb.hr, Kristian.Zoricic@fesb.hr, Tonci.Modric@fesb.hr

Sažetak - Problem optimalnog korištenja akumulacijskih bazena je posebno složen, jer proizvodnja električne energije na jednom slivu posredno utječe na potrebnu proizvodnju iz ostalih elektrana u sustavu, na troškove pogona cjelokupnog elektroenergetskog sustava, na profit itd. Treba istovremeno promatrati više akumulacijskih bazena i niz korisnika. Osnovni ulazni podaci u svim simulacijskim i optimizacijskim proračunima, su prirodni dotoci u akumulacijske bazene. Međusobne veze u pojavi određenih dotoka značajno utječu na modeliranje dotoka u proračunima. U okviru korelacijske analize ustanovljava se postojanje veze između dva niza podataka, njen oblik, jačina i smjer. Na osnovu relevantnih hidroloških podataka mjerenih u nizu godina, napravljena je korelacijska analiza prirodnih dotoka u akumulacijske bazene na različitim riječnim slivovima. Korelacijska analiza dotoka zemljopisno bliskih riječnih slivova obuhvatila je izračun Pearsonovih koeficijenata korelacije i dijagrame rasprostiranja između dotoka u akumulacijske bazene na rijeci Cetini, Krki i Zrmanji. Korelacijska analiza zemljopisno udaljenih riječnih slivova obuhvatila je izračun Pearsonovih koeficijenata korelacije i dijagrama rasprostiranja između dotoka u akumulacijske bazene na rijeci Cetini i Kupi te Cetini i Dravi. Uočene su značajne razlike u prosječnoj jakosti veza u raznim primjerima.

I. UVOD

Optimiranje rada akumulacijskih bazena je izuzetno složen i važan zadatak za niz korisnika vode te za društvo u cjelini. U elektroenergetskim sustavima sa znatnim udjelom hidroenergije planiranje rada akumulacijskih bazena utječe na moguću proizvodnju svih nizvodnih hidroelektrana na slivu. Međutim, osim elektroprivrede, postoji i niz drugih korisnika akumulacijskih bazena i vodnih tokova. Cilj planiranja rada akumulacijskih bazena je naći „optimalan“ način korištenja raspoložive količine vode nastojeći pri tome zadovoljiti više korisnika i brojna ograničenja [1], [2].

Problem optimalnog korištenja akumulacijskih bazena za potrebe elektroprivrede je posebno složen, jer proizvodnja električne energije na jednom slivu posredno utječe na potrebnu proizvodnju iz ostalih elektrana u sustavu (i hidroelektrana i termoelektrana), na troškove pogona cjelokupnog elektroenergetskog sustava, na profit itd. Dakle, treba promatrati više akumulacija te uzeti u obzir sve nizvodne hidroelektrane (HE) koje se napajaju vodom iz uzvodnih akumulacijskih bazena.

Proces planiranja obično se dijeli na kratkoročno (do sedam dana), srednjoročno (nekoliko mjeseci do godinu dana) i dugoročno (godina dana i duže). Za rješavanje problema planiranja rada akumulacijskih bazena koriste se bilo različite simulacijske metode ili pak matematički optimizacijski proračuni koji definiraju funkciju cilja i pripadna ograničenja na različite načine. U svim tim analizama cjelokupni period promatranja se dijeli u manje, jedinične vremenske periode odnosno diskretne vremenske korake unutar kojih se različite varijable i ulazni parametri ne mijenjaju. Obično se za osnovni vremenski period kod srednjoročnog i dugoročnog planiranja rada akumulacijskih bazena uzima jedan tjedan ili jedan mjesec [1] - [5].

Osnovni ulazni podaci u svim simulacijskim i optimizacijskim proračunima su prirodni dotoci u akumulacijske bazene te nizvodni pritoci na određenim mjestima. Prirodni dotoci i pritoci na određenoj lokaciji mogu se predstaviti na više načina. Najjednostavniji način je koristeći zabilježeni povijesni niz mjeren u relevantnom broju godina, a složeniji i općenitiji je korištenjem različitih prilagođenih modela ulaznih podataka. Jednostavniji je deterministički model dotoka u kojem se svi dotoci (na svim lokacijama i u svim vremenskim periodima) tretiraju kao poznate veličine, ali daljnji optimizacijski postupak zahtijeva čitav niz proračuna sa različitim mogućim scenarijima. Stohastički pristup problemu optimiranje rada akumulacijskih bazena već kroz stohastički model dotoka nastoji ugraditi neizvjesnost budućih dotoka i stohastička obilježja problema [3] - [8].

Rezultat proračuna, bez obzira na korišteni model dotoka, daje režim ispuštanja voda u svakom jediničnom vremenskom periodu na ukupno „optimalan“ način. U pravilu simulacijski i optimizacijski proračuni pretpostavljaju da u cjelokupnom periodu promatranja, npr. godini dana, suma svih ispusta iz bazena odgovara sumi dotoka uz određene gubitke zbog isparavanja, procjeđivanja itd.

O kvaliteti ulaznih podataka za simulacijski ili optimizacijski proračun ovisi i kvaliteta cjelokupnog proračuna optimalnog planiranja. Dakle, optimalno planiranje i korištenje voda na nekom slivu temelji se na sustavnom višegodišnjem prikupljanju hidroloških podataka te njihovoj analizi i obradi [9]. Međutim, riječni slivovi, iako zahvaćaju vodu iz zasebnih područja, imaju nekad slične karakteristike o kojima treba voditi računa prilikom modeliranja za potrebe proizvodnje električne energije. Međusobne veze u pojavi određenih dotoka značajno utječu na modeliranje dotoka u proračunima i složenost proračuna. U okviru korelacijske analize

ustanovljava se postojanje veze između dva niza podataka, njen oblik, jačina i smjer ne ulazeći pri tome što je uzrok, a što posljedica.

Cilj rada je, na osnovu relevantnih hidroloških podataka mjerenih u nizu godina, napraviti korelacijsku analizu između prirodnih dotoka u akumulacijske bazene na različitim riječnim slivovima. Posebno će se tretirati zemljopisno bliski riječni slivovi od onih udaljenih. Istražit će se i statistički definirati korelacijske veze među dotocima u akumulacijske bazene grupirane u manje vremenske periode. Prema jakosti veze ćemo ocijeniti da li je opravdano zanemariti neke korelacijske veze u određenom tipu proračuna. Drugim riječima, kada tražimo optimalni režim korištenje akumulacijskih bazena za rad hidroelektrana i simuliramo npr. dotoke na rijeci Zrmanji, trebamo li uzeti u obzir simulirani dotok npr. na rijeci Cetini ili pak na rijeci Dravi? Da li je realno da dotok na Zrmanji bude blizu maksimalne vrijednosti, a istovremeno na Cetini ili Dravi blizu minimalnih vrijednosti i obrnuto? Koliko su jake te veze, i koliku grešku radimo, ako ih zanemarimo prilikom simulacije?

Korelacijska analiza zemljopisno bliskih riječnih slivova obuhvatila je izračun Pearsonovih koeficijenata korelacije i oblikovanje dijagrama rasprostiranja srednjih mjesečnih dotoka za dotoke u akumulacijske bazene na rijeci Cetini, Krki i Zrmanji. Također su računati Pearsonovi koeficijenti korelacije i dijagrami rasprostiranja za srednje tjedne dotoke na rijeci Cetini i Krki. Korelacijska analiza zemljopisno udaljenih riječnih slivova obuhvatila je izračun Pearsonovih koeficijenata korelacije i oblikovanje dijagrama rasprostiranja srednjih mjesečnih dotoka za dotoke u akumulacijske bazene na rijeci Cetini i Kupi te Cetini i Dravi. Uočene su značajne razlike u prosječnoj jakosti veza u raznim primjerima.

II. KORELACIJSKA ANALIZA

Istraživanjem i kvantificiranjem povezanosti među promatranim pojavama, odnosno varijablama, bavi se korelacijska analiza [10], [11]. Neka neki događaj ima dva obilježja x i y, gdje je x mjerena vrijednost jednog, a y mjerena vrijednost drugog obilježja. Nakon n mjerenja dobit ćemo skup od n parova (x_i,y_i), i=1,2, ..., n. Budući da je riječ o uređenim parovima, nakon n mjerenja rezultat ćemo moći predstaviti sa n točaka u koordinatnom sustavu. Skup tih točaka se zove dijagram rasprostiranja. Iz rasporeda točaka otkrivamo oblik, smjer i jakost veze između obilježja x i y:

- kod funkcionalne veze točke (x_i,y_i) leže na jednoj krivulji, a ako je ta krivulja pravac govorimo o linearnoj funkcionalnoj vezi;
- kad obilježja x i y nisu vezana funkcionalnom vezom, ali dijagram ipak pokazuje tendenciju izduživanja u jednom smjeru govorimo o stohastičkoj vezi među slučajnim varijablama x i y;
- ako točke (x_i,y_i) ne pokazuju tendenciju gomilanja u nekom smjeru, obilježja x i y su nepovezana.

Kod stroge funkcionalne veze (ili deterministčkog odnosa) za svaku vrijednost x se točno zna vrijednost varijable y. Kada su slučajne varijable x i y stohastički povezane, za jednu vrijednost od x dolazi više vrijednosti od y koje su raspoređene po nekom zakonu vjerojatnosti. Stoga se pri stohastičkoj vezi za određeni x može samo odrediti interval kojem se ta vrijednost nalazi i to uz izvjesnu vjerojatnost.

Jakost stohastičke veze mjeri se koeficijentom korelacije ρ. Koeficijent linearne korelacije osnovnog skupa (populacije), ili Pearsonov koeficijent definiramo kao

$$\rho = \frac{ko\,\mathrm{var}\,ijanca(x,y)}{\sqrt{\mathrm{var}\,ijanca(x)\cdot \mathrm{var}\,ijanca(y)}} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}}. \qquad (1)$$

Vrijednost koeficijenta korelacije je $-1 \le \rho \le +1$. Za $\rho=\pm1$ govorimo o funkcionalnoj vez dviju varijablii, a za $\rho=0$ varijable su nezavisne. U ostalim slučajevima imamo stohastičku vezu među varijablama. Općenito labavijoj vezi odgovara manji $|\rho|$, a čvršćoj veći $|\rho|$.

Nepristranu procjenu koeficijenta korelacije ρ, označimo sa r. Tada je koeficijent linearne korelacije uzorka [11]

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}. \qquad (2)$$

Nepristrana procjena kovarijanca varijabli x i y, s_{xy} je

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}). \qquad (3)$$

Nepristrana procjena varijance varijable x, s_x^2 i nepristrana procjena varijance varijable y, s_y^2 su:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \qquad (4)$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2. \qquad (5)$$

U primjeni se stupanj jakosti korelacije r obično definira na slijedeći način

$\lvert r \rvert$	**Jakost korelacije**
0	nema korelacije
0 - 0,5	slaba korelacija
0,5 - 0,8	srednje jaka korelacija
0,8 - 1	jaka korelacija
1	potpuna korelacija

Ponekad se umjesto koeficijenta korelacije r koristi koeficijent determinacije r^2 $(0 \le r^2 \le 1)$.

III. KORELACIJSKA ANALIZA DOTOKA U ZEMLJOPISNO BLISKE AKUMULACIJSKE BAZENE

A. Dostupni podaci

Analizu smo proveli koristeći podatke dotoka rijeka dalmatinskih slivova i to rijeke Cetine u akumulacijsko jezero Peruća, rijeke Krke u kompenzacijski bazen Brljan uz hidroelektranu Miljacka i rijeke Zrmanje u akumulacijsko jezero Štikada [12]. Na raspolaganju smo imali:

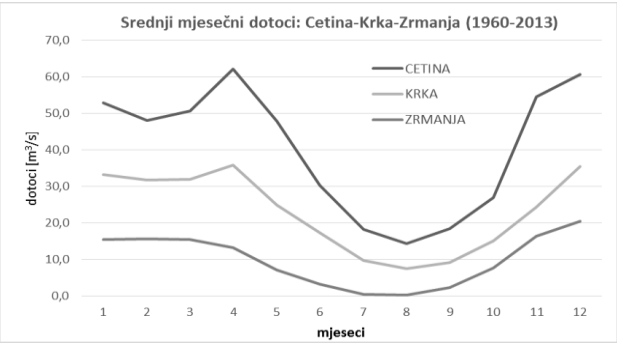
- srednje dnevne dotoke za rijeke Krku i Cetinu od 2000. do 2013. godine,
- srednje mjesečne dotoke za rijeku Krku i Zrmanju od 1950. do 1990. godine te od 1996. do 2013. godine,
- srednje mjesečne dotoke za rijeku Cetinu od 1960. do 1990. godine te od 1996. do 2013. godine.

Svi mjereni dotoci su prirodni dotoci u „gornje“ akumulacije na slivu (ili „sirovi“ podaci), odnosno nikakve uzvodne aktivnosti nemaju utjecaja na mjerene dotoke. Dotoci na pojedinom mjernom mjestu su uvijek mjereni na jednak način, kako bi se smanjila mogućnost greške u fazi prikupljanja podataka. Period od 1991. do 1995. je izostavljen iz analize obzirom da su neki podaci u ovom periodu rezultat estimacije, a ne sustavnog mjerenja kao i ostali.

B. Korelacijska analiza srednjih mjesečnih dotoka

Na slici 1 prikazani su srednji mjesečni dotoci na sva tri mjerna mjesta. Već se sa dijagrama nazire da se prirodni dotoci na ove tri rijeke dobro „prate“. Srednji mjesečni dotoci su za sva tri mjerna mjesta računati u istom periodu i to od 1960 do 1990 te od 1996. do 2013. godine kako bi bili usporedivi. U tom periodu srednja vrijednost dotoka u akumulacije je 39,7 m^3/s na rijeci Cetini, 23,1 m^3/s na rijeci Krki i 9,8 m^3/s na rijeci Zrmanji.

Nepristrana procjena Pearsonovog koeficijenta korelacije prema (2) za sve mjesece u godine izračunata je za dotoke na Cetini i Krki i Cetini i Zrmanji za period od 1960. do 1990. godine te od 1996. do 2013. godine, a za dotoke Krke i Zrmanje za period od 1950. do 1990. godine te od 1996. do 2013. Vrijednosti ovih koeficijenata po mjesecima te njihova srednja vrijednost dani su u tablici 1. Uočavaju se jake korelacijske veze u većini promatranih mjeseci.



Slika 1. Srednji mjesečni dotoci u akumulacije

TABLICA I. VRIJEDNOSTI PEARSONOVIH FAKTORA KORELACIJE ZA SREDNJE MJESEČNE DOTOKE ZEMLJOPISNO BLISKIH AKUMULACIJA

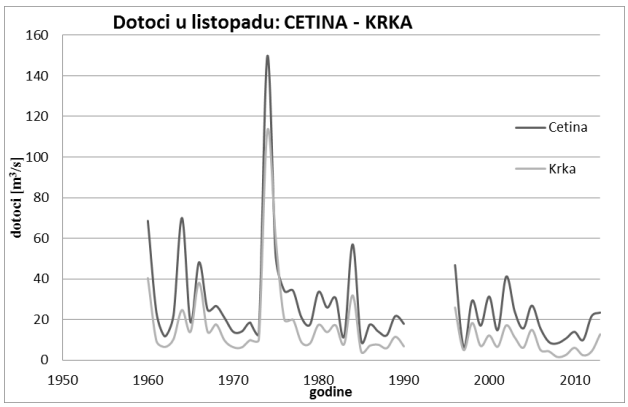
Mjeseci	CETINA-KRKA	CETINA-ZRMANJA	KRKA-ZRMANJA
I	0,8302	0,8739	0,8210
II	0,9120	0,8886	0,8191
III	0,8844	0,8221	0,8097
IV	0,8587	0,7349	0,7259
V	0,8766	0,6970	0,7076
VI	0,8412	0,8120	0,7898
VII	0,8309	0,5979	0,7381
VIII	0,8610	0,4747	0,5652
IX	0,7640	0,7260	0,6969
X	0,9433	0,9258	0,8253
XI	0,8699	0,8730	0,8054
XII	0,8892	0,8952	0,8551
r_sr god	0,8635	0,7768	0,7633

Kako bi bolje ilustrilali korelacijske veze među dotocima u akumulacijske bazene na slikama 2, 3 i 4 prikazani su dotoci u mjesecu listopadu za sva tri promatrana para. Na slikama 5, 6 i 7 prikazani su dijagrami raspršenja za iste korelacijske veze također u mjesecu listopadu. Na dijagramima su također prikazani automatski računati koeficijenti determinacije (r^2). Sva tri dijagrama raspršenja jasno pokazuju tendenciju izduživanja u jednom smjeru.

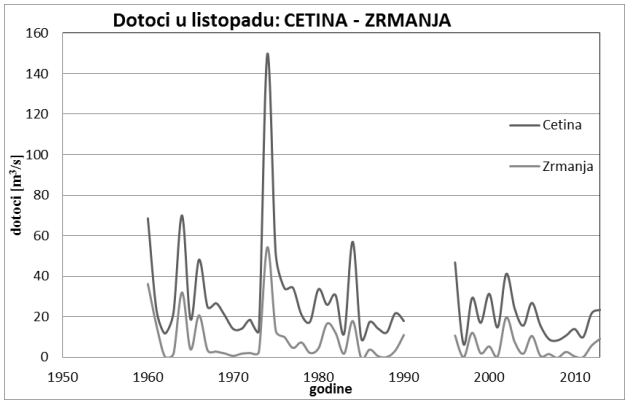
C. Korelacijska analiza srednjih tjednih dotoka

Obzirom da smo raspolagali srednjim dnevnim dotocima u uzvodne akumulacije na Cetini i Krki u istom neprekidnom periodu od 14 godina, iz srednjih dnevnih dotoka smo u prvom koraku izračunali srednje tjedne protoke, a potom izračunali nepristranu procjenu Pearsonovog faktora korelacije za ovaj uzorak prema (2). Dijagram srednjih tjednih dotoka, prikazan na slici 8, jako sliči dijagramu srednjih mjesečnih dotoka za Cetinu i Krku prikazanog na slici 1, ali uz finiju „diskretizaciju“. Osim toga, podaci srednjih mjesečnih dotoka su dobiveni iz znatno dužeg perioda promatranja od 49 godina.

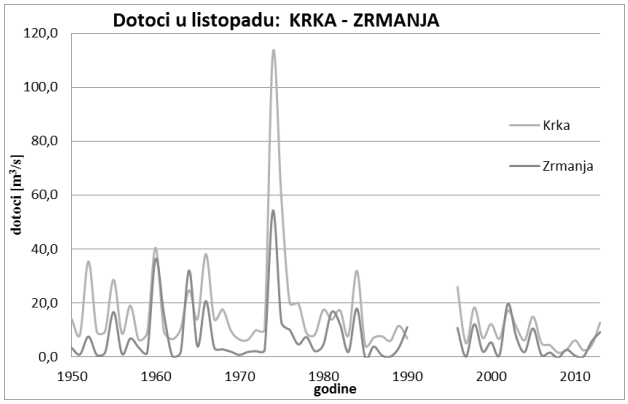
Nepristrane procjena Pearsonovog koeficijenta korelacije za svaki tjedan, prikazane grafički na slici 9, se kreću od 0,6272 u 30. tjednu do 0,9847 u 2. tjednu, prosječno 0,8728. Dakle, i tjedni koeficijenti korelacije između ova dva riječna dotoka su izuzetno visoki. Prosječna vrijednost koeficijenta korelacije srednjih tjednih dotoka je približno jednaka prosječnoj vrijednost koeficijenta korelacije srednjih mjesečnih dotoka (0,8635) iz čega se da zaključit da izbor osnovnog perioda promatranja (tjedan ili mjesec) nema znatnijeg utjecaja na iznose koeficijenta korelacije. Ipak, zaključak treba uzeti s dozom opreza jer proračun nije rađen na istom broju godina, odnosno broj promatranih parova dotoka u proračunu koeficijenta korelacije srednjih mjesečnih dotoka je značajno veći nego u proračunu koeficijenta korelacije srednjih tjednih dotoka.



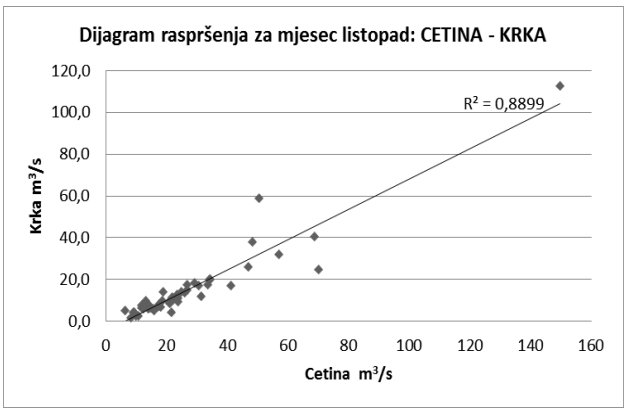
Slika 2. Srednji mjesečni dotoci u listopadu u akumulacije na Cetini i Krki



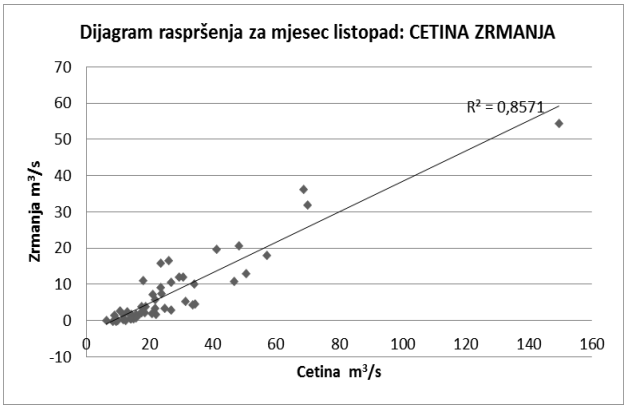
Slika 3. Srednji mjesečni dotoci u listopadu u akumulacije na Cetini i Zrmanji



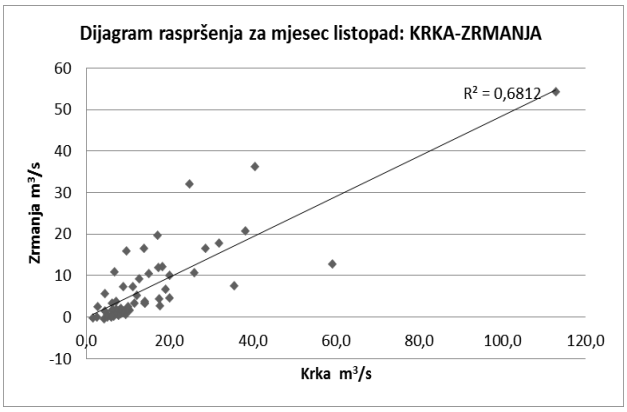
Slika 4. Srednji mjesečni dotoci u listopadu u akumulacije na Krki i Zrmanji



Slika 5. Dijagram raspršenja srednjih mjesečnih dotoka u listopadu u akumulacije na Cetini i Krki



Slika 6. Dijagram raspršenja srednjih mjesečnih dotoka u listopadu u akumulacije na Cetini i Zrmanji



Slika 7. Dijagram raspršenja srednjih mjesečnih dotoka u listopadu u akumulacije na Krki i Zrmanji

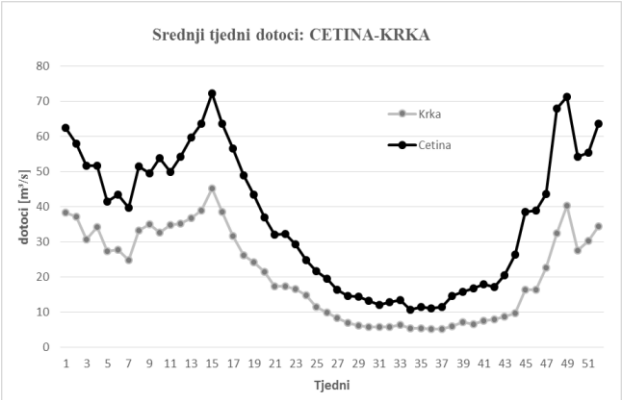
IV. KORELACIJSKA ANALIZA DOTOKA U ZEMLJOPISNO UDALJENE AKUMULACIJSKE BAZENE

A. Dostupni podaci

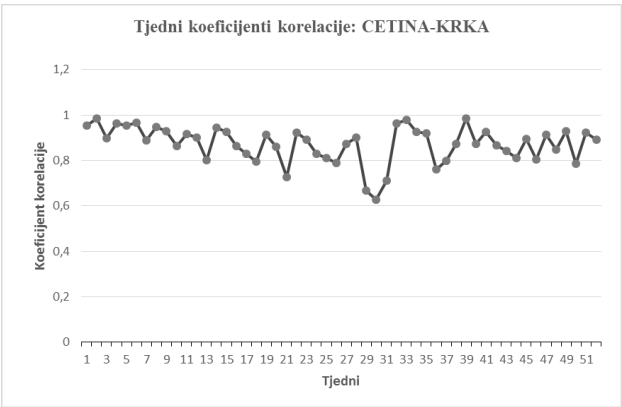
Uz podatke o dotocima na rijeci Cetini, Krki i Zrmanji za korelacijsku analizu dotoka u zemljopisno udaljene akumulacije koristili smo srednje mjesečne dotoke za rijeke Dravu i Kupu od 1950. do 2013. godine. Dok su dalmatinske rijeke Cetina, Krka i Zrmanja dio jadranskog sliva, rijeke Kupa i Drava su dio crnomorskog sliva.

Drava je rijeka kišno-ledenjačkog režima s malom vodnom količinom zimi i velikom krajem proljeća i početkom ljeta zahvaljujući velikoj akumulaciji vode u snijegu i ledu [12] i [13]. Dotoci u rijeku Dravu su mjereni na ulazu u RH, i nisu potpuno nezavisni obzirom na niz uzvodnih HE u Austriji i Sloveniji. Međutim, to su većinom protočne HE, s malim utjecajem na nizvodni režim.

Dotoci u rijeku Kupu mjereni su na zahvatu HE Gojak gdje se dotoci Zagorske Mrežnice i Ogulinske Dobre koriste za proizvodnju električne energije i ispuštaju u rijeku Kupu. Mjereni dotoci su potpuno nezavisni od uzvodnih aktivnosti na riječnom slivu.



Slika 8. Srednji tjedni dotoci u akumulacije na Cetini i Krki



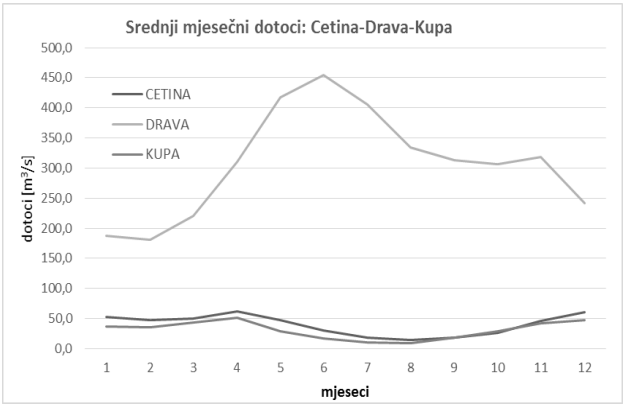
Slika 9. Nepristrana procjena koeficijenata korelacije tjednih dotoka rijeka Krke i Cetine

B. Korelacijska analiza srednjih mjesečnih dotoka

Napravljena je korelacijska analiza dotoka u rijeku Cetinu sa dotocima u rijeku Dravu i dotoka u rijeku Kupu. Iz analize je izostavljen period od 1950. do 1960 i 1991. do 1995. zbog nepotpunih podataka na rijeci Cetini.

Na slici 10 prikazani su srednji mjesečni dotoci na sva tri mjerna mjesta. Slika baš nije pregledna obzirom da je srednji godišnji dotok Drave u mjerenom periodu na promatranom mjernom mjestu (309,5 m^3/s) značajno veći od srednjih godišnjih dotoka Cetine (39,7 m^3/s) i Kupe (30,8 m^3/s) u istom periodu na promatranim mjernim mjestima, ali se ipak uočava da npr. Cetina i Drava imaju „suprotan“ režim.

Nepristrana procjena Pearsonovog koeficijenta korelacije prema (2) za sve mjesece u godine izračunata je za dotoke na Cetini i Dravi te Cetini i Kupi za period od 1960. do 1990. godine te od 1996. do 2013. godine, a za dotoke Drave i Kupe za period od 1950. do 1990. godine te od 1996. do 2013. Vrijednosti ovih koeficijenata te njihova srednja vrijednost dani su u tablici II. Korelacijske veze su slabe, naročito sa rijekom Dravom. Očito je da rijeka Drava pripada sasvim drugom režimu dotoka.

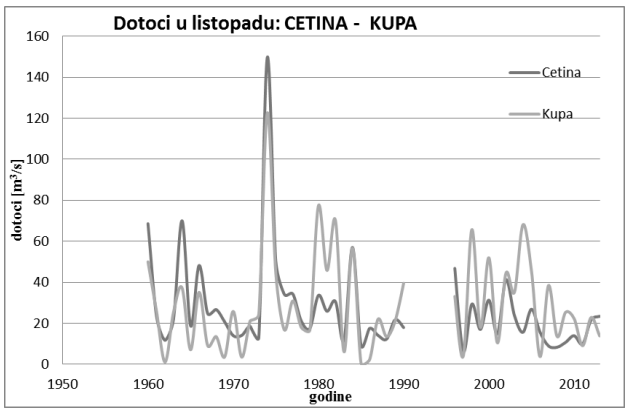


Slika 10. Srednji mjesečni dotoci u akumulacije na Cetina, Dravi i Kupi

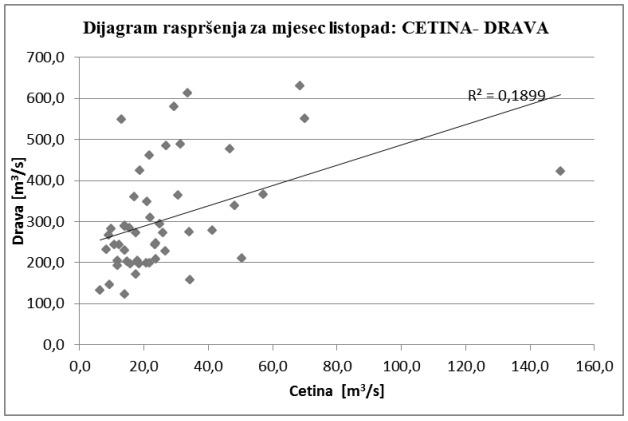
TABLICA II. VRIJEDNOSTI PEARSONOVIH FAKTORA KORELACIJE ZA SREDNJE MJESEČNE DOTOKE ZEMLJOPISNO UDALJENIH AKUMULACIJA

Mjeseci	CETINA-DRAVA	CETINA-KUPA	DRAVA-KUPA
I	0,4651	0,6079	0,5021
II	0,5345	0,6502	0,6167
III	0,6110	0,7527	0,5204
IV	0,4643	0,6921	0,4510
V	0,2267	0,5203	0,1793
VI	0,3603	0,4695	0,1313
VII	0,3688	0,2359	0,4821
VIII	0,0940	0,1873	0,2943
IX	0,1433	0,6255	0,3011
X	0,4358	0,7160	0,5255
XI	0,5610	0,4811	0,5181
XII	0,6519	0,7102	0,4037
r_sr god	0,4097	0,5541	0,4105

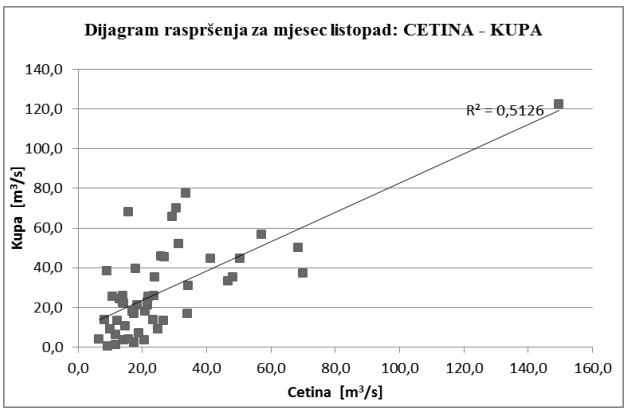
Prikaz srednjih mjesečnih dotoka u akumulacije na Cetini i Kupi u mjesecu listopadu za cijeli promatrani period prikazan je na slici 11. Iz ovog dijagrama se također vidi da dotoci na Cetini i Kupi nisu u značajnijoj korelaciji. Prikazi srednjih mjesečnih dotoka u akumulacije na Dravi, ne samo što nisu u korelaciji, već su i znatno veći od dotoka na Cetini i Kupi te nisu prikazivani takvi dijagrami. Na slikama 12, 13 i 14 prikazani su dijagrami raspršenja za sve tri korelacijske veze također u mjesecu listopadu. Uočavaju se bitno drukčiji oblici, manji faktori determinacije i veće raspršenje na ovim dijagramima nego na analognim dijagramima na slikama 5, 6 i 7.



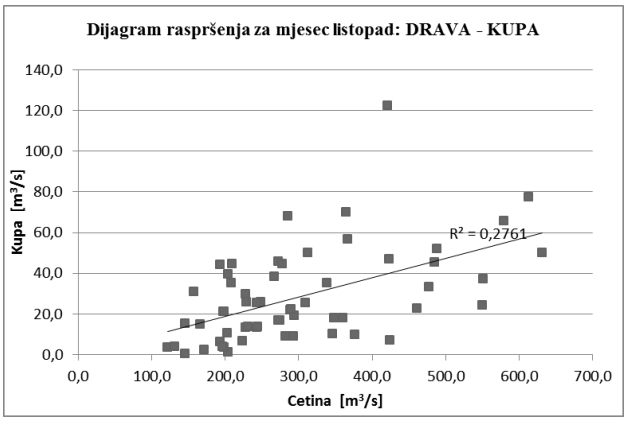
Slika 11. Srednji mjesečni dotoci u listopadu u akumulacije na Cetini i Kupi



Slika 12. Dijagram raspršenja srednjih mjesečnih dotoka u listopadu u akumulacije na Cetini i Dravi



Slika 13. Dijagram raspršenja srednjih mjesečnih dotoka u listopadu u akumulacije na Cetini i Kupi



Slika 14. Dijagram raspršenja srednjih mjesečnih dotoka u listopadu u akumulacije na Dravi i Kupi

V. ZAKLJUČAK

Napravili smo korelacijsku analizu između prirodnih dotoka posebno za geografski bliske, a posebno za geografski udaljene akumulacijske bazene, a sve na različitim riječnim slivovima. Izračunali smo jakost korelacijski veza u analiziranim slučajevima.

Dobiveni rezultati su u skladu sa očekivanjima. Pokazali smo da geografski bliski akumulacijski bazeni, iako imaju različita područja zahvaćanja oborinskih voda, imaju vrlo slične režime dotoka. S druge starne, geografski udaljeni bazeni se ponašaju različito.

Izračunate vrijednosti korelacijskih veza jasno ukazuju da prilikom modeliranja ili simuliranja dotoka u akumulacijske bazene treba vodit računa o korelacijskim vezama među njima. Prije izgradnje dobrog modela dotoka, treba ispitati korelacijske veze među dotocima na različitim bazenima. Ako su veze slabe, dotoke možemo modelirati nezavisno. Tamo gdje su korelacijske veze jake, nezavisno modeliranje dotoka nije realno.

LITERATURA

[1] B. Udovičić, „Elektroenergetika“, Školska knjiga Zagreb, 1983.R. A. Wurbs, „Modeling and analysis of reservoir system operations,“ Prentice Hall PTR, USA, 1996.

[2] D. P. Loucks and E. van Beek, „Water resources systems planning and management - An introduction to methods, models and applications“ – ISBN 92-3-103998-9 – © UNESCO 2005

[3] M. Pereira, N. Campodónico and R. Kelman, „Long-term hydro scheduling based on stochastic models“, *Proceedings of EPSOM'98*, Zurich, September 23-25, 1998.

[4] A Gjelsvik, B. Mo. and A Haugstad, „Long- and medium-term operations planning and stochastic modelling in hydro-dominated power systems based on stochastic dual dynamic programming“, *Handbook of Power Systems I*, Energy Systems, pp 33-55, Springer 2010.

[5] E. Sutlović, S. Čujić Čoko, I. Medic, “Characteristics of basin inflows - A statistical analysis for long-term/mid-term hydrothermal scheduling”, *Proceedings of SDEWES 2013*, September 2013, Dubrovnik, Croatia

[6] S. Bošnjak, E. Sutlović, „Simulacija dotoka autoregresivnim modelom prvog reda”, 37. Međunarodni ICT skup - MIPRO 2014, pp. 1838-1843, Opatija, svibanj 2014.

[7] I. Perišić, E. Sutlović: Statistička analiza funkcije vjerojatnosti prirodnih dotoka, 37. Međunarodni ICT skup - MIPRO 2014, pp. 1848-1853, Opatija, svibanj 2014.

[8] K. Zoričić, E. Sutlović: Korelacijska analiza dotoka 37. Međunarodni ICT skup - MIPRO 2014, pp. 1844-1847, Opatija, svibanj 2014.

[9] R. Žugaj, „Hidrologija“,Udžbenik Sveučilišta u Zagrebu, Zagreb, 2000.

[10] I. Pavlić, „Statistička teorija i primjena“, Tehnička knjiga, Zagreb, 1988.

[11] N. Šakić, „Korelacijska i regresijska analiza“, Sveučilišta u Zagrebu, Fakultet strojarstva i brodogradnje, Zagreb, 2010.

[12] Elektroprojekt d.d., „Hidroelektrane u Hrvatskoj“, Hrvatska elektroprivreda, Zagreb, 2000.

[13] Danijel Režek, “Hidroelektrane na Dravi”, GRAĐEVINAR vol. 55, br 11/2003, pp. 647-653

PLC/SCADA systems in automation control design for individual quick freezing process in cooling tunnels

Goran Jagetić, Marko Habazin, Tomislav Špoljarić
University of Applied Sciences - Department of Electrical Engineering, Zagreb, Croatia
g.jagetic@gmail.com, mhabazin18@gmail.com, tomislav.spoljaric@tvz.hr

ABSTRACT - Individual quick freezing process is a complex cooling process that takes place in cooling tunnels and is used for dynamic thermal processing of certain type of goods. Type of goods and temperature of goods that needs to be achieved over specific time in cooling tunnel determine the complexity of cooling process. Complex cooling process is therefore divided into two types of regulation. Temperature regulation, the inferior type, is used for control of the cooling and defrosting parts of the system according to predefined room temperature. Superior time-cycle regulation is used to control various cycles of operation by defining its time frames. These cycles together form a period of time in which the cooling system is appropriately used for goods' freezing.

This paper describes further development in automation control design for complex cooling process in cooling tunnels. This design includes selection of appropriate control devices (temperature/pressure regulators, PLC device, HMI/SCADA), definition of LADDER program solution for dynamic time-cycle control, connection and configuration of HMI/SCADA system used for defining the number/duration of time cycles of cooling and defrosting operations and monitoring the system conditions.

I. INTRODUCTION

Simple cooling process [1] is thermal process used in refrigeration systems for maintaining room temperature via temperature regulator and at least one temperature probe. Temperature is achieved and maintained by correct switching of refrigeration system elements – compressor, electromagnetic valve, evaporator fans and condenser fans. This basic type of operation is not time-dependent and is most commonly used in small scale cooling systems with required temperature above 0°C.

When required room temperature needs to be maintained below 0°C, refrigeration system needs defrosting for evaporator section due to icing. This is done by switching the electrical heaters that need to be inserted into the evaporator section or, when electrical defrost is not available, by defrosting with hot gas. Defrosting element operation (heater or valve) is done separately from other refrigeration system elements (compressor, fans, etc.).

When defrost is needed, control design takes form of one cycle regulation [1] thus making it time dependent. A repeating cycle of operation consists of two periods of operation: refrigeration period and defrosting period. Refrigeration period is longer and in it refrigeration elements (compressor, condenser and evaporator fans) are switched according to room temperature probe/probes. Defrosting period is much shorter and in it defrosting elements (electrical heaters/hot gas valve) are switched according to evaporator temperature probe or by time (off-delayed timer function of a regulator). Fig. 1 shows simplified block diagram of a small scale refrigeration system operated by one-cycle regulation (refrigeration with defrost).

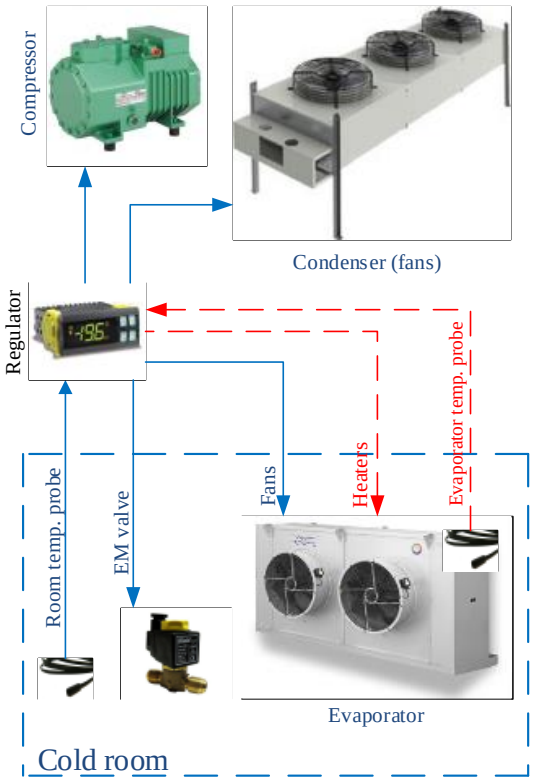


Figure 1. Block diagram of a simple cooling process

On a fig. 2 a time flow diagram of a simple cooling process is shown with one cycle control design (refrigeration with defrost period). A repeating time cycle duration is described with simple equation:

$$T_c = T_r + T_d, \qquad (1)$$

where T_r is a refrigeration period duration and T_d is a defrost period duration. When refrigeration process is active, temperature regulation according to room temperature probe is active. When defrost process is active, temperature regulation according to evaporator temperature probe is active.

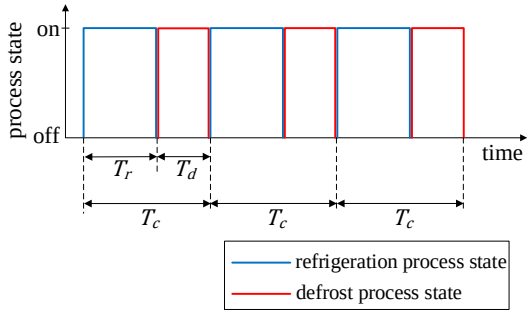


Figure 2. Time flow diagram of a simple cooling process

The complexity of a described control system may increase in larger cooling systems by adding also the pressure control in suction line (for sequential switching control of the compressors in a multiset) and discharge line (sequential switching control of condenser fans). Most of today's temperature regulators have a two cycle control, which also increases the complexity by defining durations of two different refrigeration periods, while defrost duration is the same in each cycle, but is also user defined. All these additions still make the cooling process simple.

II. Complex Cooling Process and Individual Quick Freezing

When defining complex cooling process, one needs to take into account that room temperature is not the main parameter. Room temperature is intermediary parameter for achieving the main parameter - temperature of goods that needs to be achieved over time. This type of processing certain type of goods requires specific definition of time cycles involved in process [3]. By specific definition it is assumed that each cycle has its duration defined by end user. It means that each operation period, whether it is refrigeration or defrost period, has also its duration defined in each of the cycles. A number of cycles is also defined by end user.

Because there are a lot of types of goods that can be cooled or frozen over time [5], control system design for one system needs to have a possibility for end user to define number of operating cycles and various durations of that cycles, thus making it possible for one system to process different type of goods. This type of design makes the system time-dependent, which requires more than one control device to maintain the process on a desired level of efficiency.

On fig. 3 a time flow diagram of one such complex cooling process is shown. This process consists of three refrigeration periods of different durations, and three defrost periods of same duration. Processing time is defined with T_p:

$$T_p = T_{r1} + T_{r2} + T_{r3} + 3 \cdot T_d, \qquad (2)$$

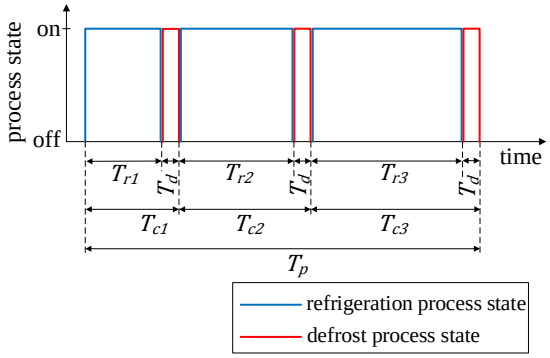


Figure 3. Time flow diagram of a three-cycle complex cooling process

Processing time is a duration of a whole process and it can be repeated. But in individual quick freezing it is a duration of a freezing process of goods, therefore it is not repeated until the next goods' exchange.

There is a possibility for defrost and refrigeration elements of a system to act at the same time for a brief moment when system goes from one state to another. That may be unwanted in larger systems from perspective of rational electric energy consumption and unwanted protective elements' switching states. In that case pause periods are introduced. Pause periods maintain the system inoperative for short period (about 5 to 10 seconds) in order the switching to be done properly.

With this in mind, processing time for n number of cycles can be generally described as:

$$T_p = \sum_{i=1}^{n} (T_{pr,i} + T_{r,i} + T_{pd,i} + T_{d,i}), \qquad (3)$$

where for i-th cycle terms are:

- $T_{r,i}$ – refrigeration period,
- $T_{d,i}$ – defrost period,
- $T_{pr,i}$ – pause period before refrigeration,
- $T_{pd,i}$ – pause period before defrost.

Complex cooling process may be implemented in control design of a small scale refrigeration system (temperature control only with switching of compressors and fans), as well as in larger scale refrigeration systems' control designs with pressure and temperature control. One such design is described in this paper. This proposed design is implemented to work with large scale refrigeration system in a freezing tunnel. Cooling process that takes place in a tunnel is also known as individual quick freezing.

Individual quick freezing is a refrigeration process of a complex nature that manages predefined

type of goods. Generally, the term or its abbreviation IQF relates to fast freezing process of solid foods such as various types of vegetables, meat, fish or fruit while trying to preserve virtually most of the properties of the parent foodstuffs. A great and fast reduction in room/tunnel temperature reduces rapidly the temperature of goods. This halts the activities of microorganisms that cause goods to deteriorate and generate toxic substances. Additionally, the enzymatic and biochemical reactions of the digestive enzymes which are found in all living tissues and whose activities continue even after death are also greatly reduced. In goods frozen with IQF process there is no clumping together and there aren't any pieces of grains. Goods remain individual separate pieces. Hence the term "individual" [6].

There are several types of individual quick freezing processes [4], and the one described in this paper is called individual quick freezing in air-blast freezing tunnels. The goods intended to be frozen are placed on trays. These trays are placed inside the tunnel, where freezing process takes place for a specified amount of time. Afterwards, the packages are removed and usually taken to cold rooms intended for storage.

III. Automation Control Design of a Freezing Tunnel

Devices used in IQF control design can be separated in two groups – devices used for time regulation and devices used for temperature/pressure regulation. Devices used for time regulation are PLC and SCADA/HMI devices, and for temperature and pressure regulation are used various types of specific microcontroller devices.

Freezing tunnel described in this paper has a carrying capacity of 30 tons of different types of fresh fish. Cooling system for that freezing tunnel has two screw type compressors, two evaporators (three fans each) and two condensers (three fans each). Compressor motors are operated by part winding start contactor combination, and all fans are operated by frequency converters. For time regulation Eaton's EASY® relay device Easy800 with additional analog I/O module and XV touch panel is chosen. For temperature/pressure regulation chosen devices are: Carel's pRACK® microcontroller device for compressor and condenser control and MPXpro® microcontroller device for evaporator, electromagnetic and thermoexpansion valve operation.

Compressor control operates according to suction pressure level, temperature and pressure of oil in compressor and cooling medium temperature. Condenser control operates according to discharge pressure level. Since the pRack device manages compressor, the only signal needed is digital input for start/stop. According to that signal and to several probes inside the compressor, the pRack device manages the switching of the compressor.

Evaporator control with MPXpro® operates according to inlet and outlet temperature of the evaporator during refrigeration period. In defrost period electric heaters are switched according to evaporator temperature probe. MPXpro® controller manages also the opening of electromagnetic valve and sequential opening of the thermoexpansion valve.

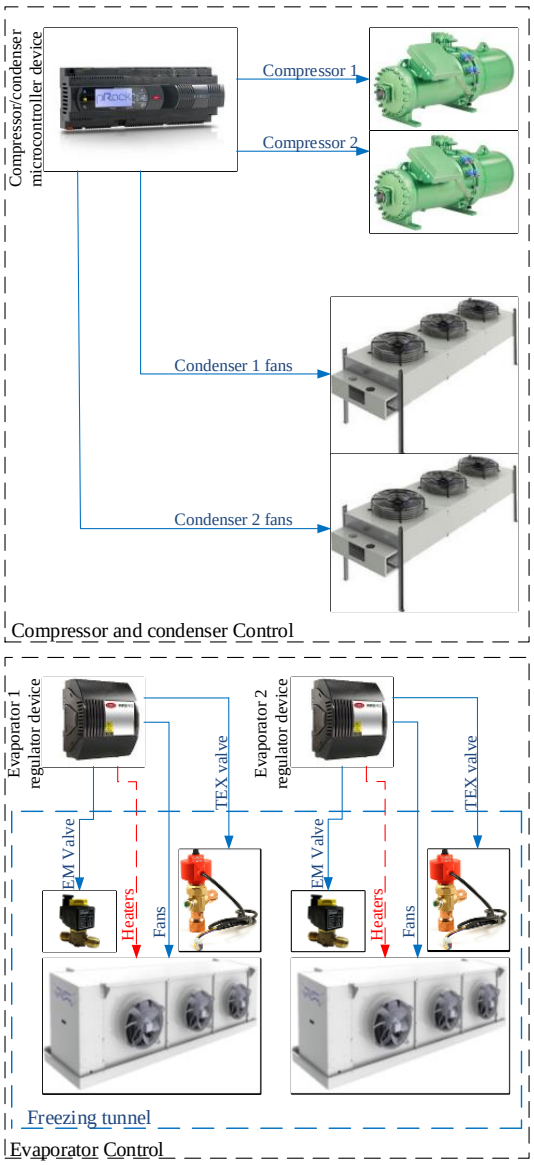


Figure 4. Simplified temperature/pressure control design for individual quick freezing process example

Time cycle control operates according to initial values of number of cycles and duration of cycles entered by end user. This provides desired variability in defining the parameters of processing time by end user. With this the process becomes dependable on goods' type and desired quality of freezing process.

When refrigeration period is active, every refrigeration element of the system works according to its regulator. Compressor is active if pressure

control demands switching on, and if protective devices are not in fault mode: if oil temperature and pressure are in defined limits, cooling medium temperature is also in defined limits and motor protective switch is not tripped. Condenser fans are switched on if discharge line pressure control demands activation and if protective devices are not are not in fault mode. In similar fashion, evaporator fans are operated by evaporator control device and its predefined temperature. All fans have a frequency converters for fine regulation of temperature and pressure.

In defrost period only electric heaters are switched on. This is done according to temperature probe in evaporator connected evaporator control device. In pause period all system components are switched off.

IV. PROGRAM SOLUTION WITH LADDER LOGIC

Complex cooling process is defined with variable number of cycles. Each cycle is defined with sequence pause-refrigeration-pause-defrost.

Minimum number of cycles in a process are three cycles and maximum number are seven cycles. Defrost and pause period are the same for all cycles and are user defined. Refrigeration period in each cycle is separately defined after selecting the number of cycles.

Before programming the system for a cooling purpose starts, it is necessary to make a list of variables. The list includes the physical and virtual I/O variables used later in the program and, and is shown in table 1.

TABLE I. LIST OF VARIABLES IN LADDER PROGRAM

Device	I/O signal	Name
EASY 822-DC-TCX (Main base)	I1	Local start/stop
	I2	Local/remote
	I3	Compressor 1 active
	I4	Comp. 1 motor protection fault
	I5	Comp. 1 oil protection fault
	I6	Evaporator 1 fan mot. prot. fault
	I7	Condenser 1 fan mot. prot. fault
	I8	Compressor 2 active
	I9	Comp. 2 motor protection fault
	I10	Comp. 2 oil protection fault
	I11	Evaporator 2 fan mot. prot. fault
	I12	Condenser 2 fan mot. prot. fault
	Q1	Refrigeration period activation
	Q2	Defrost period activation
	Q3	Pause period activation
	Q4	End of process
	QA1	Speed reference - evaporator fans
EASY 411-DC-ME (Analog module)	AI1	Evaporators' inlet temperature
	AI2	Evaporators' outlet temperature
	AI3	Suction line pressure
	AI4	Discharge line pressure
	AI5	Actual evaporator fan speed
	AI6	Actual condenser fan speed
	AO1	Speed reference - condenser fans
	AO2	(Optional for compressor speed ref.)

The code programming of the process is done in EASY Soft-Pro 6 software package [2]. It is used for communication setup, definition of variables and writing the code of the process. For the purpose of easier understanding, the code is sampled into several main sections.

Communication protocol known as easyNET® network (serial interface RS232) has been chosen for purpose of sending/receiving data between HMI interface (end user) and PLC device.

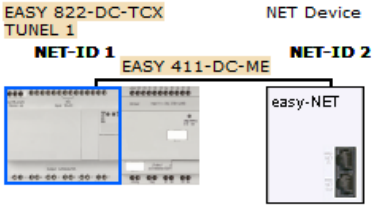


Figure 5. Hardware selection for freezing tunnel project design in easySOFT pro v6.9

First part of the circuit diagram includes getting a number of cycles from easyNET® and transferring it to a memory location in easy822 (block GT01). Comparator blocks CP01 and CP02 are used for checking whether the number of cycles is appropriately selected. If the check condition is fulfilled, a conditional marker M10 is set. This is shown on fig. 6. Based on the user input the program takes number of cycles from the SCADA.

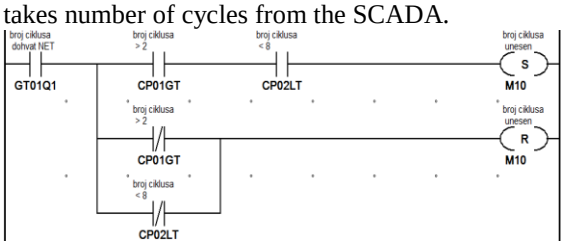


Figure 6. Program section 1 – acquirring a number of cycles

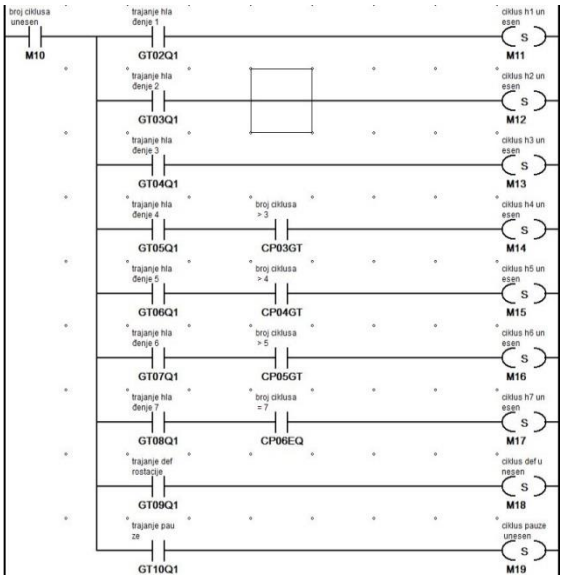


Figure 7. Program section 2 – acquirring a duration of cycles

On fig. 7 a data acquisition from SCADA (GT03-GT08 function blocks) concerning duration of refrigeration periods in each cycle, as well as fixed defrost and pause periods, is shown.

Digital input data acquisition is done by putting bit values to easyNET® network (blocks 4SN01 - 4SN10) according to digital input states (I03-I12), as shown on fig. 8.

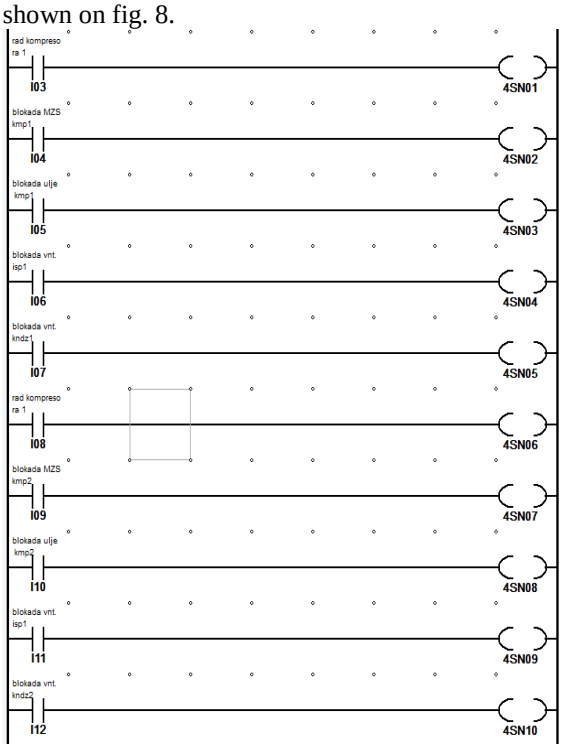


Figure 8. Program section 3 – digital input acquisition

Analog input data acquisition is done by 100ms sample flashing timer. Each time a timer contact makes/breaks, certain data is transferred via "put value to NET" block (PT01-PT06) from input to easyNET® memory location in form of variable MW (marker word with 16bit length), as shown on fig. 9.

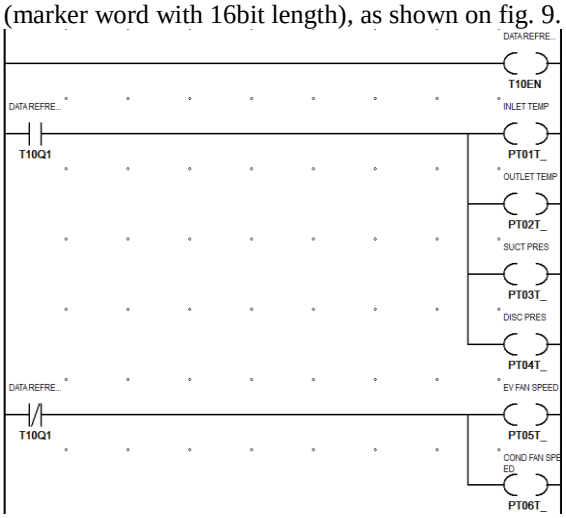


Figure 9. Program section 4 – analog input acquisition

In each cycle, data of different refrigeration period duration needs to be transferred. This is done by using data block which transfers data from memory location where GT block put the duration of refrigeration and transfers it to a memory location of a refrigeration timer.

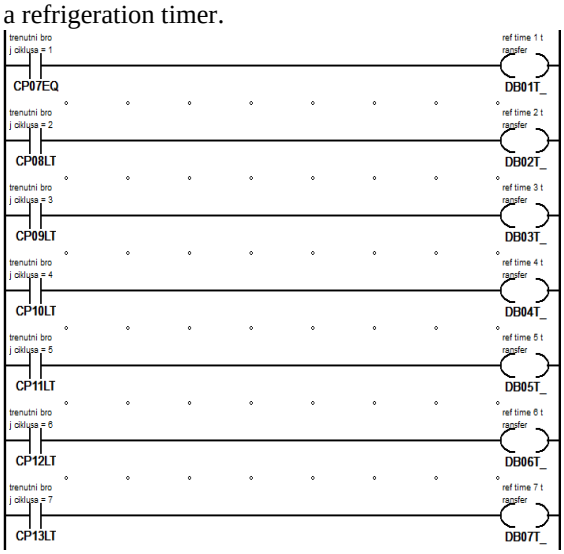


Figure 10. Program section 5 – refrigeration time transfer

User input on SCADA determines the start of the process (4RN01 contact). It is also possible to start the process manually from the electronic cabinet using the 0-1 switch on the door of the cabinet. Start is defined with rising edge marker M05 that is enabled on two additional conditions: refrigeration times for first three cycles have been set (M06) and defrost and pause periods have been set along with number of cycles (M07). Also, finishing the process depends on value in counter of cycles (C01) or local switch position (I01).

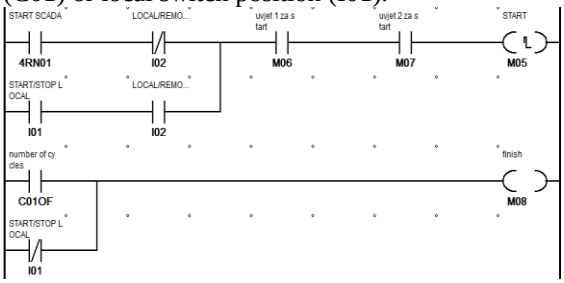


Figure 11. Program section 6 – conditions for starting and finishing of the process

Starting and repeating condition of the process is shown on fig. 12. Timer starts counting on a falling edge of a pulse given by starting (M05) or repeating cycle marker (M09). In this instance a counter C01 counts up.

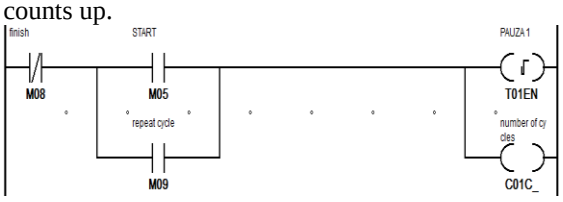


Figure 12. Program section 7 – starting the first timer of the process and repeating the cycle

When timer T01 is done with counting, similar process starts with refrigeration timer T02. This

process repeats with pause timer T03 and defrost timer T04. After T04 is done, a marker M09 for repeating the cycle is set and the process repeats in next cycle if a counter C01 contact allows it.

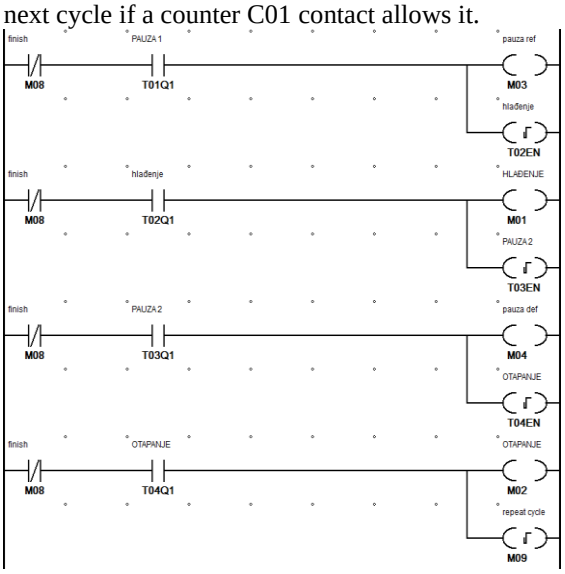


Figure 13. Program section 8 – pause, refrigeration and defrost timers counting

When the process is finished, a timer and counter are reset, as is shown on fig. 14.

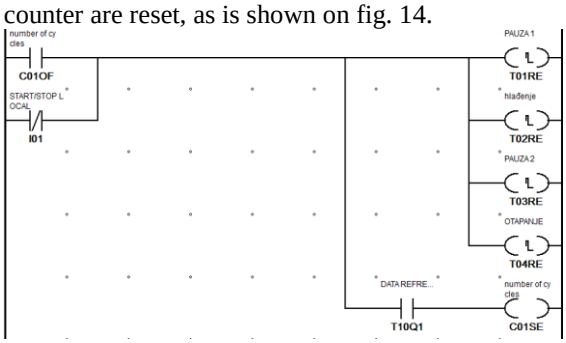


Figure 14. Program section 9 – reset of timers and counter upon completion of the process

V. HMI / SCADA System Connection

In this section, a RS232 serial connection between Easy822 PLC device and HMI device is observed. Other ways to ensure PLC-HMI connection include RS485 and Ethernet protocols. This cooling process is designed for general purpose and with minimum cost and primarily for that reason RS232 communication is chosen. Also, with this type of communication no additional converter modules are needed.

In this paper, proposed SCADA system is designed to work in a cooling system with three freezing tunnels. Control design with three PLC devices is selected to show the modular behavior of a proposed three tunnel control system. In this case each tunnel's control system is presented as one module of that design. Every tunnel is controlled by one EASY relay with additional analog module. Eaton's XV-152 touch panel is selected as a HMI device. Touch panel is connected with PLC devices through serial RS232 communication cable. This cable is used for connection of a panel and first PLC device. All three PLC devices are connected via Eaton's easyNET® network. A program for selected HMI device is made in Eaton's Galileo® software package. Described cooling system design with SCADA is shown on fig. 15.

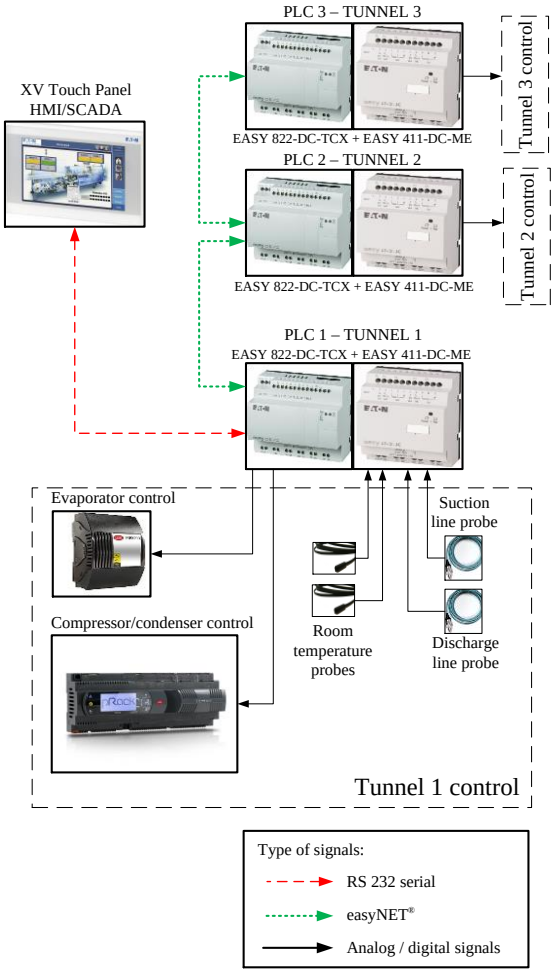


Figure 15. Observed cooling system design

Data used by SCADA system consists of input signals listed in table II and output signals listed in table III.

TABLE II. List of Input Signal Variables in SCADA

Device	Input signal	Name
XV-152 Touch panel	I1	Local start/stop
	I3	Compressor 1 active
	I4	Comp. 1 motor protection fault
	I5	Comp. 1 oil protection fault
	I6	Evaporator 1 fan mot. prot. fault
	I7	Condenser 1 fan mot. prot. fault
	I8	Compressor 2 active
	I9	Comp. 2 motor protection fault
	I10	Comp. 2 oil protection fault
	I11	Evaporator 2 fan mot. prot. fault
	I12	Condenser 2 fan mot. prot. fault
	AI1	Evaporators' inlet temperature
	AI2	Evaporators' outlet temperature
	AI3	Suction line pressure
	AI4	Discharge line pressure

TABLE III. LIST OF OUTPUT SIGNAL VARIABLES IN SCADA

Device	Output signal	Name
XV-152 Touch panel	Q1	Pause period activation
	Q2	Refrigeration period activation
	Q3	Defrost period activation
	Q4	End of process
	QA1	Speed reference - evaporator fans

Communication with easyNET® protocol manages up to 8 connected devices, and the last element of the network needs to be terminated with a bus termination resistor. Example of a wiring concept of such network is shown on fig. 16.

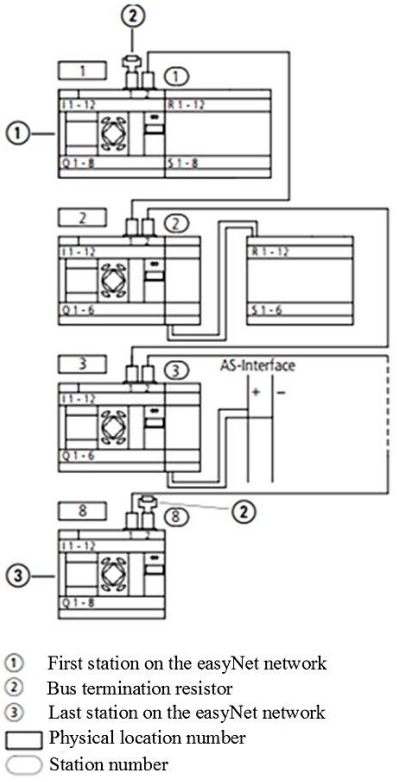


Figure 16. Wiring concept of an easyNET® network with up to 8 different Easy devices [4]

Two masks of SCADA design are shown. First mask is shown on fig. 17 and is used for process monitoring of the entire system. In this mask a selection of one of the tunnels is possible.

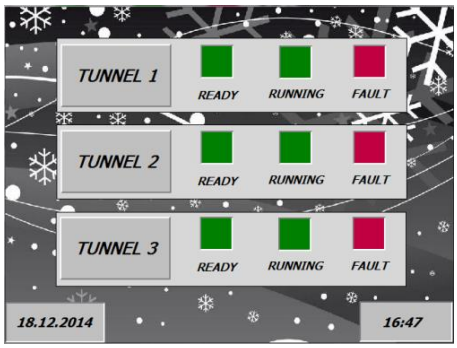


Figure 17. First mask of SCADA design

By selecting, end user can gain access to other masks used for monitoring and control each of the tunnels. One such mask (monitoring the tunnel 1) is shown on fig. 18.

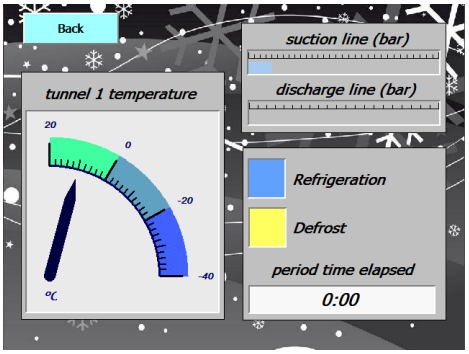


Figure 18. Monitoring mask of SCADA design for freezing tunnel 1

When connecting by easyNET® several Easy800 stations operated by the same LADDER diagram and one HMI device, a precaution needs to be taken when defining tags in SCADA. When linking a certain signal with its graphic representation on SCADA, e easyNET®'s station number from which the signal originates comes first. Then, the address of a signal is defined. In table IV several examples for signal tags from tunnel 1 are listed.

TABLE IV. EXAMPLES OF ADRESSING SIGNALS IN LADDER AND SCADA

Address name in LADDER diagram	Address name in Galileo (SCADA)
I1	NET1.I1
I3	NET1.I3
I4	NET1.I4
I12	NET1.I12
QA4 (MW95)	NET1.MW95

VI. CONCLUSION

Individual quick freezing is a complex cooling process that requires certain additions in cooling control design. Standard temperature/pressure regulators control all the refrigeration elements in a system. Due to dynamic nature of process over time additional regulation device is introduced. A PLC device controls the appropriate timing of each cycle in a process.

Possibility to define different cycles along with their number was introduced via SCADA system design running on a HMI device. Monitored values such as temperature, pressure and time of the current cycle are visible by SCADA. This gives the end user access to define the process itself, as well as visual control of the process that is currently running.

Finally, a three tunnel connection is introduced as an example to show the enhancement potentialities of modular design with several small PLC devices. Although one more complex PLC device could be selected and programmed for use

with three tunnels, this approach was selected due to its modular design and program simplicity: one simpler LADDER program can be downloaded in up to eight different smaller PLC devices and a network of tunnels can be easily set through this approach. According to requirements of a predefined three tunnel network a SCADA system was enhanced to work with three cooling tunnels simultaneously.

With correct selection of equipment a described type of automation control design for freezing tunnels can be accessible, used for general purpose and easy to maintain.

REFERENCES

[1] Špoljarić T., Fruk M., Vujisić G., "Program Solutions For The Complex Cooling Process Via LADDER Logic", 37th International Convention, MIPRO, Opatija, 2014.

[2] easy800 User Manual – 4th edition, Moeller, 2005.

[3] J. Garden-Robinson J., Food Freezing Guide, revised edition, NDSU, 2011.

[4] From Fresh to Frozen Fresh: An Explanation of Today's Quick Freezing Processes – Consumer Article, National Frozen & Refrigerated Food Association, www.nfraweb.org, 2009.

[5] Becker B.R., Fricke B. A., "Freezing Times of Regularly Shaped Food Items", University of Missouri, Kansas City, February 1999.

[6] Pruthi J. S., Quick Freezing Preservation of Foods: Foods of plant origin, Allied Publishers, 1999.

Sustav grijanja pogona za sušenje papira

Robert Bukvić, Goran Malčić i Danijel Maršić
Tehničko veleučilište u Zagrebu/Elektrotehnički odjel, Zagreb, Hrvatska
goran.malcic@tvz.hr

Sažetak - S obzirom da je u današnje vrijeme potražnja papira sve veća, proizvođači su prisiljeni postavljati sve više granice u smislu proizvodnje i poboljšanja kvalitete papira. Tako u papirnoj industriji veliku ulogu vodi sama izgradnja papirnog stroja koji je u potpunosti automatiziran pomoću programljivih logičkih kontrolera (eng. *Programmable Logic Controllers, PLC*). U radu je prikazan vrlo bitan segment papirnog stroja koji služi za sušenje papira prilikom proizvodnje. Prikazan je tehnološki proces i sustav automatizacije te nadzorno-upravljačko rješenje sustava.

I. UVOD

U današnjem svijetu papir igra veliku ulogu u našim životima zbog svakodnevnog korištenja bilo to za pisanje, tiskanje, čitanje, pakiranje, novčano plačanje, svakodnevnu higijenu i još mnoge druge svrhe. Zbog velike potražnje papira kroz povijest inženjeri su bili prisiljeni pronaći način kako da se ljudima omogući veća potrošnja papira. Izumom tiskarskog stroja započinje nova era moderne proizvodnje papira gdje prvi puta u povijesti francuz Luis Rober 1799. godine u proizvodnju uvodi prvi parni stroj pomoću kojega se mogla proizvoditi beskonačna papirna traka. Pravi napredak u masovnoj industrijskoj proizvodnji papira je započeo tek u prvoj polovici dvadesetoga stoljeća [1].

Ovaj rad se bazira na jednom od važnijih dijelova papirnog stroja u kojemu je prikazan sustav grijanja za dio pogona koji je zadužen za sušenje papira tzv. "Dryer section". Cilj rada je prikazati jedno od modernih rješenja sustava grijanja pogona za sušenje papira kompanije Voith koja je jedna od vodećih kompanija u papirnoj industriji. Pritom je prikazan pristup razvoju nadzorno-upravljačkih rješenja takvih sustava upotrebom programskog rješenja PCS7 kompanije Siemens.

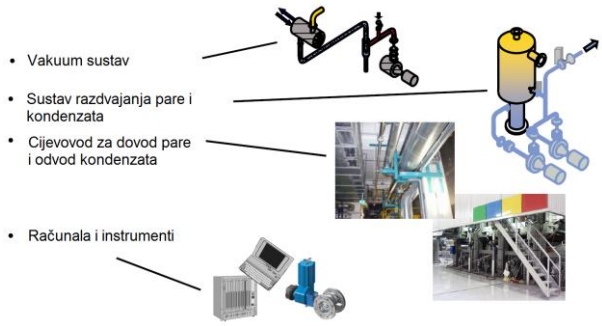
II. TEHNOLOŠKI PROCES SUSTAVA GRIJANJA POGONA ZA SUŠENJE PAPIRA

Sustav grijanja pogona za sušenje papira je vrlo bitan segment u papirnoj industriji jer ovisno o sustavu grijanja koji se koristi u postrojenju, može se doprinijeti maksimalnom učinku sušenja papira uz minimalnu potrošnju energije.

Svrha pogona za sušenje papira je povećati udio suhog sadržaja papirnog tkiva između 90 i 98% pomoću isparavanja. Sušenje papira je proces prijenosa toplinske mase kako bi toplina bila prenesena od izvora topline do papira po principu isparavanja vode. Ovisno o vrsti papirnog stroja odnosno vrsti papira primjenjuju se različiti prijenosi topline i načini sušenja papira.

Najčešći princip sušenja papira je pomoću pare i kondenzata na način da se vodena para kondenzira na unutarnjoj površini stijenke cilindara kroz koje prolazi papirno tkivo. Toplina se prenosi kroz papirno tkivo na način da se zagrijava isparavanjem vode. Stopa prijenosa topline iz pare do stijenki cilindara ovisi o brzini protoka kondenzata. Brzina protoka uglavnom ovisi o brzini papirne mašine, ali i o količini kondenzata kojeg cilindar može primiti te o promjeru samoga cilindra. Brzina papirnog stroja za vrijeme radnog procesa iznosi oko 1000 m/min.

Sustav grijanja pogona za sušenje papira u većini slučajeva nalazi se ispod postrojenja papirnog stroja te se sastoji od sustava za vakuum, sustava za razdvajanje pare i kondenzata, cjevovoda za dovod pare i odvod kondenzata te računala i instrumenata u polju koji služe za upravljanje kompletnim sustavom (sl.1.).



Slika 1. Osnovni segmenti sustava grijanja pogona za sušenje papira

Glavne značajke sustava za grijanje pogona za sušenje papira su:

- odvodnjavanje kondenzata iz cilindara za sušenje papira.
- opskrba parom cilindra za sušenje papira.
- utvrđivanje suhoće papira nakon pogon za sušenje.
- kontrola profila vlage u smjeru papirnog stroja.

Sustav grijanja podijeljen je prema toplinskim skupinama (eng. *Heating groups, HG*):

- kontrola profila vlage u smjeru papirnog stroja.
- HG 1 - 1. pred sušenje (eng. *1st pre-dryer*)
- HG 2 - 2. pred sušenje, glavna skupina (eng. *2nd pre-dryer, master group*)
- HG 3 - 3. završno sušenje (eng. *3rd after dryer*)
- HG 4 - 4. završno sušenje (eng. *4th after dryer*)

Sustav grijanja za pogon sušenja papira opremljen je sa jednim stupnjem kaskadnog sustava koji se nalazi u dijelu predsušenja i obuhvaća toplinsku skupinu 1, dok se

toplinska skupina 2 te toplinske skupine 3 i 4 koje se nalaze u završnom dijelu sušenja opskrbljuju direktno svježom parom koja pristiže iz kotlovnice.

Svrha kaskadnog sustava je iskoristiti akumuliranu paru dobivenu iz spremnika prve grupe i kondenzat koji je nastao prilikom gubitka pritiska te koji se korsiti u sljedećoj manje grijanoj toplinskoj skupini. Ovako nastala para se odvaja od kondenzata u pojedinačnim spremnicima tvz. separatorima te se dovodi do sljedeće toplinske skupine dok se kondenzat dostavlja u spremnik za pohranijvanje (eng. *Collecting tank*) preko razlike tlaka ili preko pumpe za kondenzat.

III. STRUKTURA SUSTAVA AUTOMATIZACIJE

Kompletna infrastruktura računalne mreže bazirana je na ethernet i profibus komunikaciji što je u većini slučajeva standardan način komunikacije u industrijskim postrojenjima (sl.2.).

Računalne stanice za upravljanje i nadgledanje sustava:

- WinCC serveri (eng. *WinCC Servers - Server 1, Redundant Server 2*)
- inženjerske stanice (eng. *Engineering Stations*)
- operatorske stanice (eng. *Operator stations - OS1, OS2, OS3, OS4*)
- OnQ serveri (eng. *OnQ PCs / OnV Monitoring / Profilmatic / Measurement Server*)

Računalne stanice za upravljanje i nadgledanje sustava međusobno su povezane preko mrežnog preklopnika trećeg sloja (eng. *Level 3 switch*) putem ethernet mrežne komunikacije brzine 100 Mbit/s kao i sa programirljivim logičkim kontrolerima koji se nalaze u mrežnoj strukturi. Ethernet mreža se pritom dijeli na:

- procesnu mrežu (eng. *Process Net*)
- terminalnu mrežu (eng. *Terminal Net*)

Terminalna mreža služi za komunikaciju između svih računalnih stanica te printera dok procesna mreža služi za komunikaciju inženjerskih stanica, WinCC servera te OnQ servera sa programirljivim logičkim kontrolerima putem komunikacijskog procesora CP 443-1. Daljnja komunikacija prema distribuiranim periferijama koje se nalaze u polju odvija se putem profibus računalne mreže preko PLC komunikacijskog procesora CP 443-5 [2].

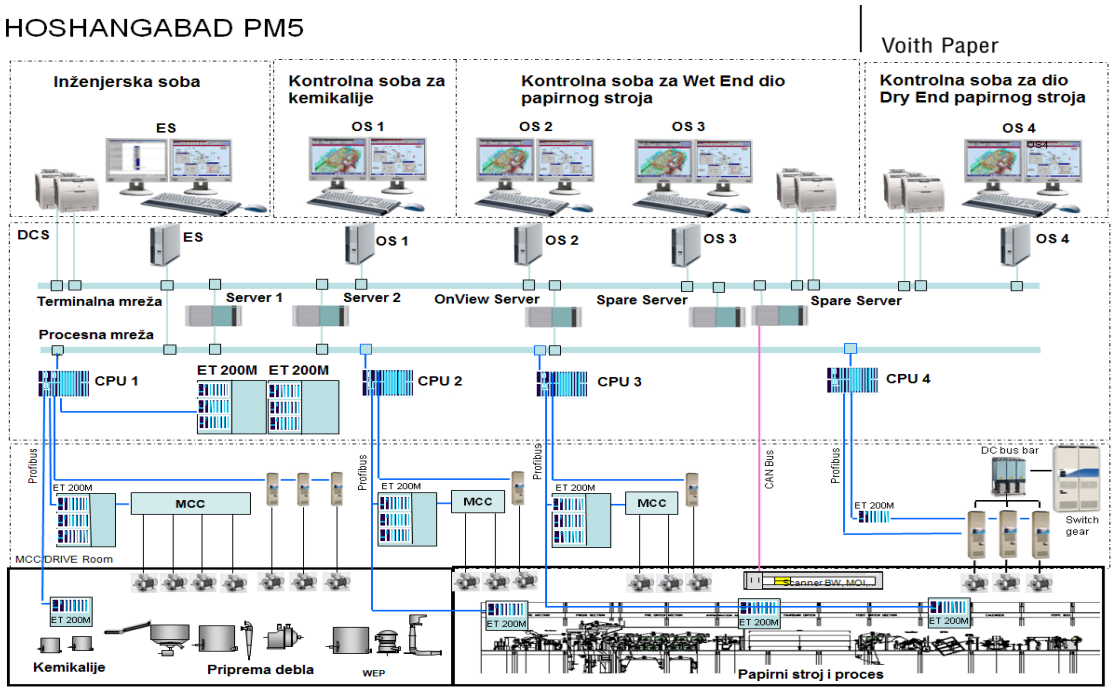
IV. PROGRAMSKO OKRUŽENJE PCS7

Kao programski alat za izradu sklopovske konfiguracije i programa za upravljanje procesom koristi se Siemensovo programsko okruženje PCS7, koji je uz Siemens Step 7 često korištišteni programski alat u industrijskim postrojenjima [3]. PCS7 je sustav centralnog inženjeringa gdje niz alata maksimalno pojednostavljuje te pruža bržu i kvalitetniju izradu projekata. Svi inženjerski alati su dio PCS7 programskog okruženja za izradu programskih aplikacija, konfiguriranje skolopvlja te komunikacijskih funkcija, a pozivaju se iz centralnog projektnog preglednika Simatic Manger (sl. 3.).

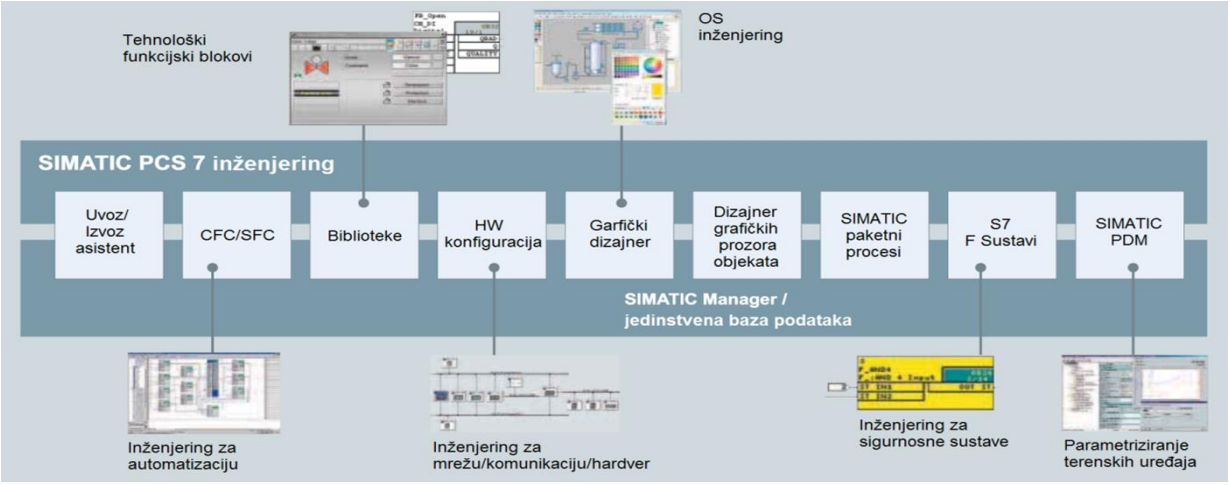
Programiranje PLC uređaja izvodi se prema zahtjevima krajnjeg korisnika te prema funkcijskom opisu procesa i funkcijskim planovima danima od izvođača projekta. PCS7 sadrži programske pakete koji se mogu programirati u tekstualnom ili grafičkom obliku kao što su:

- CFC (eng. *Continuous Function Chart*)
- SFC (eng. *Sequential Function Chart*)
- SCL (eng. *Structured Control Language*)

U ovom slučaju program se izvodi preko CFC grafičkog načina programiranja koji se inače koristi kao programski paket za dizajniranje biblioteka (eng. *Library*), automatizacijske logike, blokada (eng. *Interlocks*) i kontrolnih procesa.



Slika 2. Struktura računalne mreže sustava automatizacije



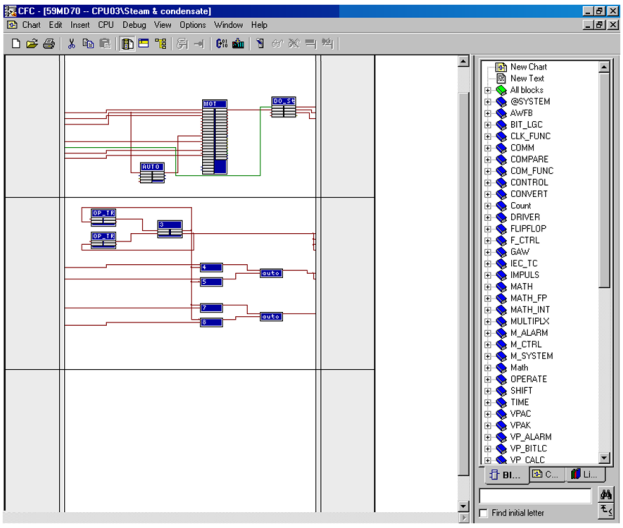
Slika 3. Struktura programskog okruženja PCS7

Program se kao što je već spomenuto izrađuje preko CFC grafičkog načina programiranja u CFC editoru. CFC petlja se može otvoriti i izraditi u bilo kojem od preglednika, ali radi bolje preglednosti postrojenja i programiranja korisiti se iz preglednika po postrojenju. Prije dodavanja CFC petlje na nivou CPU03 dodijeljuje se hijerahijska mapa sa nazivom sekcije papirnog stroja.

U CFC editoru koji je prikazan na sl. 4. sa desne strane nalazi se katalog sa bibliotekama koje sadrže:

- funkcije (eng. *Functions, FC*)
- funkcijske blokove (eng. *Functions blocks, FB*)

Funkcije i funkcijski blokovi ubacju se iz kataloga u CFC editor koji se sastoji od particije sa šest prozora za izrađivanje programa gdje svaki prozor sa lijeve i desne strane sadrži pregled ulaznih i izlaznih signala (sl. 4.).



Slika 4. Osnovni dijelovi CFC editora

Programsko rješenje je izrađeno na temelju službene dokumentacije tvrtke Voith koja sadrži:

- funkcije (eng. *Functions, FC*)
- funkcijski opis, FDS (eng. *Function description, FDS),*
- funkcijske planove, FUP (eng. *Function plan, FUP*),
- ulazno izlaznu listu signala, U/I lista (eng. *Input output list, I/O list*).

V. IZRADA UPRAVLJAČKOG PROGRAMA

Prije samog unošenja programa potrebno je odrediti u kojem ciklusu će se izvoditi pojedini dijelovi programa, definirati tablicu sa ulazno izlaznim signalima (eng. *Symbol table*) te izraditi predloške CFC petlji na temelju kojih se izrađuje upravljački program.

Ciklus za izvođenje programa odabran je prema funkciji koja se obavlja u procesu pa je tako za razmatrani sustav grijanja određeno periodičko izvođenje programa sa blokovima OB32 i OB33. Blokovi za periodičko izvođenje programa koriste se za izvođenje regulacijskih petlji, motora te prekida u realnom vremenu. U ovom slučaju CFC petlje koje sadrže motore, upravljačke ventile te grupno pokretanje postavljene su u ciklus OB33 koji vrši izvođenje programa u 500ms. CFC petlje koje vrše mjerenja u procesu postavljene su u duži ciklus OB32 koji vrši izvođenje programa u 1s. CFC petlje postavljene u ciklus OB33 postavljene su iz razloga jer dolazi do bržih promjena određenog događaja u procesu te zbog kraćeg perioda izvođenja omogućuje se brža reakcija. CFC petlje koje su postavljene u ciklusu OB32 služe samo za prikaz mjerenih vrijednosti iz procesa za što je dovoljan ciklus od 1 sekunde.

Nakon definiranog cilkusa za izvođenja programa potrebno je definirati ulazno izlaznu tablicu signala koja služi kao komunikacija između sklopovlja i programske podrške. U/I tablica signala pripremljena je od strane projektanata koji su zaduženi za detaljno planiranje sklopovlja, ali u praksi se često dogodi da dana dokumentacija ne odgovara u potpunosti softveru koji je potrebno izraditi pa je tablicu potrebno detaljno provjeriti i ispraviti ukoliko se ne podudara sa trenutnim stanjem. U/I tablicu signala potrebno je eksportirati (eng. *Export*) iz sustava Comos koji služi za detaljan inženjering. U/I tablica signala se eksportira kao excel datoteka te ista importira u PCS7 tablicu simbola koja se nalazi na nivou programa u PCS7 sučelju. Za importiranu U/I tablicu potrebno je provjeriti U/I adrese sa već konfiguriranom sklopovskom konfiguracijom te naziv svakog signala sa U/I listom signala.

Zbog opsežnosti programa u nastavku rada nisu detaljno opisane CFC petlja programa već je prikazan pristup izradi CFC predložaka koje je potrebno izraditi prije samog kodiranja na kojima se baziraju ostale CFC petlje. CFC predlošci su podijeljeni prema sljedećem:

- elektromotorima
- upravljačkim ventilima
- senzorima

Prije ubacivanja blokova definira se pozicija određenih blokova po prozorima CFC editora radi lakšeg pregleda programa. Prozori su definirani na sljedeći način:

- 1. prozor za blokove ulaznih signala
- 2. prozor za blokove izlaznih signala (osim u slučaju motora gdje se nalazi u 4. prozoru)
- 4. prozor za glavne funkcijske blokove VP_MOT, VP_PID, VP_MMON
- 5. prozor za blokove blokade i praćenja izlazne vrijednosti, te u slučaju motora za logiku za automatsko upravljanje.
- ostali prozori su za ostalu logiku po vlastitom izboru korisnika

Prilikom ubacivanja blokova potrebno je obratiti pozornost na raspored blokova koji se vidi u desnom uglu svakog bloka u zelenom kvadratiću gdje je prikazan ciklus i pozicija bloka. Pozicija bloka se definira prema rasporedu izvršavanja što je jako bitno prilikom izvođenja programa. Tako se blokovi ulaznih signala postavljaju na prvu poziciji, blokovi izlaznih signala na zadnju poziciju, a glavni funkcijski blokovi između ulaznih i izlaznih blokova u ovisnosti o redoslijedu ostalih blokova.

A. Elektromotor

Za izradu CFC petlje pumpe za kondenzat na spremniku kondenzata definiraju se blokove koji su potrebni za nadzor i upravljanje. Elektromotor ima ulogu pumpe koristi 2 ulazna i 1 izlazni signal za komunikaciju sa PLC uređajem. Potrebno je osigurati upravljanje i nadzor putem nadzornih zaslona te povezati logiku koja je zadužena za automatsko pokretanje pumpe. Za izradu CFC predloška za pumpe korišteni su standardni blokovi iz kataloga.

B. Upravljački ventil

Softverske izvedbe upravljačkih ventila baziraju se na istoj logici koja sadrži PID regulator koji reguliranu izlaznu vrijednost šalje na izlaz upravljačkog ventila. Upravljački ventil koji je upravljan preko PID regulatora koristi 1 ulazni i 1 izlazni signal dok u slučaju gdje PID upravlja sa dva upravljačka ventila koriste se 2 izlazna signala.

U slučaju da ulazna procesna vrijednost dosegne ili prekorači maksimalnu i minimalnu vrijednost, na nadzornom ekranu pojavljuje se alarm. Vrijednosti za alarm definiraju se na sljedećim ulazima bloka PID:

- LIM_HH za alarm visokog prioriteta PV Alarm High
- LIM_LH za alarm niskog prioriteta PV Warning High
- LIM_LL za alarm niskog prioriteta PV Warning Low
- LIM_LH za alarm visokog prioriteta PV Alarm Low

VI. ALGORITAM UPRAVLJANJA SUSTAVOM

Kako bi algoritam upravljanja sustavom ispravno radio potrebno je osigurati sljedeće uvjete:

- instrumentalni i radni zrak,
- paru,
- cilindri za sušenje u pripremi za pokretanje,
- otvoreni ručni ventili i ventili na pumpama,
- zatvoreni ventili za odvodnjavanje na separatorima,
- puštena voda za hlađenje za vakuum kondenzator,
- zadane vrijednosti upisane.

Nakon što su svi potrebni uvjeti osigurani može se krenuti sa pokretanjem sustava grijanja.

A. Dovod pare kroz glavnu dovodnu cijev

Osiguranu paru koja dolazi iz kotlovnice potrebno je propustiti do sustava za grijanje tj. do razdjelnika za paru (eng. *Steam Distributor*) putem dva upravljačka ventila PCV01 i PCV02 koji koriste tzv. Split range način kontrole pritiska. Zbog velikog promjera dovodne cijevi za paru koriste se dva upravljačka ventila od kojih je PCV02 premosni ventil (eng. *Bypass valve*) koji je ujedno manji od glavnog ventila PCV01 te je montiran na zaobilaznu cijev manjeg promjera. Zbog velike temperature i pritiska pare osigurano je sporije otvaranje ventila PCV02 te nakon što se on u potpunosti otvori, tek tada se u potpunosti otvoriti ventil PCV01 kako ne bi došlo do naglog porasta temperature u cijevima i hidraulučkog udara ukoliko je negdje prisutan kondenzat. Otvaranjem ventila PCV01 povećava se temperatura u cijevi, a time i pritisak kojega je potrebno smanjiti na 4.5 bara. Povećanjem temperature preko 100°C otvara se ventil TCV01 koji je upravljan preko PID regulatora koji tako propušta konenzat kako bi postepeno ohladio paru tj. smanjio pritisak.

B. Dovod pare kroz glavnu dovodnu cijev

Nakon što se dobio željeni pritisak od 4.5 bara kreće se sa modom Odzračivanje (eng. *Dearation*). Prije samog odzračivanja sustava svi regulatori i motori se postavljaju u automatski mod rada. Pritiskom na tipku AUTO na nadzornom ekranu šalje se signal na na sve blokove regulatora i motora, a pritiskom na tipku DEARATION sustav počinje sa odzračivanjem. Određeni PID regulatori postavljaju se u mod praćenja tj. upravljani ventili se u potpunosti otvaraju za vrijeme odzračivanja i pokreće se pumpa 59MD60 kako bi se odzračio vakuum sustav. Neposredno nakon pokretanja odzračivanja postepeno se počinje otvarati ventil glavnog ventila za paru kako ne bi došlo do hidrauličkog udara i oštećenja cilindara unutar toplinskih skupina.

C. Održavanje topline sustava

Nakon završetka odzračivanja sustav se postavlja u mod Održavanja topline za pred sušenje i završno sušenje (eng. *Keep warm*). Za vrijeme moda Održavanje topline prekida se mod praćenja PID regulatroa te počinje kontrola sa izlaznim vrijednostima regulatora pritiska i diferencijalnog pritiska. Putem upravljačkih ventila pritiska regulirano se propušta para u cilindre na pogonskoj strani stroja dok se sa diferencijalnim

upravljačkim ventilima pritiska odvodi kondenzat koji se nakuplja u cilindrima prema separatorima.

D. Radni mod sustava

Dolaskom papira do pogona za sušenje papira te pritiskom na tipku OPERATION prelazi se u radni mod (eng. *Operation*) gdje počinje proces sušenja papira. Prelaskom u ovaj mod počinje upravljanje sa zadanim radnim vrijednostima i faktorima koje su postavljene od strane procesnog inženjera putem nadzornog ekrana. Kako bi se izbjeglo ručno postavljanje moda operacije po dolasku papira predloženo je da se automatski osigura postavljanje sustava u mod operacije. Predložena ideja bila je da se osigura postavljanje foto ćelije na početku pogona za sušenje papira kako bi se po dolasku papira i prelasku preko foto ćelije dobio signal kojim bi se programski osiguralo automatsko postavljanje u radni mod. Ista logika vrijedi i za završno sušenje papira.

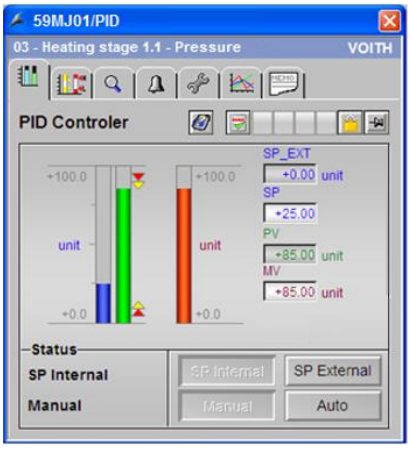
VII. IZRADA NADZORNOG PROGRAMA

Kako bi se omogućila komunikacija između operatera i procesa potrebno je izraditi nadzorni program za upravljanje procesom HMI (eng. *Human Machine Interface*) putem operaterskog računala (eng. *Operator Station*) ili panela na dodir (eng. *Touch Panel*). Nadzorni program za upravljanje procesom izrađuje se putem WinCC aplikacije (eng. *Windows Control Centre*) koja je sastavni dio Simatic programskog paketa. WinCC aplikacija služi za izradu nadzornih ekrana, procesnih alarma te komunikaciju između PLC uređaja i operativnog sustava koja je definirana tagovima. Sama aplikacija integrirana je u sučelju PCS7 te se projektima izrađenima u WinCC aplikaciji pristupa direktno preko PCS7 sučelja. Slike nadzornih ekrana izrađuju se u grafičkom dizajneru (eng. *Graphic Designer*) koji se nalazi u lijevom dijelu WinCC preglednika. U grafičkom dizajneru za sliku nadzornih ekrana dodjeljuju se elementi kojima se upravlja, procesne veličine (eng. *tags*) kojima se objekti povezuju sa PLC uređajem i alarmi za greške u procesu.

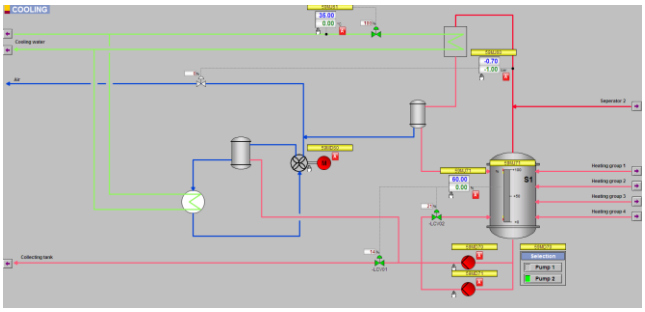
Elementi i alarmi koji se dodijeljuju na radnu površinu nalaze se u predlošku PCS7 Typicals iz kojeg se vrlo jednostavno dodjeljuju na radnu površinu nadzornog zaslona koji se izrađuje, po principu kopiraj i zalijepi (eng. *copy/paste*). Po završetku slike nadzornog ekrana elementi se povezuju sa procesnim veličinama kojima se potom prema potrebi dodjeljuju alarmi. Alarmi su podijeljeni po klasama prioriteta te se takvi pojavljuju na nadzornim ekranima u ovisnosti o greškama koje se javljaju iz procesa.

Klikom na pojedini element zaslona operateru se otvara pripadajući grafički prozor tog elementa. Slika 5. prikazuje jedan takav primjer grafičkog prozora elementa VP_PID koji služi za upravljanje i detaljan prikaz stanja PID regulatora.

Slika 6. okvirno prikazuje koncept nadzornog zaslona sustava za hlađenje pare i kondenzata sa svim pripadajućim grafičkim elementima. Zaslon omogućuje operateru upravljanje procesom u kojemu se hlađenjem pare plinovitog agregatnog stanja dobiva kondenzat u vakuum kondenzatoru.



Slika 5. Grafički prozor elementa VP_PID



Slika 6. Nadzorni zaslon sustav za hlađenje pare i kondenzata

VIII. EKSPLOATACIJSKE KARAKTERISTIKE

Kako različiti pritisci pare u cilindrima za sušenje uzrokuju različite temperature na površinama cilindra koje utječu na karakteristike sušenja papira, bitno je nadgledati temperaturu površine cilindara koja je definirana krivuljom zasićenja koja daje ovisnost temperature u odnosu na pritisak pare. Iz tog razloga potrebna je zasebna kontrola pritiska za svaku toplinsku skupinu.

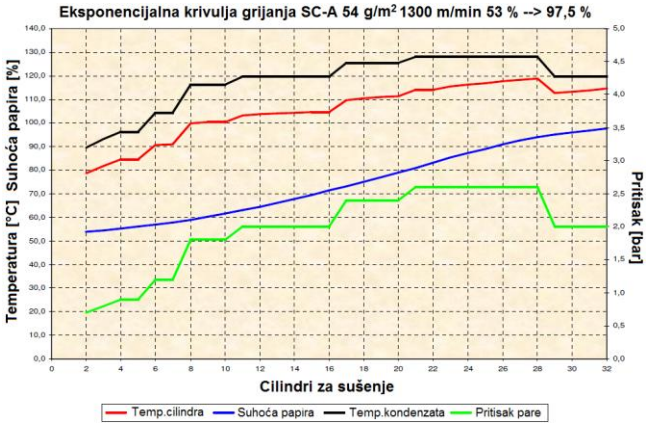
Odnos između pritiska pare, temperature kondenzata, temperature cilindara i završnog proizvoda procesa suhoće papira prikazan je eksponencijalnom krivuljom grijanja. Kako nije bilo moguće dobiti kompletnu dokumentaciju projekta, eksponecijalna krivulja iz koje se vide eksploatacijske karakteristike preuzeta je iz proizvodnje novinskog papira tvrtke SCA (Svenska Cellulosa Aktiebolaget) koja koristi isti sustav grijanja samo sa više cilindara koji su također podijeljeni po toplinskim skupinama.

Eksponecijalna krivulja prikazana je na sl. 7. te sadrži sljedeće eksploatacijske karakteristike:

- težina papira - 54 g/m^2
- brzina proizvodnje - 1300 m/min
- suhoća papira - 53 - 97,5 %
- temperatura cilindara - 79 - 119,5 °C
- temperatura kondenzata - 90 - 128 °C
- pritisak pare - 0,7 - 2,6 bara

Prikazana eksponecijalna karakteristika prikazuje sušenje papira težine 54 g/m^2 kroz cilindre za sušenje pri brzini proizvodnje od 1300 m/min. Plava krivulja suhoće papira prikazuje eksponecijalni rast od 53 % suhoće do

97,5 % prilikom prolaska papira kroz cilindre. Kako papir prolazi kroz cilindre i toplinske skupine tako se povećava pritisak pare (zelena krivulja), a time i temperatura cilindra (crvena krivulja) i konedenzata (crna krivulja). Na početku grafa se vidi kako su cilndri na početku u prvim toplinskim skupinama manje grijani te kako sadržaj papira primiče kraju, cilindri u zadnjim toplinskim skupinama se sve više griju što i rezultira suhoću papira.



Slika 7. Eksponencijalna krivulja grijanja

Ovakav princip sušenja papira se koristi da bi se dobila što bolja kvaliteta papira, izbjeglo oštećenje i pucanje papira te postepeno izvukla vlaga iz sadržaja papira. Iz tog razloga kao što je već spomenuto ranije svaka toplinska skupina ima zasebno upravljanje gdje svaki upravljački ventil ima zadanu vrijednost koja se povećava prelaskom i manje grijane toplinske skupine u veću.

IX. ZAKLJUČAK

Svakodnevna prisutnost papirnih proizvoda u današnjem svijetu označuje sastavni dio našeg kulturnog razvoja i vrlo bitan segment u modernom društvu. Današnja proizvodnja papira u svijetu iznosi otprilike oko 400 milijuna tona na godinu te predviđeni porast proizvodnje procjenjuje se na oko 500 milijuna tona do 2020. godine . Serija Hainan PM 2 tvrtke Voith je trenutno najveći papirni stroj na svijetu čija je širina 11,8 metara, a dugačak je 600 metara. Sadrži 87 cilindara za sušenje papira te proizvodi oko 4500 tona na dan .

Tehnologija koja se danas koristi kao sustav grijanja pogona za sušenje papira iznimno je bitan segment papirnog stroja. Sustav grijanja pogona za sušenje koji je prikazan u ovom radu osim minimalne vlažnosti papira od svega 2 do 5 % u konačnom proizvodu, omogućuje u potpunosti automatiziranu odvodnju kondenzata, smanjenu konzumaciju pare i smanjeni gubitak kondenzata. Nakon pokretanja sustava, sustav počinje sa automatskim režimom rada gdje je iz prakse poznato da operateri zaduženi za sustav grijanja imaju potrebu samo nadgledati proces grijanja za vrijeme proizvodnje.

Prikazan sustav grijanja je referenca za ostale sustave grijanja gdje će se sa velikom pouzdanošću u budućnosti primjenjivati u novim papirnim postrojenjima. Ovakav sustav grijanja ostavlja mjesta za nadogradnju u budućnosti gdje će se dolaskom novih tehnologija i istraživanjima unaprijediti postojeći sustav kako bi se smanjila vlažnost, potrošnja energije proizvodnje i troškovi.

LITERATURA

[1] Robert Bukvić, *Sustav grijanja pogona za sušenje papira*, diplomski rad, Tehničko veleučilište u Zagrebu, Politehnički specijalistički diplomski stručni studij, Specijalizacija elektrotehnika, Elektrotehnički odjel, Zagreb, 2014.

[2] Službena dokumentacija projekta Hoshangabad PM5, Voith

[3] H. Berger:Automating with STEP 7 in LAD and FBD, 3rdrevised edition, Publicis Corporate Publishing, Berlin and Munich, 2005.

[4] H. Berger:Automating with SIMATIC, 3rdrevised edition,Publicis Corporate Publishing, Berlin and Munich, 2006.

Uređaj za testiranje baterijskih ćelija 18650

Alen Jakoplić, Danijel Maršić i Goran Malčić
Tehničko veleučilište u Zagrebu/Elektrotehnički odjel, Zagreb, Hrvatska
goran.malcic@tvz.hr

Sažetak – Prilikom izrade baterijski napajanih elektroničkih sklopova, uvjek se postavlja pitanje vremenskog ograničenja rada tako izvedenog sklopa. Kako bi dobili točne vrijednosti, za pojedine baterijske ćelije potrebno je izvršiti njihovo testiranje odnosno kontrolirano prazniti i mjeriti njihov napon i struju. Ako se želi saznati kapacitet ćelije potrebno je istu prvo napuniti, te onda prazniti određenom strujom i mjeriti vrijeme, a dobivene rezultate treba na odgovarajući način prikazati korisniku. U radu je prikazan razvoj prototipa uređaja za testiranje baterijskih ćelija pomoću Arduino sustava. Opisane su sve sastavne komponente uređaja i objašnjen je rad sklopa kao cjeline. Testiranje se provodi na baterijskim ćelijama normiranih brojem 18650 koje su danas jedne od najkorištenijih.

I. UVOD

Arduino je *open-source* platforma bazirana na sklopovlju i programskom paketu koji je fleksibilan i jednostavan za korištenje. Može se koristiti za razvijanje interaktivnih objekata očitanjem stanja raznih prekidača i senzora, te upravljanje raznim indikatorima, aktuatorima i ostalim fizičkim izlazima [1].

Li-Ion baterija pripada skupini punjivih baterija. Prednost ovakvih vrsta baterija je vrlo velika gustoća energije koja se može pohraniti u njima, triput veći nazivni napon od napona koji je u baterijama na bazi nikla (3,7 V prema 1,2 V), nema potrebe za periodičnim pražnjenjem odnosno može ih se puniti u bilo kojem stanju napunjenosti i dugi životni vijek koji može biti od 500 do 1000 ciklusa.

Slabost litij-ionskih baterija je što nisu pogodne za pražnjenje jakom strujom, mogu stariti i kad ih se ne koristi te se u određenim okolnostima mogu pretjerano zagrijati pa čak i eksplodirati. Vrlo su osjetljive na prepunjavanje kao i na pretjerano pražnjenje, zbog čega svaka komercijalna baterija ima elektroniku koja rješava takav problem [2].

Kada baterija u sebi sadrži više baterijskih ćelija spojenih u seriju kako bi se povećao izlazni napon potrebno je koristiti sve ćelije istoga kapaciteta kako bi se one istovremeno ispraznile i napunile. Ako jedna ćelija ima manji kapacitet od ostalih onda će se prije isprazniti i tako doći u opasno niske napone ili prije napuniti, te čak stvarati rizik od požara. Kako bi se to izbjeglo baterije se balansiraju ili nadziru. Nadziranje svake pojedine ćelije omogućuje prekid punjenja ili pražnjenja kada jedna od baterija dođe do kritičnog napona.

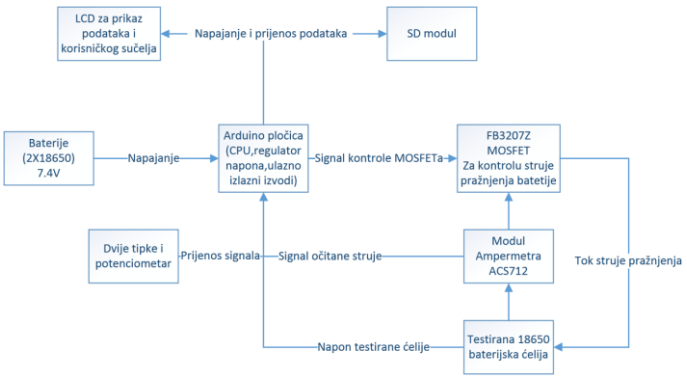
Rješenje tog problema je balansiranje. Postoji gornje i donje balansiranje. Gornje balansiranje je u širokoj uporabi radi veće pouzdanosti. Radi na principu da se na baterijske ćelije koje prve dođu do napona na kojem su pune spoji mali teret koji ih održava na tome naponu. Gornje balansirane baterije prazne se dok god jedna od ćelija ne dođe do napona na kojem se smatra praznom i onda pražnjenje mora prestati kako se ne bi oštetila ta ćelija. Ostale ćelije neće se isprazniti do kraja.

Donje balansiranje izvodi se samo jednom i nije potreba nikakva dodatna elektronika za nadzor baterija. Radi na principu da se sve ćelije isprazne na napon na kojemu se smatraju praznima, te onda sve zajedno pune i koriste. Pošto su spojene u seriju, kroz sve ćelije prolazi ista struja pa se sve pune na isti kapacitet u mAh. Pune se sve dok jedna od ćelija ne dođe do napona na kojemu se smatra punim.

Trenutno je najprodavanija baterijska ćelija oznake 18650 te joj cijena iz tog razloga neprestano pada. Prodaja elektroničkih sklopova s baterijskim napajanjem je također u porastu jer je tržište upravo preplavljeno takvim sklopovima (eng. gadgets). U ovom radu je opisan uređaj koji kontolira stvarne vrijednosti baterijskih ćelija 18650 koje nalazimo na tržištu.

II. IDEJNO RJEŠENJE UREĐAJA

Osnovni cilj je razviti uređaj za testiranje baterijskih ćelija oznake 18650 koji bi radio kao samostalan uređaj koji ne mora biti povezan s računalom ili napajanjem iz električne mreže. Uređaj za testiranje baterijskih ćelija 18650 potreban je kako bi se izmjerile energetske mogućnosti baterijskih ćelija odnosno utvrdio kapacitet ćelije, te padovi napona pod pojedinim opterećenjima. Potrebno je kontrolirano prazniti baterijske ćelije mjerenjem struje i napona ćelije te vremena koje je potrebno da se ona isprazni, te na kraju omogućiti prikaz i spremanje izmjerenih podataka. Centralni dio uređaja je ATmega mikrokontroler smješten na Arduino pločici (Sl.1.).



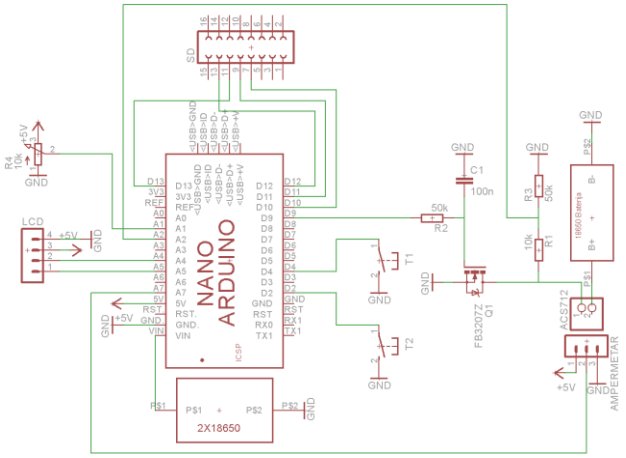
Slika 1. Idejna blok shema uređaja

Baterijske ćelije 18650 testiraju se mjerenjem njihovog kapaciteta u mAh ili Wh, mjerenjem maksimalne struje koju baterija može dati i mjerenjem pada napona na bateriji prilikom određenog opterećenja. Kada se sastavlja baterija načinjena od više ćelija potrebno je znati kapacitet svake pojedine ćelije kako bi se dobro rasporedila spremljena energija, te energetske mogućnosti ćelija kako bi se osiguralo da su prikladne za davanje željenog napona, struje ili snage. Nove ćelije obično dolaze s popisom parametara iz tvornice pa nije potrebno ponovno mjerenje, ali kod reciklaže starih baterija ili kod baterija s nepoznatim parametrima, novo mjerenje je iznimno bitno. Moguća je i provjera performansi baterija nakon nekog vremena korištenja (provjera pada kapaciteta).

Uređaj za testiranje 18650 ćelija radi na principu kontroliranoga pražnjenja baterija pomoću MOSFET (eng. *metal – oxide – semiconductor field effect transistor*) tranzistora na kojeg je montiran hladnjak kako bi se disipirala energija prethodno pohranjena u baterijama. Bateriju prvo treba napuniti na propisani napon kako bi se dobilo što točnije očitanje. Prije početka mjerenja potrebno je unijeti željenu struju pražnjenja. Struja pražnjena se određuje ovisno o budućoj uporabi baterije.

III. SKLOPOVSKO RJEŠENJE

Centralni dio cijeloga uređaja je *Arduino Nano* kompatibilna pločica na kojoj se nalazi ATmega mikrokontroler u kojem se izvodi program koji upravlja cijelim uređajem. Mikrokontroler ima mogućnost analogno digitalne i digitalno analogne pretvorbe koja je potreba kako bi uređaj radio. Na Arduino nano pločici nalazi se još utor za USB vezu koji se koristi za programiranje i prijenos podatke te čip koji omogućuje komunikaciju sa ATmega mikrokontrolerom putem USB sučelja.



Slika 2. Električna shema uređaja

A. *Kontrola izlazne struje baterije*

Izlazna struja baterije kontrolira se MOSFET tranzistorom tipa FB3207Z. S obzirom da je MOSFET naponom upravljani tranzistor kojemu vodljivost ovisi o naponu između nožica G i S u ovom se slučaju taj napon regulira mikrokontrolerom kako bi se MOSFET doveo u aktivno područje rada.

ATmega mikrokontroler analogni izlaz na svojim nožicama realizira signalom pulsno širinske modulacije (eng. *Pulse Width Modulation, PWM*). Kako bi se izlazni PWM signal mikrokontrolera mogao koristiti za upravljanje MOSFET tranzistorom potrebno ga je prvo izgladiti. To je realizirano nisko propusnim filtrom prvoga reda kojeg je potrebno podesiti na što nižu frekvenciju kako bi što bolje filtrirao izmjeničnu komponentu signala.

B. *Mjerenje struje baterije*

Kako bi mikrokontroler mogao aktivno regulirati struju baterije potrebno je mjeriti samu struju baterije. Mjerenje struje baterije izvedeno je modulom za mjerenje struje pomoću Hallovog efekta. Hallova sonda osjeti magnetsko polje čija jačina ovisi o struji koja teče kroz vodič. Ovaj modul za mjerenje struje koristi ACS712 ELCTR-20A čip koji je pogodan za rad s mikrokontrolerom. Maksimalni doseg čitanja struje je 20 A, te daje signal od 100 mV/A.

C. *Mjerenje napona baterije*

Mjerenje napona baterije izvedeno je uporabom samog ATmega mikrokontrolera. Analogni ulazi mikrokontrolera rade tako da mjere napon od 0 do 5V referenciran masom i napajanjem mikrokontrolera. Napon baterije mjeri se tako da se njena negativna strana spoji na masu, a pozitivna na analogni ulaz mikrokontrolera.

D. *Disipiranje snage baterije*

Baterija se tijekom mjerenja kapaciteta mora prazniti zbog čega je potrebno ugraditi hladnjak jer MOSFET tranzostor ne može disipirati dovoljno energije baterije pretvorene u toplinsku energiju. Kako bi se mogle testirati relativno velike struje i snage ugrađen je aktivni hladnjak koji pomoću ventilatora forsira strujanje zraka kroz sam hladnjak te time povećava njegovu disipaciju. Ventilator hladnjaka može se programski isključiti prilikom mjerenja manjih snaga kako mi se uštedjelo na energiji baterija uređaja.

E. *Korisničko sučelje i prikaz podataka*

Navigacija kroz korisničko sučelje realizirana je pomoću potenciometra i dvije tipke (jedna za odabir, a druga za povratak). Položajem potenciometra odabire se što prikazati na ekranu prilikom mjerenje, te se njime unose različiti parametri (mjerenja struja, napon i sl.). Prikaz korisničkog sučelja i podataka realiziran je pomoću 16x2 LCD modula koji ima LED pozadinsko osvjetljenje. LCD modul sa ATmega komunicira pomoću modula za I2C komunikaciju pa je broj nožica potreban za rad modula smanjen sa 8 na 4.

F. *Spremanje podataka testiranja*

Spremanje podataka za detaljniju analizu realizirano je uporabom SD Arduino modula.(eng. *Secure Digital*) koji omogućuje jednostavno povezivanje Arduino sustava sa SD karticom. Podaci testiranja spremanju se u *.csv* tip datoteke na SD karticu koje se nakon testiranja može spojiti sa računalom. Na taj način mogu se vidjeti podaci testiranja i generirati grafovi koji sadrže iznose mjerene struje i napona u vremenu te ostale podatke.

G. *Napajanje uređaja*

Napajanje uređaja realizirano je pomoću dvije baterije tipa 18650. Pritiskom na sklopku za uključivanje sustava one su spojene na 5 V regulator napona koji se nalazi na Arduino pločici te na ventilator hladnjaka ako je spojen. Regulator dalje napaja ostale komponente. Potrošnja uređaja je oko 50 mA bez ventilatora i 150 mA sa ventilatorom.

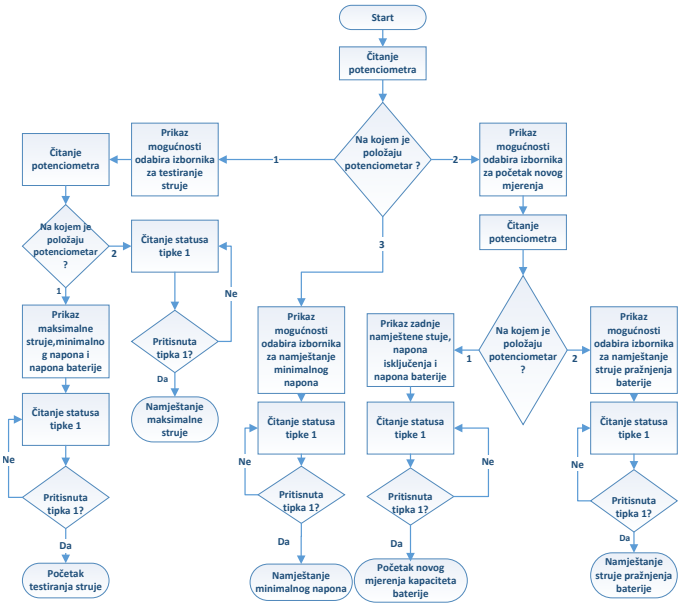
Dvije baterije 18650 koje napajaju uređaj spojene su u seriju tako da daju nazivni napon od 7,4V te imaju kapacitet od 2000 mAh. Uređaj je moguće napajati i preko 5V USB konektora koji se nalazi na Arduino pločici. Pritom treba pripaziti da napon bude točno 5V kako bi senzori mogli ispravno raditi.

IV. PROGRAMSKO RJEŠENJE

Programsko rješenje uređaja je izrađeno u Arduino aplikaciji. U ovom dijelu su ukratko opisane sve osnovne funkcije uređaja izvedene u programskom rješenju.

A. *Mjerenje kapaciteta baterije*

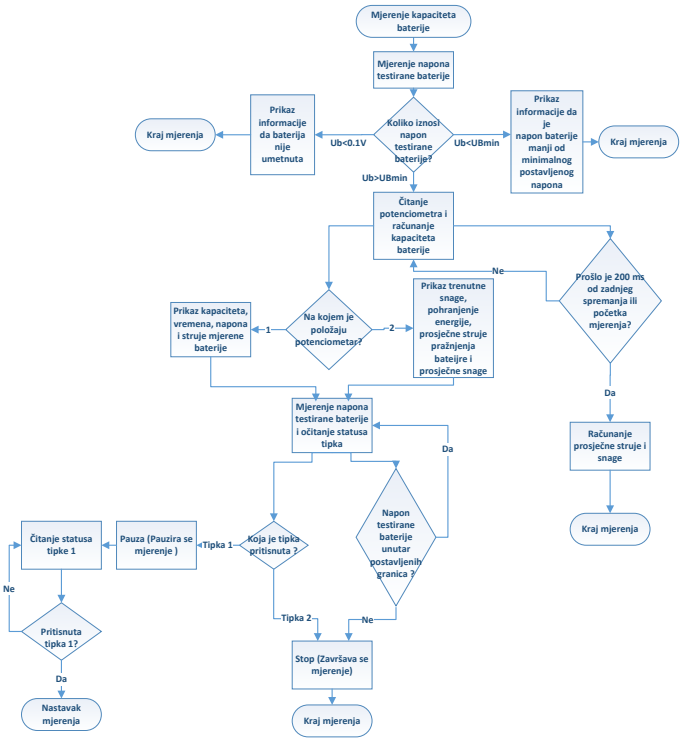
Prije početka mjerenja kapaciteta baterije potrebno je po potrebi namjestiti osnovne parametre kao što su struja kojom se baterija prazni odnosno napon do kojega se prazni. Pritiskom na tipku 1 u prikazu parametara postavljenih prvo se provjerava napon testirane baterije, te ako taj napon nije unutar postavljenih granica, mjerenje se odmah završava (Sl. 3.).



Slika 3. Dijagram toka potprograma korisničkog sučelja

Ako je napon unutar postavljenih granica tada počinje mjerenje te se svakih 200 milisekundi računa prosječna struja i snaga (Sl. 4.). Taj se iznos koristi za proračun kapaciteta baterije u mAh.

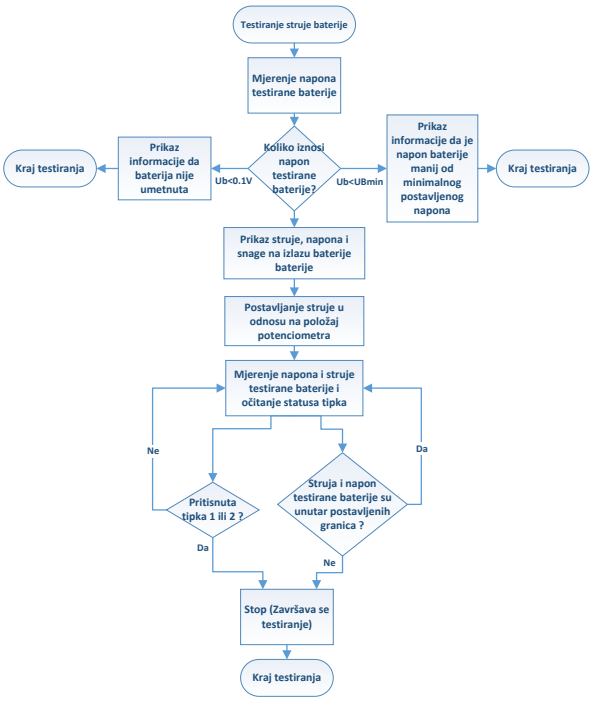
Mjerenje završava kada napon padne na postavljeni minimalni napon baterije. Svi izmjereni parametri se tijekom mjerenja spremaju na SD karticu.



Slika 4. Dijagram toka potprograma za mjerenje kapaciteta baterije

B. *Testiranje struja baterije*

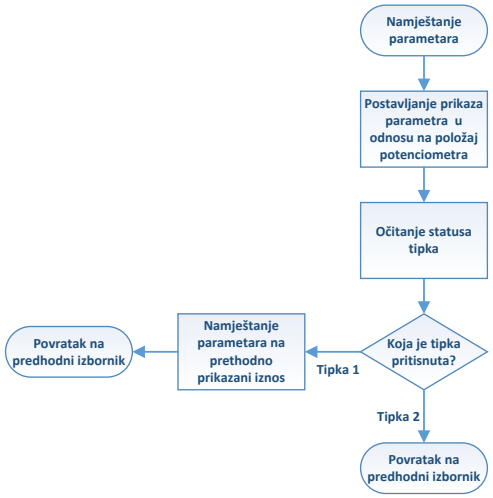
Prije samog početka testiranja struje baterije potrebno je po potrebi namjestiti maksimalnu struju pražnjenja i minimalni napon baterije za što je zadužen potprogram prikazan na Slici 3. Nakon toga se očitavaju napon i struja baterije te se ovisno o stanju baterije struja postavlja na željeni iznos. Ta se petlja vrti sve dok ručno ne prekinemo testiranje pritiskom na bilo koju tipku, dok napon ne padne ispod dozvoljenoga ili struja prijeđe postavljenu graničnu vrijednost (Sl. 5.).



Slika 5. Dijagram toka potprograma za testiranje struje baterije

C. *Namještanje i spremanje parametara*

Ako se neki parametar želi namjestiti korištenjem korisničkog sučelja prvo je potrebno doći do njegovoga prikaza u sučelju te pritisnuti tipku 1. Nakon ulaska u izbornik namještanja parametara čita se položaj potenciometra te u odnosu na njegov položaj ispisuje se iznos parametra na ekranu. Ako se pritisne tipka 1 parametar se postavlja na prikazanu vrijednost i prikaz se vraća na prethodni izbornik (Sl. 6). Ako se taj parametar sprema u EEPROM memoriju prvo ga je potrebno preračunati u vrijednost od 0 do 254 korištenjem funkcije "map()". Parametar se preračunava u vrijednost od 0 do 254 zato što je u EEPROM memoriju moguće spremiti samo broj unutar tih granica. Unutar EEPROM memorije moguće je spremiti i broj 255, ali on se koristi za provjeru je li išta spremljeno u memoriju. Ako se pritisne tipka 2, parametar ostaje na svojoj prethodnoj vrijednosti te se bez promjene parametara izvršava povratak na prethodni izbornik.



Slika 6. Dijagram toka potprograma za namještanje parametara

Prilikom svakog pokretanja očitava se EEPROM memorija. Ako u određenim mjestima u memoriji nije spremljen nikakav podatak tj. očitana vrijednost iznosi 255, taj podatak postavlja se na unaprijed određenu vrijednost (maksimalna struja baterije na 2,5 A i minimalni napon na 3 V). Ako je u očitanim mjestima u memoriji spremljen neki podatak isti se preračunava u spremljenu vrijednost.

D. *Mjerenje vremena*

Za mjerenje vremena u ovom rješenju korišten je sat (eng. *clock*) ATmega mikrokontrolera koji mjeri vrijeme otkada je mikrokontroler uključen. Kako bi se očitalo to vrijeme koristi se naredba "millis()". U programu nije ključno vrijeme proteklo od trenutka uključenja mikrokontrolera nego vrijeme proteklo od trenutka početka mjerenja. Prilikom početka mjerenja varijabla "previousMillis" izjednačuje se s trenutnim vremenom te se nadalje ona uspoređuje s tim trenutnim vremenom mikrokontrolera. Kada je njihova razlika jednaka odabranom intervalu (primjerice 1000 milisekundi kako bi se mjerila sekunda) izvršava se određena funkcija (kod mjerenja vremena sekunda se poveća za 1). Ova metoda mjerenja vremena ne koristi se samo za mjerenje vremena mjerenja kapaciteta baterije nego i za neke druge funkcije. Kada se prikazuju parametri mjerene baterije oni se ne prikazuju u realnome vremenu nego svakih 200 milisekundi koristeći ovu metodu mjerenja vremena. Pokazuju se svakih 200 milisekundi kako bi se izbjegla mutna slika radi prebrzih promjena brojeva.

V. EKSPLOATACIJSKE KARAKTERISTIKE UREĐAJA

A. *Testiranje struja baterije*

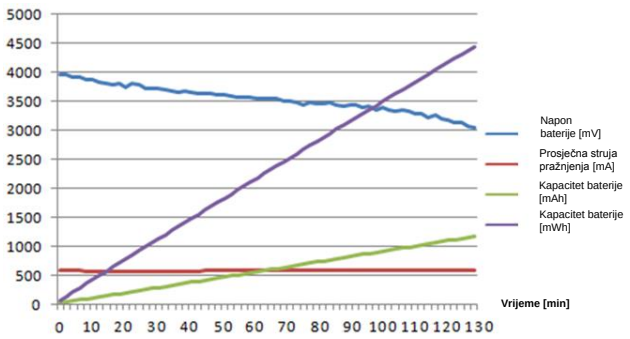
Uređajem se mogu testirati maksimalne izlazne struje 18650 baterijske ćelije te pad napona na određenoj struji. Potrebno je podesiti minimalni napon do kojega se želi opterećivati baterijska ćelija te maksimalna struja testiranja. Ovi parametri namještaju se kako se nehotice ne bi oštetila ćelija. Kada se pokrene mod rada "Test struje" potenciometrom je moguće regulirati izlaznu struju ćelije, a na ekranu mogu se pratiti njen iznos, napon i izlazna snaga baterijske ćelije. Usporedbom napona prije opterećenja određenom strujom i tijekom opterećenja određenom strujom može se ustvrditi pad napona na ćeliji uslijed opterećenja tom strujom.

B. *Donje balansiranje*

Ako se baterija želi donje balansirati potrebno je napon svih njenih ćelija spustiti na napon na kojemu se one smatranju praznima. Pomoću "Testa struje" ćelija moguće je kontrolirano prazniti baterijsku ćeliju. Ako se donja granica napona baterije postavi na napon na koji se namjerava balansirati baterija tada mod "Test struje" omogućuje spuštanje napona ćelije na taj napon budući da se struja automatski postavlja na nulu kada se dostigne taj napon. Kada baterija dostigne donji napon potrebno je malo pričekati da se napon na bateriji stabilizira, što obično znači da se i malo digne, te onda ponovno pustiti struju da se napon spusti. Taj postupak potrebno je ponavljati dok god se napon ne stabilizira na željenom naponu. Ako se struja postavi na veću od nule taj će se postupak automatski odvijati.

C. *Rezultati testiranja*

Na Sl. 7. je moguće vidjeti rezultate testiranja jedne baterijske ćelije oznake 18650 provedenog ovim uređajem. Korištena je mogućnost spremanja parametara ispitivanja na SD karticu kako bi se nakon testiranja generirao graf za pregled ispitivanja i očitanje rezultata. Uređaj je spremio podatak da je početni napon testiranje baterije bio 4176 mV. Dakle baterija je bila puna prije početka ispitivanja kako bi to i trebalo biti da bi se dobili što točniji rezultati.



Slika 7. Rezultati testiranja baterijske ćelije oznake 18650

Moguće je uočiti da je napon baterije u prve dvije minute pao na 4,0 V. Nakon toga napon je relativno konstantno padao do 120. minute nakon čega je počeo brže padati. Na 3,0 V prekinuto je mjerenje kao što je bilo podešeno u postavkama uređaju. Pritom je prosječna struja bila oko 550 mA, te da je kapacitet u mAh i mWh linearno rastao s vremenom kao što se i očekuje.

VI. ZAKLJUČAK

U radu je prikazan razvoj uređaja koji uspješno testira kapacitet i performanse baterijskih ćelija oznake 18650. Ćelije je moguće testirati pod raznim opterećenjima, na ekranu nadzirati parametre koji se testiraju, te nakon testiranja generirati graf kako bi se vidjelo ponašanje ćelije u vremenu.

Kontrola struje pražnjenja baterije izvedena je njenim mjerenjem i kontroliranim povećanjem vođenja MOSFET tranzistora dok ona ne dođe na željeni iznos. Cijelo to vrijeme mjeri se napon testirane ćelije kako on ne bi pao ispod minimalnog napona ćelije, te kako bi se vidjelo ponašanje napona u odnosu na struju.

Prilikom izrade uređaja do problema je došlo je kada je ponestalo memorije na ATmega mikrokontroleru. Program je trebalo što efikasnije napisati i broj funkcija se trebao smanjiti samo na osnovne potrebne za ispravan i koristan rad uređaja kako bi program stao na memoriju mikroprocesora.

Nedostaci uređaja su spremanje više različitih parametara u memoriju zbog nedostatka iste i mogućnost kalibriranja uređaja pomoću korisničkog sučelja. Nedostaju i neke funkcije uređaja kao što je testiranje kapaciteta baterije pražnjenjem konstantnom snagom. Rezolucija spremanja podataka na SD karticu je samo dvije minute radi nedostatka mogućnosti generiranja grafa s više parametara u Excelu zbog čega bi bilo praktičnije koristiti MATLAB. Uređaj je ograničen na testiranje samo baterijskih ćelija oznake 18650 jer samo one fizički stanu u uređaj, a i sve funkcije napravljene su s njima u vidu pa se tako ne mogu testirati struje pražnjenja veće od 15A i baterije čiji naponi ne iznose oko 4V.

Rezultati mjerena pokazuju ponašanje napona ćelije u vremenu tijekom konstantnoga opterećenja i na kraju mjerenja njen kapacitet u mAh i mWh. Uvidom u te podatke može se ustvrditi kako i za što koristiti tu ćeliju. Testiranjem struje vidi se pad napona ćelije pod određenim opterećenjem i je li ta ćelija pogodna za određeno opterećenje.

Sklopovska i programska rješenja iz ovoga rada mogu se koristiti za daljnji razvoj toga uređaja ili pak za promjenu primjene istoga. Može se dodati mikrokontroler s više memorije kako bi se dodale funkcije. Moguće je i dodavanje funkcije istovremenog testiranja većeg broja ćelija, te punjenje istih prije mjerenja kako bi očitanje bilo što točnije. Potrebno je korištenje nekih drugih aplikacija za generiranje grafova što bi omogućilo stvaranje grafova veće rezolucije radi detaljnije analize ćelije.

LITERATURA

[1] Alen Jakoplić, *Uređaj za testiranje 18650 baterijskih ćelija*, završni rad, Tehničko veleučilište u Zagrebu, Stručni studij elektrotehnike, Elektrotehnički odjel, Zagreb, 2015.

[2] http://batteryuniversity.com/learn/

[3] www.arduino.cc

Layout Optimizations to Decrease Internal Power and Area in Digital CMOS Standard Cells

J. Innocenti[1, 2, 3], F. Julien[1], J.M. Portal[2], L. Lopez[1], Q. Hubert[2], P. Masson[3], J. Sonzogni[1], S. Niel[1] and A. Regnier[1]
[1]STMicroelectronics Rousset, Rousset, France
[2]IM2NP-UMR CNRS 7334 / Aix-Marseille University, Marseille, France
[3]EpOC / Nice Sophia-Antipolis University, Biot, France
jordan.innocenti@st.com

Abstract - This paper presents several layout optimizations in order to decrease both, internal power and area of digital standard cells. Since D flip flop (Dff) represents a large part of digital area, post-layout simulations have been performed on this cell designed with advanced design rules and lower widths. Results show that internal power of Dff is reduced by 20 % during rising and falling transitions of output Q. Moreover, the area of Dff is reduced by 20 % for similar clock to Q propagation delays. Indeed, a new process has been developed and I_{ON} current is improved by 15 % and 50 % for respectively NMOS and PMOS transistors which permits to reduce the width of MOSFETs.

Keywords – Low power, standard cells; D flip flop, design rules, drive current enhancement, strain silicon

I. INTRODUCTION

Nowadays, many applications based on microcontrollers require more processing speed and less dynamic and static consumption to increase battery life-time of mobile applications such as tablets, smartphones, and laptops. The most effective way to decrease consumption of digital CMOS Integrated Circuits (ICs) is to reduce the supply voltage V_{DD} and all capacitances contributors such as routing and MOSFETs intrinsic capacitances [1], [2]. Moreover, in order to reduce the cost by microcontroller, marketers and designers are very interested in decreasing the surface area filled by the digital logic of the circuit to improve their density of integration. The aim of this paper is to study the possibility to reduce both, internal power and area of standard cells by remaining on same technological node.

This work focuses on low voltage MOSFETs devices used in standard cells and manufactured in conventional e-NVM (embedded non-Volatile Memory) CMOS 80 nm process with compressive Shallow Trench Isolation (STI) oxide, gate oxide thickness of 2.1 nm, dual doped polysilicon gate, pocket implants, and compressive Contact Etch Stop Layer (CESL).

This paper is presented as fallow: section 2 will present several layout improvements which could be used in standard cells and section 3 will show internal power and area gains obtained on D flip flop (Dff) designed with new optimized design rules. Finally, last section will give some concluding remarks.

II. LAYOUT OPTIMIZATION STUDY

A. Hammer Head Removal

In order to decrease internal power and area in standard cells, the minimum width (W_{MIN}) of transistor have to be used as far as possible, especially when timing is not predominant. However, with conventional 80 nm design rules, designers have to insert an active Hammer Head (HH) in source and drain regions (S/D). This solution increases the length of the transistor by 28 % due to the distance gate to active, as shown in Fig. 1. Hense, minimal width is not currently used in standard cells as demonstrated on classical MCU digital circuit in Fig 2. 24 % of PMOS and 35 % of NMOS transistors are designed with a higher dimension in order to avoid an increase of the standard cells area with HH layout.

In this part, we are studying the possibility to remove HH while maintaining a high level of process robustness. For that, chain resistance of more than 12 000 contacts in series is measured on several devices designed with different active widths (W_{ACT}), as presented in Fig. 3.

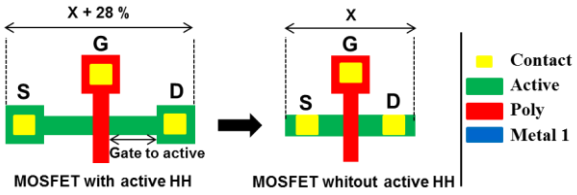


Fig. 1: Layout of MOSFETs with or without active hammer head

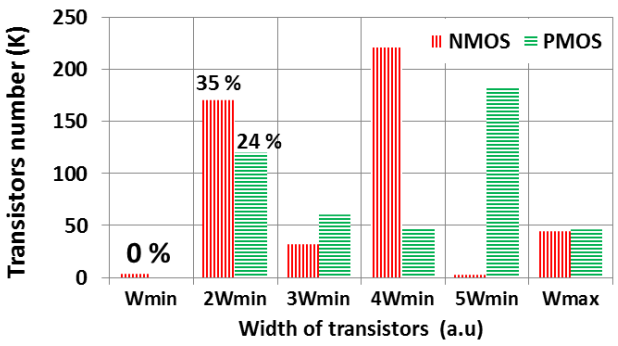


Fig. 2: MOSFETs distribution of classical digital circuit

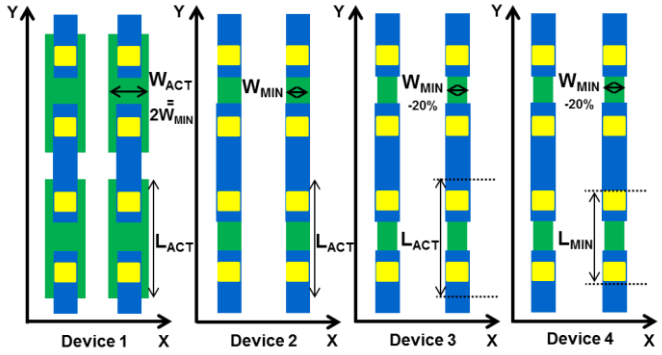


Fig. 3: Layouts of different contact chain resistances

In device 1, no HH is necessary to contact S/D regions because $W_{ACT} = 2W_{MIN.}$ However, HH has to be inserted in devices 2, 3 and 4 with lower $W_{ACT} = W_{MIN-20\%}$. Device 4 will be the most aggressive layout without contact enclosure in W and L directions. As depicted in Fig. 4, the contact resistance increases when W_{ACT} is reduced due to the increase of the active finger resistance. No effect of distance L_{ACT} is observed between device 3 and device 4 without any contact enclosure. As a consequence, the contact resistance is stable without outliers regardless of W_{ACT}. Moreover, a similar study was done with overlay matrix trials in W and L directions in order to prove the manufacturability.

In order to estimate the intrinsic change induced by the modifications, threshold voltage (V_T), saturation (I_{DSAT}), linear (I_{DLIN}) and drain leakage (I_{OFF}) currents have been measured. Poly HH is also removed (Fig. 5). I_{DSAT} currents of all devices are aligned, no change is observed between NMOS and PMOS with or without HH as shown in Fig. 6. I_{OFF} current is also plotted in Fig. 7 and remains unchanged.

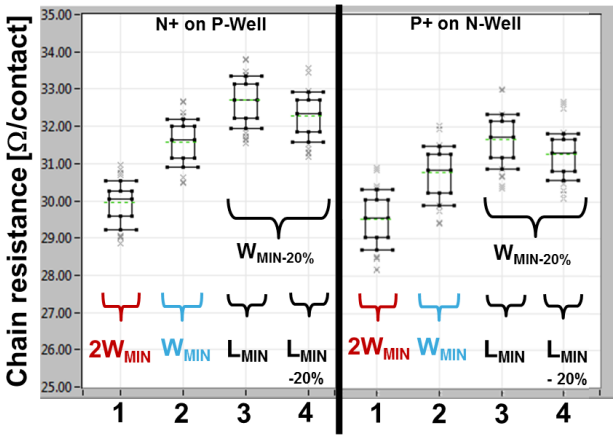


Fig. 4: Contact resistance of devices on N-well and P- well actives

Fig. 5: Layout of devices with or without active and poly HH

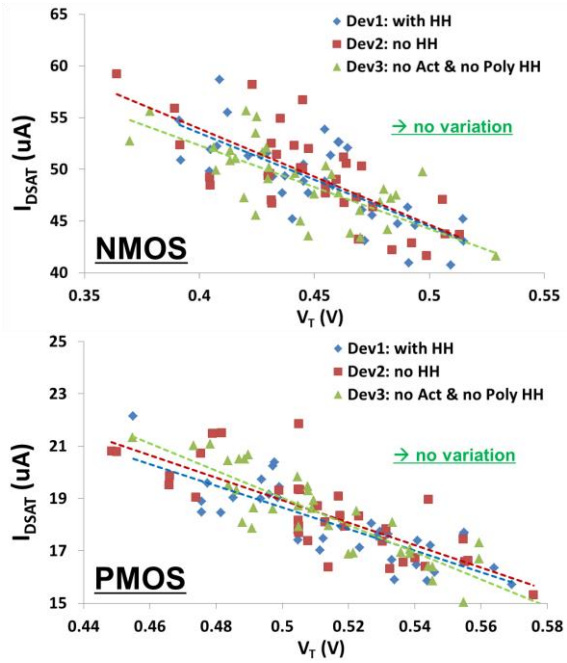


Fig. 6: $I_{DSAT}(V_T)$ for different layouts

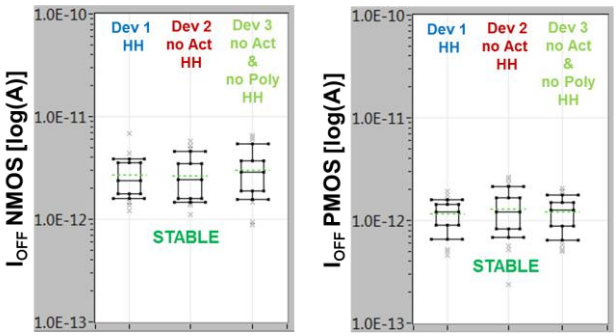


Fig. 7: Leakage currents (I_{OFF}) of NMOS (left) and PMOS (right)

Then, CMOS inverters Ring Oscillators (ROs) [4] are used to confirm the dynamic behavior of different layouts presented in Fig. 8. The clock frequency (F_{RO}) detailed in (1) is measured and compared between reference RO1 with HH and others ROs without HH. Where n is the number of CMOS inverters, I_{DYN} the dynamic current, V_{DD} the supply voltage and T_D the period of oscillation.

$$F_{RO} = \frac{1}{2.n.T_D} = \frac{I_{DYN}}{2.n.C_{LOAD}.V_{DD}} \qquad (1)$$

As observed in Fig. 9, ROs frequency is 5 to 7 % higher thanks to the reduction of S/D diffusion capacitances. Indeed, the load capacitance (C_{LOAD}) is essentially represented by the intrinsic MOSFETs capacitance (gate and diffusion). A slight reduction of these capacitances induces a decrease of C_{LOAD} and then an increase of the frequency. Finally, frequency is slightly decreased between RO2 and RO3 due to the increase of gate to contact capacitance. In conclusion, robustness and manufacturability of HH removal have been demonstrated. It confirms the ability to use minimal width in digital standard cells.

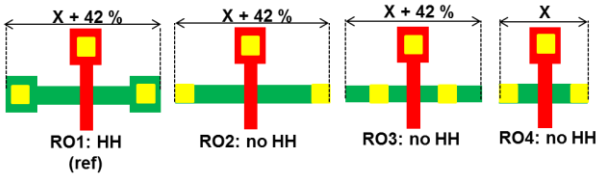


Fig. 8: MOSFETs layout used in ring oscillators

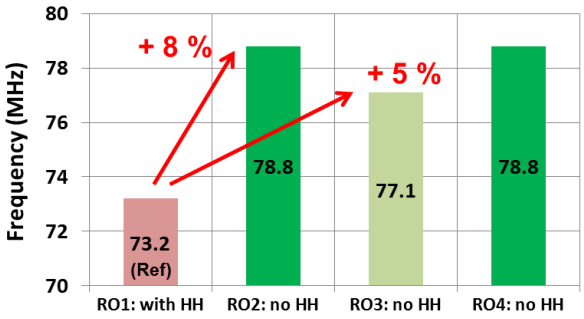


Fig. 9: Frequency gains regarding RO1

B. *Move Gate Contact over Active Area*

As previous analyse proven the ability to remove poly HH, gate contact could be directly placed over active area instead of over STI. To our knowledge, this innovative design approach has never been reported and manufactured yet. The reliability of this optimization has been demonstrated in a previous work and the main parameters of MOSFETs remain unchanged [5].

The layout with gate contact over active is presented in Fig. 10. It allows reducing significantly the height of the transistor and could be interesting in specific standard cells. I_{DSAT} versus V_T is plotted with gate contact over active and STI for typical W. no difference is observed for PMOS transistor in Fig. 11 while an insignificant reduction is observed for NMOS (- 1.5 %) in Fig. 12. It could be due to the tensile strain added on the channel during tungsten deposition. Indeed, it has been demonstrated that tensile strain in "out of plane" direction of the channel decreases the carriers mobility of NMOS and enhances the carriers mobility of PMOS [6], [7]. It means that compressive stress in the NMOS channel is more important when the gate contact is filled over active area instead of over STI. I_{OFF} current is also plotted in Fig. 13 and remains unchanged.

RO frequency has been measured and plotted in Fig. 14. All tested dies (30) on wafer are functional. As expected, the speed is slightly decreased by 1.4 % for $W = 2W_{MIN}$ and by 3.4 % for $W = W_{MIN}$, probably due to the NMOS mobility reduction.

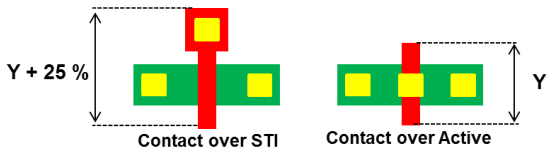


Fig. 10: MOSFETs layout with gate contact over STI and over Active

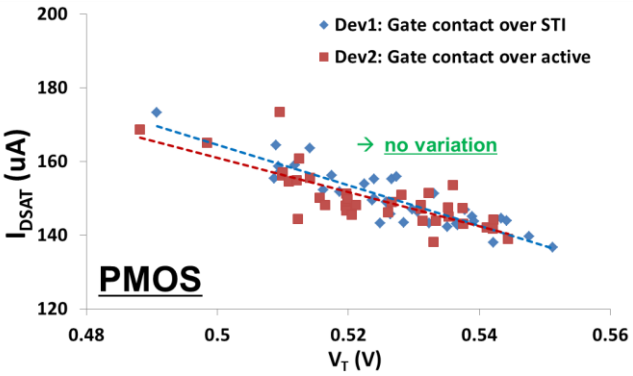


Fig. 11: $I_{DSAT}(V_T)$ for PMOS with or without contact over active

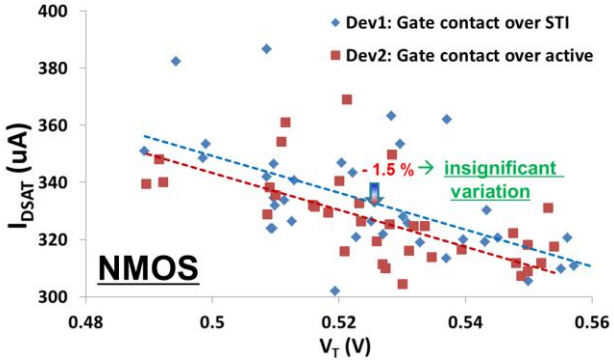


Fig. 12: $I_{DSAT}(V_T)$ for NMOS with or without contact over active

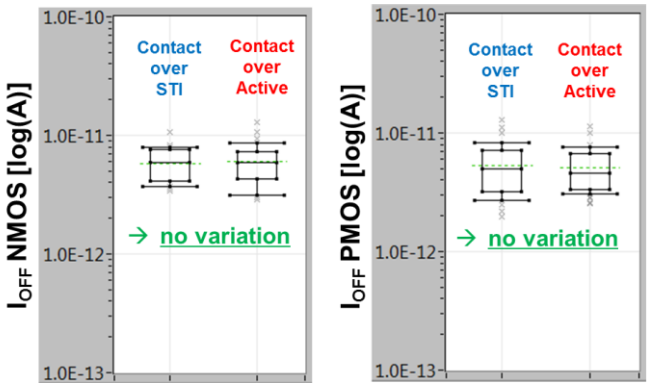


Fig. 13: Leakage currents (I_{OFF}) of MOSFETs

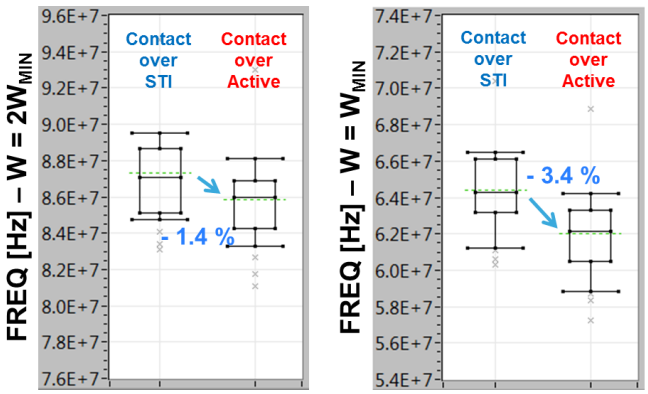


Fig. 14: ROs frequency with or without gate contact over active area

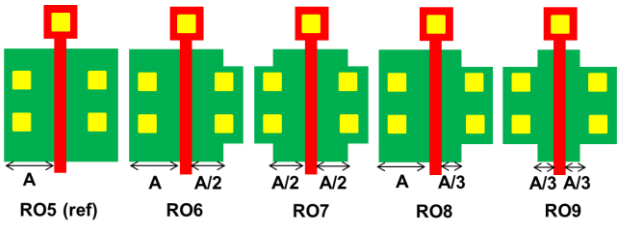


Fig. 15: MOSFETs layouts inside ROs for S/D capacitances reduction

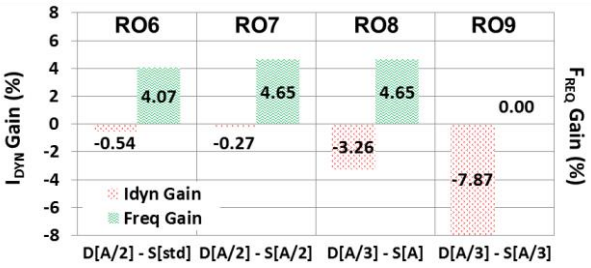


Fig. 16: Dynamic current and frequency gains regarding RO5

C. *Source/Drain Capacitances Reduction*

In order to decrease dynamic current, all capacitances contributors need to be scaled. When C_{LOAD} is mainly composed by intrinsic MOSFETs capacitance, a slight reduction of S/D capacitances allows decreasing dynamic current without impacting RO frequency. Therefore, different layouts are proposed in Fig. 15. The other benefit of this technique is to take advantage of mechanical stress coming from the STI to enhance the mobility of PMOS transistors [7]–[9]. Nevertheless, it will decrease the mobility of NMOS. As presented in Fig. 16, RO frequency is increased thanks to the reduction of diffusion capacitances. By moving active S/D distance from A to A/3, RO5 frequency is equivalent to RO1 while the dynamic current consumed is reduced by 8 % as show Fig. 16.

III. Internal Power and Area Gain

In a previous work, a new process has been developed in order to improve drive current (I_{ON}) of MOSFETs. I_{ON} is improved by 15 % and 50 % for respectively NMOS and PMOS transistors at same I_{OFF} [3]. We use 45° rotated substrate and very high tensile CESL (Contact Etch Stop Layer) to respectively enhance hole and electron mobility [6], [7]. In addition, EOT engineering has been performed and it resulted in using decoupled plasma nitridation process in order to improve MOSFETs performances through a reduced EOT and therefore a better gate electrostatic control without degrading gate leakages. With the use of different mobility boosters, active widths could be reduced in standard cells to decrease internal power and area without impacting the propagation delays.

Since, it represents 40 % of digital area in classical MCU circuit, a new D flip flop (Dff) contained 40 transistors is layouted with lower active widths and optimized design rules reminded in Fig. 17. Poly end-cap distance (E) and gate contact to active distance (S) are also reduced. Dff schematic is presented in Fig. 18. In a first trial, up to 20 % of area reduction is obtained as displayed in Fig. 19 with active and poly layers visible. It appears that routing wires (polysilicon/metal1) and gate contacts could be different between reference and new Dff.

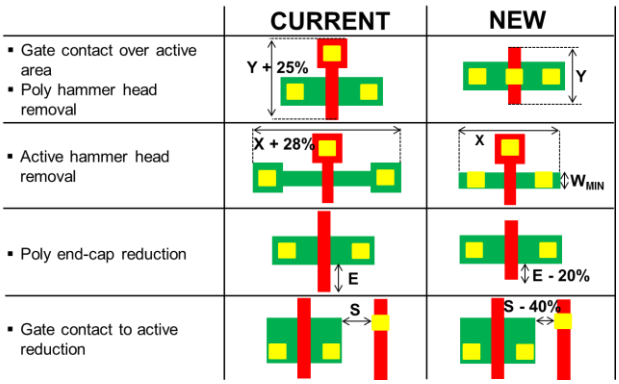


Fig. 17: Summary of optimized design rules

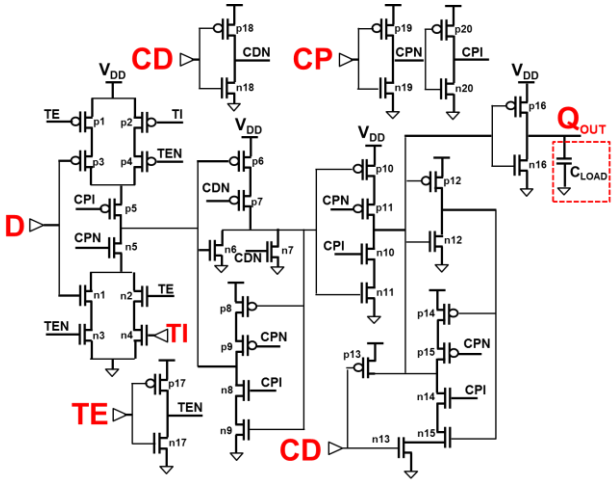


Fig. 18: Schematic of D flip flop

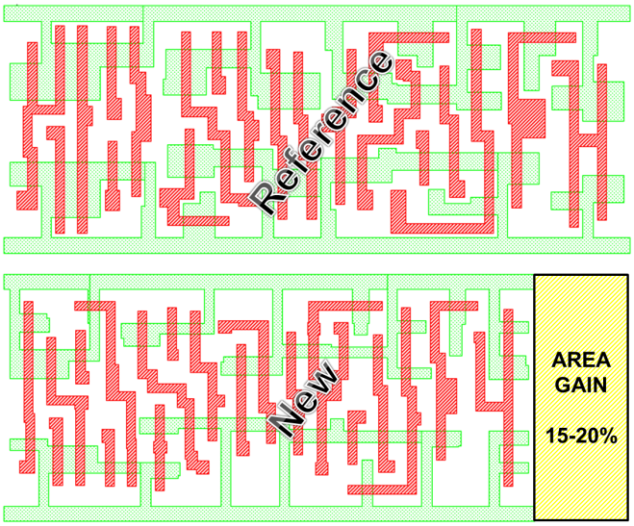


Fig. 19. Layout comparison between reference and new D flip flop

Spice post-layout simulations with RC parasitics have been performed on reference Dff with current model card and new Dff with new model card based on electrical values of new optimized process parameters (I_{DSAT}, I_{DLIN}, I_{OFF}, V_T). It appears that internal power of new Dff is reduced by almost 20 % during rising and falling edges of output Q as shown in Fig. 20. It is mainly due to the reduction of MOSFETs width by 40 % to 50 %. Moreover, clock to Q delay is reduced by 10 % for typical value of C_{LOAD} and by 5 % for maximum value of C_{LOAD} during falling edge. During rising edge, Clock to Q delay remains stable between current and new Dff (~-2%). The gains are presented in Fig. 23 and Fig. 24. These results show that propagation delays remain unchanged for all specified values of C_{LOAD} given by digital library. The propagation delay clock to Q is defined in digital library as the time interval between the input stimulus crossing the threshold of 40 % of V_{DD} for the rising signal and 60 % of V_{DD} for the falling signal. These times are reminded in displayed in Fig. 20. Internal power consumed during rising and falling transitions are calculated using (1) and (2)

$$E_{INT\_RISE} = V_{DD}.\int I_{VDD\_ALIM} - C_{load}.V_{DD}^2 \quad (1)$$

$$E_{INT\_FALL} = V_{DD}.\int I_{VDD\_ALIM} \quad (2)$$

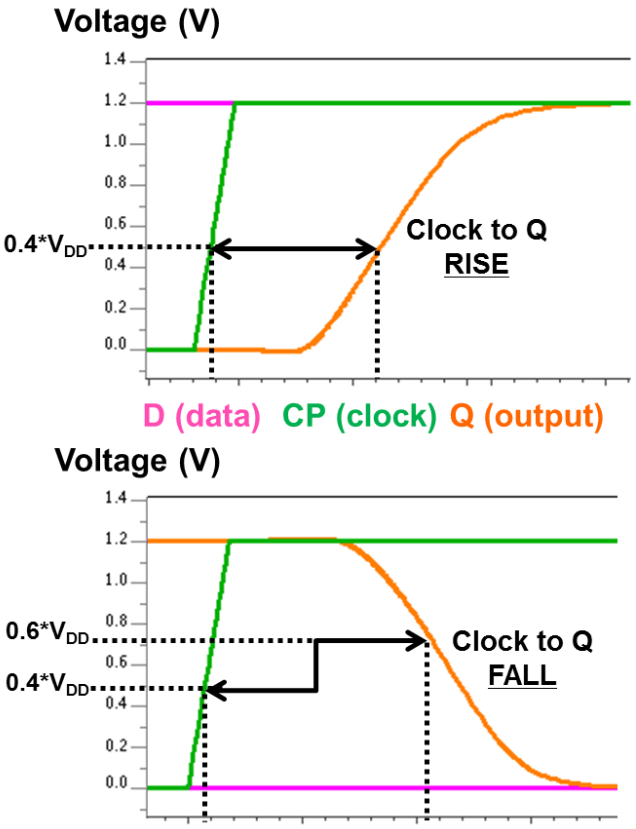


Fig. 20: Definition of clock to Q delays during rising and falling edges

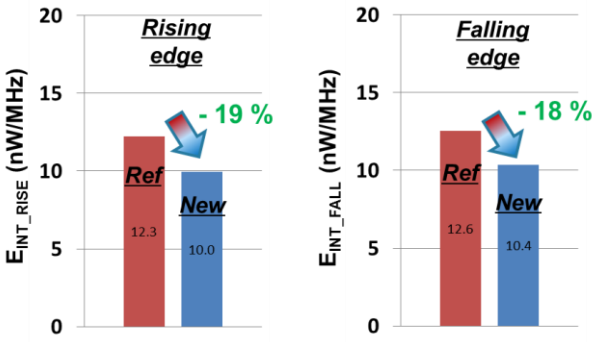


Fig. 21: Internal power gain between reference and new Dff

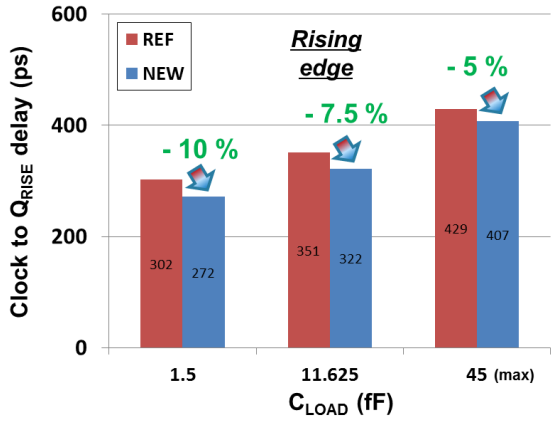


Fig. 22: Clock to Q gain during rising edge

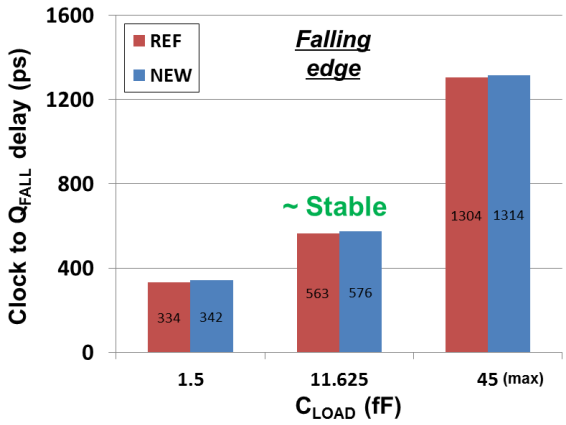


Fig. 23: Clock to Q gain during falling edge

Finally, Monte Carlo (MC) simulations have been performed in order to study the local mismatch variation between reference and new Dff. as observed in Fig. 24 and Fig. 25 for 1000 runs, the standard deviation of clock to Q delays is almost similar between reference and new Dff,. Insignificant increase (~2%) is observed during rising and falling transitions, probably due to lower transistor width. Indeed, the standard deviation of a threshold voltage ($\sigma\Delta V_T$) is inversely proportional to the transistor area [10].

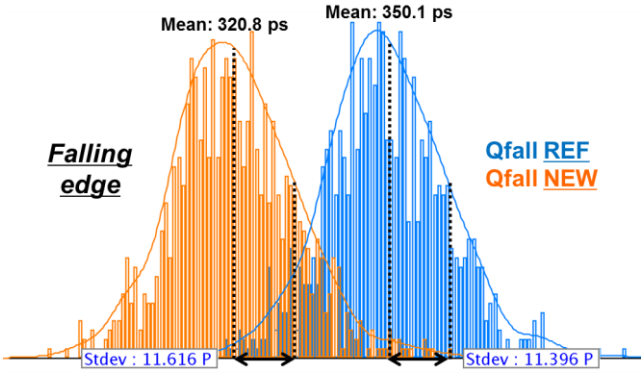


Fig. 24: Standard deviation of clock to Q during falling transition

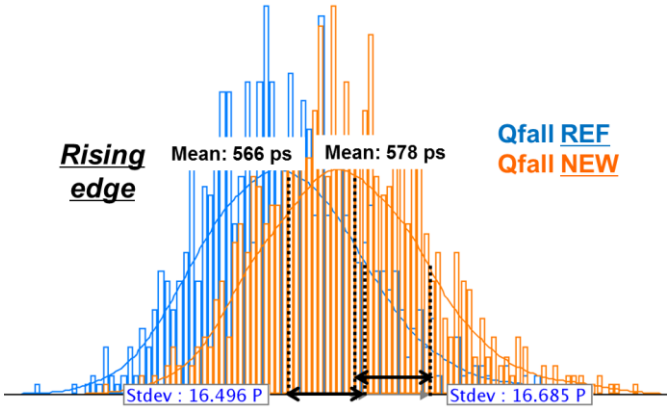


Fig. 25: Standard deviation of clock to Q delay during rising transition

IV. Conclusion

In conclusion, this work demonstrates the ability to reduce internal power of D flip flop designed in CMOS 80 nm e-NVM technology by 20 % without outliers. For that, active widths of transistor have been reduced by 40 % to 50 % and propagation delays remain unchanged thanks to the use of high mobility process features. Moreover, all layout optimizations performed on low-voltage MOSFETs used in digital standard cells were presented in this paper. For a first time, we proved the possibility to place gate contact over active area instead of STI area and remove hammer heads without any change in MOSFETs performances. All optimizations were embedded to design a new flip flop and up to 20 % in area reduction was obtained in a first trial.

Most importantly, the methodology presented in this paper could be reused with the most recent technologies in order to reduce internal power and area of digital standard cells.

References

[1] C. Piguet, *Low-power CMOS circuits: technology, logic design and CAD tools*. CRC Press, 2006, p. 10.

[2] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits A Design Perspective 2nd Edition*. pp. 209 – 216.

[3] J. Innocenti, L. Welter, F. Julien, J.-M. Portal, L. Lopez, J. Sonzogni, P. Masson, S. Niel, and A. Regnier, "Dynamic power reduction through process and design optimizations on CMOS 80 nm embedded non-volatile memories technology," *Circuits Syst. (MWSCAS), 2014 IEEE 57th Int. Midwest Symp.*, pp. 897–900, 2014.

[4] M. Mandal and B. Sarkar, "Ring oscillators : Characteristics and applications," *Indian J. Pure Appl. Phys.*, vol. 48, no. February, pp. 136–145, 2010.

[5] M. Carmona, L. Lopez, F. Julien, J. Ogier, L. Beauvisage, Q. Hubert, and D. Goguenheim, "Study of gate contact over active area," in *SBmicro In press*, pp. 5–8.

[6] C. K. Maiti and T. K. Maiti, *Strain-Engineered MOSFETs*, vol. 28. CRC Press, 2012, p. 320.

[7] S. E. Thompson, M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, T. Ghani, G. Glass, T. Hoffman, C.-H. Jan, C. Kenyon, J. Klaus, K. Kuhn, Z. Ma, B. Mcintyre, K. Mistry, A. Murthy, B. Obradovic, R. Nagisetty, P. Nguyen, S. Sivakumar, R. Shaheed, L. Shifren, B. Tufts, S. Tyagi, M. Bohr, and Y. El-Mansy, "A 90-nm Logic Technology Featuring Strained-Silicon," *IEEE Trans. Electron Devices*, vol. 51, no. 11, pp. 1790–1797, Nov. 2004.

[8] R. Bianchi, G. Bouche, and O. Roux-dit-Buisson, "Accurate modeling of trench isolation induced mechanical stress effects on MOSFET electrical performance," *IEDM*, pp. 117–120, 2002.

[9] M. Miyamoto, H. Ohta, Y. Kumagai, Y. Sonobe, K. Ishibashi, and Y. Tainaka, "Impact of reducing STI-induced stress on layout dependence of MOSFET characteristics," *IEEE Trans. Electron Devices*, vol. 51, no. 3, pp. 440–443, 2004.

[10] M. J. M. Pelgrom and A. P. G. Welbers, "Matching Properties of MOS Transistors," no. 5, pp. 1433–1440, 1989.

A Simple Clock-Fault Detection Analog Circuit for High-Speed Crystal Oscillators

A. Gamet[1,2], Y. Bacher[2], S. Meillère[1], P. Le Fevre[2], N. Froidevaux[2]
[1] IM2NP-UMR CNRS 7334, Aix-Marseille University, 13013 Marseille, France
[2] STMicroelectronics, 190 av Célestin Coq, ZI Peynier-Rousset, Rousset, France
Arnaud.gamet@st.com

Abstract - This paper presents a simple architecture for clock-fault detection in high-speed applications. The overall principle consists in converting a possible error of time to a logic voltage level. When a high voltage level is present at the output, a reliable clock is detected whereas a low voltage level implies a clock error. This detection system is intended for all System-on-Chip such as microcontrollers which use external clock from 4 MHz to 50 MHz. The proposed circuit is realized in CMOS 40 nm process technology. Simulation results prove the suitability of the structure and its integration on silicon is strongly considered by clock error detection in integrated circuits.

Keywords— Clock-fault detection, High-Speed Circuits, Clock error, Peak detector, Crystal oscillator

I. INTRODUCTION

Nowadays, microcontrollers and more generally System-on-Chip (SoC) are characterized by theirs high level of integration, and theirs very low power consumption. The increase of frequency in these chips is also a point to be noticed so as to drive high-speed applications. The clock is one of the predominant elements in a microcontroller. Indeed, it allows the processor to latch the data at the desired frequency and synchronize communication. So, a very well quality of the input clock is required in order not to impact the entire clock system distribution. A Phase-Locked Loop (PLL) is usually used to generate stable divided frequencies to synchronize signals in the core [1]. A clock default could have very bad effects on the overall performances of the microcontroller and that is why it is absolutely imperative to monitor the input clock.

Clock references for microcontrollers can be internal or external. A clock multiplexer tolerates a clock switching [2]. To avoid some potential clock errors, it exist some clock fault detection techniques to inform that the clock is not decent. Some of these techniques are mentioned in the literature for many field of application. Triple modular redundancy (TMR) [3] and N-tuple modular redundancy (NMR) [3][4] are commonly designed in fault tolerant computer systems. Considering the dimensions of integrated circuits, these architectures are too area-consuming. Most common techniques use the properties of some elementary components such as capacitor and MOS transistors to realize the clock detection [5][6]. The main advantages of this implementation are low area, low consumption and very easy integration.

The clock fault detection architecture presented here is based on the time-to-voltage converter as already employed [5][7]. The paper is organized as follows. Section II presents the general architecture of the system. The circuit implementation with its major features and some simulation results are described in section III and IV respectively. Conclusion is drawn in the last section.

II. ARCHITECTURE

Different ways allows providing clock in microcontrollers: Internal clock based on electrical phase-shift such as RC oscillators which can operate in a wide frequency range and external clock coming from mechanical resonant devices such as crystal and ceramic resonator. The external clock can be also originated from another chip. This paper is only focused on high-speed clocks from 4 MHz to 50 MHz. The simplified representation of the clock detection system is depicted in Fig. 1. It is composed of a multiplexer and the clock detector which is enabled with the bit EN_CKD.

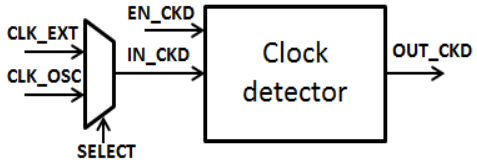


Fig. 1 Simplified representation of the clock detection system

The clock source is determined by programming the selection bit SELECT. The type of clock depends on the application. It is determined by a combination of features including power consumption, operating frequency, accuracy, cost and environmental requirements. The multiplexer enables choosing between an external clock CLK_EXT and the oscillator clock CLK_OSC. A flag OUT_CKD = '1' is generated when a clock faster than 4 MHz is present whereas OUT_CKD = '0' for the other cases of input signal. In an ideal world, the node IN_CKD is a perfect full swing square wave with 50 % duty cycle. In reality, the received clock can be damaged

for many reasons. Interferences, cross-talk, noise coupling, jitter noise, bad electrical connections on the external board or a design problem are the main causes of clock defaults. Few examples of received clocks are shown in Fig. 2. Other types of undesired clocks such as clock deletion in low and high half-cycle or in full period exist. In these cases, a recovery circuit is used in addition to the clock detector to reassemble the clock [8].

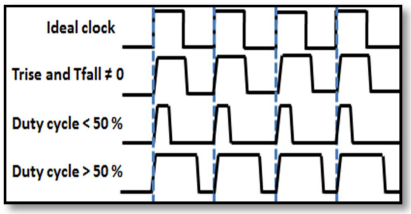


Fig. 2 Example of received clocks

So, the functionality of the clock detector is mandatory in order to inform the core that the input clock is defaulting. In the case where mechanical resonant oscillator with a good clock signal is snatched out of the board, the clock detector is able to detect it with no consequence on the clock distribution network.

III. Circuit Implementation

A. Block diagram of the proposed circuit

The circuit received an analog clock in input. Firstly, the signal is digitized to pull up to logic '1' or pull down to logic '0'. Then, the clock is filtered to delete the continuous current component in case of an input continuous voltage. A level detector is added to compare a charged value to a threshold voltage. Finally, a decision is done in function of the detection level result. So, the input analog clock is converted to voltage level equals to logic '1' when a clock is present, or logic '0' when there is no clock. The block diagram of the circuit is presented in Fig. 3.

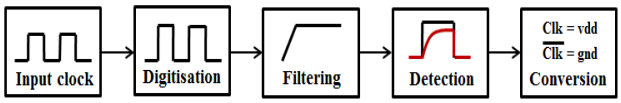


Fig. 3 Block diagram of the proposed circuit

B. Detection circuit

The complete detection circuit is made of elementary components with 12 transistors which form 3 buffers, 1 diode, 2 resistances and 2 capacitances. The circuit is composed of two basic analog functions including a high-pass filter CR and an envelope detector. The operating of the different functions for this detection circuit is explained in the following. The circuit is shown in Fig. 4.

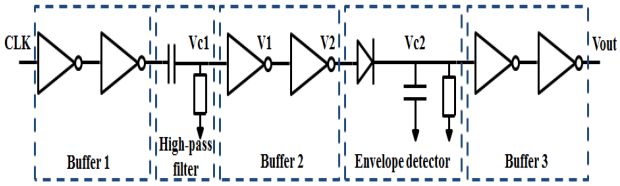


Fig. 4 Detection circuit

C. Envelope detector

The envelope detector circuit is usually used for amplitude demodulation. The circuit is composed of three passive components, a diode D, a resistance R, and a capacitance C. The circuit is represented in Fig. 5.

During the high half-cycle of the input signal, D is forward-biased and C quickly loads. When the low half-cycle occurs, D is reverse-biased and C slowly unloads in R. The time constant $\tau = RC$ must be quite high to ensure a slow unload of C through R between two high half-cycle. So, R and C values must be chosen to be functional on the whole frequency range. The circuit response to a 4 MHz square wave input with a rise time and a fall time equal to 100 ps is shown in Fig. 6.

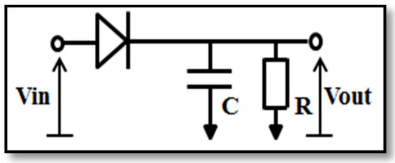


Fig. 5 Envelope detector circuit

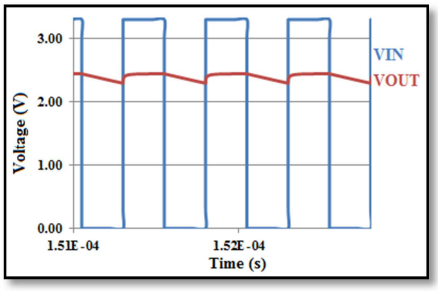


Fig. 6 Response of the envelope detector circuit

D. High-pass filter

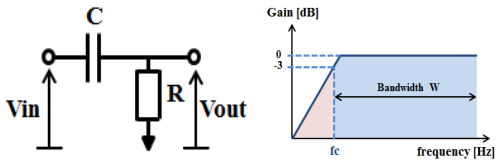


Fig. 7 High-pass filter with its frequency response

A 1st order passive high-pass filter composed of a series capacitance C and a resistance R tied to ground is used to block any DC input signal. In our case, it enables blocking input DC voltage upper to the threshold voltage of the diode present in the envelope detector circuit. At low frequencies, the capacitance acts as an open circuit because its reactance is very high. The filter blocks all input ac signals until the cut-off frequency f_c which is defined by R and C values (1).

$$f_c = \frac{1}{2\pi RC} \tag{1}$$

To be compliant with specifications, the cut-off frequency is fixed at a bit fewer than 4 MHz. The high-pass filter response with a 4 MHz and 48 MHz square wave input signal is shown in Fig. 8 and Fig. 9. The output waveform consists in two spikes, one positive on

rising edge and one negative on falling edge. The higher the input frequency will be the output signal will look like to the input signal because the capacitance has no time to be unloaded in the resistor.

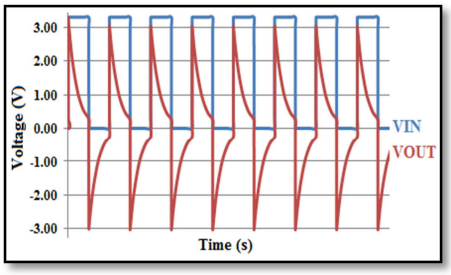


Fig. 8 CR response with 4 MHz square wave input

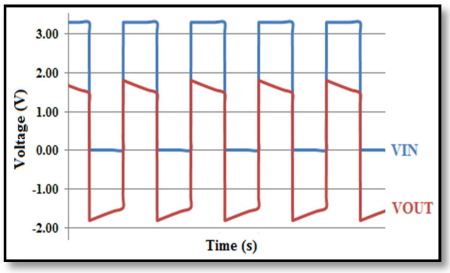


Fig. 9 CR response with 48 MHz square wave input

E. Clock buffering

Three buffers are present in the complete circuit of the clock detector as shown in Fig. 4. Each one has its own property. The first buffer is largely inserted to improve the input signal strength in case of damaged input clock. The aim is to be sure to pull up a high half-cycle to the power supply and to pull down a low half-cycle to the ground. The role of the second buffer is to amplify the output of the high-pass filter. Since the series capacitance cuts the DC current, the signal Vc1 must be amplified. As Vc1 can be a negative voltage, the threshold voltage of the second buffer is shifted by increasing the NMOS dimensions. Consequently, the threshold voltage is lower than Vdd/2 as shown in Fig. 10. In our case, a good trade-off is found at about Vdd/4. This value allows operating the circuit for all process cases and an input voltage range from 1.8 V to 3.6 V.

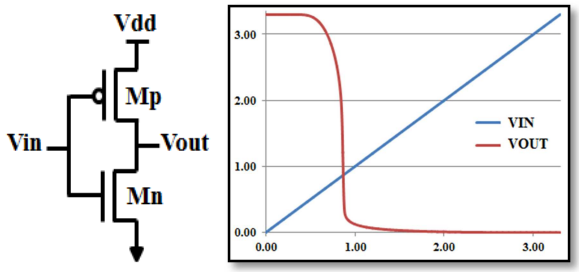


Fig. 10 CMOS inverter with its characteristics

The third buffer is placed after the envelope detector circuit. It allows ensuring a decent output level. It also provides an output signal intended for the core of the microcontroller in order to ensure the clock synchronization in the system or not.

IV. Simulation Results

The whole circuit is implemented in CMOS 40 nm process technology. All simulations are done with Eldo simulator. The voltage range is from 1.8 V to 3.6 V and three temperatures are tested. The process cases include slow, typical, and fast transistor models. The circuit is able to operate on the frequency range from 4 MHz to 50 MHz. The simulation results are shown below to demonstrate the good operating of the circuit in case of typical clock, with low and high half-cycle.

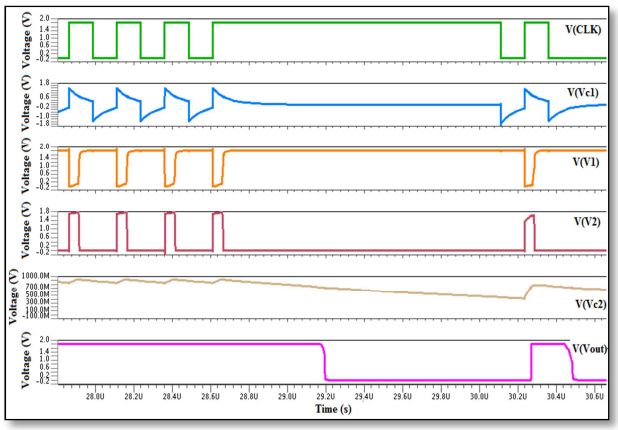


Fig. 11 Waveforms for a long high cycle with a 4MHz input clock at 1.8V in typical conditions

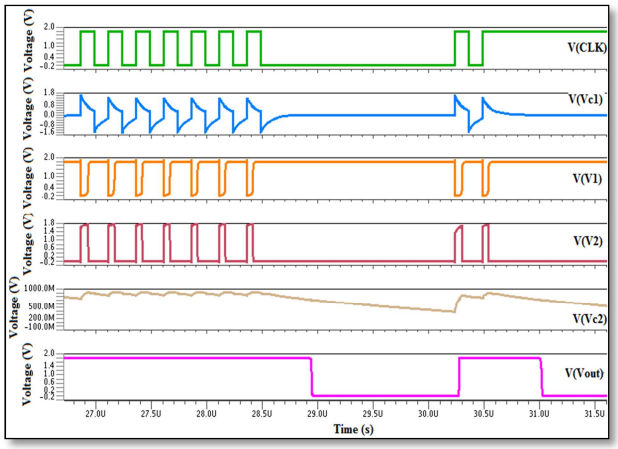


Fig. 12 Waveforms for a long low cycle with a 4MHz input clock at 1.8V in typical conditions

In both cases, when a long high cycle or a long low cycle occurs, V1 is pulled up to ‘1’. The node Vc2 slowly decreases and when the voltage is enough low, it reaches the threshold voltage of the following buffer. Then, the conversion is done. There is no clock anymore and Vout is equal to ‘0’.

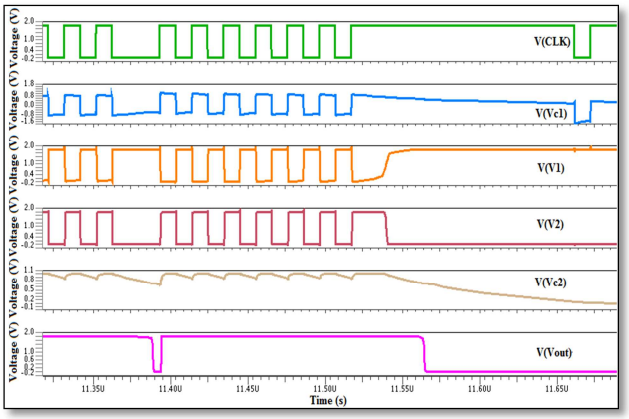


Fig. 13 Waveforms with a 48MHz input clock at 1.8V in typical conditions

The circuit is designed to address all types of mechanical resonator. The frequency range of these components for high frequency applications is about from 4 MHz to 48 MHz. So, the case with a 48 MHz input clock is shown in Fig. 13. The results seem similar to the previous result with a 4 MHz input clock. A long high state level is detected as a default by the circuit. If the clock is regenerated, then the output voltage switches its state to '1'.

The Fig. 14 shows the case where the input is very slow. The output reacts to a rising edge of the input. As the clock stays unchanged, Vc2 decreases and the output voltage goes back to '0'.

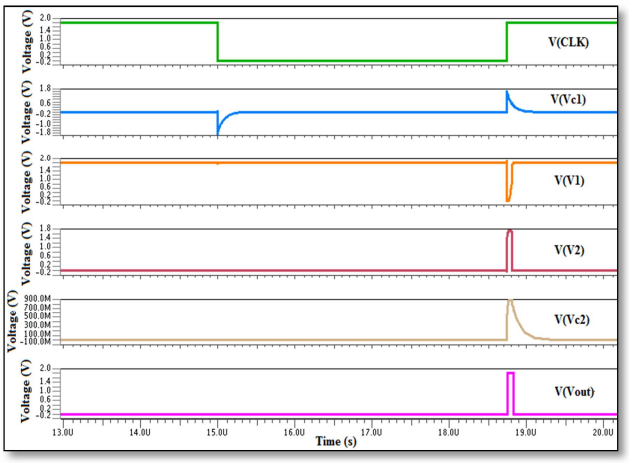


Fig. 14 Waveforms in case of slow input clock at 1.8V in typical conditions

All process cases have been simulated and the circuit operates quite well. The min and max DC consumption is written in the conclusion. The major limitation of this circuit is the presence of the diode which implies a voltage drop on Vc2. It would be a problem if the supply voltage is highly decreased. Nevertheless, the circuit offers a good trade-off between the consumption of area and the specifications.

V. Conclusion

In this paper, an analog circuit for clock detector has been introduced in 40 nm process CMOS technology. The principle is based on amplitude demodulation with an envelope detector circuit. Simulation results show the suitability of the circuit. An error of time is converted to a voltage level. The proposed circuit presents some advantages including very low area-consuming and very low power consumption in particular. It operates from 1.8 V to 3.3 V between -40 °C and 125 °C. The circuit is able to detect a clock default for the frequency range from 4 MHz to 50 MHz. The architecture covers the main clock defaults and can be associated with an integrated recovery circuit.

Table 1 Summary of the main clock detector performances

Supply voltage [V]	1.8		3.6	
Process	Slow		Fast	
Frequency range [MHz]	4	48	4	48
DC consumption @ -40°C [µA]	19.76	36.65	101.52	121.10
DC consumption @ 25°C [µA]	17.09	33.62	90.01	110.48
DC consumption @ 125°C [µA]	15.41	31.13	76.83	99.38

REFERENCES

[1] D. Cai, H. Fu, J. Ren, W. Li, N. Li, H. Yu, and K. S. Yeo, "A Dividerless PLL With Low Power and Low Reference Spur by Aperture-Phase Detector and Phase-to-Analog Converter," IEEE Trans. Circuits Syst., vol. 60, no. 1, pp. 37–50, 2013.

[2] H. Wang, Y. Zhang, X. Li, L. Chen, Z. Wen, K. Zhang, and M. Wang, "A Configurable Fault-Tolerant Glitch-Free Clock Switching Circuit," no. 2, pp. 537–540, 2013.

[3] K. S. Siddiqui and M. A. Baig, "FRAM based TMR (Triple Modular Redundancy) for Fault Tolerance implementation," Inf. Commun. Technol., pp. 1–5, 2011.

[4] P. G. Depledge, "Fault-tolerant computer systems," IEEE Proc. Reliab., vol. 128, no. 4, pp. 257–272, 1981.

[5] C. Yu, "A Clock Fault Detection Circuit for Reliable High Speed System by Time-to-Voltage Conversion," 2009 Second Int. Symp. Electron. Commer. Secur., pp. 283–286, 2009.

[6] W. T. Ang, H. F. Rao, C. Yu, J. Liu, I. Wey, A. A. Wu, H. Zhao, and J. Chen, "A Clock-fault Tolerant Architecture and Circuit for Reliable Nanoelectronics System," Des. Technol. Integr. Syst. Nanoscale, pp. 186–191, 2007.

[7] H. Hui, Y. Changhong, C. Kerning, and S. Lingling, "A Novel Clock-fault Detection and Self-recovery Circuit Based on Time-to-voltage Converter," Int. Conf. Commun. Circuits Syst., pp. 1204–1207, 2008.

[8] S. Kim and K. Cho, "Design of 2.5Gb/s non-PLL-type All-Digital Clock Recovery Circuit," ISOCC, pp. 416–419, 2010.

Performance analysis of parallel reduction and matrix multiplication on Parallella-16 heterogeneous system

Ćatić Nermin, Imamović Kanita, Jašarević Eman, Nosović Novica
Faculty of Electrical Engineering, University of Sarajevo/Department of Computing and Informatics
Sarajevo, Bosnia and Herzegovina

***Abstract*—Often one of the biggest practical challenges in parallel programming is choosing the right algorithm for a particular architecture, that is choosing the right architecture for a particular algorithm. In this paper, we discuss the differences between the behavior of one relatively simple algorithm, the reduction algorithm and somewhat more complex, the algorithm for matrix multiplication, throughout the execution of these algorithms on the unique architecture, Parallella board. In order to evaluate the performance of parallel implementation, some metrics such as speedup, have been measured through performing these algorithms both sequential and parallel. Furthermore, this analysis includes three different parallel implementations of both algorithms.**
All this has been done in order to determine exactly which type of algorithm is architecture of Parallella board designed for. The main conclusion of our experimental study is that for more complex algorithms Parallella board shows better results, because the majority of the time of total program execution time, when running simpler algorithms is spent on the communication, synchronization and other side effects that parallelization brings.

***Key words* —Parallella board, speedup, reduction, matrix multiplication**

I. INTRODUCTION

Often one of the biggest practical challenges in parallel programming is choosing the right algorithm for a particular architecture, that is choosing the right architecture for a particular algorithm. The main goal of this paper is to show how to program on Parallella board using STDCL program module as well as how to choose an appropriate algorithm for this purpose. This paper describes two types of algorithms: reduction algorithm and algorithm for matrix multiplication. The importance of the reduction algorithm is that it can serve as a replacement for the whole class of algorithms in which one scalar result has to be calculated over a range of values [1].

The implementation of the reduction algorithm on the Parallella board has been made in three different ways. The first implementation is a sequential implementation of this algorithm and it was made on the Parallella main ARM processor using C programming language. The second implementation presents a parallel implementation that is not optimized written in OpenCL. The third implementation represents optimization of the naïve approach used in the second implementation.

Besides the reduction algorithm, this paper gives both sequential and parallel implementation of the algorithm for matrix multiplication. The matrix multiplication algorithm is important because it is the core building block of Linear Algebra.

Both of the above mentioned algorithms were executed on Parallella board and in order to evaluate performance, their execution time was measured and compared.

II. USED ALGORITHMS

A reduction is a very simple operation that takes an array of data and reduces it down to a single element [1]. In this paper, reduction algorithm represents the implementation of simple algorithm in which single precision floating point elements are added to produce the scalar sum. The program is relatively trivial and consists of allocating memory for the data, initializing the array of data, summing the elements of the array and printing the result which is a floating point value.

Amdahl's law is used to find the maximum speedup to an overall system when only part of the system is improved [2].

$$\text{Speedup} = \frac{1}{S + \frac{1-S}{N}} \tag{1}$$

N—number of parallel processing elements
S—the fraction of sequential execution that is inherently sequential

Applying Amdahl's law to reduction algorithm, the following calculations have been made:

- Percentage of sequential code execution (S) = 44,3%
- Percentage of parallel code execution (1-S) = 55,7%
- Number of parallel processing elements (N) = 16

$$\text{Maximum speedup} = \frac{1}{0.443 + \frac{0.557}{16}} = 2.09 \tag{2}$$

The product C of two matrices A and B is defined as

$$C_{ik} = A_{ij} \times B_{jk} \tag{3}$$

where j is summed over for all possible values of i and k [3]. The dimensions of the matrices must satisfy

$$(n \times p)(p \times m) = (n \times m) \tag{4}$$

The algorithm for matrix multiplication consists of four parts. The first part is allocation of the storage for matrices that are being multiplied. Then, the matrices operands and result matrix are initialized. After initialization, the calculation is performed producing the result matrix. The last part of this algorithm is printing out the result matrix.

Applying Amdahl's law to the matrix multiplication algorithm, the following calculations have been made:

- Percentage of sequential code execution (S) = 2,1%
- Percentage of parallel code execution (1-S) = 97,9%
- Number of parallel processing elements (N) = 16

$$\text{Maximum speedup} = \frac{1}{0.021 + \frac{0.979}{16}} = 12.302 \quad (5)$$

III. ALGORITHMS PARALLELIZATION

The Parallella board consists of dual core ARM central processor and 16-core Epiphany coprocessor and represents a heterogeneous computer system [4]. When programming on Parallella board this unique architecture must be taken into account. Therefore, throughout writing a parallel implementation of both algorithms, reduction algorithm and matrix multiplication algorithm three basic requirements for code changes must be included. First, the code to execute on coprocessor must be written. This code is called Kernel code. The second change is referred to allocating memory that is shareable between the ARM host and the Epiphany coprocessor. Third change is coordination of activities on coprocessor. Allocating memory and activity coordination is placed in what is called Host code, because it is executed on Parallella's central processor, that is host processor. So, in conclusion, in order to run any algorithm in parallel on Parallella board the code must be divided in two parts. First one is Kernel code that will be executed on Parallella's Adapteva Epiphany 16-core coprocessor. Second one is the Host code. It is executed on Zynq dual-core host processor.

Below, it will be explained how to parallelize both reduction and matrix multiplication algorithm, that is how to write their Kernel and host Code.

A. Reduction algorithm—Naïve implementation

Kernel code represents the part of sequential code from chapter 2 in which the calculation is performed, modified and customized to perform in parallel. Kernel code has three arguments: number of elements of data vector, data vector and temporary array used to store the partial sums. The idea of the Kernel code is that the work of adding numbers is divided as evenly as possible to compute partial sums, simultaneously. The final result is calculated by adding up these partial sums.

Host code represents the part of sequential code from chapter 2 in which the allocation of the memory and vector initialization has been done, except that this time, memory that is shareable between the ARM host and the Epiphany co-processor is allocated. Also, in host code it must be ensured that the host memory is synchronized with the Epiphany device prior to its use by our kernel. After all this is done, the execution of Kernel code has been forked and block on the host until all of the co-processor operations are complete.

B. Reduction algorithm—First approach to optimization

First question when programing Parallella is not how one can organize the parallelism, but rather how the organization of the parallelism needs to be done, considering Parallella's architecture.

In naïve implementation when we defined the index range used to launch threads, we asked for n threads executed in local workgroups of 16 to match the number of cores on the Epiphany processor. In order to change this, there are some code modifications need to be made.

The code changes include "folding" loops into the kernel so that each of the coprocessor cores executes approximately 1/16th of the calculation within a single kernel. There is no guarantee that the balancing of work is fully equal, unless the total number of elements in the array is divisible by 16. However, a division of work is as close to even as much as it is possible for those numbers not divisible by 16.

C. Reduction algorithm—Second approach to optimization

With both naïve and first approach of reduction algorithm parallel implementation, there is a significant overhead to manage work items, because full array must be forwarded to each work unit every time. Therefore, the second approach to optimization introduces implementation in which local memory of cores is used instead of device-shareable memory is used.

D. Matrix multiplication algorithm—Naïve implementation

Similar to reduction algorithm, naïve implementation of matrix multiplication algorithm represents the part of sequential code from chapter 2 in which the calculation is performed, modified and customized to perform in parallel.

Kernel code has six arguments: matrices operands and result matrix, and their dimensions. The idea of the Kernel code is to simultaneously calculate the result matrix elements.

E. Matrix multiplication algorithm—First approach to optimization

To optimize matrix multiplication, we must assure that for every memory movement, we execute as many flops as possible. So, first approach to optimization matrix multiplication algorithm represents a code modification that calculates one row instead of one element of the result matrix simultaneously.

F. Matrix multiplication algorithm—Second approach to optimization

Each element of C (result matrix) in a row uses the same row of A, and the same columns of B (A and B are matrix operands). So, the second approach to optimization of matrix multiplication algorithm represents a code

modification in which matrix A rows and matrix B columns are copied into core's local memory. This prevents the overhead of pulling it from global memory for each computation of each element of the result matrix.

All implementations of both algorithms were executed on Parallella board. The results obtained during their execution are presented in chapter 4.

IV. MEASUREMENT AND RESULT PRESENTATION

Figure 1 shows comparison of the time of execution for four different implementations of the reduction algorithm described in chapter 3 (serial implementation, naïve parallelization, first approach, and second approach to optimization).

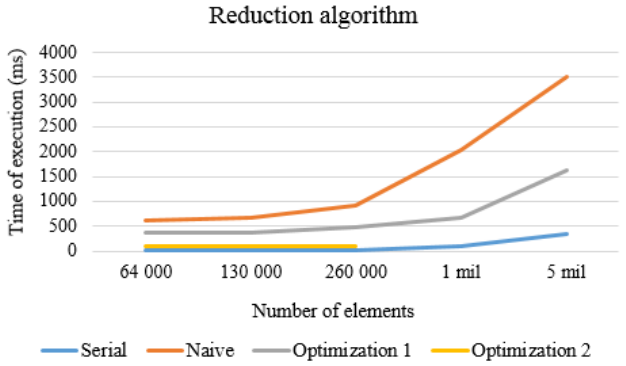


Fig. 1 Comparison of the relationship between the time of execution and number of elements in an array for four different implementations of reduction algorithm

Figure 2 shows the relationship between sequential and three different parallel implementations of reduction algorithm, that is, it shows how the time of execution is getting closer to sequential execution time with every optimization that has been made. It shows how with every optimization, parallel implementation achieves higher speedup comparing to the previous one, but is still slower than sequential implementation due to the large overhead.

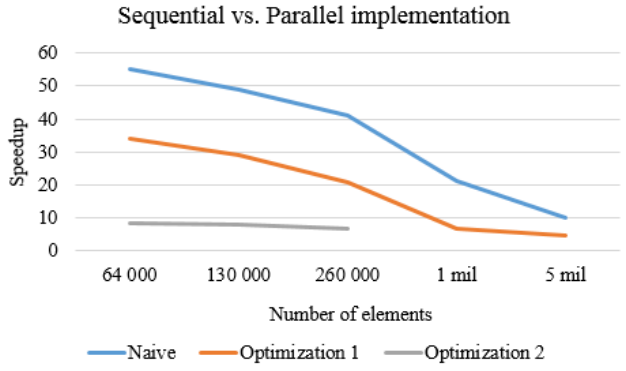


Fig. 2 Comparison of speedup achieved with three different parallel implementations of reduction algorithm using five different lengths of an input array

A. Reduction algorithm - Naïve implementation

Naïve parallel implementation of reduction algorithm is much slower than sequential implementation due to the huge overhead in establishing communication between the host processor and the Epiphany coprocessor. Numbers that will be added to the final sum need to be forwarded to each core of coprocessor. Regardless of the architecture, the goal is to employ as many processors as possible to work simultaneously. After cores have performed the addition, they return the results to the host, which performs the final summation of values stored in temporary variables. Load balancing is poor because all cores are waiting for the final result is placed in temporary variable in the host. As shown in Figure 1, tests were carried out for a different number of elements.

B. Reduction algorithm—First approach to optimization

As presented in Figure 1, this optimization of kernel achieved speedup compared to the naïve implementation, but is still slower than sequential implementation due to the communication overhead.

C. Reduction algorithm—Second approach to optimization

Unlike previous implementations of the algorithm, in this kernel there is less communication between coprocessor and host memory. Therefore, the performance of this implementation is better than other kernel implementations, but is still slower than sequential. This balance of work is better than the previous implementations because processors do not lose time to copy the elements that will not be used. Each core of Epiphany coprocessor has 32 KB of local memory, and float numbers occupy 4B, so the maximum number of float elements that can be stored in each core's local memory is approximately 8.000 elements [4]. In Parallela-16 board, this is a memory constraint, because only 130.000 elements can be stored in Parallela's local memory, and therefore additional speedup can not be achieved.

Figure 3 shows comparison of the time of execution for four different implementations of matrix multiplication algorithm described in chapter 3 (serial implementation, naïve parallelization, first approach, and second approach to optimization).

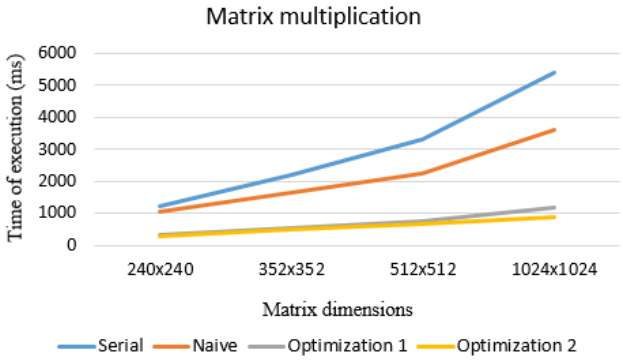


Fig. 3 Comparison of the relationship between the time of execution and matrix dimensions for four different implementations of matrix multiplication algorithm

Figure 4 shows the relationship between sequential and three different parallel implementations of matrix multiplication algorithm. It shows how with every

optimization, parallel implementation achieves higher speedup comparing to the previous one and to the sequential implementation.

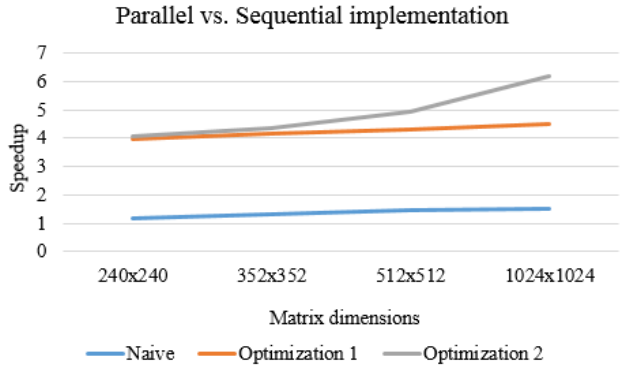


Fig. 4 Comparison of speedup achieved with three different parallel implementations of matrix multiplication algorithm using four different matrix dimensions

D. Matrix multiplication algorithm—Naïve implementation

Since the algorithm of matrix multiplication contains high level of parallelism, a naïve parallelization achieved small speedup in comparison to the sequential implementation. Each of the cores gets indexes of row and column that will be calculated, and after the calculation is performed, core stores the result in a temporary variable. As in the naïve implementation of reduction algorithm, there is a large overhead due to communication between host and coprocessor. This balance of work is not satisfactory because the individual working items are waiting for others to finish the work if the number of elements is not divisible by 16. Execution time and speedup are shown in the Figure 3.

E. Matrix multiplication algorithm—First approach to optimization

In this kernel optimization, a significant speedup is achieved in comparison to the naïve and sequential implementation of the matrix multiplication algorithm. Every Epiphany core calculates entire row of the matrix that is multiplied, not a single item as it was the case in naïve kernel function. A communication overhead is reduced due to the less communication between host and cores. Yet, every core still takes elements from global memory which is significantly slower than local memory.

F. Matrix multiplication algorithm—Second approach to optimization

Every working element row and column that will be calculated are stored in core's local memory. Since every core on Parallella board has 32 kB of local memory, and knowing the fact that floating point number takes 4B of memory, a maximum of 8,000 elements can be copied to the core's local memory [5]. As the data are closer to the processor, the time required for fetching data processing is shorter. This approach showed the best results for the matrix multiplication algorithm.

G. Comparison of results of reduction algorithm and matrix multiplication algorithm

For matrices of a size less than 32x32, using parallel matrix multiplication algorithms described in chapter 3, Parallella doesn't achieve any speedup in comparison to the sequential implementation, due to the large overhead. As the number of elements increases, so does the speedup, as expected. The limitation of Parallella 16 board is the size of the local memory. On newer versions of Parallela, local storage is higher as well as the number of processors, which allows larger speedup to be achieved.

If one compares the speedup and achieved execution time for reduction and matrix multiplication algorithms for the same number of elements, it can be seen that it is better to use Parallela board for matrix multiplication algorithm because the algorithm has a higher amount of parallelism that can be achieved, in comparison to the reduction algorithm.

In the study, we concluded that Parallella board consumes an average of 5.7 W. The maximum energy consumption reached was 6.1 W.

V. Conclusion

This paper presents a parallel implementation of two algorithms, reduction algorithm and matrix multiplication algorithm. The performance of these parallel implementation is evaluated through comparing speedup (time of execution).

Common feature, for both algorithms described in this paper, is that speedup which can be achieved is directly proportional to the total number of elements. As the number of elements increases, speedup increases.

However, parallel implementations of an algorithm for matrix multiplication significantly reduced the computing time both compared to an implementation on a single-core processor and compared to parallel implementation of reduction algorithm. Reduction algorithm showed no significant speedup compared to its sequential implementation due to additional costs of overhead.

Therefore, the main conclusion of our experimental study is that for more complex algorithms such as matrix multiplication algorithm Parallella board shows better results. (It is said that the algorithm for matrix multiplication is more complex than reduction algorithm because it contains a larger proportion of the code that is possible to parallelize.) The main reason for that is that majority of the time of total program execution time, when running simpler algorithms, such as reduction algorithm, is spent in the communication, synchronization and other side effects that parallelization brings.

A. Guidelines for future work

Additional speedup can be achieved using newer versions of Parallella board with Epiphany coprocessor that has 64 instead of 16. That speedup can be achieved because there are more work units available (64) to simultaneously carry

out mutually independent and demanding tasks such as matrix multiplication.

In addition, on newer version of Parallella board capacity of the local memory is larger. As the conclusion of this paper earlier presented total number of elements is directly proportional to speedup. And if there is additional memory, larger total number of elements can fit in local memory. Therefore, higher speedup can be achieved.

Another optimization that can be done to parallel implementations of two algorithms presented in this paper is loops unrolling.

REFERENCES

[1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest and Clifford Stein, Introduction to Algorithms, MIT Press, 2001

[2] http://15418.courses.cs.cmu.edu/spring2014/lecture/progbasics

[3] http://mathworld.wolfram.com/MatrixMultiplication.html

[4] https://www.parallella.org/

[5] http://www.adapteva.com/parallella-board/

Building Network-enabled Smart Sensors and Actuators

Ivan Pavić*, Josip Puškar*, Ivan Soldo*, Ivan Spasić*, Hrvoje Džapo*, Dražen Ćika**

* Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** Research and Engineering Center d.d.,Zagreb, Croatia
E-mail: drazen.cika@rec-global.com

Abstract –The advent of the new Internet-of-Things (IoT) paradigm in last few years redefined the ways how the small embedded systems can enhance smart sensors and actuators to provide new possibilities in ubiquitously interconnected global world. Although there is a rising tendency in modern low-power embedded system design to provide some kind of IoT functionality out-of-the-box due to the availability of network-aware microcontrollers and system-on-chip (SoC) solutions, there is also a great need to upgrade legacy solutions to fit them into this new paradigm shift. In this article we describe a simple generic framework for upgrading legacy devices, which was built upon the state-of-the-art low cost and low power off-the-shelf hardware components, and open source software. The solution can be easily adapted for various types of smart sensors and actuators. We describe a common set of requirements for building such systems, design decisions reasoning, comparison and justification of HW/SW components choice, and sample demonstration of adaptation and upgrade of the chosen commercial equipment, a device for bubble production at children fun events and similar applications.

I. INTRODUCTION

The embedded devices based on microcontrollers have been in use for many years, and most devices today have some sort of intelligence in a form of software built in, at least in part of functionality. It could be concluded that all those devices were capable of communication with their environment in elegant and transparent manner, but this is not always the case. The devices designed and put into operation before the proliferation of low power microcontrollers and low cost wireless interfaces are very often isolated when it comes to the communication with external world. There may be various interfaces available for users or system maintenance, but those are often based on the wired infrastructure such as the legacy RS232 port, USB connection, Ethernet port and similar. It is not always an easy task to enable such devices to communicate with their environment over the Internet. Older devices which have not been designed based on the microcontroller or microprocessor could be completely cut out of the Internet of Things (IoT) future driven by most silicon vendors and technology adopters today [1].

At the lowest level of the system design for the typical IoT application in most cases there is a requirement for some computationally capable low power microcontroller. Usually requirement is also long autonomy from a single battery power supply, embedded network interface (wired or wireless, typically Ethernet, Wi-Fi, and Bluetooth), embedded communication protocol stack, persistent data storage, and some scalable real-time operating system (RTOS). Moreover, there is typically a need to provide access from a smart client device with rich user interface, such as tablets or smart phones. This is a typical scenario for upgrading legacy systems to provide Internet connectivity, remote monitoring, data logging, crowd sensing etc. [2].

In this article the possibility is explored to enable such legacy devices, even those originally containing no firmware and digital electronics, to exchange the information with their environment using the wireless communication and Internet. In other words, the attributes ‘smart’, and ‘networked’ are added to some low cost equipment possibly having only few simple functions.

II. MICROCONTROLLER PLATFORM

The STMicroelectronics© microcontroller (MCU) used on the development board chosen for this demonstration project belongs to the STMF32F4 family, which is based on the ARM Cortex M4 core. The STM32F4 Discovery evaluation board is shown in the Figure 1. It is used for this project because of the low cost, expandability and convenient form factor enabling easy design of the necessary carrier and expansion board with power supply and wireless and I/O interfaces. The carrier board custom designed for the purpose of this project, and populated with the STM32F4 Discovery and all interfaces is shown in the Figure 2.

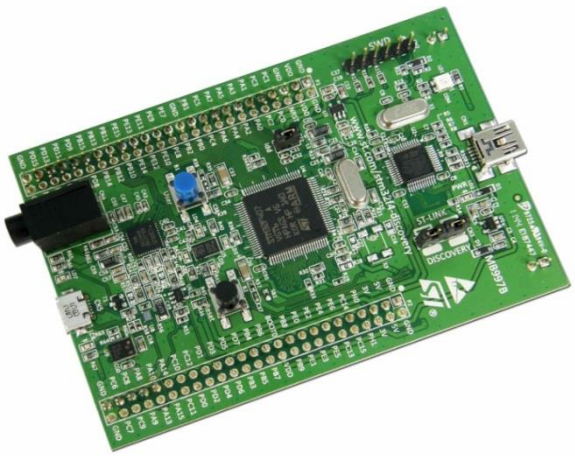


Figure 1. STM32F4 Discovery evaluation board

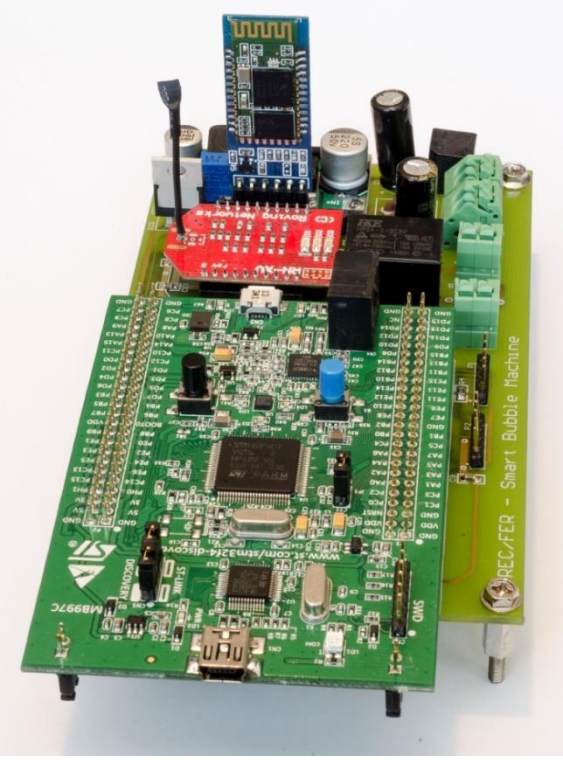


Figure 2. Carrier board with power supply and I/O interfaces

The specific DSP instructions and extensions available on Cortex M4 platform have not been used here, so that some other low power microcontroller core like the ARM Cortex M0 could also be used. Key factors for the platform selection are often device availability and support by the tool vendors, manufacturers and community. The STM32F4 platform is well supported and easily available on the market. It also has the Ethernet interface controller and more Flash and RAM resources than Cortex M0 based derivatives, which makes it more suitable for RTOS based applications with TCP/IP, web server functionality and other more resource demanding tasks.

The software (firmware) for the STM32F4 Discovery board is implemented using various peripheral and system core libraries. Functions for controlling and interaction with peripheral devices, such as GPIO pins, USART interface and I2C interface are written using the Standard Peripheral Drivers library supplied by the STMicroelectronics©, but any other functionally equivalent library could be used.

The example project device described in this article has two actuators for which the output control is implemented:

- AC motor relay is driven by the general purpose input / output (GPIO) pin which is configured as output
- DC motor speed is controlled using pulse width modulation (PWM), therefore appropriate timer channel had to be configured for PWM.

The Real-Time Clock (RTC) is the clock oscillator with support needed to keep track of the current time. Although RTCs are often used in personal computers, servers and embedded systems, they are also present in almost any electronic device that requires accurate time keeping. Use of the separate RTCs avoids the necessity for microprocessor or microcontroller to run all the time just in order to keep time, so the sleep mode of operation is possible if needed. RTC can be realized as external chip from the same or different manufacturer, or it could be available within the silicon of the main MCU, which is the case in this project. RTC on the STM32F4 series offers additional features like the calendar, the alarm, the periodic wakeup unit, tamper detection, and support for the time stamp and calibration applications. RTC uses external oscillator and Static RAM (SRAM) memory which are powered by Li-Ion battery connected to dedicated MCU pins. Configuration and initialization of RTC is done using the Standard Peripheral Library.

III. WIRELESS INTERFACES

Bluetooth is the wireless technology standard for exchanging data over short distances, usually between fixed and / or mobile devices. It is mainly used for peer to peer communication, but could also be useful for building personal area networks (PANs). It was originally conceived as a wireless alternative to the RS-232 data cables.

Bluetooth can connect several devices while overcoming their synchronization problems. There is large number of ready-made Bluetooth modules on the market, which simplifies the design of the device. One of the most popular modules is the HC-05 [3]. It offers serial (USART) interface for communication with MCU, and configuration is done using AT commands. The module itself is not suitable for direct soldering on the through-hole PCB, so the custom carrier board was designed for the project using only seven pins as an interface, as shown in the Figure 3.

Wi-Fi is a local area wireless technology that allows an electronic device to participate in computer networking using 2.4 GHz UHF and 5 GHz SHF ISM radio bands.

Wi-Fi devices can connect to a network resource such as the Internet via a wireless network access point. Such access point (or hotspot) has a range of about 20 meters indoors and a longer range outdoors. Hotspot coverage can comprise an area as small as a single room with walls that block radio waves, or as large as many square kilometers across multiple overlapping access points.

Wi-Fi has adopted various encryption technologies. The early encryption (WEP) proved to be easy to break. Higher quality protocols (WPA, WPA2) were added later. There are various pre-built Wi-Fi modules on the market, such as the WiFly from the company Roving Networks© shown in the Figure 4 and described in [4].

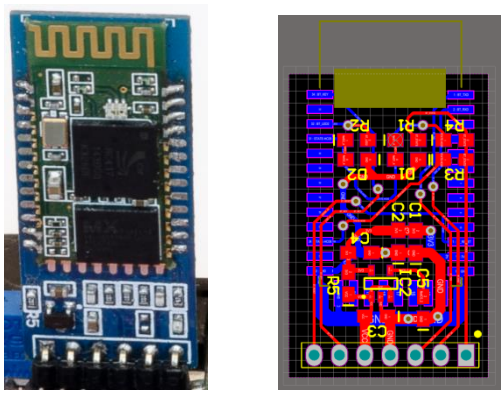


Figure 3. Bluetooth module with its custom carrier board

Figure 4. Wi-Fi module RN-XV WiFly by Roving Networks

Wi-Fi module is actually a small communication computer incorporating 802.11 b/g radio, 32 bit processor, TCP/IP stack, the real-time clock, crypto accelerator, power management unit and analog sensor interface. The module is pre-loaded with Roving firmware to simplify integration and minimize development time of the application. In the simplest configuration, the hardware only requires four connections (PWR, TX, RX and GND) to create a wireless data connection.

IV. DATA LOGGING AND FILE SYSTEM

Each field device with sensors and actuators can have requirement for data logging, so the storage space for larger amount of data might come handy. SD card functionality is implemented to fulfill such requirements, and the Secure Digital Input/Output (SDIO) Interface was chosen for this purpose and enabled using the Standard Peripheral Driver Library. The SD card connector had to be added to the custom designed carrier board, as it was not present on the basic evaluation board.

For handling and managing files the well-known and popular Chan's FatFs is used[5]. FatFs is generic FAT file system module for small embedded systems. It is written in compliance with the ANSI C standard, and is separated from the I/O layer. Since the FatFs module is a file system layer, it is completely separated from the physical devices, such as memory card, hard disk or any other type of storage device.

The low level device control module is not a part of FatFs module and is provided by implementer of the file system functionality.

V. REAL TIME OPERATING SYSTEM

When confronted with the requirements to handle several asynchronous communication tasks, while keeping the responsive user interface and handling measurement data from the sensors in the background, developers usually opt for the usage of some operating system to help partition tasks in manageable chunks of code. Also, the embedded devices often have real time response requirements, and specifically IoT devices have to behave economically and be able to go to sleep when there are no tasks to be executed left. The support of the Real Time Operating System is needed, or at least preferred.

Multitasking and the real time requirements are in the case of this application achieved by FreeRTOS [6]. FreeRTOS is the real-time operating system kernel for embedded devices. It is distributed under the GPL with an optional exception. FreeRTOS is designed to be small and simple. The kernel itself consists of only three or four C files. To make the kernel code readable, easy to port, and maintainable, the FreeRTOS is written almost entirely in C and there are only few assembly functions included where needed, mostly in the architecture-specific scheduler routines. The FreeRTOS provides functions for multiple threads, mutexes, semaphores, software timers, and other common RTOS objects. It also supports thread priorities.

The FreeRTOS based software structure of the demo application shown in this article can be seen in the Figure 5. Various colors mark tasks, interrupts, HW resources, global variables, or the RTOS objects such as semaphores or message queues. Data received over the USART is handled by interrupts in the background. Every byte is forwarded by the interrupt to corresponding message queue and then handled by the task responsible for the particular message queue content. The task checks the message queue one byte at the time and sends them to the parser function. The parser sets some control signals and semaphores, accordingly to the command parsed from the string received from the message queue.

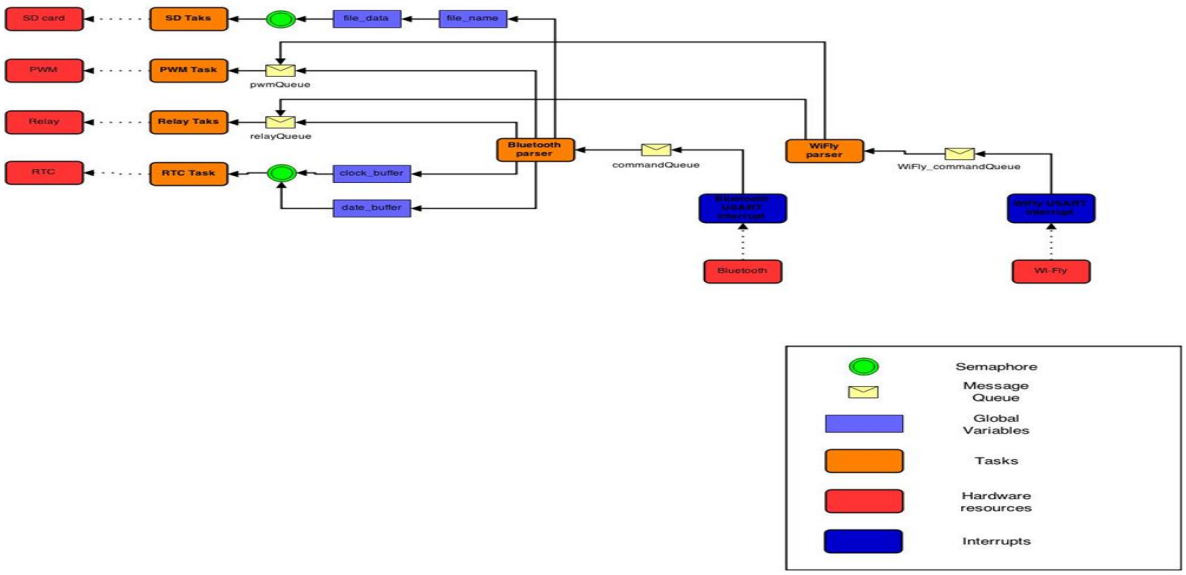


Figure 5. Firmware block diagram

VI. EXAMPLE PROJECT DESCRIPTION

The example project chosen for the demonstration purposes is called the Smart Bubble Machine. Its purpose is purely entertainment, and as its name suggest its main function is to produce bubbles. In its original form shown in the Figure 6 it is very simple device which basically can only be turned on or off, and when turned on it produces the soap bubbles in a constant stream until turned off.

The mechanics and electronics built initially into the machine by the manufacturer are simple: the transformer, some basic power supply, and AC and DC motors. Figure 7 shows the hardware components added to the bubble machine so that it could earn the 'Smart' attribute.

The basic principle of operation is very simple. The AC motor is connected directly to the 220 V AC mains power supply, and it was used to rotate the big blue perforated disc. After modification, this motor can be also started and stopped under software control using the on-board electromechanical relay. The transformer is used to transform 220 V AC to 12 V AC which is then fed to the full wave rectifier circuit. The resulting 12 V DC voltage was originally used only to continuously drive the DC motor (actually the PC CPU type ventilator) which blows air into the perforated disc, thus creating bubbles. Before exposure to the air stream, the plastic disc rotates through the liquid soap (detergent) filled into externally mounted reservoir.

After modification, the step down switching convertor module was added instead of simple basic rectifier supplying 5 V DC, and additional low drop out linear voltage stabilizer provides the 3,3 V DC voltage needed. All components needed for the project are placed on already mentioned carrier board shown also in the Figure 7. The PC type ventilator is now driven by the PWM output, so that fine control over the ventilator speed became possible and it was implemented in software. The remote control is now possible from the near distance and over the Internet using the tablet or Smartphone with custom Android application installed, or via the web based user interface from any web browser.

Figure 6. The exterior of the Bubble Machine

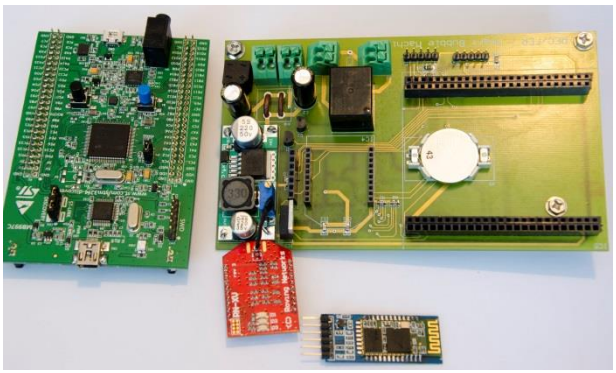


Figure 7. Components added to the Bubble Machine

VII. CONTROL SOFTWARE USER INTERFACE

The software control for the Smart Bubble on the remote side is implemented twice: as the Android application and as the Web application. Both applications enable users to monitor and control the device using simple graphical user interface, as shown in the Figure 8 and the Figure 9. Only parts of the screen are shown here due to the space constraints, and there are connection and help buttons in the bottom of the Android application screen, as well as the console terminal at the bottom of the web user interface screen.

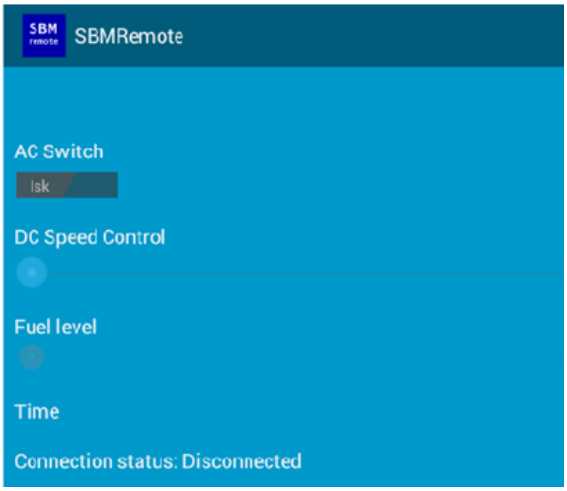


Figure 8. The Android application user interface

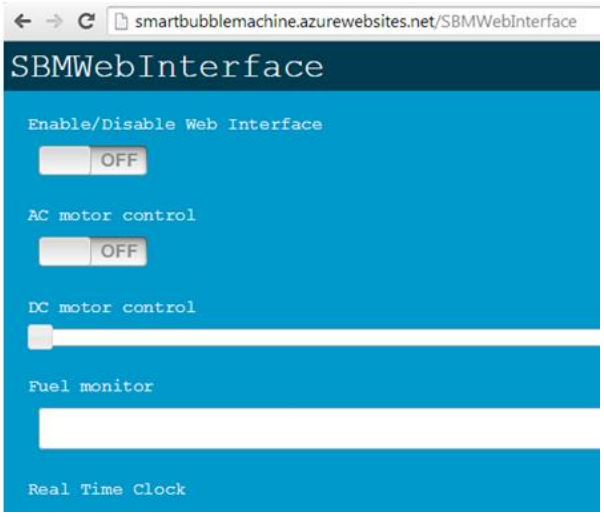


Figure 9. The Web application user interface

The Android application connects to the device using the Bluetooth interface, so that user can change additional parameters and configure device also for the web interface access. It resolves the situation in which the Wi-Fi access has to be configured for new network.

Additionally the Android application offers some debugging features, like the terminal for monitoring transmitted and received data. The terminal application and permission to change the device parameters such as Bluetooth or Wi-Fi connection passphrase are administrator privileged functions.

Parameters which can be changed are divided in three groups:

- Bluetooth connection parameters: Bluetooth module name and passphrase
- Wi-Fi module parameters: Network SSID, WEP encryption, WPA passphrase
- Real-time clock parameters: time and date

Web application is developed using the ASP.NET framework. Application runs on the Microsoft Azure web server, and interacts with the Smart Bubble device using the HTTP GET method. It also enables multiple users to control and exchange data with the device at the same time. This is done using the SignalR notification hubs and jQuery Ajax library. Application has console which is updated every time change is made by any user. Additionally it contains debugging messages which are useful in developing or upgrading the device firmware. Both applications can't be used at the same time, since the Android application is event driven, but device must periodically send http requests to refresh. Therefore the web interface can be disabled using the *Enable/Disable*control shown in the Figure 9.

Furthermore, Android application sends and receives data independently. For example, the real-time clock data is sent every second, while sensor data such as the fluid level sensor (not implemented jet) could be sent when the sensor readings are changed. The web application keeps the TCP connection open and sends data after device requests data, and posts sensor data and real-time clock data to the web application.

V. CONCLUSION

This project is demonstration of basic principles of microcontroller usage to control and monitor device remotely using the wireless connection, actuators such as AC or DC motors, and standard off-the-shelf client device such as the tablet or smart phone.

Possibility is added for the user to remotely turn simple legacy device or machine on and off, to regulate the speed of the ventilator thus influencing the bubble production rate, and it would be easy to program the machine to operate independently and log data during operation.

It is also shown how to implement and use some widely used technologies, such as the Bluetooth, the Wi-Fi, the SD card, the Flash file system, or the RTC. The Real Time Operating System (RTOS) has been applied successfully.

Using the resources and techniques described, any legacy machine can be controlled and monitored remotely from the close range or over the internet using two different wireless options, data from sensors can be collected and sent to the server, new machine configuration can be sent and information stored on the MCU or SD card, and the machine can connect itself to the Internet based on some predefined events.

Therefore, control and monitoring can be performed without being anywhere near the machine, and without usage of any proprietary technology or components.

All modules, voltage convertors, actuator drivers and microcontroller development board itself are mounted on the custom designed PCB, and some more custom boards have been designed where necessary, or feasible.

ACKNOWLEDGMENT

Project sponsored by Research and Engineering Center d.d., Zagreb. The authors would like to thank the RESEARCH AND ENGINEERING CENTER d.d. Zagreb company (REC) for the support and resources provided during this development. The device was displayed at the REC Global stand during 2015 Embedded World Exhibition in Nurnberg, Germany [7], where it was working flawlessly for three days.

REFERENCES

[1] http://en.wikipedia.org/wiki/Internet_of_Things

[2] Wang, K.I.;Salcic, Z.; Atmojo, U.D.; Pediredla, B.; Hadi, M.;Daji, C.:IP-enabled smart sensor and actuator node for ambient intelligence systems, Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on DOI: 10.1109/ISSNIP.2013.6529793

[3] HC-05 Bluetooth Module Overview, ftp://imall.iteadstudio.com/Modules/IM120723009/D_IM120723009.pdf

[4] RN-XV WiFly module, https://www.sparkfun.com/products/10822

[5] FatFs - Generic FAT File System Module, http://elm-chan.org/fsw/ff/00index_e.html

[6] FreeRTOS - Cross Platform Real Time Operating System, http://www.freertos.org

[7] Embedded World Exhibition and Conference, https://www.embedded-world.de/en/

Turning a smartphone into structured light 3D scanner

M. Lelas, T. Pribanić
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
marko.lelas@fer.hr, tomislav.pribanic@fer.hr

Abstract: As they became a very power full hardware platform, available and attractive for a very high number of users, smartphones are receiving more and more attention recently, being an important commercial field but a popular research topic as well. In this paper, we present a method to create dense 3D models of small-scale objects. We facilitate the hardware already available on the smartphones thus turning them into practical 3D scanners without the need of additional hardware. Our algorithm is based on the weak structured light approach, using the capabilities of the HTC Evo 3D smartphone (flash, rear camera). The first reconstruction results are presented as well as the guidelines for the future work. Finally, an approach which should be able to turn every smartphone available on the market into a practical and easy to use 3D scanner is proposed.

I. INTRODUCTION

Generation of accurate 3D models has been a goal in computer vision for some time, resulting in the numerous approaches and methods for achieving flexible and accurate systems for 3D reconstruction. To name just a few of the most successful approaches, there are the active stereo – coded structured light [4], photometric stereo [7], and the passive stereo [5] systems. They all have in common that in each approach the trade-off between system complexity and reconstruction results must be made. In general, the more accurate and reliable the reconstruction results are, the more expensive and processing intensive the system becomes. Depending on the predictable field of the system application and the demand on the level of reconstruction accuracy that the application demands, one or the other approach is applied.

Mobile devices (smartphones and tablets) have strongly evolved in the last decade. This devices offer mobility combined with almost all the convenience that was previously available only with the personal computers (internet access, camera, multimedia, variety of applications…). As a result, new applications expanding the scope of the field in which mobile platforms have been utilized, are developed and the research field is now days very propulsive. The above mentioned was a strong motivation for our work, which has a final goal to provide the mean of turning popular smartphones into practical and easy to use 3D scanner, thus expanding the field of application of smartphones to the field of computer vision - 3D reconstruction. However, some platform-related limitations still remain and the most rigid hardware restriction, common to all mobile platforms, is still the power consumption limitation. In order to provide a long enough period of autonomous life of the device, applications must be power-efficient, which means that both processing and memory requirements of the application must be as low as possible.

This fact turned our research towards investigation of 3D reconstruction methods with lower processing requirements suitable for optimal utilization of the resources available on the smartphones thus leveraging the advantages of the targeted hardware platform. The proposed method is based on the weak structured light approach [1], which is relatively simple method in the means of computing and processing requirements. However, unlike in [1] we do not require an extra light source. We simply introduce smartphone flash for the similar purpose. The result is a novel method for dense 3D reconstruction of small-scale objects which is optimal in the means of efficient utilization of the resources available on the HTC Evo 3D smartphone [9].

This paper is structured as follows. Section 2 gives the insight in the computer vision methods which are in the grounds of our approach and the current state of the art in the field of 3D reconstruction on smartphones. In Section 3 our approach is described in more detail, current shortcomings and the possible improvements in the future work are pointed out. Also the results showing the efficiency of the proposed method are

shown. Section 4 concludes with a summary and an outlook on future work.

II. RELATED WORK

Field of 3D reconstruction on smartphones was all until recently relatively poorly investigated and therefore there are not many works which deal with this topic. Because of the emerging need for new applications and all the commercial potential which this area has, the situation is more and more changing. Klein and Murray [3] have presented the Parallel Tracking and Mapping (PTAM) system on a mobile phone which aims at enabling different augmented reality applications of the mobile phone. The goal to give the users the possibility to create and maintain the Augmented Reality (AR) content without the need for complex and expansive tools has also motivated authors in [2] to design an 3D reconstruction system on the smartphone based on the shape-from-silhouette approach. Won, Lee and Park [8] have presented the work where they implemented an active 3D shape aquisition method based on photometric stereo. They used a pair of smartphones which collaborated as the master and a slave where a slave was engaged to illuminate the scene (using the flash) from the appropriate viewing points while master recorded the images. In [6] authors have presented a complete on-device 3D reconstruction pipeline for mobile monocular hand-held devices, which generates dense 3D models. To make the tracking and mapping process more resilient to rapid motions and to estimate the metric scale of the captured scene, the on-device inertial sensors are utilized.

The base method for our work is the weak structured light approach which is in more detail explained in the following paragraph. The method is simple and not processing intensive. It also utilizes the resources available on most of the modern smartphones (camera, flash) and therefore it seems optimal for the application. A fact that a possibility to turn a smartphone into a practical weak structured light 3D scanner by utilizing the flash has not yet been recognized by a scientific research further motivated our work in this direction.

A. Weak Structured Light

This approach aims at constructing a simple 3D reconstruction system which requires minimal equipment and has low processing requirements but still yields a very good accuracy. The general principle consists of casting a moving shadow with a stick onto the scene, and estimating the three-dimensional shape of the scene from the sequence of images of the deformed shadow. The objective is to extract scene depth at every pixel in the image.

Figure 1: The Weak Structured Light setup [10].

The system setup consists of a point light source, a stick which is used to cast a shadow upon the scene and a camera and a processing system, as can be seen from the Figure 1.

IV. PROPOSED METHOD AND RESULTS

As it can be seen on the Figure 2, the experimental setup for the proposed method is very simple. It takes one L-shaped profile which is in our case made of two orthogonally placed plexiglass boards (otherwise used for the camera calibration), holder for the smartphone, a smartphone and a stick. There are some procedure modifications which make it possible to use just one plexiglass board as a horizontal plane at the cost of having to compute the accurate position of the light source.

The reconstruction process is performed through four main activities. Those activities are: image acquisition, shadow detection, triangulation and output model generation.

Figure 2: The experimental setup.

A. *Image acquisition*

The first step of the reconstruction process is acquisition of a video of the reconstructed scene. The smartphone is placed on the holder with the flash turned on. The video is acquired while the shadow is being casted on the scene by dragging the stick in front of the flash from one side to the other. Once the scene has been scanned with the shadow, the video acquisition is stopped. Then, the consequent frames are extracted from the video and used as an input for the shadow detection step.

B. *Shadow detection*

Shadow detection process is performed in the grey level colour space. The input images (frames extracted from the video) are scanned line by line, identifying the shadow edge pixel in each line. Assuming the ideal pinhole model of the light source, brightness distribution should be uniform throughout the whole image with a big drop in intensity only on the pixels covered by the shadow making the shadow detection process a trivial task. Due to the properties of the scene that is reconstructed and the deviation of the flash from the ideal pinhole light source model, shadow detection process is not trivial. Figure 3 presents one representative input frame and grey scale representation of the corresponding line (the line which is marked red in the input image represented by Figure 3). As it can be seen from the Figure, the pixel intensity distribution is not uniform. Therefore, the shadow is modelled by a parabolic function and the pixels in line which correlate the most with the model are identified as a "shadow pixels" (the pixels marked green in the Figure 3). Parabolic interpolation is than performed at these pixels to identify the shadow edge at sub pixel precision.

Shadow detection process is finished after all the frames of interest have been processed. The goal is to have all the pixels of the reconstructed scene identified as a part of the detected shadow, in one of the analysed frames, as this enables the dense reconstruction because every pixel detected as a shadow pixel on the reconstructed model results in reconstruction of the correspondent image 3D point at the output of the triangulation process.

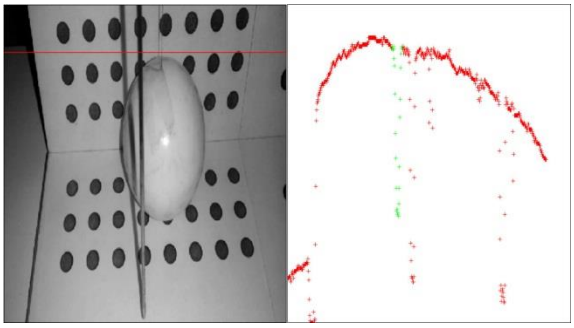


Figure 3: Shadow detection.

C. *Triangulation*

The 3D reconstruction is done by finding the intersection of the back projected line of the "shadow pixel" identified on the reconstructed object with the shadow plane. To identify the shadow plane, the pixels lying in the horizontal and vertical plane are identified. Those pixels are presented in the Figure 4 (green – vertical and blue-horizontal plane).

By fitting the identified pixels to a line, horizontal and vertical lines are reconstructed and the plane defined by those two lines is identified as a shadow plane. The "shadow pixels" which correspond to the reconstructed model (marked red in the Figure 4) are used to calculate the back projected lines to intersect with the calculated shadow plane. The points of the intersections are the 3D representations of the shadow pixels. The proposed method results in a dense 3D reconstruction as every shadow pixel yields a corresponding 3D point.

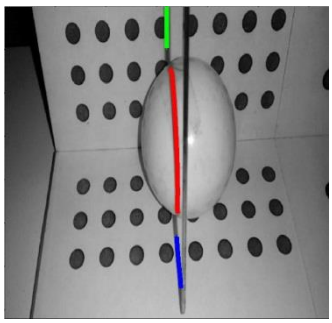


Figure 4: Shadow pixels and shadow plane.

D. Output model generation

The reconstructed 3D points generated by the triangulation process are used as an input to Delaunay triangulation. This way a mash is generated. Fitting the image texture to the mash, the final 3D representation of the reconstructed model is obtained. Figure 5 shows the result of the output generation process for the case of reconstruction of a white ball. As it can be seen a proposed method recovers successfully the circular ball shape. Moreover, the Figure 6 presents the reconstruction of a small angel figure. Even in this more demanding scenario with abrupt changes in depth, we have recovered a realistic 3D shape.

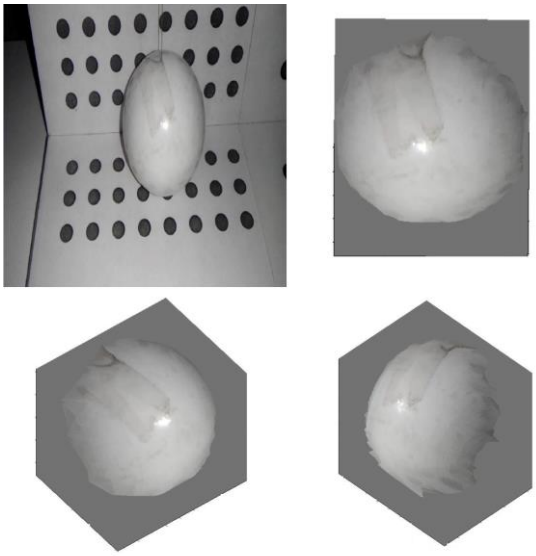


Figure 5: Results of the 3D reconstruction of the white ball. 2D view, front, left and right 3D view representation.

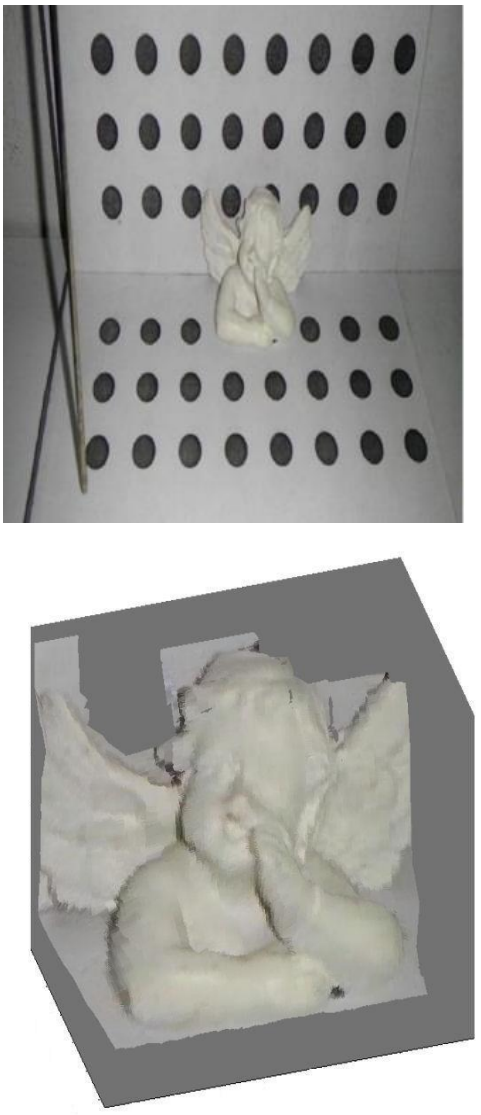


Figure 6: Results of the 3D reconstruction of the figure of angel.

V. SUMARY AND FUTURE WORK

We have shown that the ground idea of this work, turning the smartphone into a practical 3D (weak) structured light scanner using only the hardware available on the device (flesh, camera), is feasible. The presented results show that the system works well for reconstruction of the small-scale objects. The improvements in the shadow detection process, that would make it even more accurate and robust to the conditions of the environment, are in the scope of our future work.

In addition, some smartphones, like HTC Evo 3D, are equipped with two rear facing cameras, thus opening up the space for combining the aforementioned reconstruction approach with the passive stereo. In future work investigating that possibility is planned. There are also smartphones which are not equipped with the flash. For that scenario, we plan to develop practical laser add-on for the smartphone, which would project the laser line that could be used for scanning the scene and performing the 3D reconstruction using the laser line instead of the projected shadow.

REFERENCES

[1] Bouguet, J., 1999. Visual methods for three-dimensional modelling. A Thesis Submitted in partial fulfilment of the requirements of the California Institute of Technology for the degree of Doctor of Philosophy. California Institute of Technology Pasadena, California

[2] Hartl, A., Gruber, L., Arth, C., Hauswiesner, S., Schmalstieg, D., 2011. Rapid reconstruction of small objects on mobile phones. In IEEE Computer Society Computer Vision and Pattern Recognition Workshops *(CVPRW)* June 2011 Pages 20-27. IEEE.

[3] Klein, G., Murray, D., 2007. Parallel Tracking and Mapping for Small AR Workspaces. In 6th IEEE and ACM International Symposium on Mixed and Augmented Reality

[4] Salvi, J., Fernandez, S., Pribanic, T., Llado, X., 2010. A State of the Art in Structured Light Patterns for Surface Profilometry. In Pattern Recognition 43, 2666-2680.

[5] Scharstein, D., Szeliski, R., 2002. A taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. In International journal of computer vision Volume 47 Pages 7-42.

[6] Tanskanen, P., Kolev, K., Meier, L., Camposeco, F., Saurer, O., Pollefeys, M., 2013. Live Metric 3D Reconstruction on Mobile Phones. In IEEE International Conference on Computer Vision (ICCV) December 2013 Pages 65-72. IEEE.

[7] Vlasic, D., Peers, P., Baran, I., Debevec, P., Popović, J., Rusinkiewicz, S., Matusik, W., 2009. Dynamic Shape Capture using Multi-View Photometric Stereo. In Proceedings of ACM SIGGRAPH Asia Volume 28 Issue 5, December 2009. ACM New York.

[8] Won, J. H., Lee, M. H., Park, I. K., 2012. Active 3D shape acquisition using smartphones. In IEEE Computer Society Computer Vision and Pattern Recognition Workshops (CVPRW) June 2012 Pages 29-34. IEEE.

[9] GSMARENA (2014) *GSMARENA: HTC EVO 3D.* Available from: http://www.gsmarena.com/htc_evo_3d-3901.php [Accessed 10 November 2014]

[10] California Institute of Technology *Jean-Yves Bouguet's Web Page*. Available from: http://www.vision.caltech.edu/bouguetj/index.html [Accessed 11 November 2014]

A Heuristics Approach to Hamiltonian Completion Problem (HCP)

Hermina Petric Maretic*, *student, PMF,* and Ante Grbic*, *student, PMF*
*Faculty of Science, University of Zagreb, Zagreb
hpetricmaretic@gmail.com
antegrbic8@gmail.com

***Abstract*—The Hamiltonian completion problem is an NP-hard problem which consists of finding the smallest number of additional edges to make the graph Hamiltonian. This minimal number of extra edges is defined to be the Hamiltonian Completion Number (HCN). This is a well known problem but rarely analyzed in academic circles. We study the problem on random sparse graphs and compare results of determining HCN with different heuristic approaches. We developed a modification of the Ant colony optimization (ACO) and show that, while the standard ACO presents a better approach to this problem than observed Genetic and Immunological algorithms, the modified ACO provides the best results by far.**

I. Introduction

HAMILTONIAN completion problem (*HCP*) is a problem that has been addressed to a limited extent in scientific literature. It has applications in assigning frequencies to multiple transmitters in a geographic area [1] and in distributed computing (assigning processes to processors). Let $G = (V,E)$ be an undirected, non-weighted graph. The Hamiltonian completion problem consists of finding a minimal number of edges $d \in \mathbb{N}$ to add to graph G to make it Hamiltonian. This minimal number of extra edges is also denoted as Hamiltonian Completion Number (HCN). A Hamiltonian graph, also called a Hamilton graph, is a graph possessing a Hamiltonian cycle, a cycle which visits each vertex in the graph exactly once.

The problem was presented in 1970. by Boesch, Chen and McCugh and by Goodman and Hedetniemi independently. The problem is clearly NP-hard in general case (since its solution gives an answer to the NP-complete problem of determining whether a given graph has a Hamiltonian cycle [2]). For some special cases, like trees, cactus graphs and series-parallel [3], there are polynomial time and even linear time [4] algorithms. However, to our knowledge there have still not been found acceptable solutions, including approximate algorithms, in general cases. In 2012. David Gamarnik and Maxim Sviridenko [5] have computed an asymptotic solution to the HCP for random sparse graphs $G = (V,\frac{c}{n})$, where $c < 1$, $n = |V|$ is the number of edges in the graph and each edge is selected independently with probability $\frac{c}{n}$. We will use this class of graphs in our analysis.
This paper will present a detailed overview of different heuristic approaches to solving the HCP and introduce a novel approach, called Twins Ants Colony algorithm. The novel approach is presented in detail and tested thoroughly in comparison with standard approaches. Tests prove the Twins Ants Colony algorithm provides much better results than the observed standard algorithms.

The paper is organized as follows. In Section II, we explain the genetic and immunological algorithm approaches to the problem. Section III focuses on the Ant colony optimization approach, with a special accent on the original Twins ant colony optimization algorithm and its justification. Section IV describes the testing procedures and explains the results. Section V provides concluding remarks and directions for future work.

II. Genetic and immunological algorithm approach

Genetic and immunological algorithms are used widely in approaching NP-hard problems. These algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural phenomena. In this section, we give a short overview of basic genetic and immunological algorithm approaches as applied to the HCP.

A. Genetic algorithm

The solution representation is a permutation of vertices without the vertex 1, which we use as the first vertex by default. Setting one vertex at a fixed position simplifies the implementation of crossover operators. It is important to emphasize that this does not reduce the number of solutions, i.e. each solution is a cycle and the "first" position doesn't exist in a cycle, so fixing one vertex only prevents different rotations of the same cycle to appear as different solutions.
For the termination condition, we used a predefined number of generations so that the algorithm converges.
The algorithm begins with a randomly generated population, with the size of $numIndividuals$:

B. Genetic algorithm with a greedy operator

A possible improvement can be achieved with a greedy operator similar to those often used in solving the Traveling

Data: Randomly generated population of individuals
Result: Population from which we take the one with least edges added
while *(termination condition is not reached)* **do**
1) Give scores to individuals.
2) Select two best rated individuals and store them in $newPopulation$. Use the Roulette wheel selection to select $numIndividuals - 2$ and store them in $newPopulation$.
3) Use the crossover operator, as proposed in [6], to cross selected individuals.
4) Use inversion mutation to inverse individuals in $newPopulation$.
5) Possibility of mutation decreases in time.
6) Replace initial population with $newPopulation$. Repeat until the termination condition is reached.

end

Algorithm 1: Genetic algorithm with a crossover operator

salesman problem. The main idea is that edges appearing in both parents are more likely to be advantageous. Thus, the operator keeps the intersection of two parent cycles and sets it as the basis for their children, while the rest of the children edges are generated randomly. Unfortunately, this variant didn't provide a significant improvement and, for that reason, we have omitted it from further discussion and test results.

C. Immunological algorithm

The solution representation and the termination condition are similar to those already described in the genetic algorithm. The algorithm begins with a randomly generated population, with the size of d:

Data: Randomly generated population of size d
Result: Population from which we take the one with least edges added
while *(termination condition is not reached)* **do**
1) Clone every antibody k times (we now have $(k+1) \times d$ antibodies).
2) Each clone undergoes the process of hypermutation.
3) Count affinity of old antibodies and hypermutated clones. Build the new population out of d best individuals.

end

Algorithm 2: Immunological algorithm

As it will be shown later on, both of these algorithms have had difficulties finding optimal solutions in bigger graphs, i.e. graphs with a 100 vertices or more. There is a naturally imposed improvement which consists of not generating our initial population randomly, but in an educated fashion, e.g. using the Twins Ant Colony algorithm to create an initial population of solutions.

III. Ant Colony Optimization approach

A. Ant colony algorithm

Ant colony optimization is an algorithm inspired by natural behavior of ants. Ants, while searching for food, travel through a terrain (graph) spreading on different paths and leave traces of pheromones on the paths that prove successful or possibly lead to a source of food. After a while, they all "converge" to the same path. We can observe HCP accordingly and notice that since the solution is represented as a cycle, every vertex is a good enough starting point for ants. Thus, all ants start at vertex 1 and follow the algorithm:

Result: Cycle with the least HCN from all the ants who participated in the algorithm
while *(termination condition is not reached)* **do**
1) Every ant finds a path:
   while *(the path is not complete)* **do**
   a) Let $V = \{v_{i1}, ..., v_{in}\}$ be the set of all unvisited neighbors of the vertex v_i.
   b) If $V \neq \emptyset$, connect to a neighbor j with probability ${}^{c_j}/_{\sum_{k \in V} c_k}$, where c_k is the strength of pheromone scent on the edge (v_i, v_{ik}).
   c) If $V = \emptyset$, randomly choose a vertex from all unvisited vertices and connect.
   end
2) Compare results for all ants.
3) The best ant leaves its pheromones on all edges that it used and that exist in the graph.
4) Pheromones evaporate in the entire graph.

end

Algorithm 3: Ant colony optimization algorithm for HCP

B. Twins Ant colony optimization algorithm

A clear flaw of the ACO approach to this problem is that it relies heavily on randomness. Next algorithm, as a greedy modification of the standard ant colony algorithm, represents a novel approach to solving HCP for sparse graphs. The basic idea is that every ant comes with a twin ant which helps construct an optimal cycle by constructing the cycle a from a series of smaller paths. The first ant is greedy, he starts from the lowest degree vertex connects to vertices using the same probabilities as before. When he stops, the twin ant tries to "fill in" the path that his brother has created, making sure no new connected components have emerged. It can be shown that the twin ant indeed justifies the greedy approach.
In the following algorithm description, the word "team" is used to denote both ants (brothers) which are building one path:

Proposition. *The twin ant justifies the greediness of the algorithm.*

Proof. Let $G = (V, E)$ be an arbitrary graph and $x = (x_1, e_1 ..., x_n, e_n)$ an optimal cycle, where $\{x_1, ..., x_n\}$ are vertices and $\{e_1, ..., e_n\}$ edges in the cycle.

We define $E^{'} := E \cap \{e_1, ..., e_n\}$ and observe connected components $C_1, ..., C_m$ in graph $G' = (V, E')$. It is clear that

Result: Cycle with the least HCN from all the teams who participated in the algorithm

while *(termination condition is not reached)* **do**

1) Every team finds a path:
2) Start from a lowest degree vertex and make it the base.
   while *(the path is not complete)* **do**
   a) Let $V = \{v_{i1}, ..., v_{in}\}$ be the set of all unvisited neighbors of the current vertex i.
   b) If $V \neq \emptyset$, connect to a neighbor j with probability $c_j / \sum_{k \in V} c_k$, where c_k is the strength of pheromone scent on the edge (i, v_{ik}).
   c) If $V = \emptyset$, stop at current vertex i_1. The twin searches for more edges (as in step b), starting from the base.
   d) When the the twin stops at a vertex, let's call it i_2, add the path from i_1 to i_2 to the cycle under construction (a).
   e) The team jumps to an unvisited vertex with the lowest degree. This is done by adding a new, nonexisting edge to the cycle. This is the new base.
   end
3) Compare results for all teams.
4) The best team leaves its pheromones on all edges that it used and that exist in the graph.
5) Pheromones evaporate in the entire graph.

end

Algorithm 4: Twins Ant colony optimization for HCP

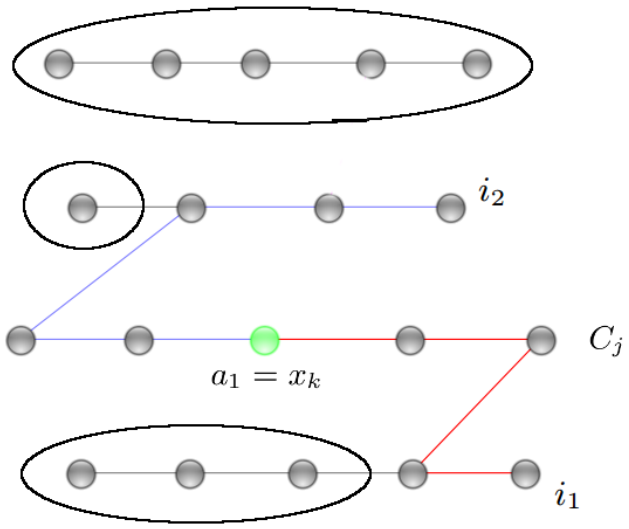


Fig. 1: One step in the Twins ant colony algorithm. There are four connected components from the modified graph $E^{'}$, where C_j is the third one, and each component is represented by one line. Grey edges present edges in $E^{'}$, red ones present edges passed by Ant 1 and blue edges present edges passed by the Twin. All of those edges exist in the original graph E. Ant 1 starts at $a_1 \in C_j$ and goes through red edges. The twin then starts at a_1 and goes through blue edges. After this step, the path from i_1 to i_2 (red and blue edges) will be added to the cycle a. We can see that after that, the entire component $C_j \subseteq a$ and there is the same number (4) of unconnected components (the current one and the 3 circled ones) as before the ants passed, in the graph $E^{'}$.

every component C_i is a linear graph. We will show that the algorithm can construct a cycle $a = (a_1, ..., a_n)$ for which $HCN(a) \leq HCN(x)$.

Let $a_1 = x_k$ be a vertex that belongs to $C_j = (x_{k-s}, ..., x_k, ..., x_{k+l})$ and we set $HCN(a) = 0$. Now Ant 1 can choose the vertices towards x_{k+l}: $x_{k+1}, ..., x_{k+l}$. If Ant 1 has unvisited neighbours left, it connects to the vertex x_{k_j} in component $C_{j_1} = (x_{k_{j_1}-s_{j_1}}, ..., x_{k_{j_1}}, ..., x_{k_{j_1}+l_{j_1}})$. At this point, the ant can choose vertices towards one end of the component, $x_{k_{j_1}-1}, ..., x_{k_{j_1}-s_{j_1}}$.

It is important to see that this has not increased the number of connected components N_c. Namely, the set of the vertices that we haven't included in a, $\{x_{k_{j_1}+1}, ..., x_{k_{j_1}+l_{j_1}}\} =: C'_{j_1}$, makes at most one connected component, which replaces C_{j_1}. Let us notice that this point is valid also in the case where $C_{j_1} = C_j$. In that case $x_{k_j} = x_{k-r}$, where $x_{k-r} \in \{x_{k-s}, ..., x_{k-1}\}$. Following the same procedure as above, the set of vertices not included in a will be $\{x_{k-r+1}, ..., x_{k-1}\} =: C'_j$, which makes at most one connected component, while replacing C_j.

Once again, if the ant can move forward, it connects to a vertex in another component and repeats the same procedure. When Ant 1 has no more unvisited neighbors to connect to, Ant 2 starts from the base x_k.

Ant 2 can now connect to $x_{k-1}, ..., x_{k-s}$. Doing this, we have included the entire component C_j to the path a. At this point, Ant 1 either can't connect to new vertices, or it connects to a vertex in another component and continues with the similar procedure as Ant 1.

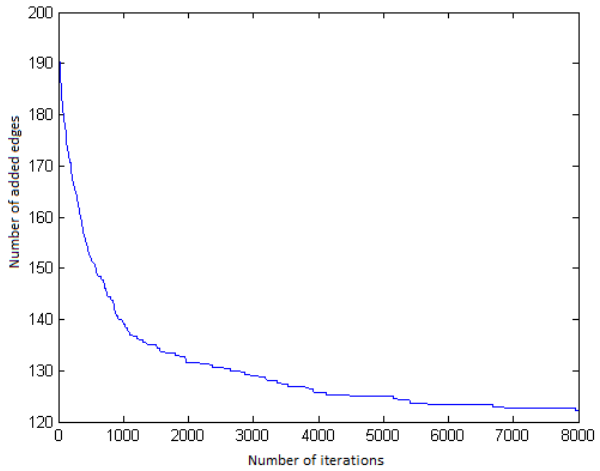
When Ant 2 stops, $C_j \subseteq a$, and $N_c \leq m$. Adding a new edge $e \notin E$, we connect to component C_i which increases $HCN(a)$, $HCN(a) = HCN(a) + 1$ and decreases N_c, $N_c = N_c - 1$. Analog to the last step, at the end of this one, we will have $C_i \subseteq a$, and $N_c \leq m - 1$.

Continuing the procedure inductively, in at most $m - 1$ steps, we will have $HCN(a) \leq m - 1$, $N_c = 1$ and $V \subseteq a$. Adding one more edge, we'll have created a Hamiltonian cycle with $HCN(a) \leq m$. But, $HCN(x) = |C_1, ..., C_m| = m$, so $HCN(a) \leq HCN(x)$.

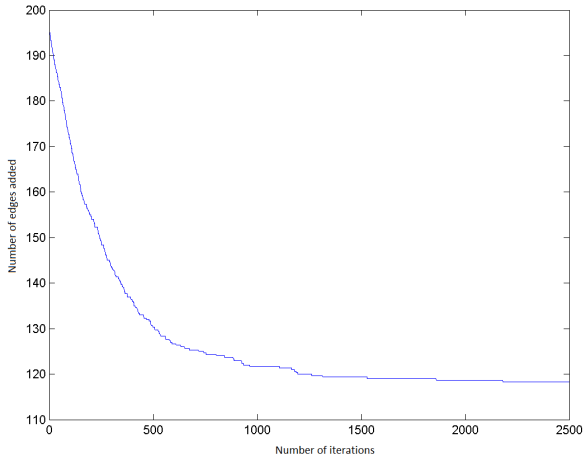
□

IV. Testing and Results

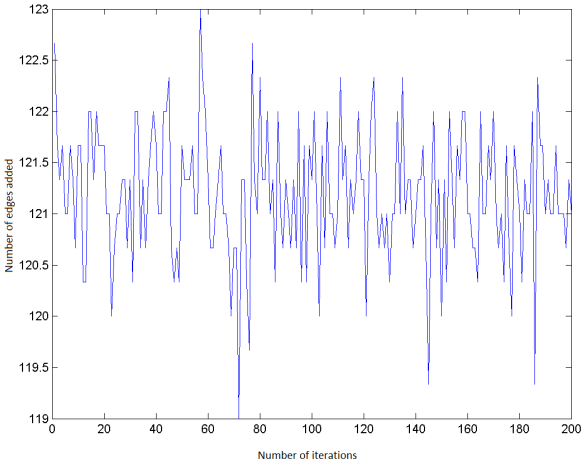
As we mentioned earlier - in our testing we have considered only graphs $G = (V, \frac{1}{n})$, where $\frac{1}{n}$ denotes the possibility that an edge will be included in a graph or not. We can see that this sort of graph is actually an Erdős-Rényi random model graph [7]. We have implemented all of our algorithms and GUI in MATLAB®. Graphs have been stored as incidence matrices. Also, we used compressed sparse format for storing graphs. We used operations for I/O with sparse matrices from



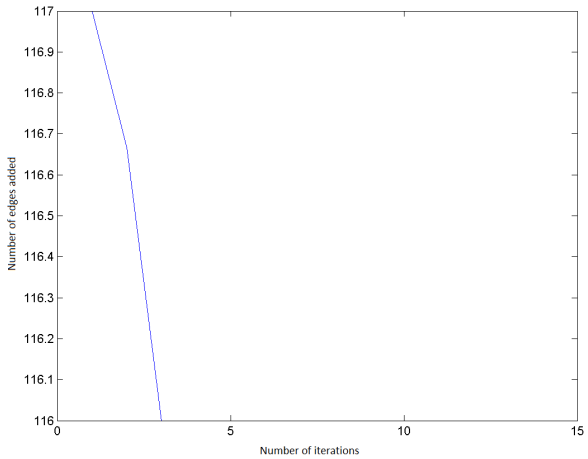
(a) Genetic algorithm, duration: 40.01 sec, best solution: 122.233



(b) Immunological algorithm, duration: 100.01 sec, best solution: 118.33



(c) Ant colony algorithm, duration: 24.14 sec, best solution: 118



(d) Twins Ant colony algorithm, duration: 1.17 sec, best solution: 116

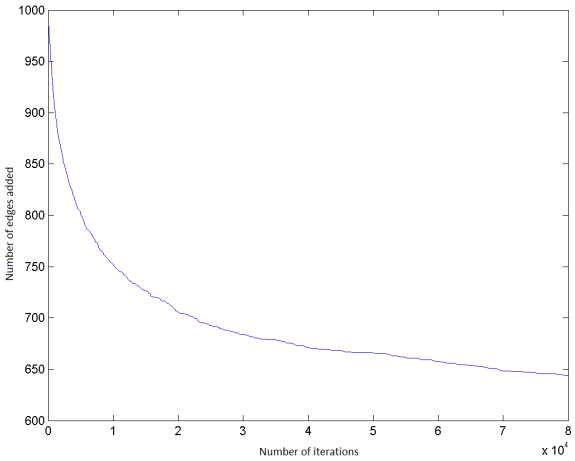
Fig. 2: The testing results for randomly generated graphs with 200 vertices

http://math.nist.gov/MatrixMarket/mmio/matlab/.

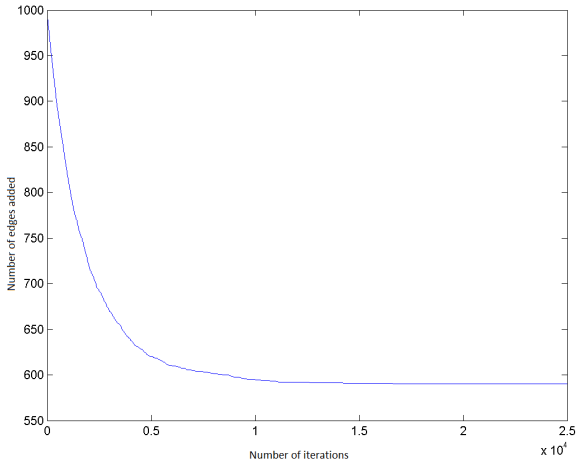
Algorithms have been tested on 30 instances of randomly generated graphs of order 200 (results on Figure 2) and on instances of order 1000 (results on Figure 3). Each testing has been repeated three times with the same parameters and each result represents the average of all three tested cases. Figure 4 shows the results of all 4 algorithms on 10 different graphs in a column-fashion. Obviously, the Twins Ant Colony algorithm achieves best results. The computer that was used for computation had a Intel©Core i5 2,3 GHz processor and 6GB RAM.

A. Some remarks on results and testing

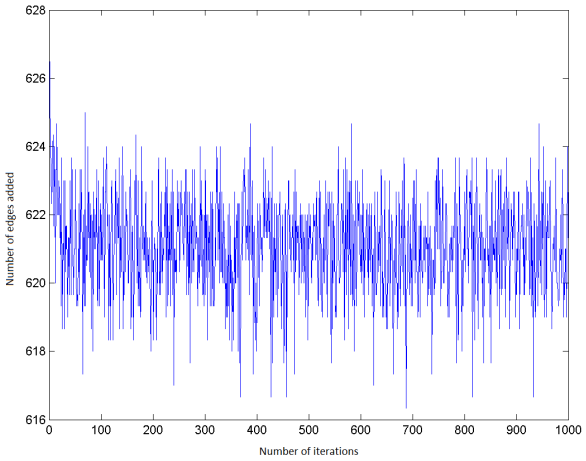
- The Genetic algorithm gave disappointing results, not converging for a fairly long time and not providing satisfying solutions as presented in Figure 2a and Figure 3a. Various experimenting with parameters and time vs. result analysis resulted in concluding the best elitism parameter is 8 (of 64).
- The Immunological algorithm seems to approach this problem significantly more successfully than the Genetic algorithm and provides fairly acceptable solutions, but is still not time feasible. Results can be seen on Figure 2b and Figure 3b.
- The Ant colony algorithm gives acceptable solutions when tested on smaller graphs compared to other algorithms (Figure 2c), but is not very successful in bigger graphs (Figure 3c). The obvious solution divergence is a result of the excessive randomness used in the algorithm, which we have discussed before.
- The Twins Ant colony algorithm provides best solutions in both smaller and bigger graphs (see Figure 2d and Figure 3d). Moreover, it proves to be the fastest out of the tested algorithms by far. As the testing was conducted on randomly generated graphs, we are unable to confirm the optimality of the Twins Ant colony algorithm solutions.



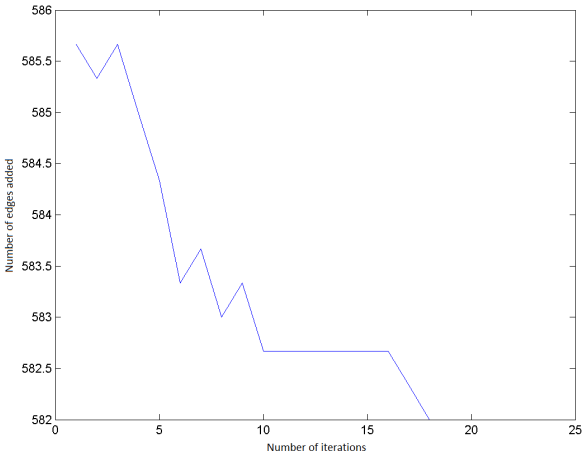
(a) Genetic algorithm, duration: 5136.6 sec, best solution: 664



(b) Immunological algorithm, duration: 3045.6 sec, best solution: 590



(c) Ant colony algorithm, duration: 2399.5 sec, best solution: 611



(d) Twins Ant colony algorithm, duration: 24.51 sec, best solution: 582

Fig. 3: The testing results for randomly generated graphs with 1000 vertices.

Nevertheless, as all other algorithms seem to converge to solutions with equal or greater Hamiltonian completion number, we can be optimistic about the nature of our solution.

V. Conclusion

Genetic algorithm proved to be too slow and imprecise for approaching this problem. Immunological algorithm has shown an improvement in terms of performance and considerable improvement in precision. Ant colony algorithm proved to be better than aforementioned algorithms but the real improvement is achieved in the Twins Ant colony algorithm. Adding the greedy component to the algorithm improves accuracy and reduces the time required to achieve convergence by narrowing the solution space. The twin ant ensures that an optimal solution persists in the solution space. It would be interesting to see how this approach could be applied to similar heuristics used in solving NP-hard problems.

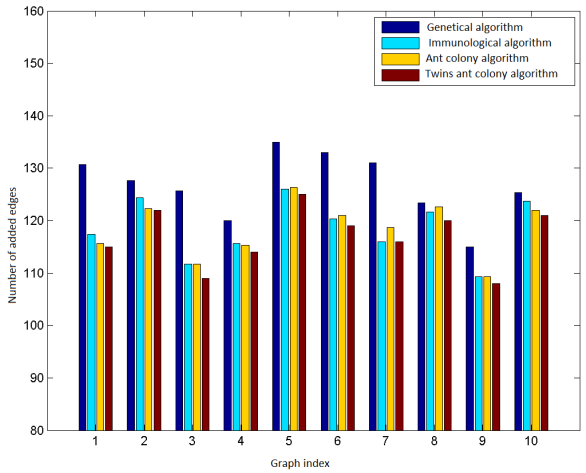


Fig. 4: Final solutions of algorithms when tested on 10 different graphs with 200 vertices. The x-axis is labeled ”Graph index” and y-axis is ”Number of added edges”. We can see that the best results are achieved with the algorithm.

REFERENCES

[1] D. Franzblau and A. Raychaudhuri, "Optimal hamiltonian completions and path covers for trees, and a reduction to maximum flow," Tech. Rep. 02, 2002.

[2] R. G. Michael and S. J. David, "Computers and intractability: a guide to the theory of np-completeness," *WH Freeman & Co., San Francisco*, 1979.

[3] K. Takamizawa, T. Nishizeki, and N. Saito, "Linear-time computability of combinatorial problems on series-parallel graphs," New York, NY, USA, Tech. Rep. 3, Jul. 1982. [Online]. Available: http://doi.acm.org/10.1145/322326.322328

[4] A. Agnetis, P. Detti, C. Meloni, and D. Pacciarelli, "A linear algorithm for the hamiltonian completion number of the line graph of a tree," *Information Processing Letters*, vol. 79, no. 1, pp. 17–24, May 2001. [Online]. Available: http://dx.doi.org/10.1016/S0020-0190(00)00164-2

[5] D. Gamarnik and M. Sviridenko, "Hamiltonian completions of sparse random graphs," *Discrete applied mathematics*, vol. 152, no. 1, pp. 139–158, 2005.

[6] L. Davis, "Job shop scheduling with genetic algorithms," Hillsdale, NJ, USA, Tech. Rep., 1985. [Online]. Available: http://dl.acm.org/citation.cfm?id=645511.657084

[7] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

Security of End User Wireless Networks in Zagreb Area

Jasmin Redžepagić*, Danijel Studen *, Vladimir Gavranić*, Alberto Teković**
* University College for Applied Computer Engineering
** VIPNet d.o.o.
danijel.studen@racunarstvo.hr

Abstract - Every network has its operator. The goal for this paper was to show how much configuration was performed by the operators after wireless networks were put in place. We have used a method called war driving to map out the locations and settings being broadcasted from the multitude of Wi-Fi devices in the Zagreb area. The main driver was to point out how much of the scanned devices are left at their factory settings and the security consequences that might follow should they remain as such.

A. Adoption and availability of Internet in Croatia

According to Eurostat, in data released for 2014, 69% of Croatian domestic population has logged on to the Internet during the past three months preceding the poll. European 28 country average is 78%, and Croatia is in the lower third of the table, according to the results [1].

Broadband penetration is also below European average. Only 68% of households have broadband Internet [2]. Although this is below average, broadband can be considered widespread when talking about urban areas. We base this assumption on relatively slow adoption of broadband in rural areas, which either have no access due to lack of infrastructure, or are unable to afford high speed Internet or any Internet connection at all.

For an average user, high speed Internet is readily available, at moderate cost.

B. Technologies used

Today almost all devices use some form of wireless communication technology. Technology most commonly used is Wi-Fi.

By itself, Wi-Fi [3] is trademark by Wi-Fi Alliance, and is used to denote equipment that is fully compliant with standard. Currently, it covers IEEE 802.11a, and IEEE 802.11b/g/n standards [4].

C. Encryption standards

In context of wireless security, we are dealing with three distinct encryption standards, all of them defined on 802.11 networks.

In order of protection provided, they are WEP, WPA and WPA2. Network can also be configured as unprotected, but this is strongly discouraged.

WEP is a protocol that is very easy to bypass. There are numerous very detailed documents explaining and showing step by step how to obtain the required password for the network intrusion [5]. This is a major problem because all exploits are well documented and no previous knowledge about how or why it works is needed. Given a set of tools and specific instructions it takes only a few minutes to circumvent WEP. Once the required password is successfully acquired, and connection to the network is established by obtaining an IP address, access to all the available resources in the network is possible. Given the combination of high success rate and public knowledge this is the most commonly used attack vector in the wireless networks area. The ISPs claim that they keep the devices operating on WEP for "backwards compatibility" but since WEP is considered obsolete[1], and all the devices available on the market today are capable of supporting higher grade encryption, this encryption method should not be used at all.

WPA is a standard designed to replace WEP on same hardware, or to be more precise, to enable old devices running WEP to be more secure. This was a major requirement, since there were a huge number of devices already on the market that had to be supported. Keys that are used are 128 bits, although it is possible to use 256bits keys they are not usually configured. WPA, sometimes also called TSN (Transition Security Network) uses RC4 standard for encryption.

WPA2 is a different beast and is currently the standard used in wireless encryption. It uses AES for encryption, a very strong algorithm enabling high security. Pre-shared key (PSK) configuration is the most common in both personal and enterprise users.

I. WHAT IS SECURITY?

As defined by Institute for Security and Open Methodology [6], Security is "a form of protection where a separation is created between the assets and the threat. This includes but is not limited to the elimination of either the asset or the threat. In order to be secure, the asset is removed from the threat or the threat is removed from the asset. "

Safety is defined as "A form of protection where the threat or its effects are controlled. In order to be safe, the controls must be in place to assure the threat itself or the

[1] WEP is officially retired since 2004 by Wi-Fi Alliance.

effects of the threat are minimized to an acceptable level by the asset owner or manager".

Using these definitions as guidelines, we decided to define that end-user networks are assets needing protection, and that a threat is any third party not explicitly authorized to use the network.

Threats can take many forms, we focused only on those coming from "outside" through wireless networks.

Looking from the security perspective networks can be separated into different groups.

Unsecure networks are those that have no configured encryption, and allow unsecure logon. Even from an average user standpoint they are least usable, since all the modern operating systems by default warn the user if he tries to connect to such a network. Leaving a network in such a state means that all information on the network is available, and all network shares are visible to any computer, unless configured otherwise.

Absolutely no wireless network should be left open unless we intend to provide open access to the network, but even then other ways of protecting the information on the network should be implemented.

Weak-security networks are currently the biggest problem if you look at them from a defense standpoint. In this context, our primary concern is with WEP enabled networks, since they use protocol that was proven unsecure a long time ago. Its insecurity is so deep that all sources dealing with wireless security recommend avoiding using WEP. Exploits and tools to crack and avoid WEP are available for years, and a dedicated attacker can crack and use a WEP protected network in a matter of minutes.

Networks that use WPA provide stronger security, since availability of dedicated cracking applications is limited [7, 8]. There are known ways to crack a WPA protected network, but they are extremely slow, and require advanced tools and knowledge outside of a casual hacker. WPA is mainly used by older equipment, and is hardware compatible with WEP devices. All new devices support both WPA and WPA2.

Strongest networks are those utilizing WPA2. Although there are rumors that WPA2 can be hacked, there are no known vulnerabilities that will provide entrance other than brute force types of attack, and even those are extremely complicated due to the nature of the algorithm.

II. EQUIPMENT USED

For the data capture process we used a Linux system running Kismet [9], a popular application for wireless network scanning.

System was installed on a laptop, configured on a mobile platform, a car that was deployed across the city of Zagreb.

For time and geo referencing we used an external GPS receiver connected to the laptop. Receiver was made by Garmin, and provided precise data on each scanned network.

External wireless card made by Alpha networks, model AWUS036NHR has been used. Card uses USB connection, and has an external antenna connector that was used to connect omnidirectional antenna with 8dBi gain. Antenna was placed inside the vehicle, next to the rear window, away from any source of interference. This has assured enough flexibility to emulate real life scenario of a user with a laptop trying to connect to a wireless network.

Kismet was run as a server, with visual interface running only to verify correct operation while (during) scanning.

Results are provided as an xml file. Kismet is capable of doing packet capture, and even running an active scan against a given network. However, for the purpose of this paper only passive methods of scanning have been used.

For a given network Kismet logged start of signal acquisition, end of signal acquisition, and data on signal power at min, max, and peak values. Location for each measurement was also logged.

From the packets captured, Kismet logs all the data it can extract - frequency of signal carrier, channel used, wireless standard used, and type and availability of security protocols used.

III. IDENTIFYING EQUIPMENT

Identifying network equipment used by a specific network is very simple with wireless devices, the same as all the rest of the network devices. Each device has a unique MAC (media access control) address. With a few clicks on the Internet it is possible to accurately find the model and manufacturer of any device given its MAC address. The easiest way is to scan through the MAC databases to check the vendor ID.

Once the device is sold to the ISP, the ISP pre-configures IP address on the device as well as the encryption method. With a little research, it is possible to identify the ISP using the pre-configured name. Sometimes the name itself is obvious as the ISP uses the name as a sort of free advertisement.

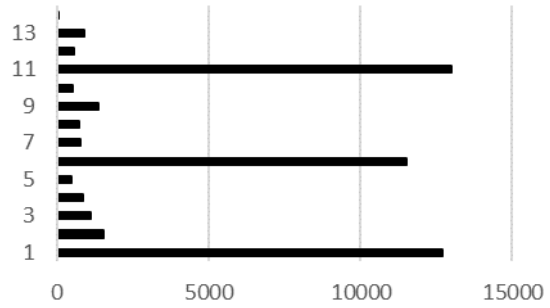


Figure 1. Distribution of Wi-Fi Channels used by channel number / number of networks using it

Once we have all the data, rest is simple. We can access the Internet and find default values for the username and password, which in turn gives us a good

probability of logging in to the network and taking it over and this is where wireless security comes into play.

From the end user standpoint, any wireless security mechanism can be seen as a way to authorize users to connect to the network itself and to ensure that communication is secured and not visible by third parties.

IV. ANALYZING DATA

Collected data has been stored in xml files, and evaluated offline, after the scan. All in all, we had around more than 700 GB of raw text data consisting of XML formatted files. Since XML is difficult to work with, we used a utility, GisKismet that is able to read Kismet formatted xml files, and transfer them to a much more usable SQL database. Data format that this particular utility uses is called SQLLite3, and is basically all in one SQL database that does not need a server, and can be used on multiple platforms. By running SQL queries, it is possible to analyze data, and even output data as distinct geo referenced points on a map, enabling us to visualize data.

For further data visualization through charts, Microsoft Excel has been used to prepare the data, and Google Earth being used to visualize georeferenced data.

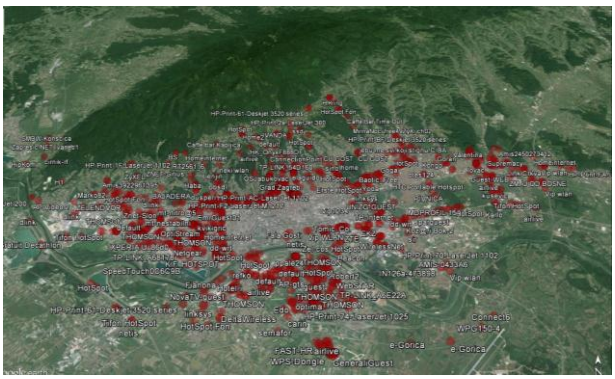


Figure 2 visualizing the data within Google Earth

Due to a huge number of individual data points, using Google Earth for this task is difficult, and extremely slow. Georeferenced data is on the other hand only needed for targeted attacks on a given network, but we logged it anyway, to ensure that all the data we collected can be verified, and to enable extraction of data subsets for latter analysis. But once the data is filtered, Google Earth visualization becomes invaluable tool in any analysis.

In total, 46332 individual networks, or network devices that identified themselves as individual networks have been scanned. It is possible that some of them are just parts of some larger network, and that by itself is another possible security problem.

Out of those networks, 1027 were configured as "cloaked". Cloaking is a security technique that hides SSID (service set identifier) from casual scanning, and can stop network from being visible on some devices. In reality, this poses almost no protection at all, since there are multiple ways that network information can be found [10]. It has to be noted that, after all, we managed to locate and note all those networks while in reality only managing to grab a few frames of useful data from any given network.

Table 1 depicts all the encryption methods used by individual networks. Each network can use more than one method of encryption, in order to accommodate different devices with different encryption capabilities.

TABLE I. ENCRYPTION TYPES

Encryption	Number	Percentage
None	1079	2,33%
WEP	7407	15,99%
WEP40	1	0,00%
WPA Migration Mode WEP104 WPA+TKIP	6	0,01%
WPA Migration Mode WEP104 WPA+TKIP WPA+PSK WPA+AES-CCM	1	0,00%
WPA Migration Mode WEP40 WPA+TKIP WPA+PSK	1	0,00%
WPA+AES-CCM	285	0,62%
WPA+PSK WPA+AES-CCM	12687	27,38%
WPA+TKIP	33	0,07%
WPA+TKIP WPA+AES-CCM	126	0,27%
WPA+TKIP WPA+PSK	9971	21,52%
WPA+TKIP WPA+PSK WPA+AES-CCM	14735	31,80%

Since network encryption is negotiated between the device and the network itself, we classified each network by its least secure network standard. In this context, WEP40 can be classified as "normal" WEP, as well as WEP104. WPA+AES-CCM denotes WPA2, since it uses most secure CCM standard, and AES encryption.

When broken down this means that most of the networks actually use two standards, WPA and WPA2, at the same time.

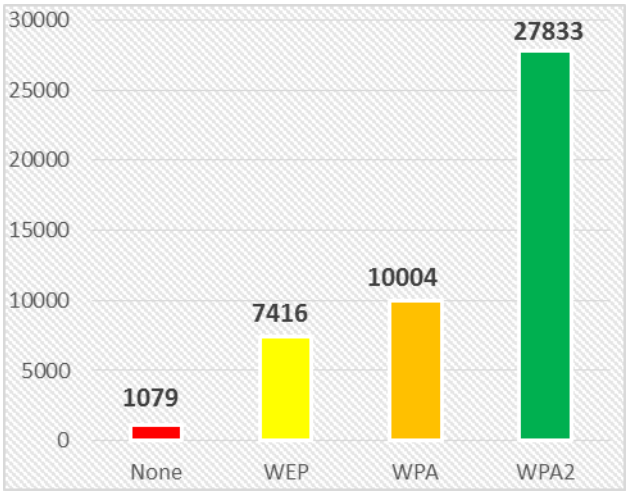


Figure 3 Distribution of network encryption based on our sample

As can be seen from the data collected, around 18% of networks use no encryption at all, or use WEP. We consider these networks as unsecure. Since these types of networks made relatively large percentage of our data sample, this points to a major problem. Users are configuring security settings that will serve only to deter casual hackers.

More than 50% of the networks use WPA2, which is good, but they are configured to allow both WPA and WPA2 connections. This means that device that connects to the network has to make sure to use stronger encryption algorithm.

One more important point to take into account are network names. All the ISPs in Croatia use similar ways of naming their networks, almost all of them using some kind of prefix, and appending a unique identifier to the name. Taking that into account, occurrences of large sets of similar names in the dataset point to default network names being used. All of this means that end users rarely change any settings on their equipment, even if it means using sometimes unusual and inconveniently complicated network names and passwords assigned by their operator.

V. ARE THE NETWORKS SECURE?

Most, if not all of the equipment we scanned during our data collection was sold and deployed by ISPs. Only devices that did not fit that description were printers, a couple of similar end user devices (Access points for extending coverage), and professional devices used in enterprise and similar networks. One notable mention is Eduroam network spanning University of Zagreb locations, open to all the students, and utilizing advanced data protection and traffic monitoring way beyond the scope of this paper. During the scan we also noticed a small number of point to point links, but their relatively small number (a couple of hundred) compared to the data sample enables us to ignore them as statistically insignificant.

VI. CONCLUSION

Most of the network equipment that was encountered during the scan was sold or leased from local telecom operators, as was concluded by network names and MAC addresses used. Network configuration follows similar pattern, newer network devices being configured with stronger encryption than the older one, probably due to large part of configuration being done from the operators' side, not involving end users.

A lot of networks still use weak encryption or no encryption at all, but the situation is improving as new equipment rolls out.

Equipment that uses Wi-Fi technology but comes preconfigured without any encryption poses a large problem for the security.

All this points out to a dire need for large scale campaign to educate end users on benefits and problems that new technology such as Wi-Fi has.

Another point is that right now operators more or less have all the control over end users network, since the cannot rely on users being in control of their own networks. This should be fixed by education, since users need to be able to at least recognize possible threats to their network and their data, and be able to take appropriate actions.

REFERENCES

[1] http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_ci_id_h&lang=en

[2] http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=isoc_ci_id_h&lang=en

[3] http://windows.microsoft.com/is-is/windows/what-is-internet-service-provider#1TC=windows-7

[4] http://standards.ieee.org/findstds/standard/802.11-2012.html

[5] http://www.wikihow.com/Break-WEP-Encryption

[6] OSTMM v3 – page 22, http://www.isecom.org/research/osstmm.html

[7] http://null-byte.wonderhowto.com/how-to/hack-wi-fi-cracking-wpa2-psk-passwords-using-aircrack-ng-0148366/

[8] https://www.youtube.com/watch?v=knllpZF508k

[9] www.kismet.org

[10] http://nirsoft.net/utils/wifi_information_view.html

Learning from student thesis products: examples and experiences

M. Vučićević, S. Predanić, T. Horvat
University of applied sciences, Department of Electrical Engineering, Zagreb, Croatia
martina.vucicevic@tvz.hr; spredanic@tvz.hr; thorvat3@tvz.hr

Abstract - Products life cycle goes through several different development stages. One of the tasks set before engineers is to start from an idea and build it up all to the final product. This process is taught to students through seminars and graduation thesis. In education environment it is possible, and desirable, to divide the process so that students can easily acquire necessary knowledge and skills. Experiences at Department of Electrical Engineering of University of applied sciences proved that electrical and electronic products resulting of graduation thesis are also useful in the teaching process of current students, either by direct use in laboratory work or by student work on product enhancement. This paper will present student products and applications implemented in classroom, and their possible improvements.

I. INTRODUCTION

In the education process student learning is often divided between being a passive listener to new knowledge in ex-cathedra lectures, and a hands-on active work during various types of work like laboratory exercises or problem solving in the classroom. Most often the active work is preferred as students can remember just 10% of what they see, but more than 90% of what they do. So, at the University of applied sciences in Zagreb, several different approaches were taken on the problem of making an active classroom. From making special circuits and models for use in laboratory, to using professional software for simulating real electronic projects.

That approach to learning proved successful and is currently the way which most teachers take as the best one. But making models and examples is often time consuming tor the teachers, while at the same time it constrains student creativity. A different approach of giving the students the opportunity of making teaching materials proved as a better system in both engaging the students and giving them a feeling of pride and independence.

In this article a few examples of previous teaching models and examples will be described, and a few of student projects will be shown. A continuous evolve of the projects is possible due to knowledge sharing between student generations, with teacher as a mentor and a facilitator in giving access to knowledge rather than boring teaching from afar.

II. CURRENT MODELS

Current models and examples for teaching in which students are engaged in a continuous work can be summarized in two different categories:

- software elements in which a simulation or planning of a real system is made
- hardware elements in which students can engage themselves in a tactile work with the elements and the surroundings (additional tools like measuring equipment or just trying to "break" things).

Although the software systems have an advantage of being democratic, as each student can take the software home and install it for home usage (e.g. if they are more interested and involved), the end result sometimes feel empty due to its virtual nature. Such extensively used systems is Labcenter Proteus, which is often used in courses of digital electronics and microcontrollers. [1] This software allows the user in rapid prototyping and designing a circuit, in making firmware for such circuits and to simulate not only the student made system, but to simulate various tools which would normally take a lot of space and money. As Labcenter Proteus is often used in professional work, students get a lot of practical knowledge, with the only disadvantage of using a completely virtual system.

Hardware elements are much better in use, where students have a real circuit board, box, or some other part of equipment. Such elements are used extensively throughout the University of applied sciences in Zagreb, from Electrical engineering with special circuits for big motors, over the automation courses where simulators for industrial system were made [2], all to communication systems and digital circuits where special boards are made.

Figure 1 PLC's State Simulator

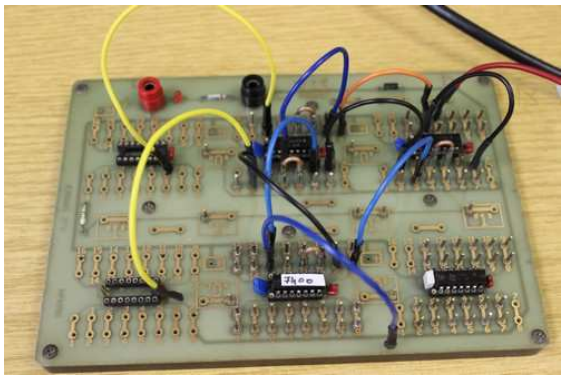


Figure 2 Circuit board used in the laboratory exercises

The option of seeing the systems, of understanding them, and using them in several possible ways, gives the students a much more realistic feel to expectations about their learning and knowledge level, and their future professional life.

Both of these approaches are starting to be even more combined in systems that use material tactile simulators and measuring computer software which allows fast and easy analysis of the given data. University of applied sciences uses National Instruments LabVIEW (as a software part) and MyDAQ (as a hardware acquisition system) as an easy-to-use integration system. Such models are then used in both ex-cathedra teaching [3], and as laboratory exercises and examples. But as mentioned earlier, making a good model is a tedious task which takes a lot of time and often requires teachers' expertise in the subject and their abilities to improvise and design simple but effective learning systems.

The idea which has taken off is to make teaching models as the necessary proof of independent work in showing the ability to solve engineering problems. Such work is expected for students' graduation thesis. By making models for future students in their thesis projects, students deepen their current knowledge while making the best possible product in the fear of being embarrassed in front of their peers and feeling proud when they do well. The projects often don't look as a polished product because it's in their intended nature to be changed, improved, and destroyed if necessary, for the better learning experience of future students.

Figure 3 Traffic light simulator

Figure 4 Model of a house, simulating heating and cooling

III. EXAMPLE OF A STUDENT PRODUCT

One of extensively changed and used student model is a model of an elevator. Started as a student graduation thesis, it is used for demonstration in several different academic areas:

- as an example for systems which use electric motor power to generate force
- as an example for digital electronics
- as an system which includes one programmable microcontroller, but also has an option of connecting other microcontrollers systems
- as a system which can be controlled through different physical simulators/controllers and through computer based (LabVIEW) simulators.

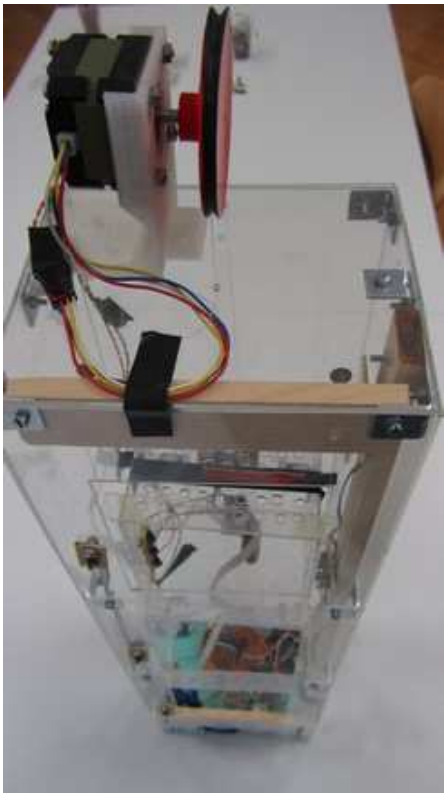


Figure 5 Elevator model

The elevator was conceived primarily as a programmable microcontroller tool, but its use widened after the student who prepared the system made it modular and easy to repair, change and improve. With student's thesis as a starting point, different simulations in LabVIEW and Proteus were made, also by students. Such simulations are then used in laboratory exercises, and are given as homework to students, in which they can change the software in the microcontroller and in some cases, completely change how the elevator works. This gives everyone a chance in the tactile part of learning, when they see the real product of their work. As elevator is a common thing in their lives, they can relate in how it should work and can propose improvements.

Proposed improvements are tested in simulations (if possible) and documented for the future use. In cases where the improvements are physical, an option to reverse the changes is necessary. During the use of the elevator, such reversing was seldom, as students have great ideas and are motivated in improving the system. Such examples are 3D printed motor bracket and the pulley which proved like a good change, while the students learned to use a 3D printer and the necessary software. Another example is the change in motors, where a DC motor was changed to a stepper motor for finer control of the elevator - this made an additional change in a simpler motor driver circuit. Due to such flux of this system, it doesn't looks like a finished product which is by itself an advantage as students are not afraid of working on it, even during the laboratory exercises.

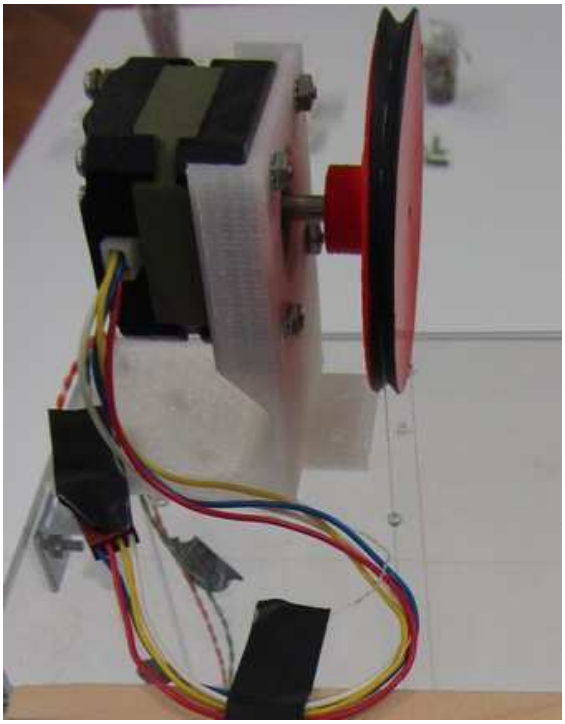


Figure 6 Parts of the elevator - printed on the 3D printer

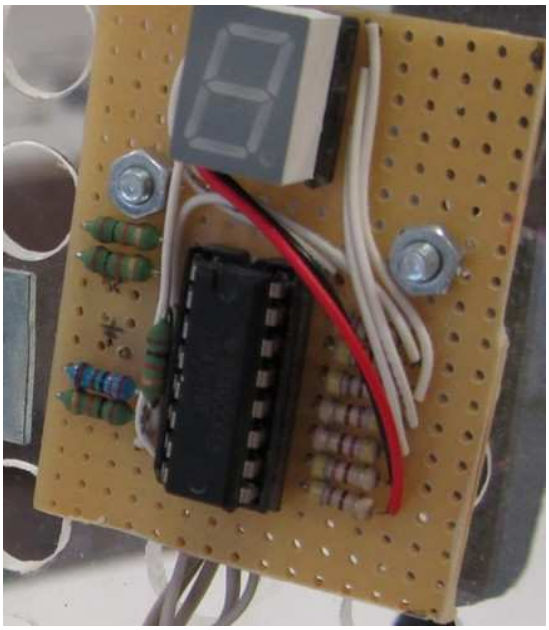


Figure 7 Parts of the elevator – 7 segment display

Some parts of the elevator are simple circuits normally thought separately from the connecting circuits. Usual way of learning about them was only on special circuit boards or on protoboards, but in the elevator these systems are integrated into one unity. Such an example is the use of a 7 segment display to show the floor on which the elevator is currently, which is controlled through a BCD decoder chip, which is in turn controlled by a microcontroller or a MyDAQ acquisition system. Such an arrangement is obvious for a professional, but sometimes students have problems with integrating simple systems into a complex system.

The elevator has an Arduino microcontroller for its primary control. Arduino is one of the most popular microcontroller systems for novices. In the elevator this microcontroller board can be reprogrammed and the programmer (student) can change the behaviour of the elevator logic. The program is first checked in Proteus simulation, to ensure that it is working correctly. The elevator is modular so a change in the electronic brain (the microcontroller) is easy, as most wires are already prepared for connection on other boards or systems (like MyDAQ).

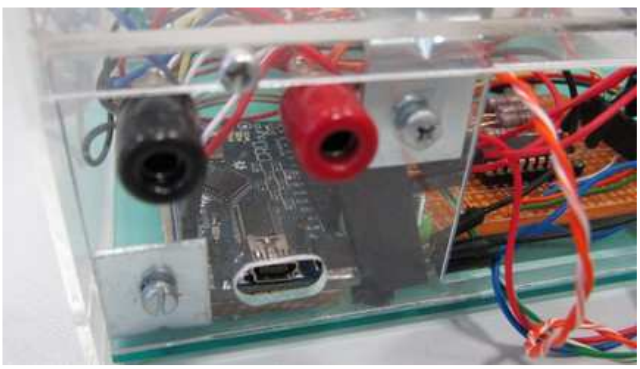


Figure 8 Parts of the elevator – Arduino board

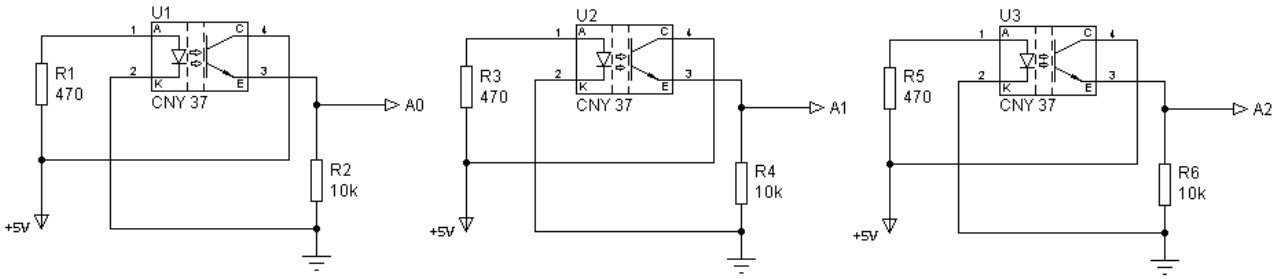


Figure 9 Schematic for elevator position detection using optocouplers

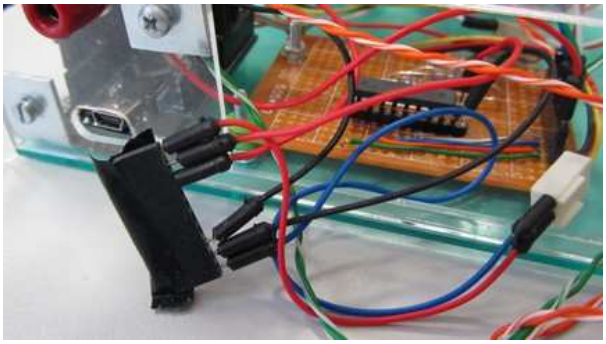


Figure 10 Parts of the elevator – ease of access for modular changes

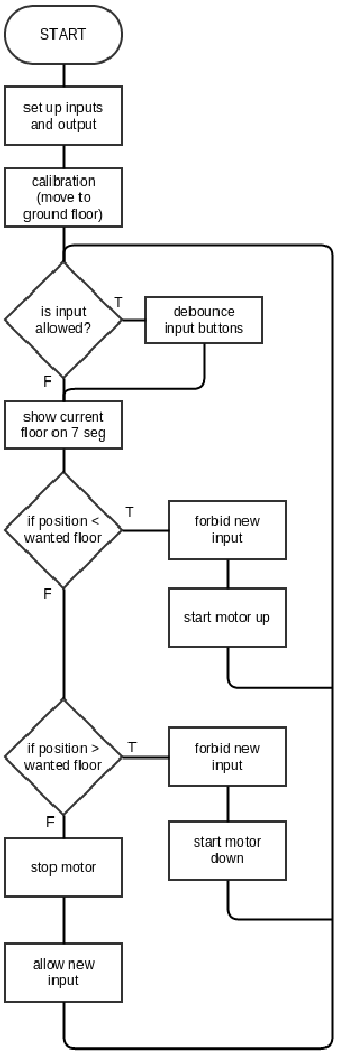


Figure 11Elevator model flowchart

IV. FUTURE MODELS

Previous models proved successful, and there are ideas for future models made as student thesis. As models with several different systems are more practical in multidisciplinary teaching, the future models will incorporate such ideas.

First interesting model is a model of a parking ramp, which is already used as a practical simulation in software. Such a system consists of several parts: the sensors for a car, sensors for the ramp, the motor, and a display for showing the number of parked cars. Parking ramp system covers most of the ideas used in elevator, but it can be made smaller and more practical for use by students. Due to its low complexity it can be made in several copies, so several students could work in parallel.

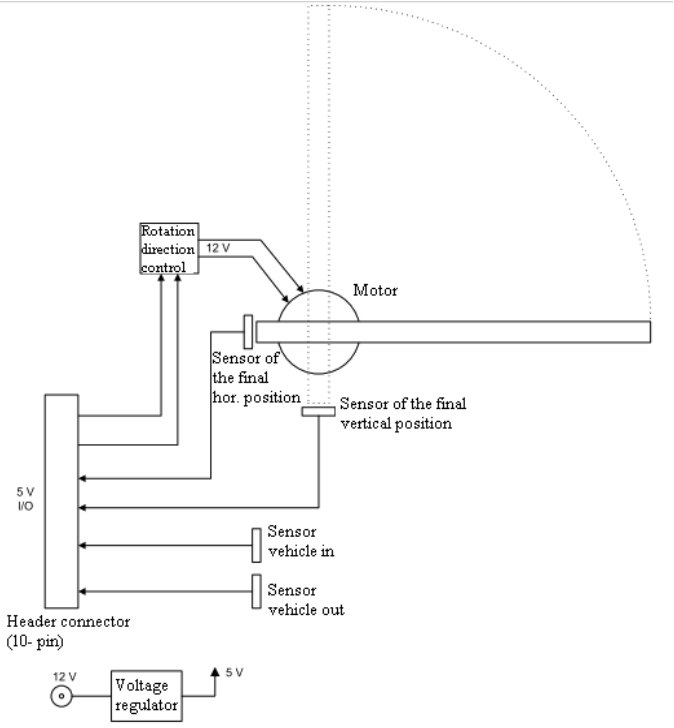


Figure 12 Basic schematic for the parking ramp model

Another interesting model is the use of different sensors with a microcontroller and which has an option to send data wireless. Using one of standard communication protocols, a portable sensor system can be used for different tasks. Although the sensory nodes are already easily obtained, a modular system should be used in teaching environment to prepare students for future IoT (Internet of Things). Currently available system are often very integrated, so modularity and a connection to a software like LabVIEW would help students in achieving higher understanding of sensory nodes. Modular also must give an option in changing the communication protocols: connecting to mobile phones via Bluetooth, to Wi-Fi networks via additional modules and to other possible communication protocols (like Xbee, nrf24 or even to 3G or 4G networks).

Some ideas come from students themselves, but they often take their models home as they grow attached to them. Their graduation thesis project becomes their pride and joy.

V. CONCLUSION

Models of different real systems are often used in teaching, and the problem of making a model is left to the teacher. At University of applied sciences Zagreb the making of a real life model is given as an opportunity to students, as their final thesis. Student work has to be checked by the mentor during the making of the model, but even a model which is badly put together helps in making the next version or gives clues in how to improve the original. This practice cannot be used in all courses as some are still reliant on the teachers' ability to imagine and to make an exemplary system, but many teachers are changing the paradigm of learning and are trying to show the students of what they are capable of, even with having low resources or technical options. This paper shows some examples of previous view on models (software and hardware), and gives an example of a different type of multipractic model (in this case, an elevator), which was made by a student.

Models made this way which are used in lectures and laboratory work are constantly improving due to student engagement, good documentation which is usable for knowledge sharing (it's made out of students final thesis) and due to new ideas which are coming from a pool of young perspective people, eager to leave their mark on their University.

REFERENCES

[1] D. Ćika, D. Grundler: Proteus Virtual System Modelling used for microcontroller education; MIPRO; 2010.

[2] G. Malčić, M. Fruk, D. Josić: PLC's State Simulator; MIPRO; 2005.

[3] T. Horvat, Lj. Cvitaš: Poboljšanja u primjeni računalom vođenog demonstracijskog stola u nastavi; MIPRO; 2013.

[4] F. Trkeš: Maketa lifta upravljana mikroupravljačem; Bachelor thesis at the Department of Electrical Engineering, University of applied sciences; 2014

Extracting Domain Knowledge by Complex Networks Analysis of Wikipedia Entries

Neven Matas*, Sanda Martinčić-Ipšić**, Ana Meštrović**
*Faculty of Humanities and Social Sciences, University of Rijeka, Croatia
**Department of Informatics, University of Rijeka, Croatia
nmatas@ffri.hr, smarti@uniri.hr, amestrovic@inf.uniri.hr

Abstract - In this paper we describe a complex networks analysis of Wikipedia. We construct 10 different networks from Wikipedia entries (articles) related to the chosen domain. The goal of the experiment is to extract domain knowledge in terms of identifying entries that are centrally positioned and entries that are strongly related. We apply complex networks analysis on all acquired networks and examine the networks' structure. We employ centrality measures in order to find centrally positioned entries in the network. Furthermore we identify communities and find which entries are densely connected according to the network structure.

I. INTRODUCTION

Complex networks exhibit specific topological features, such as high clustering coefficients, small diameters, a power-law degree distribution, community structure, one or several giant components, hierarchical structures, etc. Two important classes of complex networks that can be further differentiated are small-world networks [1] with small distances and high clustering coefficients as main properties and scale-free networks [1] which can be characterized by a power-law degree distribution.

Wikipedia can be modelled as a complex network in a way that Wikipedia entries are nodes, and links between two nodes are established if there is a hyperlink between these two entries. Early attempts to quantify Wikipedia using complex networks analysis were focused only on network structure of linked Wikipedia entries. In [2] Zlatić et al. present an analysis of Wikipedias in several languages as complex networks. They show that many network characteristics (degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths and triad significance profiles) are common to Wikipedias in different languages and show the existence of a unique growth process. The same authors studied Wikipedia growth based on information exchange in [3]. In [4] an analysis of the statistical properties and growth of Wikipedia is presented. Pemble and Bingol [5] have constructed two complex networks out of English and German Wikipedia corpora and analyzed conceptual networks in different languages.

The other research direction is focused on content found on Wikipedia and analyses Wikipedia as a (domain) knowledge network. In Fang [6] they first extract a specific domain knowledge network from Wikipedia (specifically, four domain networks on mathematics, physics, biology, and chemistry) and then carry out statistical analysis on these four knowledge networks. Also, they show that MathWorld and Wikipedia Math share a similar internal structure. In [7] Masucci et al. extract the topology of the semantic space and measure the semantic flow between different Wikipedia entries. They further analyze a directed complex network of semantic flow. In [8] the results of semantic language networks analysis are presented in general.

Motivated by the second approach that studies Wikipedia as a knowledge network, we wanted to study how the network structure is related to domain knowledge. The goal of our experiment was to extract centrally positioned entries in the network and analyze how these entries are related to domain knowledge and are some more important than other. In the second part of the experiment the task was to extract entries that belong to the same community and check whether they are semantically related.

In our previous research, we have already analyzed Wikipedia as a complex network [9], but by constructing a network of syllables. Also, we examined the structure of Croatian language networks in [10,11,12]. In [12,13] we applied network measures for a keyword extraction task. In all our previous experiments we were focused solely on language structure and this is our first attempt to analyze semantic relations in a network.

In the second section we present key measures of complex networks involved in network structure analysis. In the third section we describe data sources and network construction principles. In the fourth section we present the results. Finally, the fifth section contains a conclusion and possible directions for future research.

II. NETWORK STRUCTURE ANALYSIS

In this section we review some of the most important network measures [14]. Every network has an N number of nodes and a K number of links. The degree

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

of a node *i* is the number of links with which the node is connected, k_i. Considering the fact that we are working with directed networks, we must specify two types of degrees: the in-degree, k_i^{in}, corresponding to the number of incoming links and the out-degree, k_i^{out}, equal to the number of outgoing links for any particular node *i*. The average degree of the network is:

$$< k > = \frac{2K}{N}. \tag{1}$$

For the directed networks we omit multiplication by 2. In the further equations we assume that the network is directed and that the total possible number of links is equal to $N(N\text{-}1)$. For every two connected nodes *i* and *j*, the number of connections lying on the path between them is represented as d_{ij}, and so d_i is the average distance of a node *i* from all other nodes, and it is obtained by:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{2}$$

For the next two measures, if a network contains more than one component, we consider the largest component. The average shortest path length between every two nodes in a network is:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}. \tag{3}$$

And the maximum distance results in the network diameter, *D*:

$$D = max_i d_i. \tag{4}$$

The clustering coefficient is a measure which defines the presence of connections between the nearest neighbours of a node. And so, c_i (the clustering coefficient) of a node is a fraction between the number of edges E_i that exist between that k_i and the total possible number:

$$c_i = \frac{2E_i}{k(k-1)}. \tag{5}$$

The average clustering coefficient of a network is defined as the average value of the clustering coefficients of all nodes in a network:

$$C = \frac{1}{N} \sum_i c_i. \tag{6}$$

Density of a network is a measure of network cohesion defined as the number of observed links divided by the number of total possible links:

$$d = \frac{K}{N(N-1)}. \tag{7}$$

Degree centrality of a node *i* is the degree of that node. It can be normalised by dividing it by the maximum possible degree $N-1$:

$$dc_i = \frac{k_i}{N-1}. \tag{8}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of shortest paths from node *j* to node *k* and let $\sigma_{jk}(i)$ be the number of those paths that pass through the node *i*. The normalised betweenness centrality of a node *i* is given by:

$$bc_i = \frac{\sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{9}$$

Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest distances between a node and all other nodes. Let d_{ij} be the shortest path between nodes *i* and *j*. The normalised closeness centrality of a node *i* is given by:

$$cc_i = \frac{N-1}{\sum_{i \neq j} d_{ij}}. \tag{10}$$

Modularity measures the quality of the network partition in the communities. The modularity of a network partition is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities. Let e_{ij} be the fraction of edges in the network that connect vertices in group *i* to those in group *j*, and let $a_i = \sum_j e_{ij}$. Then the modularity can be calculated using following equation:

$$Q = \sum_{i=1}^{k} \left(e_{ii} - a_i^2 \right). \tag{11}$$

The degree assortativity coefficient measures the tendency of nodes in a network to connect to nodes similar to themselves. The coefficient lies between -1 and 1 and it is quantified via the Pearson correlation. Positive *r* values indicate a correlation between similar-degree nodes. Let q_k and q_j be the distribution of the degree of out-edges that do not connect to the other node in question, e_{jk} the joint probability distribution of q_k and q_j, and σ_q^2 the variance of the distribution. Then we can calculate the assortativity coefficient using the following equation:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}. \tag{12}$$

On the meso-scale level complex networks analysis includes a community detection task [15]. Communities, in this sense, are groupings of densely interconnected nodes within a network. In other words,

nodes in a community have a greater amount of connections amongst each other than with other nodes in the network. Several algorithms are used for community detection such as hierarchical clustering, Girvan-Newman's algorithm, minimum-cut method, etc. One of the most efficient is the Louvain method [16], a greedy optimization method that optimizes the modularity of a network's partitions. The number of communities (*Nc*) represents the amount of such groupings found within a network.

III. NETWORK CONSTRUCTION

For the purpose of our experiment we collect entries from Wikipedia and construct networks related to the domain. Our intention was to construct two types of networks: level 2 networks and level 4 networks. We construct level 2 networks by starting with a chosen seed entry (e.g. "*Complex network*" or "*Data*"), storing all the hyperlinks to related entries from the seed entry's text (level 1) and proceeding to extract the hyperlinks from all the entry pages taken from the original entry (level 2). Analogously, we construct level 4 networks by taking the first 10 hyperlinks from a given entry page and proceeding to repeat the task three times, arriving at level 4. We limit the hyperlinks to the first 10 due to the computational complexity at the same time having in mind that the most general hyperlinks are usually at the beginning of the entry's text.

Therefore, the first task is the construction of a web scraping program which would extract hyperlinks from a Wikipedia entry's text. The hyperlinks are extracted using a Python package for HTML parsing called Beautiful Soup which parses the HTML structure of a given HTML document into a parse tree. By navigating the tree we locate the tag ID which corresponds to article content ("mw-content-text") and proceed to extract the hyperlinks which themselves are found within paragraph (<p>) tags and finally inside link (<a>) tags in that section of the page. Finally, each network is stored in an edge list in the following format: "entry title" \t "linked entry title". We had some difficulties with processing non-ASCII script and hyperlinks that weren't connected to other documents (citations, in-page references, etc.), but we managed to avoid those by checking the data during the extraction process.

In our directed network, each entry's title represents a node and it is connected to other entries hyperlinked in its text, again represented as network nodes. We construct a total of 10 domain networks for five chosen seed entries: "*Byte*", "*Complex network*", "*Computer science*", "*Data*" and "*Programming language*". The naming scheme includes the level of a specific network in its name (e.g. the level 2 network for "Byte" is BT2). Since we consider unweighted networks, we dismiss double links. This, along with the fact that some entries do not contain 10 hyperlinks resulted in our level 4 networks having less than 10^4 expected edges. We use Python and the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [17].

IV. RESULTS

In this section we present the results of our measuring described in section 2, such as average degree $\langle k \rangle$, average path distance L, diameter D, average clustering coefficient C, density d, modularity Q, number of communities (N_c) and degree assortativity coefficient r. We also present the most central nodes (according to the three centrality measures) and communities in networks detected by using the Luvain algorithm.

In Table I. we present estimated global network measures. There are certain differences between measures for level 2 and 4 which are evident upon closer inspection. For instance, level 4 networks have significantly larger average path lengths, diameters, assortativity coefficients, often a significantly larger number of detected communities and slightly larger average degrees. The modularity measure and density are comparable between the two, whilst level 2 networks show larger clustering coefficients.

For comparison with random networks, the table also includes two measures for equivalent random networks (Erdös-Renyi random graphs) – the average shortest path length ($L_{ER} = \ln N / \ln \langle k \rangle$) and the average clustering coefficient ($C_{ER} = \langle k \rangle / N$). The results show that the complex networks we have constructed have a significantly higher average clustering coefficient than their Erdös-Renyi random graph counterparts. This, in addition with a relatively small average shortest path length L led us to conclude that we are dealing with small-world networks as described by Watts and Strogatz [20]. For the purposes of this comparison we treat the networks as undirected.

Moreover, a distinctly high modularity coefficient Q (higher than 0.7 in all but one network, as visible in Table I.) shows a clear tendency towards community clustering of nodes present in the networks. We did not observe any strict rule governing community size across networks, although level 2 networks have an understandably smaller N_c which we contributed to the very construction principle as described in section 3.

Figure 1 and Figure 2 show the degree distribution plots for level 2 and level 4 networks corresponding to the "computer science" and "complex network" entries respectively. The observed difference between the level 2 and 4 networks is believed to be due to the method of network construction.

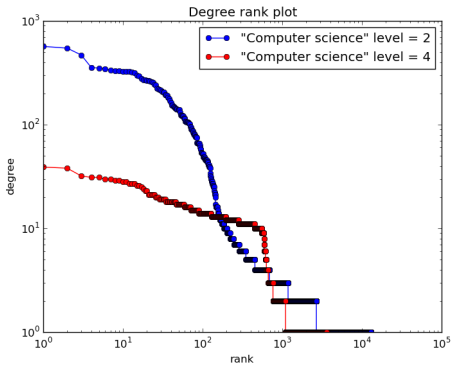


Figure 1 - Degree distribution for CS2 & CS4

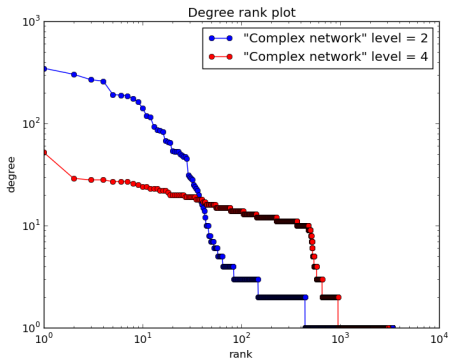


Figure 2 - Degree distribution for CN2 & CN4

TABLE I. GLOBAL NETWORK MEASURES CALCULATED FOR ALL 10 NETWORKS

Measure	"Byte"		"Complex network"		"Computer science"		"Data"		"Programming Language"	
Network	BT2	BT4	CN2	CN4	CS2	CS4	DT2	DT4	PL2	PL4
Number of nodes (N)	3945	3632	3405	3070	12881	3630	2297	3658	7467	3965
Number of edges (K)	5112	5611	4132	5008	18852	5851	2630	5531	13933	6215
Average degree ($<k>$)	1.296	1.545	1.214	1.631	1.464	1.612	1.145	1.512	1.145	1.612
Avg. shortest path (L)	3.693	6.834	3.198	9.218	3.417	6.277	3.086	6.369	3.127	6.277
Avg. shortest path (L_{ER})	8.693767	7.2662195	9.168408	6.791134	8.8088393	7.0022451	9.340827	7.4144377	10.763658	7.0776521
Diameter (D)	9	15	6	22	7	14	7	14	6	22
Average clustering coefficient (C)	0.06	0.021	0.043	0.024	0.074	0.019	0.043	0.019	0.082	0.021
Average clustering coefficient (C_{ER})	0.000657	0.0008508	0.0007131	0.0010625	0.0002273	0.0008882	0.000997	0.0008267	0.0003067	0.0008131
Density (d)	0.0003	0.00042	0.00035	0.00053	0.00011	0.00044	0.00049	0.00041	0.00025	0.0004
Modularity (Q)	0.778	0.776	0.794	0.763	0.725	0.771	0.828	0.779	0.594	0.78
Number of communities (N_c)	17	32	17	21	23	27	18	31	19	30
Degree assortativity coefficient (r)	-0.592	-0.048	-0.521	0.021	-0.491	-0.028	-0.561	-0.048	-0.468	-0.059

After the analysis on the global level, we analyse the networks on the local level in terms of centrality measures. Tables II. and III. show lists of top ten entries according to the three centrality measures for the two seed entries: "*Computer science*" and "*Programming language*". We analyse the degree centrality, betweenness centrality and closeness centrality. For the degree centrality we treated the network as undirected. For each centrality measure and domain there are two lists of entries, one for level 2 networks and another for level 4 networks. We noticed that the lists for level 2 networks consist of entries that are semantically related to the seed entries ("*Computer science*" or "*Programming language*") in a way that might be ascribed as belonging to a hierarchy. This is especially evident for the closeness centrality measure. For example, the list of top ten entries according to the closeness centrality for the seed entry "*Computer science*" contains other scientific domains (theoretical computer science, mathematics, artificial intelligence, physics, engineering) and for the seed entry "*Programming language*", list contains some prominent programming languages (C, Java, Perl, Python, C++).

TABLE II. TOP TEN ENTRIES IN THE „*COMPUTER SCIENCE*" NETWORKS (CS2, CS4) REGARDING THE THREE CENTRALITY MEASURES: DEGREE CENTRALITY, BETWEENNESS CENTRALITY AND CLOSENESS CENTRALITY

	Degree centrality		Betweenness centrality		Closeness centrality	
	CS2	CS4	CS2	CS4	CS2	CS4
#1	*human*	*mathematics*	***computer science***	*computer science*	***computer science***	*computer science*
#2	*university of cambridge*	*cell (biology)*	*computer*	*information*	***mathematics***	*information*

#3	***philosophy***	*computer science*	***mathematics***	*protein*	***theoretical computer science***	*science*
#4	*industrial revolution*	*computer*	***artificial intelligence***	*science*	*computer*	*cell (biology)*
#5	*gottfried wilhelm leibniz*	*information*	***philosophy***	*algorithm*	***artificial intellgience***	*mathematics*
#6	***physics***	*protein*	*human*	*logic*	***philosopy***	*ancient greek*
#7	***eletrical engineering***	*organism*	*gottfried wilhelm leibniz*	*organism*	***physics***	*latin*
#8	***artificial intelligence***	*dna*	*algorithm*	*cell (biology)*	*human*	*computing*
#9	***mathematics***	*computer program*	***theoretical computer science***	*computing*	*gottfried wilhelm leibniz*	*algorithm*
#10	*alan turing*	*philosophy*	***physics***	*mathematics*	*engineering*	*bit*

TABLE III. TOP TEN ENTRIES FOR THE „*PROGRAMMING LANGUAGE*" NETWORKS (PL2, PL4) REGARDING THE THREE CENTRALITY MEASURES: DEGREE CENTRALITY, BETWEENNESS CENTRALITY AND CLOSENESS CENTRALITY

	Degree centrality		Betweenness centrality		Closeness centrality	
	PL2	PL4	PL2	PL4	PL2	PL4
#1	*history of computing hardware*	*mathematics*	*programming language*	*programming language*	*programming language*	*programming language*
#2	*internet*	*computer*	*computer*	*computer*	***c (programming language)***	*ancient greek*
#3	*english language*	*computer science*	***c (programming language)***	*software engineering*	*computer programming*	*computer*
#4	*computer*	*physics*	*compiler*	*computing*	***java (programming language)***	*mathematics*
#5	***c (programming language)***	*set (mathematics)*	*english language*	*computer science*	***perl***	*arithmetic*
#6	***python (programming language)***	*greek language*	*computer program*	*algorithm*	*compiler*	*science*
#7	*university of manchester*	*logic*	*internet*	*message*	*computer program*	*greek language*
#8	***perl***	*language*	***perl***	*communication*	***python (programming language)***	*physics*
#9	***programming language***	*central processing unit*	***python (programming language)***	*machine*	*control flow*	*latin*
#10	***php***	*electronics*	***java (programming language)***	*function (mathematics)*	***c++***	*computer science*

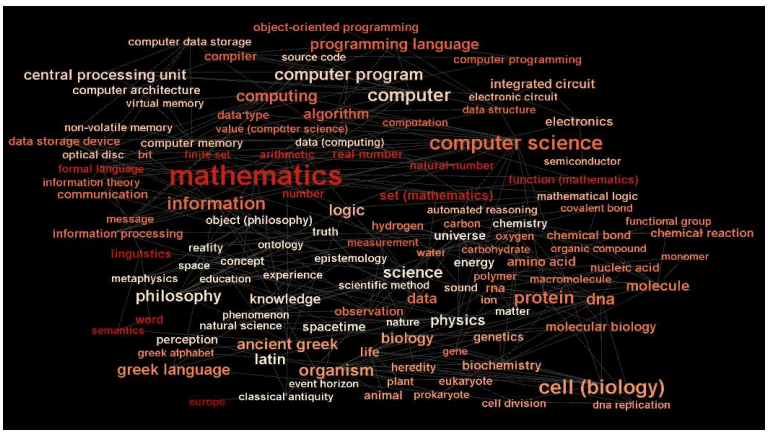


Figure 3 – Communities in the CS2 network

In the second part of the experiment we analyse communities in all 10 networks in order to explore which entries are grouped together. Figure 3 shows most significant entries from the CS2 network grouped into communities. Different communities are presented in different colours. For example, entries related to the mathematics domain (*mathematics*, *number*, *set*, *function*, *real number*, etc.) are in the red-coloured community; entries related to the computer science domain (*computing*, *algorithm*, *compiler*, etc.) are in the orange-coloured community; entries that are related to the biology domain (*cell*, *organism*, *gene*, etc.) are in the light-orange coloured community and entries that are related to the philosophy domain (*reality*, *concept*, *knowledge*, etc.) are in the white-coloured community. It can be observed that entries grouped into communities are more closely semantically related than entries from different communities. The results are similar for other networks; semantically related entries are grouped into communities much more than entries that are not semantically related.

V. CONCLUSION

In this paper we present our initial attempt to study Wikipedia as a complex network. We extract parts of Wikipedia related to 5 chosen seed entries. We construct 10 different networks using two different principles of construction. Then we analyse the global structure of all networks. We show that all networks have similar properties: a high average clustering coefficient in comparison to the random networks, small distances, low density and community structure. From these global measures we may conclude that all 10 networks extracted form Wikipedia are small-world networks. These results are in line with previous studies of Wikipedia as a complex network.

Furthermore, we explore semantic relations in the constructed networks. We use network centrality measures to extract entries in the networks that are significant according to the network structure. Three centrality measures are employed for this task: degree centrality, betweenness centrality and closeness centrality. It can be observed that for level 2 networks centrality measures obtain good results (especially closeness centrality). Among top ten entries according to the closeness centrality are entries that are semantically related to the domain. This can be useful for modelling taxonomy or domain ontology. Furthermore, semantically related entries are grouped into communities more often that entries that are not semantically related.

These findings can be partially explained as a consequence of network construction rules employed in this experiment. However, these preliminary results suggest that Wikipedia is well organised and its structure can be captured and explored by a complex networks approach. In future work we plan to extract a broader section of Wikipedia and explore its potential as a knowledge network. We will study the domain knowledge extraction possibilities and perform the evaluation of the results.

VI. REFERENCES

[1] A. L., Barabasi and R. Albert, "Emergence of scaling in random networks", Science, 286, pp. 509–512, 1999.

[2] V. Zlatić et al., "Wikipedias: Collaborative web-based encyclopedias as complex networks," in Physical Review E 74.1 (2006): 016115.

[3] V. Zlatić and H. Štefančić, "Model of wikipedia growth based on information exchange via reciprocal arcs," EPL (Europhysics Letters) 93.5 (2011): 58005.

[4] G. Caldarelli, A. Capocci, V. Servedio, L. Buriol, D. Donato, and S. Leonardi, "Preferential attachment in the growth of social networks: the case of Wikipedia," in APS Meeting Abstracts, vol. 1, p. 33003. 2006.

[5] F. C. Pemble and H. Bingol, "Complex Networks in Different Languages: A Study of an Emergent Multilingual Encyclopedia", Unifying Themes in Complex Systems, 1, 612, 2008.

[6] Z. Fang, J. Wang, B. Liu, W. Gong, "Wikipedia as domain knowledge networks: domain extraction and statistical measurement," in Proceedings of international conference on knowledge discovery and information retrieval (KDIR 2011), Paris, Oct 2011

[7] A.P. Masucci, A. Kalampokis, V. Martínez Eguíluz, and E. Hernández-García. (2011) "Wikipedia Information Flow Analysis Reveals the Scale-Free Architecture of the Semantic Space," PLoS ONE 6(2): e17333. doi:10.1371/journal.pone.0017333

[8] J. Borge-Hothoefer and A. Arenas, "Semantic Networks: Structure and Dynamics," in Entropy, 12, pp. 1264-1302, 2010.

[9] K. Ban, I. Ivakić, and A. Meštrović, "A preliminary study of Croatian language syllable networks." In Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th, pp. 1296-1300. IEEE, 2013.

[10] D. Margan, S. Martincic-Ipšic, and Ana Meštrovic, "Network Differences Between Normal and Shuffled Texts: Case of Croatian," in Complex Networks V, pp. 275-283. Springer International Publishing, 2014.

[11] D. Margan, A. Meštrović and S. Martinčić-Ipšić, "Complex networks measures for differentiation between normal and shuffled Croatian texts," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on, pp. 1598-1602. IEEE, 2014.

[12] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, "Toward Network-based Keyword Extraction from Multitopic Web Documents," in Proceedings of 6th International Conference on Information Technologies and Information Society (ITIS2014), Šmarješke toplice, pages 18-27, Slovenia, 2014.

[13] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "Toward Selectivity Based Keyword Extraction for Croatian News," in CEUR Proceedings (SDSW 2014), Vol. 1301, pages 1-14, Riva del Garda, Trentino, Italy, 2014.

[14] M. E. J. Newman. "Networks: An Introduction", Oxford University Press, 2010.

[15] S. Fortunato, "Community detection in graphs," in Physics Reports 486, no. 3 (2010): 75-174.

[16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment 2008, no. 10 (2008): P10008.

[17] A. A. Hagberg, D. A. Schult and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, 2008.

LaNCoA: A Python Toolkit for Language Networks Construction and Analysis

Domagoj Margan, Ana Meštrović
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {dmargan, amestrovic}@uniri.hr

***Abstract*—In this paper we describe LaNCoA, Language Networks Construction and Analysis toolkit implemented in Python. The toolkit provides various procedures for network construction from the text: on the word-level (co-occurrence networks, syntactic networks, shuffled networks), and on the subword-level (syllable networks, grapheme networks). Furthermore, we implement functions for the language networks analysis on the global and local level. The toolkit is organized in several modules that enable various aspects of language analysis: analysis of global network measures for different co-occurrence window, comparison of networks based on original and shuffled texts, comparison of networks constructed on different language levels, etc. Text manipulation methods, like corpora cleaning, lemmatization and stopwords removal, are also implemented. For the basic network representation we use available NetworkX functions and methods. However, language network analysis is specific and it requires implementation of additional functions and methods. That was the main motivation for this research.**

I. Introduction

The study of graphs and networks plays an important role in various research domains. The advent of the computer age increased the interest in the large real-world networks that are studied as complex networks. These networks exhibit specific topological properties (high clustering coefficient, small diameter, community structure, one or several giant components, hierarchical structure, heavy tail degree distribution, etc.). Various classes of complex networks have been analyzed, such as for example: technological networks, biological network, information networks or social networks [31]. One possible class includes language networks as well.

Various construction rules may be applied in order to construct a network from the text. The usual way is to construct networks of word co-occurrences [14], [23], [24], [28], [35] or syntactic networks [1], [11]–[14], [21], [22]. There are also experiments with shuffled (randomized) networks [10], [18], [21], [28], [29]. Furthermore, syllables networks [4], phoneme networks [2] or semantic networks [9] can be constructed as well. Additionally all these networks can be constructed as undirected or directed, unweighted or weighted. In most cases the best way to represent the text is to chose directed and weighted variant of the network [24]. There are various software, tools and packages designed and developed for the task of complex networks analysis: Gephi [6], NodeXL [17], SNAP [3], Cytoscape [34], NetworkX package [16] for Python, igraph package [15] for C, Python and R. All these tools enable calculating standard global network measures (such as average clustering coefficient, average shortest path length, diameter, average degree, degree distribution, density, modularity, assortativity, etc.) and the local network measures (different centrality measures: degree, betwenness, closeness, eigenvector, etc.). Some of the tools (for example Gephi, NewtorkX, iGraph, SNAP) provide network analysis on the meso-scale level with implemented algorithms for community detection. Also there are some tools designed and focused only on one aspect of the network analysis, for example GraphCrunch [19] for graphlet analysis, GRAAL [30] for graph and network alignment or FANMOD [38] for motif analysis.

However, there is no software specialized for tasks of language network construction and analysis. Our main motivation was to implement a simple toolkit that provides various language network construction possibilities and suitable network analysis functions. Furthermore this toolkit can be used for various NLP applications, such as for example keyword extraction task [7], [8], [37] or text type classification [27].

The LaNCoA toolkit is focused mainly on the language networks construction task which includes various methods for the corpora manipulation (text preprocessing and cleaning, lemmatization, shuffling procedures and preparation for the language networks construction) and procedures for generation of various word-level and subword-level networks directly from the given corpora. The toolkit also enables complex network analysis in terms of calculating all important global and local network measures, network and text content analysis, and various plotting data possibilities. To some extent, the LaNCoA toolkit uses existing functions from the NetworkX Python package as a basic foundation for some more specific network construction and analysis tasks. Furthermore, there are some measures important for weighted and directed networks that are not implemented in the standard network-manipulation packages which are therefore implemented in the LaNCoA toolkit, such as the selectivity measure, network reciprocity, network entropy, inverse participation ratio, and link overlap measures.

The paper is structured as follows. In Section II we describe the language networks. In Section III we give a short overview of the complex networks analysis task. Then we present the LaNCoA toolkit in Section IV and we describe LaNCoA toolkit applications in Section V. We give a conclusion remarks

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

in Section VI.

II. Complex Network Analysis Task

A complex network is modeled as a graph G. A graph $G = (V, E)$ consists of a collection of vertices, or vertex set, V and a collection of edges, or edge set, E. In the complex network approach vertices are called nodes and edges are called links. The study of networks can be classified in three levels: global (macro-scale) level, meso-scale level and local (micro-scale) level.

The study at the macro level attempts to understand the global structure of a network. At this level, relevant parameters are average degree, degree distribution, average path length, average clustering coefficient, density, modularity, assortativity, etc. An the meso-scale level the interaction between nodes at short distances are studied. This includes community detection or analysis of small subgraphs such as motifs or graphlets. At the micro level the study is focused on the behavior of single nodes. Identification of the important nodes in the network using different centrality measures or just deterring degree, strength, clustering coefficient or betweenness and other parameters of a single node. In [31] a detailed overview of all network measures and formulas is given.

III. Language Networks

Written, as well as spoken language can be modeled via complex networks where the lingual units (e.g. words) are represented by vertices and their linguistic interactions by links. Language networks are a powerful formalism to the quantitative study of language structure at various language sublevels. Complex network analysis provides mechanisms that can reveal new patterns in complex structure and can thus be applied to the study of patterns that occur in the natural languages. Thus, complex network analysis may contribute to a better understanding of the organization, structure and evolution of a language.

On the word-level, text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links [14]. The interactions can be derived at different levels: structure, semantics, dependencies, etc. On the subword-level, syllable or grapheme networks can be constructed, where nodes can be represented by syllables or graphemes, while their dependencies (e.g. positions of syllables within words or graphemes within syllables) are links [4].

The properties of the co-occurrence networks are derived from the word order in texts [14], [23], [24], [28], [35]. Commonly they rise from the simple criterion such as co-occurrence of two words within a sentence, or text; or as co-occurrence of words within the given co-occurrence window. In the networks where the linkage is limited to the sentence borders during the construction, the sentence boundary can be considered as the window boundary too.

The syntactic networks are constructed using syntactic dependencies relations. Syntactic dependencies between words are formally expressed by dependency grammar (e.g. set of productions (rules) in a form of a grammar). The dependency grammar is used to present the syntactic relationships from sentence in a form of syntactic dependency tree. The properties of the syntactic networks are analyzed in [1], [11]–[14], [21], [22]. The results suggests that modelling human language using syntactic networks is important for language analysis because not all of the properties of the text structure are captured within co-occurrence networks.

For the purpose of better understanding of the language structure, one approach to address the questions of the word order in the language is to compare networks constructed from normal texts with the networks from randomized or shuffled texts [10], [18], [21], [28], [29]. Networks constructed from such shuffled texts are commonly regarded as shuffled networks.

Syllable and grapheme networks are important for studying structure of a language at the subword-level. In the syllable networks, nodes are represented by syllables and a link between two syllables can be established if they belong to the same word or if they are neighbors in the word. [4]. The same principle applies to the grapheme networks, where two graphemes are linked if they co-occurre as neighbors within a word or a syllable.

Figure 1 present different construction rules for stated language network types.

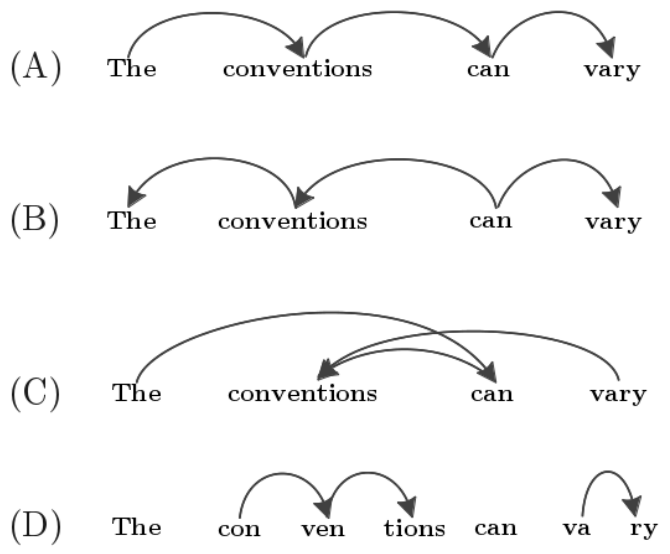


Fig. 1. Language networks construction rules presented on one toy-example sentence "The conventions can vary.": (A) Co-occurrence network, (B) Syntactic dependency network, (C) Shuffled network, (D) Syllable network

IV. The LaNCoA Toolkit Overview

LaNCoA is free and open source software licensed under the General Public License version 2. The source code of a working version is available for download from the official GitHub repository at https://github.com/domargan/LaNCoA.

All of the LaNCoA functionalities work for any of the languages written in any set of graphemes based on the letters of the classical Latin alphabet (Latin script). Only Latin script languages are supported (commonly used by about 70% of the world's population).

The LaNCoA toolkit is implemented in Python programming language. Python is an excellent tool for scanning and manipulating textual data and also provides various packages and libraries for scientific computations and data visualization. One of those is the popular NetworkX package, designed for exploration and analysis of complex networks and network algorithms. Our goal was to utilize basic NetworkX functions to develop extra procedures suitable for language network construction and analysis. We have also implemented functions for calculation of some non-standard general complex network measures. We have based our plotting procedures on proven quality matplotlib library for producing visualization figures.

Toolkit is divided into six modules that enable various aspects of language and text corpora analysis: i) corpora manipulation, ii) language networks generation, iii) single language network analysis, iv) multiplex language networks analysis, v) content analysis, and vi) data plotting. Modules are grouped into two main parts: network construction and network analysis. Modules provide procedures for tasks such as corpora cleaning, utilization of different network construction principles, analysis of global and local network properties, comparison of networks based on original and shuffled corpora, comparison of networks constructed on different language levels, etc. The generalized architectural structure of our toolkit is visualized and presented in Figure 2. In this section we describe each module's function individually.

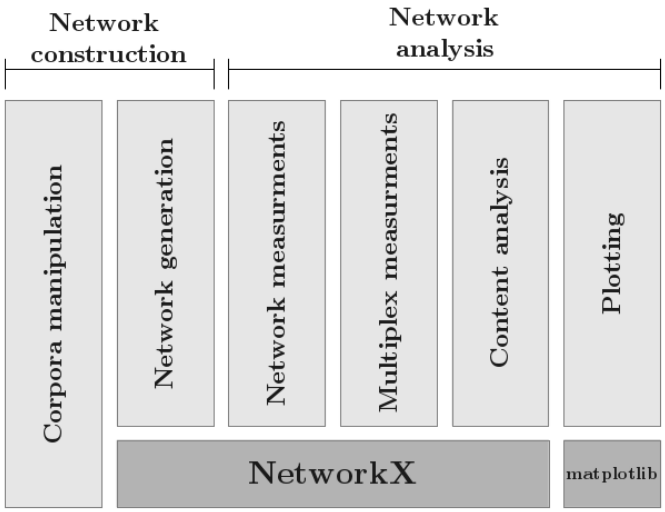


Fig. 2. LaNCoA architecture

A. Network Construction

Network construction part of our toolkit consists of two modules: the corpora manipulation module and the network generation module.

1) Corpora Manipulation Module: Corpora manipulation module can be used for several tasks with focus on various functions used to manipulate textual corpora. All of the tasks are optional and can be performed independently by the user's choice. Implemented LaNCoA functions are the following:

a) Corpus cleaning and Unicode normalization: It is important to have clean and high quality corpus before the process of the network generation, so the user can utilize LaNCoA to clean the corpus from the unwanted characters or data. LaNCoA also supports the usage of unclear and "dirty" textual data, but the noise-cleaning of the corpus is recommended, since it can greatly reduce the risks of badly constructed networks or network pollution. All UTF-8 characters which are not defined as letters or numbers of the classical Latin alphabet can be removed from the textual corpus. On the other hand, any set of those UTF-8 characters (or all of them) can be kept (preserved) in the text by users' choice. In addition, the NFKD unicode normalization [39], [40] of all Latin script letters can optionally be performed directly in the process of corpus cleaning. Compatibility decomposition replaces the code points of a base letter into a single precomposed letter. For example, unicode letters 'ć' or 'š' can be normalized into 'c' and 's' characters.

b) Removal of stopwords from a corpus: Stopwords are a list of the most common, short function words which do not carry strong semantic properties, but are needed for the syntax of language (pronouns, prepositions, conjunctions, abbreviations, interjections,...). Examples of stopwords are: 'is', 'but', 'and', 'which', 'on', 'any', 'some'. Stopwords from any language based on the Latin script can be removed by providing adequate text file containing the list of stopwords.

c) Lemmatization of a corpus: Lemmatization is the process by which single words are reconducted to their citational form. For instance the word 'networks' is converted into its standard form 'network'. Lemmatization, along with the morphological analysis, is the foundation of all the processes involved in language normalization. Lemmatization can be performed for any language based on the Latin script by providing adequate text file containing the list of all word form-lemma pairs, since the lemmatization in LaNCoA is based on the find-and-replace principle.

d) Text shuffling: Co-occurrence complex networks properties are derived from the word order in texts. Commonly, the shuffling procedure randomizes the words in the text, transforming the text into the meaningless form. Shuffling procedures destroy the sentence and text organization in a way that the standard word-order and syntax of the text is eradicated. As expected, the typical word collocations and phrases are completely lost, as well as the forms of the morphological structures and local structures of words' neighborhood. We implemented two different shuffling principles: shuffling on the sentence level and shuffling on the whole text level. The vocabulary size, word and sentence frequency distributions stay preserved in both shuffling procedures. Additionally, in the sentence level shuffling approach, the sentence structure of the text and the number of words per sentence are also preserved. In the text-level shuffling, the original text is randomized by shuffling the words and punctuation marks over the whole text. This approach also changes the number of words per sentence.

2) Network Generation Module: LaNCoA's network generation module can be used for the generation of complex language networks directly from corpora or from other language networks. It can be used for several independent tasks of building networks on word and subword-level Networks can be generated as weighted or unweighted, as well as directed or undirected. All generated networks can be saved for later

use in the standard edgelist file format. Implemented functions are the following:

a) Co-occurrence networks generation: The co-occurrence window m_n of size n is defined as a set of n subsequent words from a text. Within a window the links are established between the first word and $n-1$ subsequent words. Words are also linked according to the optional usage of specified delimiters (e.g. punctuation marks). In the networks where the linkage is limited to the sentence borders during the construction, the sentence boundary is then the window boundary too. In the networks without delimiters, words are linked within a given co-occurrence window regardless of being in different sentences. Standard approach is to limit the co-occurrence window size within the sentence delimiters, but a user may or may not specify any type of delimiters (any UTF-8 character). The weight of the link between two nodes is proportional to the overall co-occurrence frequencies of the corresponding words within a co-occurrence window. Co-occurrence networks can be generated directly from the raw text data that does not necessary conform to rules of grammar or orthography. Three steps in the network construction for a sentence of 6 words, with usage of the delimiters, for the co-occurrence window sizes $n = 2$ and $n = 6$ are shown in Figure 3.

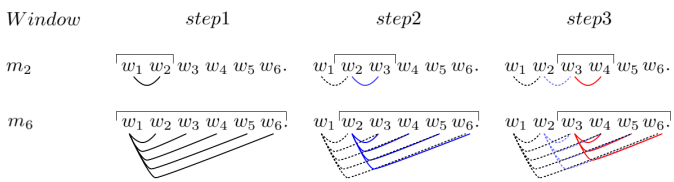


Fig. 3. An illustration of 3 steps in a network construction with a co-occurrence window m_n of sizes $n = 2$, and $n = 6$. $w_1 ... w_6$ are words within a sentence

b) Syntactic networks generation: The syntactic structure of language is captured through syntactic dependency relations between pair of words in a sentence: the head word the governor of relationship and the dependent word - the modifier. Syntactic dependencies between words are formally expressed by dependency grammar which is used to represent the syntactic relationships from sentence in a form of syntactic dependency tree. The sentences boundaries are preserved, since the syntactic dependency is inherent to the sentence. The weight of the link between two nodes is proportional to the overall frequencies of the corresponding words within a syntactic dependency tree. User must provide a corpus in a form of syntactic dependency treebank file written in the CoNLL-X format.

c) Syllable and grapheme networks generation: The syllable networks are constructed from the co-occurrence of syllables within words. Syllable list is obtained from the dictionary file already containing syllabified words. The weight of the link between two syllables is proportional to the overall frequencies of the corresponding syllables co-occurring within words from a text. Syllable networks can be generated directly from the raw text corpora. The structure of grapheme networks depends on a existing network of syllables. Two graphemes are linked if they co-occur as neighbours within a syllable. The weight of the link between two graphemes is proportional to the overall frequencies of the corresponding graphemes co-occurring within syllables from a syllable network.

d) Word-list subnetwork and word-ego subnetwork generation: Two types of word-level subnetworks can be generated from existing co-occurrence or syntactic networks: word-list network and word-ego network. Word-list network is a simple subnetwork based on a provided list of words. Specified nodes (words) and corresponding links between them are extracted from the original network. Word-ego network is a subnetwork of neighbors centered at one specified node (word) within a given radius. These subword-level networks can be generated to examine the networks of semantic importance, word's predecessors, successors, or entire word neighborhood of keywords within a given radius.

B. Network Analysis

Network analysis part of the LaNCoA toolkit consists of four modules: single network analysis module, multiplex network analysis module, content analysis module and data plotting module.

1) Single network analysis module: Single network properties can be analyzed on global and local scale. LaNCoA uses some calculation methods implemented in the NetworkX Python package. These include standard basic network features, for e.g., the average path length, diameter and radius, global and local clustering coefficient, network transitivity, and network density. In addition to NetworkX's procedures used for calculation of classic network properties, LaNCoA provides several other procedures for calculation of non-standard network measures, such as selectivity and inverse participation ratio (both available for directed and undirected networks), calculation of network entropy based on the degree, strength, selectivity and inverse participation distributions, and network reciprocity.

2) Multiplex network analysis module: LaNCoA provides some simple functions for analysis of multiplex networks. Multiplex network is network in layers, and with connections between layers; the interconnections between layers are only between a node and its counterpart in the other layer (the same node). This module enables overlap analysis of two different separated networks consisted of the same sets of nodes. Implemented functions, for example, enable calculation of the Jaccard distance of two different networks, as well as the total and total weighted link overlap measures.

3) Content analysis module: Content analysis implies the examination of the text corpora's content (e.g. role of the individual words) by using the complex network environment. This module provides several ways of text analysis by calculating simple network statistics, such as the top n words, syllables or graphemes with the largest number of different individual neighbors, calculation of the most frequent word-pair relations, the distance between given words in the network environment, or calculation of the centrality measures for all of the words within a corpus.

4) Data plotting module: LaNCoA provides functions for plotting of the network's data by utilizing the methods from the matplotlib Python library. Users can generate various 2D figures based on the calculated network measures. Such figures

include rank plots for the directed (in- and out-) or undirected degree, strength and selectivity distribution values of multiple networks on the same scale, as well as the degree, strength and selectivity histograms and scatter plots. It is also possible to generate the plots describing the dynamic growth of a network regarding the number of connected components, presenting the ratio of newly 'read' unique words (or syllables or graphemes) and the corresponding number of components in a given point of time in the process of co-occurrence network construction.

V. The LaNCoA Toolkit Applications

We have used the LaNCoA toolkit in several of our experiments where we have worked with the language networks.

In [24] we presented the results of our first experiment with the Croatian co-occurrence language networks. In this experiment we constructed 30 different co-occurrence networks, weighted and directed, from the corpus of literature, containing 10 books written in or translated into the Croatian language. We examined the change of network structure properties by systematically varying the co-occurrence window sizes, the corpus sizes and removing stopwords. We used the LaNCoA toolkit for all these network construction tasks.

In [5] we compared Croatian, English and Italian language networks based on the same five books. We performed lemmatized and non-lemmatized network construction with and without stopwords using the LaNCoA toolkit.

In another experiment [25] we addressed the problem of Croatian text complexity by constructing the linguistic co-occurrence networks form normal texts and shuffled text. In this experiment we have tested whether complex network measures can differentiate between normal and shuffled texts. We employed various methods from the LaNCoA toolkit for calculating the network measures and generating various plots in order to find the differences between two classes of networks. In [26] we extended this research by introducing additional shuffling procedure: the sentence-level shuffling procedure and by introducing a node selectivity as a new complex network measure. All shuffling procedures and network construction tasks were performed with the LaNCoA toolkit.

Furthermore, we used the LaNCoA toolkit for various experiments with the selectivity measure. In [36] we compared language networks from Croatian literature and blogs. In [7], [8], [37] we analysed the potential of the selectivity measure for the keyword extraction task. We also used the LaNCoA toolkit for the Croatian language networks construction for the purposes of the network motif analysis of Croatian literature performed in [33]. In [27] we used methods from the LaNCoA toolkit to generate 150 different weighted and directed networks and to calculate local and global network measures used in the task of text classification.

VI. Conclusion

In this paper we presented an overview of the LaNCoA toolkit for language networks construction and analysis. Currently, its basic functionalities rely on the corpora manipulation and language network construction methods implemented in the two separate modules. Another set of modules provide methods for the network analysis task. These modules employ some of the basic methods that already exists in the NetworkX package. However there is a set of functions for the network analysis not covered by the standard network-manipulation packages. Among them are certain functions that deals with the measures for the weighted and directed networks. These functions are of special interest for the language networks analysis and we implemented them in our toolkit.

The LaNCoA toolkit is in the early stage of development and there is still place for major improvements, especially in the network analysis tasks suited for the language networks. However, we managed to use this toolkit successfully in all of the language network experiments that we performed. For the future work, we plan to implement simple and robust user interface. In addition, we would like to develop some more specific language-oriented network analysis functions and also make improvements in the existing code whenever it is possible.

References

[1] O. Abramov, and A. Mehler, "Automatic language classification by means of syntactic dependency networks," Journal of Quantitative Linguistics, vol. 18, no. 4, pp. 291-336, 2011.

[2] S. Arbesman, S.H. Strogatz, and M.S. Vitevitch, "The structure of phonological networks across multiple languages," Chaos, vol. 20, no. 3, pp. 679-685, 2010.

[3] D. A. Bader, and K. Madduri, "SNAP, Small-world Network Analysis and Partitioning: an open-source parallel graph framework for the exploration of large-scale networks," In Proceedings of the International Symposium on Parallel and Distributed Processing, IEEE, pp. 1-12, 2008.

[4] K. Ban, I. Ivakić, and A. Meštrović, "A preliminary study of Croatian language syllable networks," In Proceedings of the 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, IEEE, Croatia, pp. 1296-1300, 2013.

[5] K. Ban, S. Martinčić-Ipšić, and A. Meštrović, "Initial comparison of linguistic networks measures for parallel texts," In Proceedings of the 5th International Conference on Information Technologies and Information Society, Slovenia, pp. 97-104, 2013.

[6] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," In Proceedings of the 8th International AAAI Conference on Web and Social Media, pp. 361-362, 2009.

[7] S. Beliga, and S. Martinčić-Ipšić, "Node selectivity as a measure for graph-based keyword extraction in Croatian news," In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, 2014, pp. 8-17.

[8] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "Toward selectivity based keyword extraction for Croatian news," In Proceedings of the Workshop on Surfacing the Deep and the Social Web, CEUR, Vol. 1301, Italy, pp. 1-14, 2014.

[9] J. Borge-Hothoefer, and A. Arenas, "Semantic networks: structure and dynamics," Entropy, vol. 12, pp. 1264-1302, 2010.

[10] S. Caldeira, P. Lobao, R. Andrade, A. Neme, and V. Miranda, "The network of concepts in written texts," The European Physical Journal B - Condensed Matter and Complex Systems, vol. 49, no. 4, pp. 523-529, 2006.

[11] R. F. i Cancho, V. Sol, and R. Khler, "Patterns in syntactic dependency networks," Physical Review E, vol. 69, no. 5, 2004.

[12] R. F. i Cancho, "The structure of syntactic dependency networks: insights from recent advances in network theory," Problems of Quantitative Linguistics, pp. 60-75, 2005.

[13] R. F. i Cancho, A. Mehler, O. Pustylnikov, and A. Daz-Guilera, "Correlations in the organization of large-scale syntactic dependency networks," TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pp. 65-72, 2007.

[14] M. Choudhury, and A. Mukherjee, "The structure and dynamics of linguistic networks," Dynamics on and of Complex Networks, pp. 145-166, 2009.

[15] G. Csardi, and T. Nepusz, "The igraph software package for complex network research," InterJournal Complex Systems, vol. 1695, no. 5, pp. 1-9, 2006.

[16] A. Hagberg, P. Swart, and D. Chult, Exploring network structure, dynamics, and function using networkx, Technical report, Los Alamos National Laboratory (LANL), 2008.

[17] D. Hansen, B. Shneiderman, and M. A. Smith, Analyzing social media networks with NodeXL: Insights from a connected world, Morgan Kaufmann, 2010.

[18] M. Krishna, A. Hassan, Y. Liu, and D. Radev, "The effect of linguistic constraints on the large scale organization of language," arXiv preprint arXiv:1102.2831, 2011.

[19] O. Kuchaiev, A. Stevanovi, W. Hayes, and N. Pržulj, "GraphCrunch 2: software tool for network modeling, alignment and clustering, " Bioinformatics, vol. 12, no. 1, pp. 24, 2011.

[20] W. Li, "Random texts exhibit zipf's-law-like word frequency distribution," IEEE Transactions on Information Theory, vol. 38, no. 6, pp. 1842-1845, 1992.

[21] H. Liu, and F. Hu, "What role does syntax play in a language network?," Europhysics Letters, vol. 83, no. 1, pp. 18002, 2008.

[22] H. Liu, and C. Xu, "Can syntactic networks indicate morphological complexity of a language?," Europhysics Letters, vol. 93, no. 2, pp. 28005, 2011.

[23] H. Liu, and C. Jin, "Language clustering with word co-occurrence networks based on parallel texts," Chinese Science Bulletin, vol. 58, no. 10, pp. 1139-1144, 2013.

[24] D. Margan, S. Martinčić-Ipšić, and A. Meštrović, "Preliminary report on the structure of Croatian linguistic co-occurrence networks," In Proceedings of the 5th International Conference on Information Technologies and Information Society, Slovenia, pp. 89-96, 2013.

[25] D. Margan, S. Martinčić-Ipšić and A. Meštrović, "Network differences between normal and shuffled texts: Case of Croatian," Springer, Studies in Computational Intelligence, Complex Networks V, vol. 549, 2014, pp. 275-283.

[26] D. Margan, A. Meštrović, and S. Martinčić-Ipšić, "Complex networks measures for differentiation between normal and shuffled Croatian texts," In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, Croatia, IEEE, pp. 1819-1823, 2014.

[27] D. Margan, A. Meštrović. M. Ivašić-Kos, and S. Martinčić-Ipšić, "Toward a Complex Networks Approach on Text Type Classification," 6th International Conference on Information Technologies and Information Society, Slovenia, 2014.

[28] A. Masucci, and G. Rodgers, "Network properties of written human language," Physical Review E, vol. 74, no. 2, pp. 026102, 2006.

[29] A. Masucci and G. Rodgers. "Differences between normal and shuffled texts: structural properties of weighted networks". Advances in Complex Systems, vol. 12, no. 1, pp. 113-129, 2009.

[30] T. Milenković, and N. Pržulj, "Uncovering biological network function via graphlet degree signatures," Cancer Informatics, vol. 6, pp. 257273, 2008.

[31] M. E. J. Newman, Networks: An Introduction, Oxford University Press, 2010.

[32] W. De Nooy, A. Mrvar, and V. Batagelj, Exploratory social network analysis with Pajek. Cambridge University Press, Vol. 27, 2011.

[33] H. Rizvić, S. Martinčić-Ipšić, and A. Meštrović, "Network motifs analysis of Croatian literature," In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, pp. 2-7, 2014.

[34] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," Genome Research, vol. 13, no. 11, pp. 2498-2504, 2003.

[35] R. V. Sole, B. Corominas-Murtra, S. Valverde and L. Steels, "Language networks: Their structure, function, and evolution," Complexity, vol. 15, no. 6, 2010, pp. 20-26.

[36] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, "Comparison of the language networks from literature and blogs," In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, IEEE, Croatia, pp. 1824-1829, 2014.

[37] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, "Toward network-based keyword extraction from multitopic web documents," In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, pp. 18-27, 2014.

[38] S. Wernicke, and F. Rasche, "FANMOD: a tool for fast network motif detection," Bioinformatics, vol. 22, no. 9, pp. 1152-1153, 2006.

[39] Python Software Foundation. "Unicode Database Documentation," [Online]. Available: https://docs.python.org/2/library/unicodedata.html#unicodedata.normalize [Accessed: April 10, 2015].

[40] The Unicode Consortium. "Unicode Standard Annex #15: Unicode normalization forms," http://unicode.org/reports/tr15/#Norm_Forms [Accessed: April 10, 2015].

Parallelization of Two Popular Parallel Sorting Algorithms on GPU Using OpenCL

Azinović Dejan*, Hasanspahić Aldin*, Hasić Haris*, Novica Nosović*
*Faculty of Electrical Engineering/Department for Computer Science, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
dazinovic1@etf.unsa.ba, ahasanspah1@etf.unsa.ba, hhasic2@etf.unsa.ba, nnosovic@etf.unsa.ba

***Abstract* — This paper presents a comparative analysis of the two most commonly used parallel sorting algorithms: Bitonic Sort and Parallel Radix Sort. Many problems require efficient sorting of large amounts of data and that is why sorting has been, and still is, an important discussion topic in computer science. Traditionally, sorting algorithms are implemented on CPUs, but the performance of the CPU doesn't increase as steadily as until about 10 years ago. It has been proven that the complexity of comparison based sorting algorithms can be no lower than O(*nlogn*). The computing power of parallel devices is rapidly increasing, which indicates a new approach towards improving the efficiency of sorting. The benchmark in this paper will show whether or not any speedup can be gained from more compute power, and if so, can it outweigh and by how much, the overhead caused by a higher complexity of the sorting algorithm. This will be analyzed by comparing the performance of *OpenCL* kernels of the two mentioned parallel sorting algorithms with the sequential implementation of the Quick Sort algorithm. The hardware used for this experiment is an *AMD R9 270* GPU and an *Intel i5-4460* CPU.**

***Keywords* – Sorting algorithms, *OpenCL*, GPGPU**

I. Introduction

Sorting algorithms, nowadays, are a very common problem in computer science and a very often discussed topic mainly because the overall improvement of sequential sorting algorithms has stopped in the last few decades. By following the trend in overall computing, the most obvious answer lies in parallel computing, but is it really possible to divide the sorting process into individual independent pieces fit for parallel execution? After much work and effort that has been put into the development of such parallel versions of sorting algorithms, it has been proven that sorting really can be "divided and conquered", which resulted in the creation of a very large number of sorting algorithms that show the advantages of parallel processing. But how much speedup can there really be achieved ?

As mentioned before, the best sequential sorting algorithm today (without using special properties of the elements) has the complexity *O(nlogn)* [6]. Generally speaking, by using a parallel algorithm the *optimal* performance that can be achieved on *n* processors is [6]:

oc – Optimal Complexity
stc – Sequential Time Complexity
nop – Number of processors

$$oc = \frac{stc}{nop} = \frac{O(n\,logn)}{n} = O(log\,n) \qquad (1)$$

Equasion (1) does not take into account the overhead spent on communication between processors and this makes, generally, a *realistic O(logn)* parallel sorting algorithm hard to achieve. Nevertheless, any improvement made represents a step towards the goal: faster sorting.

Today, there are a lot of proposed algorithm models for parallel sorting, but the focus of this paper lies on the two most commonly used:

- Bitonic Sort
- Parallel Radix Sort

It is very hard to determine which algorithm is the best for which particular architecture taking into consideration the vast variety of parallel architectures and the fact that there is no theoretical model that can be applied to predict the performance of an algorithm on different architectures. This makes experimental studies, like this paper, more important for the behavior and performance evaluation of different sorting algorithms on different parallel machines.

The *AMD R9 270* has 20 compute units of which each has 64 ALUs. Each ALU can be considered a processing element in *OpenCL* terminology. [4] As a result there is a total of 1280 processing elements. It is important to split the work appropriately so that no processing element is idle while other processing elements do work. The GPU core works at a frequency of 975 *MHz* and the memory at a frequency of 5600 *MHz*. The single precision processing power of the card is about 2.3 *TFLOPs*. Since the card uses a 256-*bit* wide bus, this results in a maximum bandwidth of 179.2 *GB/s*. [1] Whether this bandwidth given the card's compute power will be a bottleneck in the sorting algorithms will depend on the algorithm's implementation and this will reviewed later in more detail.

The other important hardware component in the experiment is the *Intel i5-4460* CPU. It has four cores which work at a clock rate of 3.2 *GHz*. Each of the cores has 64 *KB* of L1 cache and 256 *KB* of L2 cache. On top of that there is a shared L3 cache of 6 *MB*. [2] Caching will play an important role in the sequential implementation of the sorting algorithm. As will be shown in the next section, Bitonic Sort's time complexity is $O(nlog^2n)$. Quick Sort has a time complexity of only *O(nlogn)*. Given that Bitonic Sort will be executed on 1280 more threads and that the clock rate ratio between the GPU and CPU is 9.75/32 = 0.304 a speedup of at most 1280*0.304/*logn* = 390/*logn* is expected where *n* is the number of elements sorted. Radix Sort's complexity, in the best case scenario, is *O(kn)* [7] and it is expected that it gives us the best performance.

The expected performance for each of the three algorithms is shown in *Figure 1* :

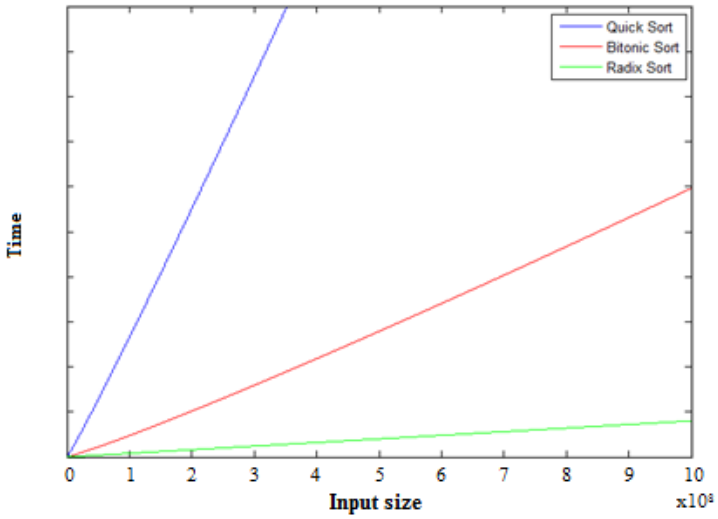


Fig. 1. Expected performance of the mentioned sorting algorithms

It should be repeated that this is only the ideal case that represents the maximum speed up. What can be achieved in practice is different.

II. Parallel Sorting Algorithms

In this section, the Bitonic Sort and Parallel Radix Sort algorithms will be described.

A. Bitonic Sort

Bitonic Sort is a well-suited algorithm for parallel architectures. It works by creating bitonic sub-sequences in the original array, starting with sequences of size 4 and continuously merging the sub-sequences to generate bigger bitonic sub-sequences. Finally, when the size of this sub-sequence is the size of the array, it means the entire array is a bitonic sequence. By repeated applications of the bitonic split, a bitonic sequence can be converted into a monotonic sequence and thus a sorted array is achieved. The time complexity of the Bitonic Sort algorithm is $O(nlog^2n)$.

Next, the algorithm is going to be described in more detail. For clarification, the term bitonic sequence must be well defined. A bitonic sequence is a sequence that monotonically increases (decreases), reaches a maximum (minimum), and then monotonically decreases (increases). A bitonic split is an operation on a bitonic sequence such that if $a_i > a_{i+n/2}$, the two elements are exchanged, $1 \leq i < n$. This will produce two bitonic sequences, A and B, with the property that all elements of A are lesser than any element in the sequence B. If the process would be repeated on both A and B, it would lead to four bitonic sequences such that each sequence would contain elements that are lesser than the elements in the next sequence. Repeating this process until *n/2* bitonic sequences are produced will result in a sorted array. It is clear that a bitonic sequence can be sorted in *O(logn)* steps where each step has *n* operations. This results in a complexity of *O(nlogn)* for sorting a bitonic sequence. [3]

Next up is the arbitrary sequence of numbers. Sorting an array of length 2^n is done in *n* stages. Every *i*-th stage has a total of *i* passes. Every pass does length/2 comparisons. Since the total number of passes is $O(log^2n)$ and each pass does *O(n)* comparisons, the total time complexity of the algorithm has to be $O(nlog^2n)$.

The first stage of the algorithm converts the unsorted array into segments of length 2. Every pair of adjacent segments forms a bitonic sequence. The next stage produces segments of length 4 and so on. In each pass elements at a distance of $2^{(Stage\ -\ PassOfStage)}$ are compared. The elements are exchanged depending on the sort direction for the particular block (the length of the block equals twice the distance). The first stage inverts the direction of sorting after every successive pair of elements. The second stage does this after two pairs of elements, the third after four pairs and so on. After n stages the whole array is a bitonic sequence.

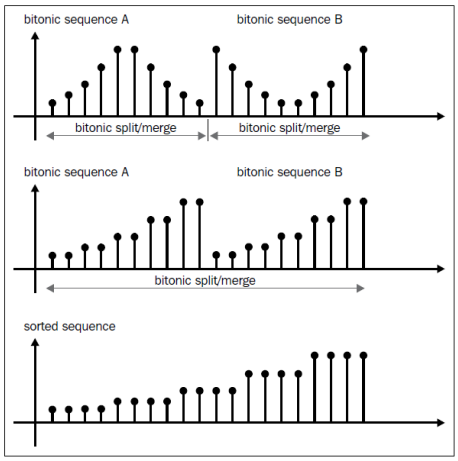
Now it's time to examine the OpenCL implementation of the Bitonic Sort algorithm. The variable *pairDistance* is the distance between the pair of elements that are being compared in each invocation of the kernel and the variable *threadId* is the thread identificator. A block is a group of consecutive elements in the array that are being compared. The variable *blockwidth* is the width of this block. Each processing element can compute the indices of the elements in the array that need to be compared. The index of the left element can be computed according to the following formula:

$$Leftid = (threadId \ \% \ pairDistance) + (threadId \ / \ pairDistance) * blockwidth \quad (1)$$

And the index of the right element can be computed as:

$$Rightid = Leftid + pairDistance \quad (2)$$

An illustrative example of how the whole process is executed and the Bitonic sort *OpenCL* kernel pseudo code are given in *Figure 2* and the code snippet below.




```
__kernel sort(__global *data, int stage i, int pass_of_stage j, int dir)
{
        /*using values of i, j, dir – get left_Id & right_Id*/
        left_child = data [left_Id]
        right_child = data [right_Id]
        compare(left_child, right_child)
        /*copy left & right child values to data with respect to dir*/
        data [left_child] = max(left_child, right_child)
        data [right_child] = min(left-child, right_child)
}
```

Fig. 2. Bitonic Sort execution example [4] & pseudo code

It can be concluded that over 90% of the code is parallelizable. The host processor only has to pass the stage number and the pass of the stage to the processing elements. This means that nearly all of the computing time is spent executing the parallelizable code which is implemented as the kernel function.

B. Radix Sort

Radix Sort treats keys as multi-digit numbers in which each digit is an integer with a value between 0 and m, where m is the radix. [4] Given this definition a 32-bit integer could be treated as a 4-digit number with $m = 2^{32/4} = 256$. Radix Sort sorts keys by splitting them into digits and sorting the digits one at a time, starting from the least significant digit. In each pass keys are distributed into buckets according to the digit currently being processed and then collected from the buckets. This is illustrated in the following example :

Suppose an array consists of the following input elements: 72, 32, 13, 84, 95, 42, 53, 61, 14 and 15 and the base (*radix*) is 10. During the first pass the keys will be distributed into bins as follows:

(61), (72, 32, 42), (53, 13), (84, 14), (95, 15)

The keys are collected and distributed again according to the second (and in this case most significant) digit :

(13, 14, 15), (32), (42), (53), (61), (72), (84), (95)

After collecting the keys again, a sorted array will be the end result.

In an *OpenCL* implementation, keys are broken into 8-*bit* digits and sorted one digit at a time, starting with the least significant digit. This is repeated four times to complete the sorting procedure (sorting of 32-*bit* integers). Each time the following three phases are executed to sort the appropriate digit for that phase:

1) Calculate the histogram of an unsorted array
2) Pre-scan the histogram bins
3) Rank and permute the keys to get a sorted array

A detailed description of each of the phases will be given in the next section.

In the first phase, the input array is divided into blocks of N * M elements. M is the radix (256 in this case) and N is the number of work-items in a block. Ideally, each thread of the GPU would be assigned a portion of the work that needs to be completed. With this in mind, the ideal number for N for the *AMD R9 270* is N = 64.

In the second phase, the histogram bins are pre-scanned. The pre-scan operation takes a binary associative operator $\oplus$ with identity I, and a vector of n elements.[8][9] There are B * N histograms where B is the number of blocks and N the number of work-items in a block. The bins are arranged so that the 0^{th} block bin comes first, and the $(B - 1)^{th}$ block bin last. Each block's histogram bins are arranged so that the 0^{th} work-item bin comes first, and the $(N - 1)^{th}$ work item bin comes last. After the pre-scanning is completed, the histograms are passed into the third and final phase.

In the third phase, each work-item permutes its allotted 256 elements by using its pre-scanned histogram bins.

The pseudo code for the *OpenCL* kernel is given below:

```
b ← no. of bits
A← Input Data
cmp ← 1
cnt0 ← contains zero's count
cnt1 ← contains one's count
One, Zero ← Bucket Arrays
Mask ← Temporary Array

for ( i = 0 to 2b – 1)
{
    for ( j = 0 to A.size)
    {
        if (A [j] && cmp)
            cnt1 ++
            One [cnt1] ← a[j]
        else
            cnt0 ++
            Mask [cnt0] ← j
     }
    for( j = cnt0 to A.size)
        Mask [j] ← A.size – cnt0 + j
    A ← shuffle(A, one, Mask)
    cmp ← left_shift(cmp)
}
result ← A
```

In conclusion, this algorithm is embarrassingly parallel, as each bin can be sorted independently. The only sequential part of the program would be the sorting by the least significant digit. In the ideal case, as the program progresses, more and more processors would get engaged in the sorting process.

III. Main Results

As already mentioned in the introductory section, the Quick sort algorithm has been executed on the *Intel i5-4460* CPU and the Bitonic and Radix sort algorithms have been executed on the *AMD R9 270* GPU. The execution times for various input sizes are presented in the *Figure 3* :

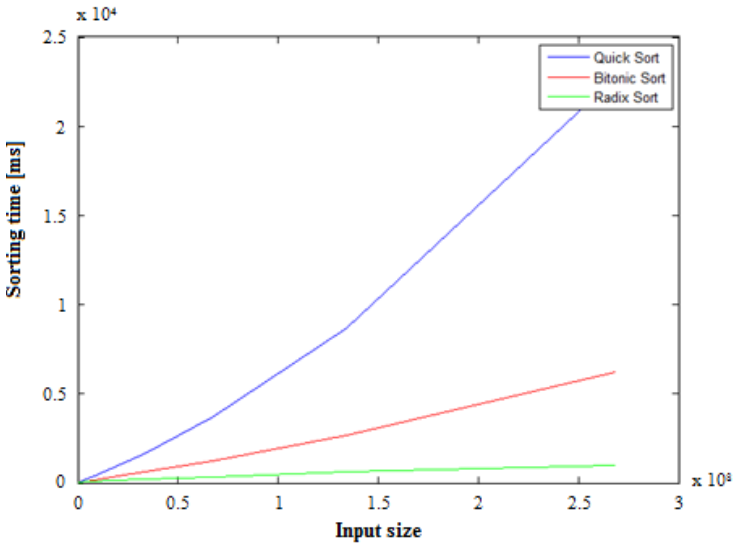


Fig. 3. Execution time results for the mentioned parallel sorting algorithms

As can be seen in the table and graph, the performance of the parallel sorting algorithm is much better than the performance of Quick sort for large input arrays. For small input arrays (up to a million elements), Quick sort performs better. The reason why this happens is because of the overhead introduced in the parallel algorithms. It takes time to split the work and distribute it to the processing elements. That overhead outweighs the gains gained from parallelization in small problems. This was expected since parallel algorithms are developed to handle large problems.

The Quick sort algorithm behaves as expected. The increase in computation time is super-linear starting from 1 million elements up. For smaller problem sizes the increase in computation time is sub-linear. This is caused by caching. As the problem size increases, a lower percentage of the data can be cached and thus it takes longer to fetch it for computation.

The same can be said about the Bitonic Sort algorithm. The difference here is that the increase in computation time becomes super-linear after a much larger problem size. The larger the problems size, the more insignificant does the overhead become. It should be noted that the speedup gained is much lower than the ideal speedup.

As already mentioned, caching reduces the sorting time when Quick sort on the CPU is executed. On the other hand, all data on the GPU is fetched from global memory. The latency of the global memory severely limits the speedup potential. The bandwidth is sufficient for all problems up to a billion elements. At that point it's the capacity of the global memory that limits further gains from parallel processing.

As expected, Radix Sort has an almost linear increase in computation time and is also the fastest among the three tested sorting algorithms. Local memory was used to increase performance. The reason why the computation time increase is only almost linear and not truly linear is the limited capacity of the local memory. There is only 64 KB of local memory available to each of the compute units. Another performance limit is imposed by the latency of the global memory. Local memory reduces that latency, but due its limited capacity data must be continuously transferred between the global memory and local memory if the problem size is too large.

IV. Conclusion and Future Work

The performance of Bitonic Sort and Parallel Radix Sort for the GPU using *OpenCL* has been analyzed and compared to the serial implementation of the Quick Sort algorithm. The conclusion is that speedup can be achieved by parallelizing sorting algorithms. The speedup depends on the number of elements that need to be sorted. As presented in the graph, the larger the array, the bigger the speedup. For arrays larger than 64 million elements Bitonic Sort performs more than three times better and Radix Sort more than ten times better than Quick Sort with some limitations that need to be taken into account. Nonetheless, there are many problems with these characteristics and the research of parallel sorting algorithms will be an important topic in the future.

Hardware is also an important factor in parallelization of the algorithms. The work must be split and distributed in a way that all processing elements receive an equal share of the work which can be a major problem in the implementation of sorting algorithms. Given the diversity of available hardware it is impossible to make a single best implementation since the best implementation is always slightly different for each hardware component. In the case of Parallel Radix Sort, it is important to split the array into blocks so that the element count divided by the radix is equal to the number of processing elements in a single compute unit of the GPU. Since the AMD R9 270 has 64 processing elements per compute unit, the element count in each block would ideally be M * 64 where M is the radix.

The biggest problem was the latency of the global memory. Bitonic sort can be further improved by making use of local memory. Radix Sort's limitation is the capacity of the local memory.

The capacity is only 64 KB and for arrays larger than 65536 elements, data needs to be swapped between the global and local memory which limits speedup. These are the biggest two problems which need to be solved in future work on this topic.

References

[1] AMD APP SDK Documentation

[2] Intel Core i5-4460 Processor Documentation http://ark.intel.com/products/80817/Intel-Core-i5-4460-Processor-6M-Cache-up-to-3_40-GHz

[3] http://facultyfp.salisbury.edu/taanastasio/COSC490/Fall03/Lectures/Sorting/bitonic.pdf Explanation of Bitonic Sort with examples by Thomas Anasosio, 2003

[4] T. Raymond, "OpenCL Parallel Programming Development Cookbook", Packt Publishing Ltd. 2013.

[5] B. R. Gaster, L. Howes, D. R. Kaeli, P. Mistry and D. Schaa, "Heterogeneous Computing with OpenCL" Revised OpenCL 1.2 Edition USA 2013.

[6] F. Silva, *Parallel Algorithms- sorting, DCC-FCUP*

[7] http://www.dcs.gla.ac.uk/~pat/52233/slides/RadixSort1x1.pdf

[8] http://cis565-fall-2012.github.io/lectures/09-26-Parallel-Algorithms.pdf

[9] https://www.cs.cmu.edu/~guyb/papers/Ble93.pdf

Mobile application for finding ATMs

Mario Lončar, Jasmin Abou Aldan, Marina Ivašić-Kos
Department of Informatics, University of Rijeka, Rijeka, Croatia
{mario.loncar, jasmin.abou.aldan, marinai}@uniri.hr

Abstract - This paper describes mobile application that gives Android users the ability to find ATMs quickly. Currently, the application offers locations of ATMs for various banks in Rijeka, Croatia. It also provides the possibility of finding the shortest path to the chosen ATM, as well as helpful tips where they can withdraw the money free of charge. The idea for this application was born after realizing that many people were wasting time looking for a place to withdraw money. This application saves time to inhabitants as well as to visitors of the city.

Keywords: mobile application, Android, geo-locations, maps, navigation.

I. INTRODUCTION

Many people waste their time looking for places where they can withdraw the money. Positions of the closest ATM are sometimes unknown to the inhabitants, especially when they are in the part of the city that do not visit often. Finding the ATM becomes even more demanding task for tourists in unfamiliar city. They want to spend as much time as possible having fun and sightseeing instead of wandering to get the money.

Touristic cities like Rijeka usually offer Wi-Fi hotspots at most frequently visited places so everyone can use internet connection. This opportunity can be used to download and use available applications on Google Play that facilitate the user to gather additional information about the destination. Below are applications that could be used for finding ATM together with their shortcomings.

AdriaGUIDE [13] Rijeka is a travel guide that provides most important information for the visitors, including ATM locator. However, after choosing a language on welcome screen, the application crashes so we could not see its functionality. Last update was on October 12, 2012.

Bankomati HR [14] offers ATM locations of all Croatian cities. On the other hand, it offers only eight banks and no navigation to chosen ATM. User interface looks out of date and there is no option for showing locations of all ATMs, instead you have to check each of them. On the bottom of the screen are ads that cannot be moved. Last update was on September 16, 2013.

Adria ATM [15] has intrusive ads on the top of action bar. When starting this application, map shows all ATMs of all banks on the list. If you want to check ATM locations of only one bank you have to uncheck banks on the list and then choose the bank you need which takes a lot of time. Application does not offer navigation, but only address of an ATM. It has implemented crowdsourcing, but crashes sometimes after adding a new ATM location. Last update was on August 2, 2013.

BankoMap [16] does not offer navigation nor banks to choose, every bank has the same default pin. It offers a small amount of ATMs. Last update was on October 12, 2010.

In order to help users finding an ATM quickly and easily, we developed Find My ATM application. It is a mobile application that offers ATM locations of all banks in Rijeka, useful advices on where the money can be withdrawn free of charge and shortest path to each ATM.

II. APPLICATION OVERVIEW

Find My ATM is mobile application developed for Android platform using Android Studio [3]. Android Studio has recently replaced Eclipse with Android Development Tools (ADT) and become the official Java integrated development environment (IDE) for developing Android applications.

It is based on Java IDE developed by JetBrains and available as an IntelliJ IDEA open source community edition [8]. The IntelliJ IDE enables fast coding by giving smart and relevant suggestions, code completion, on-the-fly code analysis, project navigation and reliable refactoring tools. It has better layout previewing and is much easier to use and user friendly than Eclipse.

The key components of the Find My ATM application are working with maps, determining the geo-location of the user and computing the shortest path to the selected ATM. Therefore, the Android package that includes map libraries was added to the Android software development kit (SDK).

A. Android mapping libraries

We have considered several mapping libraries available for Android applications [1]: Google Maps API [2], ArcGIS Runtime API [9], OSMDroid API [10] and MapsForge [11].

Google Maps Android API is the default mapView for Android, and uses the data from Google Maps. It allows presentation of maps as satellite, terrain, and hybrid as well as Street View. Vector-based tiles of Google map can be used in offline mode. Additionally, map tiles can be customized adding information and markers onto the map to indicate specific points for an application. Customized maps can be shown in online mode.

ArcGIS runtime API shows data in offline mode, by storing tiles in the TIFF, PNG or JPG storage format, or by using tile packages. Users can access high-quality geospatial data and explore raster data and aerial imagery through Image GeoServices. OSMDroid is also library of raster maps and provides using tiles for off-line access.

MapsForge is library for rendering OSM (Open Street Maps) data on the fly. It has a special optimized tile format, and good rendering performance. Nutiteq library is open source if it is used with OpenStreetMaps. It has many unique features like support for 3D, custom Map API's, offline routing and maps in any projection.

After testing different mapping APIs, we have chosen Google Maps API. The main reason was the familiarity of users to Google Map interface and functionalities since it is preinstalled on every Android device.

B. Set up of Google Maps in application

The Google Maps Android API uses OpenGL ES version 2 to render the map. If OpenGL ES version 2 is not installed, map will not appear. It is necessary to add the following in AndroidManifest.xml:

```
<uses-feature
        android:glEsVersion="0x00020000"
        android:required="true"/>
```

This notifies external services of the requirement. In particular, it has the effect of preventing Google Play Store from displaying your app on devices that does not support OpenGL ES version 2.

To use Google Maps in an application you need to create a valid Google Maps API key. The key can be used with any application that calls Google Maps API, and supports unlimited number of users. To obtain the key one should register to the Google APIs Console providing the application signature key given to application during development and the application package name. Then, a valid API key must be entered in AndroidManifest.xml file. Later it will be necessary to register application via its package in this Play Store console together with the SHA-1 fingerprint of signature key.

C. User interface

Android Studio offers advanced layout editor that allows you to drag-and-drop widgets into your layout and preview layout while editing the XML. In general, the XML vocabulary for declaring user interface elements closely follows the structure and naming of the classes and methods, where elements names correspond to class names and attribute names correspond to methods. This application has two XML layouts. First is for the main activity, which includes Google Maps fragment and *listview* for banks. Second XML layout is consisted of one *textview* that presents one row in *listview*.

An Android application typically has several activities. Each activity displays a user interface that allows the user to perform a specific task such as to select an item or view a map. To perform interactions between application components or different applications, an Intent should be used. Intent is abstract description of an operation to be performed. It is most often used for launching of activities.

Our application is launched by passing a method *startActivity*() to the system that opens a single screen with user interface showing all banks that have ATMs in Rijeka, Fig. 1. To define the list of all banks in Rijeka, data available on the website [12] was used. For locations of each ATM, geo-coordinates posted on the official documents of each of the banks were used.

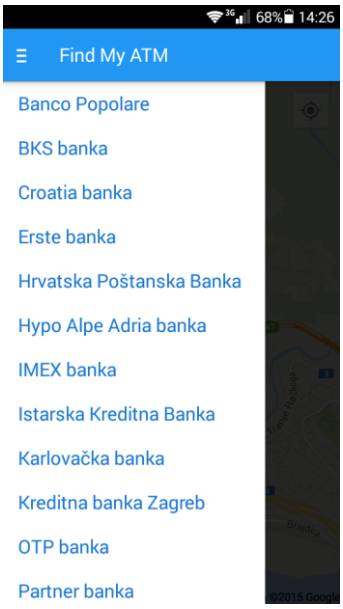


Figure 1. Navigation drawer displays available banks in Rijeka

Data about banks and their ATM coordinates and addresses is stored in JSON format database on Parse.com [6]. Parse.com offers backend services for the developers as well as SDK for a multitude of platforms. It allows accessing data through objects. When you create a new App in Parse, you can create a new class and then import data to that class. Imports can be from CSV, XLS, TXT and JSON files, or you can manually create a table in each class by defining the columns and types of the stored data. The important feature of Parse.com is the subsequent implementation of crowdsourcing. In that way, users can enter position and the address of new ATM on their own, as well as report possible wrong coordinates or address of an ATM. In order to use Parse classes and methods, a developer has to include Parse SDK files to Android Studio project within the folder libs. The following code directs *build.gradle* to look for appropriate libraries in the folder libs:

```
dependencies: {

   compile fileTree(dir: 'libs', include:
   ['*.jar'])

   }
```

To connect an application with Parse Core, in *onCreate*() method, a *Parse.initialize*() method is embedded that takes context, application ID and client key as parameters. If *ParseAnalytics*().*trackAppOpened*() and

ParseAnonymousUtils.logIn() methods are implemented in application, developer can keep track and get daily reports about usage of the application.

The main advantage of using Parese.com is that the developer does not have to think about the infrastructure, because the team responsible for Parse takes care of it. By using Parse, one can easily upgrade a database that is immediately available for use without entering the code of the application.

List of 20 banks is stored in Activity object in string array called banks_array:

```
mBankNames = getResources()
.getStringArray(com.mario.findmyatm.R.array.ba
nks_array);
```

To provide access to the data items that should be presented and to enable view for each item, an Adapter object is used. An Adapter object acts as a bridge between types of available displays of data defined in *AdapterView* object and the data that will be used for a selected view. In this application *ArrayAdapter* object is used and backed by an array of arbitrary *mBankNames* objects. The *ArrayAdapter* class expects that provided resource id reference a single *TextView* in the larger layout resource. When *TextView* is referenced, it is filled with each object in the *mBankNames* array.

```
    mDrawerList = (ListView) findViewById
(com.mario.findmyatm.R.id.left_drawer);

    mDrawerList.setAdapter(new ArrayAdapter
<String> (this,
com.mario.findmyatm.R.layout.drawer_list_item,
mBankNames));
```

D. Implementation and Customization of Google Maps

If the Internet connection is available, user can choose one bank among 20 of them on the list for which wants to see the ATM locations. In addition, an option that displays ATMs of all banks on the list is available too. To perform these operations the system uses the Intent that identifies and starts the appropriate service with method *SetUpMapIfNeeded*. The Intent can be also used to start an activity that is contained in a separate application or to communicate with a background service using method *bindService(Intent, ServiceConnection, int)*.

A method *SetUpMapIfNeeded*() sets up the map if the Google Play services APK is correctly installed and the map has not already been instantiated. If the map is not installed, a prompt for the user is shown to install/update the Google Play services APK on his device. A user can return to this fragment activity after following the prompt and correctly installing/updating/enabling the Google Play services. Since the fragment activity may not have been completely destroyed during this process, *onCreate*() method may not be called again so we should call this method in *onResume*() to guarantee that it will be called.

If the map is installed properly, data taken from Parse Core [7] is displayed at map in the form of pins (Fig. 2), each time the user selects a bank item in the list view.

Figure 2. Pins on the map present locations of ATM for a selected bank

The address and coordinates of an ATM on Parse Core look like this: *["Riva Boduli 1, 45.325389, 14.442115"]*. When user selects a bank, method that gets object identified with a given key on Parse is called. Method stores object from Parse to an *ArrayList*<String>. With foreach loop, the *ArrayList* is split by comma and its elements are stored in variables that are later used for drawing the pins on a map. The code snippet used to get data from Parse is as follows:

```
for (String s : myArray) {

    String[] s2 = s.split(",");
    String address = s2[0];
    Float lat = Float.parseFloat(s2[1]);
    Float lng = Float.parseFloat(s2[2]);
    LatLng coo = new LatLng(lat, lng);
}
```

Instead of standard Google Maps pins, we used custom pins, so that each bank uses specific pin as its logo (Fig. 2). Pins are 32 x 32 .png format images. After selecting a bank, action bar changes title with *setTitle*() method according to the bank we have chosen:

```
mTitle = title;
getSupportActionBar().setTitle(mTitle);
```

E. Implementation of Navigation Directions

To be able to provide information to the user about the closest ATM, the current location of the user should be determined.

One has to allow access to the location information in the settings in order to use My Location. Otherwise, an Alert, which leads to Location settings, will pop up as shown on the Fig. 3.

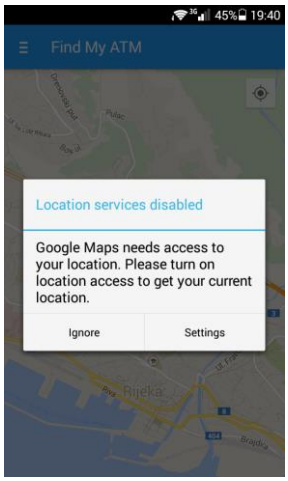


Figure 3. Location services alert

To check if Location access is enabled we have used the code snippet:

```
if(!locationManager.isProviderEnabled(Location
Manager.GPS_PROVIDER) &&
!locationManager.isProviderEnabled(LocationMan
ager.NETWORK_PROVIDER))
    //alert shows up
```

Also an Internet connection is required for using the application. If a user does not have Internet connection, an Alert which leads to Internet settings will appear (Fig. 4). We used methods from Connectivity Manager class to check whether Network Connection is available or not:

```
connectivity(){
    if(isNetworkAvailable()){
       //set up map
    }else{
       //alert shows up
    }
}
```

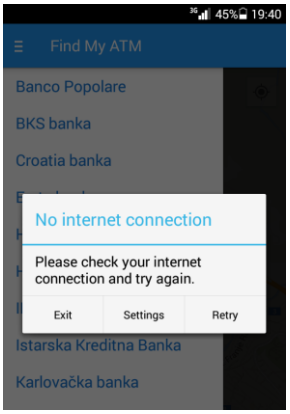


Figure 4. Connection alert

When the location information is enabled and network connection is available, the button with label My Location appears in the upper right corner and shows the location of the user (Fig. 5).

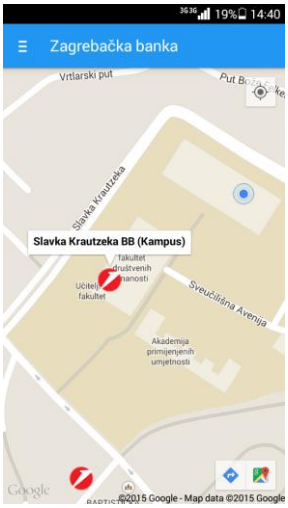


Figure 5. Location of the user is marked in the screen

Navigation options appear when one of the ATMs is selected. The navigation to a specific ATM depends on the chosen means of transport. All functions from Google Maps API are implemented so the user can select the navigated to a specific ATM by car or by public transport or on foot (Fig. 6). For the selected navigation type, several different routes are offered and time necessary to reach the selected ATM is calculated. In addition, voice control can be used to control Navigation.

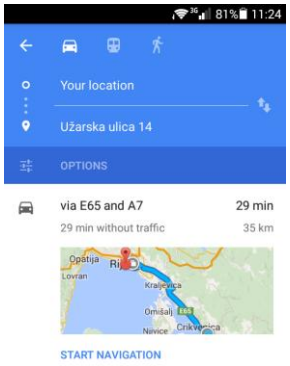


Figure 6. Navigation to the selected ATM

Furthermore, an application offers all sorts of information available for the selected ATM and advices on where you can withdraw your money free of charge (Fig. 7). Advices appear in a Dialog box that can display one, two or three buttons. Which button to implement depends on dialog's action. Method *setPositiveButton*() is used when you should accept and continue with the action. For canceling the action method *setNegativeButton*() is used. Method *setNeutralButton*() is used when the user may not want to proceed with the action, but does not necessarily want to cancel. It is possible to add only one of each button type to an

AlertDialog. To display information to the user in a dialog box we have used the *setMessage*() method.

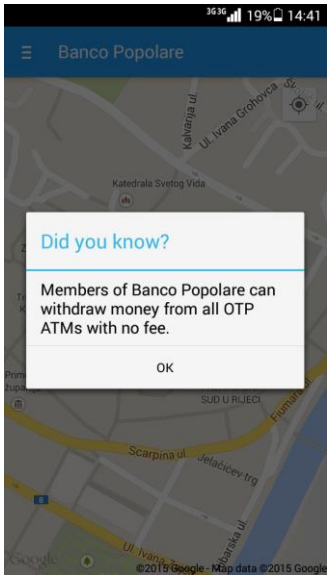


Figure 7. Tip about selected ATM

Find My ATM runs on android 4.0 or higher which covers 87% of Android devices [4]. Using lower Android API versions may affect in losing core feature set. Application requires multiple permissions defined in *AndroidManifest.xml* (Fig. 8).

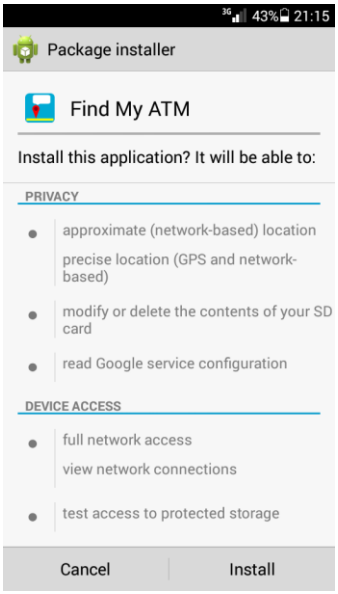


Figure 8. Permissions

We have used *Genymotion* [5] emulator for testing the application as well as physical devices. Genymotion is a fast third-party emulator that can be used instead of the default Android emulator that turned to be very slow even after installing Intel Hardware Accelerated Execution Manager.

III. CONCLUSION

We have developed the Android mobile application Find My ATM that enables finding of ATM locations quickly. The application also provides the possibility of finding the shortest path to the chosen ATM, as well as helpful tips where money can be withdrawn free of charge. The application is currently available only for various banks in Rijeka. However, an expansion on other Croatian cities is expected.

Inhabitants find this application very useful. Foreign tourists does not have an internet connection in Croatia, but major Croatian cities offer Wi-Fi hotspots on frequently visited places, so in future we will try to make an arrangement with tourist office for a free promotion of this application.

Currently, Find My ATM is in version 1.1.1 available for Android. It will be available for iOS users soon. In addition, voice search on Croatian language will be implemented so ATMs will be even easier to find. Parse.com offers an easy way to implement local data store, so we are planning to add this feature, as well as offline maps, in version 1.2. Other upgrading of the application will be made according to the user's needs and suggestions.

REFERENCES

[1] gis.stackexchange.com; questions; Geographic Information Systems; http://gis.stackexchange.com/questions; (Accessed 2014-12)

[2] developers.google.com; Google Maps Android API v2; https://developers.google.com/maps/documentation/android/; (Accessed 2014-12)

[3] developer.android.com; Tools; http://developer.android.com/sdk/index.html; (Accessed 2014-08)

[4] developer.android.com; Dashboards; https://developer.android.com/about/dashboards/index.html; (Accessed 2014-12)

[5] genymotion.com; Product; https://www.genymotion.com/#!/product; (Accessed 2014-12)

[6] Parse.com; Android guide; https://parse.com/docs/android_guide (Accessed 2014-12)

[7] Parse.com; Parse Core; https://www.parse.com/products/core; (Accessed 2014-12)

[8] IntelliJ IDEA, https://www.jetbrains.com/idea/ (Accessed 2014-12)

[9] ArcGIS Runtime SDK for Android, https://developers.arcgis.com/android/ (Accessed 2014-12)

[10] I. F. Darwin, Android Cookbook, http://androidcookbook.com/home.seam (Accessed 2014-12)

[11] MapsForge, http://www.openandromaps.org/ (Accessed 2014-12)

[12] Popis banaka u Rijeci, http://www.stednjainfo.net/lokacijeRijeka.html (Accessed 2014-12)

[13] Google Play; AdriaGUIDE Rijeka, https://play.google.com/store/apps/details?id=com.navigosistem.rijeka (Accessed 2015-4)

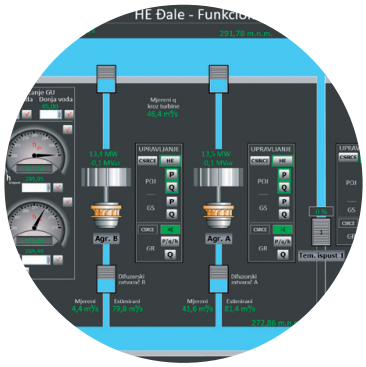
[14] Google Play; Bankomati HR, https://play.google.com/store/apps/details?id=com.Bankomati.main (Accessed 2015-4)

[15] Google Play; AdriaATM, https://play.google.com/store/apps/details?id=com.optimazadar.adriaatm (Accessed 2015-4)

[16] Google Play; BankoMap, https://play.google.com/store/apps/details?id=net.initeq.android.bankomap (Accessed 2015-4)

tradition.knowledge.responsibility.

tradicija.
znanje.
odgovornost.



KONČAR

www.koncar.hr

pouzdanost izvrsnost odgovornost

www.hep.hr

HEP

EXCHANGE IDEAS

SHARE KNOWLEDGE

BUILD SOFTWARE

www.in2.hr

in2 GROUP

INFODOM

Konzalting, BPM i ERP

Sistem integracija i implementacija

Razvoj proizvoda i aplikacija

ASP, Cloud computing i
nove tehnologije

www.infodom.hr

- Javni sektor
- Javne financije i financijska industrija
- Elektroničke komunikacije, promet i turizam
- Privatni sektor, energetika i državna poduzeća

e-Inspektor x2x Platforma e-GOP e-BAP

SPIN@ Oracle E-Business Suite SRMA

KREATIVNI LJUDI INOVATIVNA RJEŠENJA

Ericsson Nikola Tesla osigurava inovativna ICT rješenja koja unaprjeđuju život ljudi, stvaraju novu vrijednost i pozitivno utječu na okoliš.

www.ericsson.hr
www.youtube.com/user/ericssonnikolatesla

ERICSSON

BITI PRVI JE JAKO DOBAR OSJEĆAJ

PRVI DOŽIVITE OSJEĆAJ SURFANJA NA NAJBRŽOJ MOBILNOJ MREŽI

Osjetite nevjerojatnu 4G brzinu surfanja do 150 Mbit/s i to **bez naknade do 30.6.** te prvi prepoznajte novi glazbeni hit ili snimite poseban trenutak i podijelite ga s najdražima u trenu zahvaljujući najbržoj mobilnoj mreži u Hrvatskoj. Biti prvi je stvarno dobar osjećaj! Ponudu ostvarite pozivom na 0800 9000 ili u webshopu na hrvatskitelekom.hr

Ponuda vrijedi i za poslovne korisnike.

ŽIVJETI ZAJEDNO

Brzina ovisi o razini 4G signala, modelu uređaja te trenutnoj opterećenosti mreže

sponsors

golden sponsors

silver sponsors

bronze sponsors

sponsors

MIPRO
Kružna 8/II, HR-51000 Rijeka, Croatia

phone: +385 51 423 984
mipro@mipro.hr • www.mipro.hr

ISBN 978-953-233-083-0